Meta Path-Based Collective Classification in Heterogeneous Information Networks

Xiangnan Kong* Philip S. Yu*§ Ying Ding† David J. Wild†
*University of Illinois at Chicago, Chicago, IL, USA
†Indiana University Bloomington, Bloomington, IN, USA
§Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia
xkong4@uic.edu, psyu@cs.uic.edu, {dingying, djwild}@indiana.edu

ABSTRACT

Collective classification approaches exploit the dependencies of a group of linked objects whose class labels are correlated and need to be predicted simultaneously. In this paper, we focus on studying the collective classification problem in heterogeneous networks, which involves multiple types of data objects interconnected by multiple types of links. Intuitively, two objects are correlated if they are linked by many paths in the network. By considering different linkage paths in the network, one can capture the subtlety of different types of dependencies among objects. We introduce the concept of meta-path based dependencies among objects, where a meta path is a path consisting a certain sequence of linke types. We show that the quality of collective classification results strongly depends upon the meta paths used. To accommodate the large network size, a novel solution, called HCC (meta-path based Heterogenous Collective Classification), is developed to effectively assign labels to a group of instances that are interconnected through different meta-paths. The proposed HCC model can capture different types of dependencies among objects with respect to different meta paths. Empirical studies on real-world networks demonstrate that effectiveness of the proposed meta path-based collective classification approach.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Data Mining

Keywords

Heterogeneous information networks, Meta path

1. INTRODUCTION

Collective classification methods [6] aim at exploiting the label autocorrelation among a group of inter-connected instances and predict their class labels collectively, instead of independently. The dependencies among the related instances should be considered explicitly during classification process. Most approaches in collective classification focus on exploiting the dependencies among interconnected objects in homogeneous networks. However, many real-world applications are facing large scale heterogeneous information networks [5] with multiple types of objects inter-connected through multiple types links. These networks are multi-mode and multi-relational networks, which involves large amount of information. For example, Figure 1 involves a bibliographic network with five types of nodes (papers, author, affiliations, conference and proceedings) and five types of links.

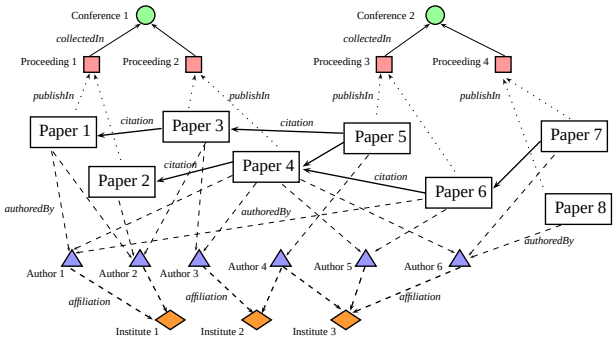


Figure 1: A Heterogeneous Information Network

In this paper, we focus on studying the problem of collective classification on one type of nodes within a heterogenous information network, *e.g.*, classifying the paper nodes collectively in Figure 1. Formally, the collective classification problem in heterogeneous information networks corresponds to predicting the labels of a group of related instances simultaneously. If we consider collective classification and heterogeneous information networks as a whole, the major research challenges can be summarized as follows:

Multi-Mode and Multi-Relational Data: One fundamental problem in classifying heterogeneous information networks is that the network structure involves multiple types of nodes and multiple types of links. For example, in Figure 1, one paper node can be linked *directly* with different types of objects, such as authors, conference proceedings and other papers, through different types of links, such as *citation*, *authoredBy*, *etc.* Trivial application of conventional methods by ignoring the link types and node types can not fully exploit the structural information within a heterogeneous information network.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Table 1: Semantics of Meta Paths (Paper Nodes)

Notation	Meta Path	Semantics
P→P	Paper $\xrightarrow{cite}$ Paper	Citation
P←P→P	Paper $\xrightarrow{cite^{-1}}$ Paper $\xrightarrow{cite}$ Paper	Co-citation
P→P←P	Paper $\xrightarrow{cite}$ Paper $\xrightarrow{cite^{-1}}$ Paper	Bibliographic coupling
PVP	Paper $\xrightarrow{in}$ Proceeding $\xrightarrow{in^{-1}}$ Paper	same proceeding
PVCVP	Paper $\xrightarrow{in}$ Proceeding $\xrightarrow{in}$ Conference $\xrightarrow{in^{-1}}$ Proceeding $\xrightarrow{in^{-1}}$ Paper	same conference
PAP	Paper $\xrightarrow{write^{-1}}$ Author $\xrightarrow{write}$ Paper	sharing authors
PAFAP	Paper $\xrightarrow{write^{-1}}$ Author $\xrightarrow{affiliate}$ Institute $\xrightarrow{affiliate^{-1}}$ Author $\xrightarrow{write}$ Paper	same institute

Heterogeneous Dependencies: Another problem is that objects in heterogeneous information networks can be linked *indirectly* through different types of relational paths. Each types of relational path corresponds to different types of *indirect* relationships between objects. For example, in Figure 1, paper nodes can be linked with each other indirectly through multiple *indirect* relationships, such as, 1) the "paper-author-paper" relation indicates relationships of two papers sharing same authors; 2) the "paper-author-institute-author-paper" relation denotes relationship between papers that are published from the same institute. Heterogenous information networks can encode various complex relationships among different objects. Thus, ignoring or treating all relations equally will loss information dependence information in a heterogeneous information network.

In this paper, we propose a novel solution, called HCC (meta-path based Heterogenous Collective Classification), to effectively assign class labels to one type of objects in the network. Different from conventional collective classification methods, the proposed HCC model can exploit a large number of different types of dependencies among objects simultaneously. We define meta path-based dependencies to capture different types of relationships among objects. By explicitly exploiting these dependencies, our HCC method can effectively exploit the complex relationships among objects.

2. PROBLEM DEFINITION

DEFINITION 1. **Heterogeneous Information Network**: A heterogeneous information network [5] is a special kind of information network, which is represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is the set of nodes, including t types of objects $\mathcal{T}_1 = \{v_{11}, \cdots, v_{1n_1}\}, \cdots, \mathcal{T}_t = \{v_{t1}, \cdots, v_{tn_t}\}$. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of links between the nodes in $\mathcal{V}$, which involves multiple types of links.

Different from conventional networks, heterogeneous information networks involve different types of objects (e.g., papers and conference) that are connected with each other through multiple types of links. Each type of links represents an unique binary relation R from node type i to node type j, where $R(v_{ip}, v_{jq})$ holds iff object v_{ip} and v_{jq} are related by relation R. R^{-1} denotes the inverted relation of R, which holds naturally for $R^{-1}(v_{jq}, v_{ip})$. Let $dom(R) = \mathcal{T}_i$ denote the domain of relation R, $rang(R) = \mathcal{T}_j$ denotes its range. $R(a) = \{b : R(a, b)\}$. For example, in Figure 1, the link type "*write*" can be written as a relation R between paper nodes and author nodes. $R(v_{ip}, v_{jq})$ holds iff author v_{jq} is one of the authors for paper v_{ip}. For convenience, we can write this link type as "$paper \xrightarrow{write^{-1}} author$" or "$\mathcal{T}_i \xrightarrow{R} \mathcal{T}_j$".

In heterogenous information networks, objects are also inter-connected through indirect links, *i.e.*, paths. For example, in Figure 1, paper 1 and paper 4 are linked through a sequence of edges: "$paper1 \xrightarrow{write^{-1}} author1 \xrightarrow{write} paper4$". In order to categorize these paths, we extend the definition of link types to "path types", which are named as *meta path*, similar to [5].

DEFINITION 2. **Meta Path**: A meta path $\mathcal{P}$ represents a sequence of relations $R_1, \cdots, R_\ell$ with constrains that $\forall i \in \{1, \cdots, \ell - 1\}, rang(R_i) = dom(R_{i+1})$. The meta path $\mathcal{P}$ can also be written as $\mathcal{P} : \mathcal{T}_1 \xrightarrow{R_1} \mathcal{T}_2 \xrightarrow{R_2} \cdots \xrightarrow{R_\ell} \mathcal{T}_{\ell+1}$, *i.e.*, $\mathcal{P}$ corresponds to a composite relation $R_1 \times R_2 \times \cdots \times R_\ell$ between node type $\mathcal{T}_1$ and $\mathcal{T}_{\ell+1}$. $dom(P) = dom(R_1)$ and $rang(P) = rang(R_\ell)$. The length of P is ℓ, *i.e.*, the number of relations in $\mathcal{P}$.

Different meta paths usually represent different semantic relationships among linked objects. In Table 1, we show some examples of meta paths with their corresponding semantics. Most conventional relationships studied in network data can naturally be captured by different meta paths. For example, the paper *co-citation* relation [1] can naturally be represented by meta path "$paper \xrightarrow{cite^{-1}} paper \xrightarrow{cite} paper$", and the co-citation frequencies can be written as the number of path instances for the meta path. Here a *path instance* of $\mathcal{P}$, denoted as $p \in \mathcal{P}$, is an unique sequence of nodes and links in the network that follows the meta path constrains. For convenience, we use the node type sequence to represent a meta path, *i.e.*, $\mathcal{P} = \mathcal{T}_1\mathcal{T}_2 \cdots \mathcal{T}_{l+1}$. For example, we use PAP to represent the meta path "$paper \xrightarrow{write^{-1}} author \xrightarrow{write} paper$". Note that for meta paths involving *citation* links, we explicitly add arrows to represent the link directions, *e.g.*, the paper *co-citation* path can be written as $P \leftarrow P \rightarrow P$.

Collective Classification in Heterogeneous Information Networks

In this paper, we focus on studying the collective classification problem on *one* type of objects, instead of on all types of nodes in heterogeneous information networks. This problem setting exists in a wide variety of applications. The reasons are as follows: in heterogenous information networks, the label space of different types of nodes are quite different, where we can not assume all types of node share the same set of label concepts. For example, in medical networks, the label concepts for patient classification tasks are only defined on patient nodes, instead of doctor nodes or medicine nodes. In a specific classification task, we usually only care about the classification results on one type of node. Without loss of generality, we assume the node type $\mathcal{T}_1$ is the target objects we need to classify. Suppose we have n nodes in $\mathcal{T}_1$. On each node $v_{1i} \in \mathcal{T}_1$, we have a vector of attributes $\boldsymbol{x}_i \in \mathbb{R}^d$ in the d-dimensional input space, and $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_1}\}$. Let $\mathcal{C} = \{c_1, c_2, \cdots, c_q\}$ be the q possible class labels. On each node $v_{1i} \in \mathcal{T}_1$, we also have a label variable $Y_i \in \mathcal{C}$ indicating the class label assigned to node v_{1i}, $\mathcal{Y} = \{Y_i\}_{i=1}^{n_1}$.

Assume further that we are given a set of known values $\mathcal{Y}_L$ for nodes in a training set $\mathcal{L} \subset \mathcal{T}_1$, and L denotes the index set for training data. $\mathcal{Y}_L = \{y_i | i \in L\}$, where $y_i \in \mathcal{C}$ is the observed labels assigned to node x_{1i}. Then the task of collective classification in heterogeneous information

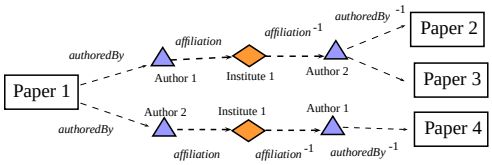


Figure 2: Path instances for $PAFAP$.

networks is to infer the values of $Y_i \in \mathcal{Y}_U$ for the remaining nodes in the testing set ($\mathcal{U} = \mathcal{T}_1 - \mathcal{L}$).

The inference problem in classification tasks is to estimate $Pr(\mathcal{Y}|\mathcal{X})$ given a labeled training set. Conventional classification approaches usually require i.i.d. assumptions, the inference for each instance is performed independently: $Pr(\mathcal{Y}|\mathcal{X}) \propto \prod_{i \in U} Pr(Y_i|\boldsymbol{x}_i)$

Homogeneous Link-based Dependency

In collective classification problems, the labels of related instances are not independent, but are closely related with each other. Conventional approaches focus on exploiting label dependencies corresponding to one types of homogeneous links to improve the classification performances, e.g., citation links in paper classification tasks, co-author links in expert classification tasks. These methods can model $Pr(Y_i|\boldsymbol{x}_i, \boldsymbol{Y}_{\mathcal{P}(i)})$. Here $\boldsymbol{Y}_{\mathcal{P}(i)}$ denotes the vector containing all variable Y_j ($\forall j \in \mathcal{P}(i)$), and $\mathcal{P}(i)$ denotes the index set of related instances to the i-th instance through meta path $\mathcal{P}$. Hence, by considering the single type of dependencies, we will have $Pr(\mathcal{Y}|\mathcal{X}) \propto \prod_{i \in U} Pr(Y_i|\boldsymbol{x}_i, \boldsymbol{Y}_{\mathcal{P}(i)})$

Meta Path-based Dependency

In heterogeneous information networks, there are complex dependencies not only among instances directly linked through links, but also among instances indirectly linked through different meta paths. In order to solve the collective classification problem more effectively, in this paper, we explicitly consider different types of meta-path based dependencies in heterogeneous information networks. Meta path-based dependences refer to the dependencies among instances that are inter-connected through a meta path.

To the best of our knowledge, meta path-based dependencies have not been studied in collective classification research before. Given a set of meta paths $\mathcal{S} = \{\mathcal{P}_1, \cdots, \mathcal{P}_m\}$, the meta path-based dependency model is

$$Pr(Y_i|\boldsymbol{x}_i, \boldsymbol{Y}_{\mathcal{P}_1(i)}, \boldsymbol{Y}_{\mathcal{P}_2(i)}, \cdots, \boldsymbol{Y}_{\mathcal{P}_m(i)})$$

$\mathcal{P}_j(i)$ denotes the index set of related instances to the i-th instance through meta path $\mathcal{P}_j$.

For each meta path, one instance can be connected with multiple related instances in the network. For example, in Figure 2, Paper 1 is correlated with Paper 2, 3 and 4 through meta path $\mathcal{P}_i = PAFAP$, *i.e.*, $\mathcal{P}_i(Paper1) = \{Paper\ 2, 3, 4\}$. Hence, by considering meta path-based dependencies, we will have
$Pr(\mathcal{Y}|\mathcal{X}) = \prod_{i \in U} Pr\left(Y_i|\boldsymbol{x}_i, \boldsymbol{Y}_{\mathcal{P}_1(i)}, \boldsymbol{Y}_{\mathcal{P}_2(i)}, \cdots, \boldsymbol{Y}_{\mathcal{P}_m(i)}\right)$

3. META PATH-BASED COLLECTIVE CLASSIFICATION

For classifying target nodes in a heterogeneous information network, the most naïve approach is to approximate $Pr(\mathcal{Y}|\mathcal{X}) \propto \prod_{i \in U} Pr(Y_i|\boldsymbol{x}_i)$ with the assumptions that all

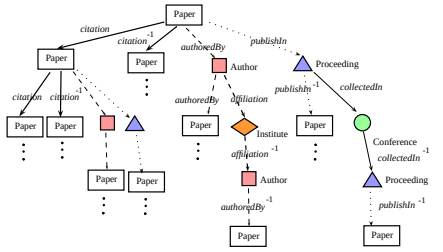


Figure 3: Dependence tree for meta paths.

instances are independent from each other. However, this approach can be detrimental to their performance for many reasons. This is particularly troublesome when nodes in heterogeneous networks have very complex dependencies with each other through different meta paths.

In this section, we propose a simple and effective algorithm for meta path-based collective classification in heterogeneous information networks. We aim to develop a model to estimate the probabilities $Pr\left(Y_i|\boldsymbol{x}_i, \boldsymbol{Y}_{\mathcal{P}_1(i)}, \cdots, \boldsymbol{Y}_{\mathcal{P}_m(i)}\right)$. We first introduce how to extract the set of meta paths from a heterogeneous information network, then propose our collective classification algorithm, called HCC (Heterogeneous Collective Classification).

We first consider how to extract all meta paths in a heterogeneous information network of bounded length ℓ_{max}. When ℓ_{max} is small, we can easily generate all possible meta paths as follows: We can organize all the type-correct relations into a prefix tree, called *dependence tree*. In Figure 3, we show an example of dependence tree in ACM conference networks. The target nodes for classification are the paper nodes, and each paper node in the dependence tree corresponds to an unique meta path, indicating one type of dependencies among paper instances. However, in general the number of meta paths grows exponentially with the maximum path length ℓ_{max}. As it has been showed in [5], long meta paths may not be quite useful in capturing the linkage structure of heterogeneous information networks. In this paper, we only exploit the instance dependences with short meta paths ($\ell_{max} = 4$).

In many really world network data, exhaustively extracting all meta paths may result in large amount of redundant meta paths, *e.g.*, $PVPVP$. Including redundant meta paths in a collective classification model can result in overfitting risks, because of additional noisy features. Many of the redundant meta paths are constructed by combining two or more meta paths, *e.g.*, meta path $PVPVP$ can be constructed by two PVP paths. In order to reduce the model's overfitting risk, we extract all meta paths that cannot be decomposed into shorter meta paths (with at least one *non-trivial* meta paths). Here non-trivial meta paths refer to the paths with lengths greater than 1. For example, in ACM conference network, meta paths like $P{\rightarrow}PAP$ can be decomposed into $P{\rightarrow}P$ and PAP, thus will be excluded from our meta path set. In Figure 4, we showed the meta path set extract process as the "Initialization" step of our proposed method. By breadth-first search on the dependence tree, our model first select shortest meta paths from the network. Then longer meta paths are incrementally selected into path set $\mathcal{S}$ until we reach a meta path that can be decomposed into shorter meta paths in $\mathcal{S}$.

After the meta path set $\mathcal{S}$ is extracted from the heterogeneous information network, we then show how to use

Input:
$\mathcal{G}$: a heterogeneous network, ℓ_{max}: maximum length.
$\mathcal{X}$: features for all instances, $\mathcal{Y}_L$: labels for training set.
L: index set for training set, U: index set for test set.
A: a base learner for local model, Max_It: maximum # of iteration.

Initialize:
- Construct the meta path set $\mathcal{S} = \{\mathcal{P}_1, \cdots, \mathcal{P}_m\}$
  Breadth first search on dependence tree
  1. If the length of meta path in current tree node is greater than ℓ_{max}, exit the BFS;
  2. If the current meta path $\mathcal{P}_j$ in current tree node cannot be reconstructed by the paths in $\mathcal{S}$, Add $\mathcal{P}_j$ into $\mathcal{S}$; Otherwise, prune the current node from BFS.

Training:
- Learn the local model f:
  1. Construct an extended training set $\mathcal{D} = \left\{(\boldsymbol{x}'_i, \boldsymbol{y}_i)\right\}$ by converting each instance $\boldsymbol{x}_i$ to $\boldsymbol{x}'_i$ as follows: $\boldsymbol{x}'_i = (\boldsymbol{x}_i, \text{PathRelFeature}(v_{1i}, \mathcal{E}_L, \mathcal{Y}_L, \mathcal{S}))$
  2. Let $f = A(\mathcal{D})$ be the local model trained on $\mathcal{D}$.

Bootstrap:
- Estimate the labels, for $i \in U$
  1. Produce an estimated value $\hat{Y}_i$ for Y_i as follows: $\hat{Y}_i = f((\boldsymbol{x}_i, \mathbf{0}))$ using attributes only.

Iterative Inference:
- Repeat until convergence or #iteration$> Max_It$
  1. Construct the extended testing instance by converting each instance $\boldsymbol{x}_i$ to $\boldsymbol{x}'_i$ $(i \in U)$ as follows: $\boldsymbol{x}'_i = \left(\boldsymbol{x}_i, \text{PathRelFeature}(v_{1i}, \mathcal{E}, \mathcal{Y}_L \cup \{\hat{Y}_i | i \in U\}, \mathcal{S})\right)$
  2. Update the estimated value $\hat{Y}_i$ for Y_i on each testing instance $(i \in U)$ as follows: $\hat{Y}_i = f(\boldsymbol{x}'_i)$.

Output:
$\hat{\boldsymbol{Y}}_U = \left(\hat{Y}_1, \cdots, \hat{Y}_n\right)$: The labels of test instances $(i \in U)$.

Figure 4: The Hcc algorithm

these meta paths to perform collective classification effectively. Conventional collective classification based on iterative inference process, *e.g.* ICA (Iterative Classification Algorithm) [4, 3], provide a simple yet very effective method for collective classification in homogeneous networks. Inspired by the success of these iterative inference methods, in this paper, we propose a similar framework for meta path-based collective classification method. This approach is called Hcc (Heterogeneous Collective Classification), summarized in Figure 4.

The general idea is as follows: we model the joint probability based on the following assumption: if instance v_{1i} and v_{1j} are not connected via any meta path in $\mathcal{S}$, the variable Y_i is conditional independent from Y_j given the labels of all v_{1i}'s related instances, *i.e.*, $\{v_{1j} | j \in \bigcup_{k=1}^{m} \mathcal{P}_k(i)\}$. Hence the local conditional probability each instance's label can be modeled by a base learner with extended *relational features* built upon the predicted Y_j's $(j \in \bigcup_{k=1}^{m} \mathcal{P}_k(i))$. And the joint probability can be modeled based on these local conditional probabilities by treating the instances as being independent.

In collective classification, each instance may be linked with different number of instances through one meta path. In order to build a fixed number of relational features for each instance, we employs *aggregation* functions to combine the predictions on the labels of related instances. Many aggregation functions can be used here, such as COUNT and MODE aggregators [3]. In this paper, we use the *weighted label fraction* of the related instances as the relational feature for each meta path. We calculate the average fraction of each label appearing in the related instances. Each related instance is re-weighted by the number of path instances from the current node, e.g., for meta path PAP, the papers that share more authors in their author lists are more likely to share similar topics than those only share one author. In detail, given an aggregation function, we can get one set of relational features from the labels of related instances for each meta path, as shown in Figure 5.

Inspired by the success of ICA framework [3] in collective classification, we designed a similar inference procedure for our Hcc method as shown in Figure 4. (1) For inference steps, the labels of all the unlabeled instances are unknown.

$\boldsymbol{x}^r = \text{PathRelFeature}(v, \mathcal{E}, \{Y_i\}, \mathcal{S} = \{\mathcal{P}_1, \cdots, \mathcal{P}_m\})$

For each meta path $\mathcal{P}_i \in \mathcal{S}$:
1. Get related instances $\mathcal{C} = \mathcal{P}_i(v, \mathcal{E})$
2. $\boldsymbol{x}^i = Aggregation\left(\{Y_j | v_{1j} \in \mathcal{P}_i(v)\}\right)$

Return relational feature $\boldsymbol{x}^r = \left(\boldsymbol{x}^1, \cdots, \boldsymbol{x}^m\right)$

Figure 5: Get meta path-based relational features

We first *bootstrap* an initial set of label estimation for each instance using content attributes of each node. In our current implementation, we simply set the relational features of unlabeled instances with zero vectors. Other strategies for *bootstrap* can also be used in this framework. (2) *Iterative Inference*: we iteratively update the relational features based on the latest predictions and then these new features are used to update the prediction of local models on each instance. The iterative process terminates when convergence criteria are met. In our current implementation, we update the variable Y_i in the $(r+1)$-th iteration (say $\hat{Y}_i^{(r+1)}$) using the predicted values in the r-th iteration ($\hat{Y}_j^{(r)}$) only.

4. EXPERIMENTS

ACM Conference Dataset: It was extracted from ACM digital library[1] in 2011. ACM digital library provides detailed bibliographic information on ACM conference proceedings, including paper abstracts, citation, author information *etc.* We extract two ACM sub-networks containing conference proceedings before the year 2011.The first subset, *i.e.*, *ACM Conference-A*, involves 14 conferences in computer science: SIGKDD, SIGMOD, SIGIR, SIGCOMM, CIKM, SODA, STOC, SOSP, SPAA, MobiCOMM, VLDB, WWW, ICML and COLT. The network structure is summarized in Figure 1, which involves five types of nodes and five types of relations/links. This network includes 196 conference proceedings (*e.g.*, KDD'10, KDD'09, *etc.*), 12.5K papers, 17K authors and 1.8K authors' affiliations. On each paper node, we extract bag-of-words representation of the paper title and abstract to use as content attributes. The stop-words and rare words that appear in less than 100 papers are removed from the vocabulary. Each paper node in the network is assigned with a class label, indicating the ACM index term of the paper including 11 categories. The task in this dataset is to classify the paper nodes based on both local attributes and the network information. The second subset, *i.e.*, *ACM Conference-B*, involves another 12 conferences in computer science: ACM Multimedia, OSDI, GECCO, POPL, PODS, PODC, ICCAD, ICSE, ICS, ISCA, ISSAC and PLDI. The network includes 196 corresponding conference proceedings, 10.8K papers, 16.8K authors and 1.8K authors' affiliations. After removing stop-words in the paper title and abstracts, we get 0.4K terms that appears in at least 1% of the papers.

Compared Methods

• Heterogeneous Collective Classification (Hcc): The proposed approach exploits dependencies based on multiple meta paths for collective classification.

• Homogeneous Collective Classification (Ica): Our implementation of the ICA (Iterative Classification Algorithm) [3] by only using homogeneous network information for collective classification. In the homogeneous information networks, only paper-paper links are used.

• Combined Path Relations (Cp): A baseline for multi-relational collective classification [2]: We first convert the

[1]http://dl.acm.org/

heterogeneous information networks into multiple relational networks with one type of nodes and multiple types of links. Each link type corresponds to a meta path in the HCC method. Then, the CP method combines multiple link types into a homogeneous network by ignoring the link types. We then train one ICA model to perform collective classification on the combined network.

• Collective Ensemble Classification (CF): Our implementation of the collective ensemble classification [2], which trains one collective classification model on each link types. We use the same setting of the CP method to extract multi-relational networks. Then we use ICA as the base models for collective classification. For the iterative inference, each model vote for the class label of each instance, and prediction aggregation was performed in each iteration. Thus this process is also called collective fusion, where each base model can affect each other in the collective inference step.

• Ceiling of HCC (HCC-ceiling): One claim of this paper is that HCC can effectively infer the labels of linked unlabeled instances using iterative inference process. To evaluate this claim, we include a model which use the *ground-truth* labels of the related instances during the inference. This method illustrate a ceiling performance of HCC can possibly achieve by knowing the *true* label of related instances.

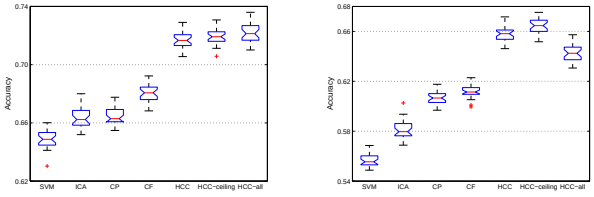
• HCC with all meta-paths (HCC-all): Another claim of this paper is that selected meta path in HCC can effectively capture the dependencies in heterogeneous information networks and avoiding overfitting. To evaluate this claim, we include a model which uses all possible meta paths with a maximum path length of 5. This method illustrates the performance of HCC if we exhaustively involves all possible path-based dependencies without selection.

We use LibSVM with linear kernel as the base classifier for all the compared methods. The maximum number of iteration all methods are set as 10.

Performances of Collective Classification: 10 times 3-fold cross validations are performed on each heterogeneous information network to evaluate the collective classification performances. We report the detailed results in Figure 6. It shows the performances of the six methods on three datasets with box plots, including the smallest/largest results, lower quartile, median and upper quartile.

The first observation we have in Figure 6 is as follows: almost all the collective classification methods that explicitly exploit the label dependencies from various aspects, can achieve better classification accuracies than the baseline SVM, which classify each instance independently. These results can support the importance of collective classification by exploiting the different types of dependencies in network data. For example, ICA outperformances SVM by exploiting auto-correlation among instances while considering only one type of links, *i.e.*, citation links. CF and CP methods can also improve the classification performances by exploiting multiple types of dependencies. Similar results have also been reported in collective classification literatures.

Then we find that our meta path-based collective classification method (HCC) consistently and significantly outperform other baseline methods. HCC can utilize the meta path-based dependencies to exploit the heterogenous network structure more effectively. These results support our claim that in heterogeneous information networks, instances can be correlated with each other through various meta paths. Exploiting the complex dependencies among the related instances (*i.e.*, various meta path-based dependencies) can effectively extract the heterogenous network structural information and thus boost the classification performance.



(a) ACM Conference-A (b) ACM Conference-B

Figure 6: Collective classification results.

We further observe that the HCC models perform comparably with the HCC-ceiling models which had access to the *true* label of related instances. This indicates that the HCC model reach its full potential in approximated inference process. In addition, HCC method with a small set of representative paths can achieve also comparable performances with HCC-all models which includes all meta path combinations with path length $\ell_{max} \leq 5$. And the performances of HCC are more stable than those HCC-all in ACM datasets. In ACM Conference-B dataset, HCC method with fewer meta-paths can even outperform HCC-all method. These results support our second claim that the heterogeneous dependencies can be captured effectively by selecting a small set of representative meta-paths and thus our HCC model can avoid overfitting than using all meta paths.

5. CONCLUSION

In this paper, we studied the collective classification problem in heterogeneous information networks. We propose a novel solution to collective classification in heterogeneous information networks, called HCC, which can effectively assign labels to a group of interconnected instances involving different meta path-based dependencies. The proposed HCC model is able to capture the subtlety of different dependencies among instances with respect to different meta paths.

Acknowledgment

This work is supported in part by NSF through grants IIS-0905215, CNS-1115234, IIS-0914934, DBI-0960443, and OISE-1129076, US Department of Army through grant W911NF-12-1-0066, and Google Mobile 2014 Program.

6. REFERENCES

[1] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

[2] H. Eldardiry and J. Neville. Across-model collective ensemble classification. In *AAAI*, San Franciscol, CA, 2011.

[3] Q. Lu and L. Getoor. Link-based classification. In *ICML*, pages 496–503, Washington, DC, 2003.

[4] J. Neville and D. Jensen. Iterative classification in relational data. In *AAAI'10 Workshop on Learning Statistical Models from Relational Data*, Austin, TX, 2000.

[5] 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*, Seattle, WA, 2011.

[6] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *UAI*, pages 482–492, Edmonton, Alberta, 2002.