HetPathMine: A Novel Transductive Classification Algorithm on Heterogeneous Information Networks

Chen Luo[1], Renchu Guan[1], Zhe Wang[1,*], and Chenghua Lin[2]

[1] College of Computer Science and Technology, Jilin University,
Changchun 130012, China
rackingroll@163.com, {wz2000,guanrenchu}@jlu.edu.cn
[2] School of Natural and Computing Sciences, University of Aberdeen, UK
chenghua.lin@abdn.ac.uk

Abstract. Transductive classification (TC) using a small labeled data to help classifying all the unlabeled data in information networks. It is an important data mining task on information networks. Various classification methods have been proposed for this task. However, most of these methods are proposed for homogeneous networks but not for heterogeneous ones, which include multi-typed objects and relations and may contain more useful semantic information. In this paper, we firstly use the concept of meta path to represent the different relation paths in heterogeneous networks and propose a novel meta path selection model. Then we extend the transductive classification problem to heterogeneous information networks and propose a novel algorithm, named HetPathMine. The experimental results show that: (1) HetPathMine can get higher accuracy than the existing transductive classification methods and (2) the weight obtained by HetPathMine for each meta path is consistent with human intuition or real-world situations.

1 Introduction

Information network is an efficient way to represent relational data in data mining tasks [1]. For example, a co-author relationship dataset can be represented as a co-author network and a web-page connection dataset could construct a WWW network. The network research has been attracted many attentions in recent years [2–5]. Among these researches, transductive classification [6] (TC) is one of the popular method to extract knowledge with the help of a small amount of labeled data [7]. Most of these algorithms are proposed for the homogeneous information networks [6, 7], however, the real world is full of heterogeneous networks which include various types of objects and relations. In this paper, we move transductive classification problem to heterogeneous networks [1, 8], which have more meaningful information than homogeneous ones.

[*] Corresponding author.

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 210–221, 2014.

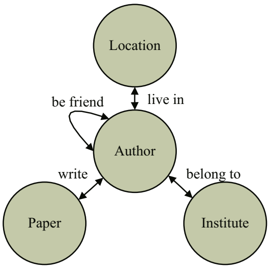


Fig. 1. Network Schema for a Sample Heterogeneous Network

Up till now, many transductive classification methods have been proposed on information networks. For example, learning with local and global consistency algorithm (LLGC) [6] utilizes the network structure to propagate the labels to the unlabeled data. For networked data, it is one of the most popular transductive classification method. The weighted-vote relational neighbor classifier (wvRN), proposed in [9], is another widespread method, which determines the label by considering all the neighbors label. All these algorithms are proposed for homogeneous networks. In 2010, Ming *et.al* proposed GNetMine[10], which is a classification method on heterogeneous networks. GNetMine assumes that all the objects in the network have the same classification criterion. However, most of the heterogeneous networks with different types of objects may have different classification criteria. For example, a network with four types objects: people, institute, location, and paper. The network schema is showed in Figure 1. Author can be classified into research area, but institute and location cannot be classified with this criterion.

In this paper, we extend the transductive classification into heterogeneous networks by utilizing the relation paths existed in the network. The relationships in heterogeneous network include not only the relations but also the relation path. Considering network in Figure 1, two authors can be connected by not only the "author-author" (friendship) relationship but also the "author-institute-author" or "author-paper-paper" relation path. Meta-path [11], proposed as a topology measure for heterogeneous information networks, depicted the different relation paths in heterogeneous networks. We use this concept to model the different relationship in heterogeneous information networks. In addition, a novel meta path selection model is proposed to calculate the different weight of each relation paths. Finally, by using the different weight of each meta path, we propose a transductive classification framework, named HetPathMine, on the heterogeneous information network.

The rest of this paper is organized as follow: some concepts used in this paper is introduced in section 2, we present the HetPathMine algorithm in section 3. Experimental result is showed in section 4. Finally, we conclude our work in section 5.

2 Background

2.1 Problem Definition

In this section, we introduce the concept of *Heterogeneous Information Network* and *class*. Then we formally define the problem of *transductive classification in heterogeneous network*. Following [8], the Heterogeneous Information Network is defined as follows:

Definition 1 (Heterogeneous Information Network [8]). *Suppose we have m types of data objects, denoted by $X_1 = \{x_{11}, ..., x_{1n_1}\}, ..., X_m = \{x_{m1}, ..., x_{mn_m}\}$, a heterogeneous information network is in the form of a graph $G = \langle V, E, W \rangle$, where $V = \bigcup_{i=1}^{m} X_i$ and $m \geq 2$, E is the set of links between any two data objects of V ,and W is the set of weight values on the links. G reduces to a homogeneous network, when $m = 1$.*

A *class* on heterogeneous information network is defined as follow:

Definition 2 (Class). *Given a heterogeneous information network $G = \langle V, E, W \rangle$, a subset of V: V' is a class in G, when $V' \subseteq V \in X_i$, and X_i is the target type for classification.*

It is pointed out that, our definition of *class* is different from the definition in [10]. In this paper, we only consider a *class* as a set of objects within the same type. It is meaningless to take objects of different type into same *class*. In addition, different type has very different semantic meaning and may have different classification criteria.

After defining the concept of class in this paper, following [10], we define the transductive classification in heterogeneous information network as follow:

Definition 3 (Transductive classification in heterogeneous information network). *Given a heterogeneous information network $G = \langle V, E, W \rangle$, suppose V' is a subset of V and $V' \subseteq V \in X_t$, where X_t is the target type for classification, and each data object O in V' is labeled with a value γ indicating which class O should be in. The classification task is to predict the labels for all the unlabeled objects $V - V'$.*

2.2 Meta Path

In [11], the author uses the concept meta-path to denote the different relations and relation paths in heterogeneous information networks. Follow [11], the meta path is defined as follow:

Definition 4 (Meta Path [11]). *Given a heterogeneous information network $G = \langle V, E, W \rangle$, A meta path P is in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} ...A_{l-1} \xrightarrow{R_{l-l}} A_t$, which defines a composite relation $P = R_1 \circ R_2 \circ ... \circ R_{l-1}$ between two objects, and $\circ$ is the composition operator on relations.*

Note that, in this paper, we only consider the meta path that $A_1 = A_t$. After defining the concept of meta path, we need to know how to measure it. Here, we introduce a topology measure, PathSim, proposed in [6], for measuring the meta-path: Given a meta path, denoted as P, the PathSim between two objects $s, t \in X_t$ can be calculated as follow:

$$S_P^{PathSim}(s,t) = \frac{2 * S_P^{PathCount}(s,t)}{S_P^{PathCount}(s,:) + S_{P^{-1}}^{PathCount}(:,t)}$$

In the above, $S_P^{PathCount}$ is a Path Count Measure [6] and it can be calculated as the number of path instances between s and t. P^{-1} denotes the inverse meta path [6] of P, $S_P^{PathCount}(s,:)$ denotes the path count value following P starting with s, and $S_{P^{-1}}^{PathCount}(:,t)$ denotes the path count value following P^{-1} ending with t.

3 The HetPathMine Framework

Transductive classification, as a relational learning [12] method, utilizes the structure of the network data to predict labels. There are two assumptions in transductive classification. First, it assumes that the network resulting from a social process often possess a high amount of influence [13]. Such that linked nodes may have a high possibility to have the same label. Second, transductive classification, as a semi-supervised learning algorithm, some pre-labeled information is obtained before the learning process. And the classification result should be consisted with the pre-labeled information [10].

In this section, we propose a novel transductive classification method on heterogeneous networks based on the meta path, named HetPathMine. The HetPathMine algorithm will consist with the two assumptions introduced before. We firstly introduce the Meta-path selection model. Secondly, we extend the transductive classification to the heterogeneous networks. Finally, the detailed steps of HetPathMine is introduced.

3.1 Meta Path Selection Model

Each meta-path can represent a relationship in the heterogeneous networks [14]. For example, in Figure 1, the co-author relationship can be represented by the meta path: “author-paper-author”, and co-institute relationship can be represented by the meta path: “author-institute-author”. If we want to classify the authors into different research area, the co-author relationship will play a very important role, while the co-institute relationship will bias the classifying process. As a result, we need to know different weight for each meta path. In [14], the author proposed an algorithm which can find the different weight of each meta-path. This process is a called meta-path selection process [14]. However, the model, proposed in [14], is too complicated and has high-computational. In [15], the author proposed a relation extraction model for multi-relational network. Similar to [15], in this paper, we propose a novel meta path selection model.

Given a set of meta paths, denoted as $\boldsymbol{P} = [p_1, p_2, ..., p_d]$, where d is the number of meta paths used in our algorithm. Then, the weight for each meta path is denoted as $B = [\beta_1, \beta_2, ..., \beta_d]$. We use a linear regression algorithm [16] to calculate B. The cost function $L(B)$ is given bellow:

$$L(B) = \left\| \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbb{R}(x_{t,i}, x_{t,j}) - \sum_{d=1}^{m} \beta_d s_{p_k}^{PathSim}(x_{t,i}, x_{t,j})) \right\| + \mu \left\| \sum_{k=1}^{d} \beta_k \right\| \quad (1)$$

where n is the number of labeled objects. $x_{t,i}, x_{t,j} \in X_t$, and X_t is the target type for classification. $S_{p_k}^{PathSim}(x_{t,i}, x_{t,j})$ is the PathSim measure for meta path p_k. $\mathbb{R}$ is a relation matrix and can be calculated as follow:

$$\mathbb{R}(x_{t,i}, x_{t,j}) = \begin{cases} 1 \; x_{t,i}, x_{t,j} \; \textit{are labeled as the same label} \\ 0 \qquad\qquad\qquad \textit{otherwise} \end{cases} \quad (2)$$

The first term of (1) ensure that the weight for each meta path can follow the pre-labeled information. The second term is a *smoothness constraint*. The weight for each term is captured by μ, a positive parameter. Then $B = [\beta_1, \beta_2, ..., \beta_d]$ can be calculated by solving the following problem:

$$B^* = \arg max_{B=[\beta_1, \beta_2, ..., \beta_d]} L(\beta) \quad (3)$$

Many useful methods can be used to solve such problem [16], in this research we use gradient descent method [16] to solve this problem. After having all the weights for each meta path, we can use them to model the transductive classification on heterogeneous information network.

3.2 Meta Path-Based Transductive Classification Model

In this section, we extend the transductive classification model to the heterogeneous network by considering different meta paths. As introduced before, there are two assumptions in transductive classification tasks: (1) objects have a tight relationship tend to have the same label. (2) The classification result should consist with the pre-labeled information. In [6], the author proposed a transductive classification model on homogeneous networks. We extend the model to heterogeneous network as follow:

$$\varrho(F) = \frac{1}{2} \sum_{k=1}^{d} (\sum_{i=0}^{n} \sum_{j=0}^{n} \beta_k W_{ij}^{k} \left\| \frac{F_i}{\sqrt{D_{ii}^{k}}} - \frac{F_j}{\sqrt{D_{jj}^{k}}} \right\|^2) + \lambda \sum_{i=0}^{n} \sum_{i=0}^{n} \|F_i - Y_i\|^2 \quad (4)$$

In the above, n is the number of objects within the target type X_i, W^k is the similarity matrix for the target type reduced by the $k-th$ meta path. F is a $n*p$ matrix, where p is the number of class, and $F(i,j)$ denotes the probability of $i-th$ object belonging to the $j-th$ class. Y is also a $n*p$ matrix which denotes the pre-labeled information in the network, D^k is a diagonal matrix with its $(i,i)-th$ element equal to the sum of the $i-th$ row of W^k. $B=[\beta_1,\beta_2,...,\beta_d]$ is calculated by Eq. (3). The first term of this cost function is the *smoothness constrain*, which follows the first assumption. The second term is the *fitting constrain*, which follows the second assumption introduced before. These two constraints are captured by the parameter λ.

Then, we can get result as follow:

$$\frac{\partial \varrho}{\partial F} = F^* - F^*(\sum_{k=1}^{m} \beta_k S^k) + \mu(F^* - Y) \tag{5}$$

where $S = D^{k^{-1/2}} W^k D^{k^{-1/2}}$.

As the process in [6], F^* can be directly obtained without iterations as follow:

$$F^* = \beta(I - aS_{com})^{-1}Y \tag{6}$$

where $a = \frac{1}{1+\mu}$, $a = \frac{\mu}{1+\mu}$ and $S_{com} = \sum_{k=1}^{m} \beta_k S^k$.

Noticing that, when the network is a homogeneous information network, then $S_{com} = S$, the formula becomes $F^* = \beta(I - aS)^{-1}Y$, which is the same as the formula in homogeneous networks. In other words, our model is a generalized form for the homogeneous one.

After getting F, we can obtain labels of all the objects $x_{t,i} \in X_t$ as follow:

$$Lable(x_{t,i}) = \max\{F(i,1), F(i,2), ..., F(i,n)\} \tag{7}$$

3.3 The Detailed Steps of HetPathMine

After having the calculation method for each relevant variable, the complete framework of HetPathMine is then summarized as the following steps:

Step1. Given a heterogeneous information network $G = \langle V, E, W\rangle$,target type X_t for classification, a set of meta paths $\boldsymbol{P} = [p_1, p_2, ..., p_d]$ and labeled information Y.
Step2. Use Eq. (2) to calculate the relation matrix $\mathbb{R}$.
Step3. Use Eq. (3) and the labeled information Y to obtain the weight of each meta-path: $B = [\beta_1, \beta_2, ..., \beta_d]$.
Step4. Use Eq. (6) to obtain the classification result matrix F^*.
Step5. Use Eq. (7) to obtain the final label of all the objects $x_{t,i} \in X_t$.

It is pointed out that, the output of HetPathMine not only contains the classification result but also the different weight $B = [\beta_1, \beta_2, ..., \beta_d]$ for the selected meta paths, B can be used in many data mining tasks [17, 18].

3.4 Complexity Analysis

In this section, we consider the time complexity of the HetPathMine. All the topological features are calculated at the beginning of our framework, the time complexity is$O(k_p n^2)$. Here k_p is the number of meta paths selected for HetPathMine, and $n = |X_t|$ is the number of target objects. For the Meta Path selection model, the time complexity is $O(t_1 \sum_m |T_m|)$. Here t_1 is the number of iterations, and m is the dimension of the training dataset. For the Meta Path-Based transductive classification model, F can be calculated directly, then time complexity is $O(1)$. Then the overall time complexity of our framework is $O(k_p n^2) + O(t_1 \sum_m |T_m|) + O(1)$.

4 Experimental Results

In this section, we represent an empirical study of the effectiveness of HetPathMine compared with several baseline algorithms and state-of-the-art algorithms based on DBLP dataset.

4.1 Datasets

We use the DBLP dataset [1] for performance test. DBLP, computer science bibliography database, which has been used in many research papers [1, 10, 19], is a typical heterogeneous information network. There are four types of objects in this network: paper, author, term, and conference. Links between author and paper defined by the relation of "write" and "written by", denoted as "$write^{-1}$". Relation between Term and Paper is "mention" and "mentioned by", denoted as "$mention^{-1}$". Relation between Paper and Conference is "publish" and "published by", denoted as "$publish^{-1}$". The network schema is showed in Figure 2.

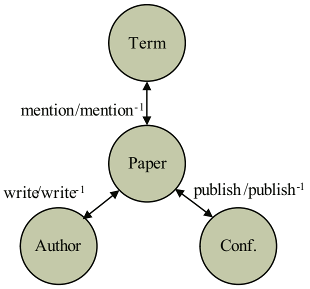


Fig. 2. Network Schema for a Sample Heterogeneous Network

We extract a sub network of DBLP, "four-area dataset [10]", which contains 18 major conferences in four areas: Data Mining, Database, Information Retrieval and Machine Learning. Each area contains several top conferences. They are KDD, ICDM, SDM and PAKDD in Data Mining, SIGMOD, VLDB, ICDE, PODS and EDBT in Database, SIGIR, ECIR, WSDM and WWW, CIKM in Information Retrieval and NIPS, ICML, AAAI and IJCAI in Machine Learning. Four sub-datasets are used in our experiment:

[1] http://www.informatik.uni-trier.de/~ley/db/

DataSet-1 top 100 authors in the DBLP within the 18 major conferences, and the corresponding papers published by these authors after 2007.

DataSet-2 top 500 authors in the DBLP within the 18 major conferences, and the corresponding papers published by these authors after 2007.

DataSet-3 top authors in the DBLP within the 18 major conferences, and the corresponding papers published by them after 2007.

DataSet-4 top authors in the DBLP within the 18 major conferences, and the corresponding papers.

Top authors means that the number of papers published by these authors are larger than the other authors. In these datasets, the term is extracted from the paper titles. The labeled information or the ground truth is obtained from the "four-area dataset [10]". The summary of these four data sets is showed in Table 1.

Table 1. Summary of the Four Sub-DataSet

DataSet	**Authors**	**Papers**	**Terms**	**Conferences**
DataSet-1	100	3212	3984	18
DataSet-2	500	8194	7561	18
DataSet-3	1479	10298	8831	18
DataSet-4	3109	12873	10239	18

We choose 3 kinds of meta-paths in the experiment: "$author \overset{write}{\rightarrow} paper \overset{write^{-1}}{\rightarrow} author$", "$author \overset{write}{\rightarrow} paper \overset{publish^{-1}}{\rightarrow} conference \overset{publish}{\rightarrow} paper \overset{write^{-1}}{\rightarrow} author$", "$author \overset{write}{\rightarrow} paper \overset{mention}{\rightarrow} term \overset{mention^{-1}}{\rightarrow} paper \overset{write^{-1}}{\rightarrow} author$". The semantic meaning of these three meta paths is showed in table 2. Note that, "$A - P - A$" is short for "$author \overset{write}{\rightarrow} paper \overset{write^{-1}}{\rightarrow} author$", "$A - P - C - P - A$" is short for "$author \overset{write}{\rightarrow} paper \overset{publish^{-1}}{\rightarrow} conference \overset{publish}{\rightarrow} paper \overset{write^{-1}}{\rightarrow} author$", "$A - P - T - P - A$" is short for "$author \overset{write}{\rightarrow} paper \overset{mention}{\rightarrow} term \overset{mention^{-1}}{\rightarrow} paper \overset{write^{-1}}{\rightarrow} author$".

Table 2. Summary of the Three Meta Paths

ID	**Meta Path**	**Description**
1	A-P-A	Tow author are co-author relationship
2	A-P-C-P-A	Two author published their paper in the same conference
3	A-P-T-P-A	Two author write papers is mentioned the same term

4.2 Methods for Comparison and Evaluation

Methods for Comparison. Two baselines are used in this paper: The first baseline is the Learning with Local and Global Consistency (LLGC) algorithm proposed in [6], which utilizes link structure to propagate labels to the rest of the network. The second baseline is Weighted-vote Relational Neighbor classifier (wvRN) proposed in [9, 20]. These two algorithms are widely used in network classification. As LLGC and wvRN is designed for homogeneous networks, we use all the three homogeneous networks reduced by the three corresponding relation paths (A-P-A, A-P-C-P-A, A-P-T-P-A) to run each algortihm.

One state-of-the-art method, GNetMine [10], is also used in our experiment. GNetMine proposed by Ming et.al. This method handles the classification problem in heterogeneous information network. For the input, we use three types of link (A-P, P-C, P-T) and three four of object (A, P, C, T). The same as HetPathMine, the pre-labeled information is only distributed for the target type objects.

Evaluation Method. We use accuracy [14] to evaluate the effectiveness of classification. The accuracy [14] measure, which is calculated as the percentage of target objects going to the correct class, is defined as follow:

$$Accuracy = \frac{\sum_{i=1}^{k} a_i}{n}$$

where a_i is the number of data objects classified to its corresponding true class, k is the number of class and n is the number of data objects in the dataset.

4.3 Result and Analysis

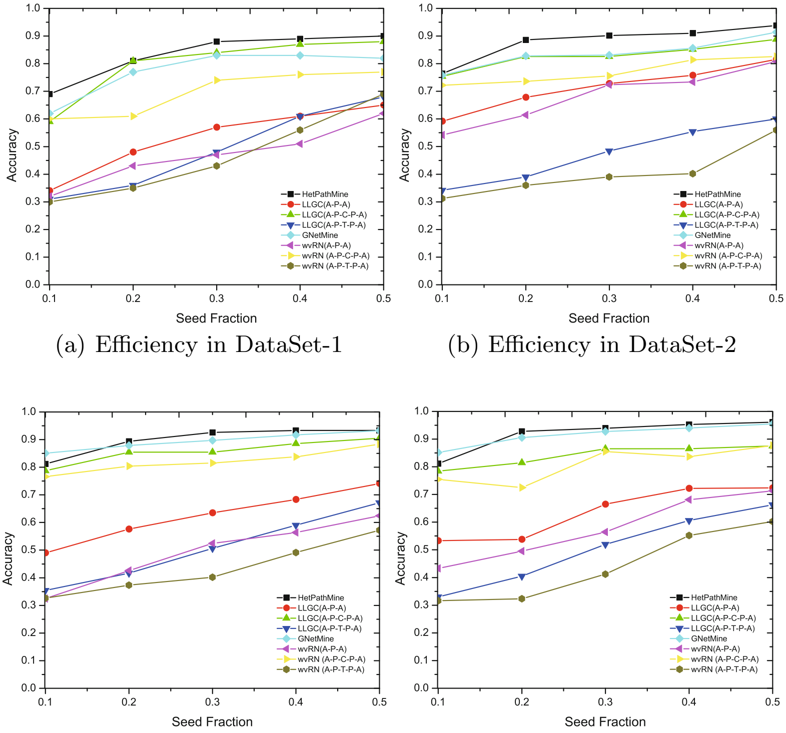
The result is showed in Figure 3. The results obtained by LLGC and wvRN are sharply different by using different meta paths. When using meta path: "A-P-A" or "A-P-T-P-A", the result is very bad. Compared with these two homogeneous algorithm (LLGC and wvRN), HetPathMine can obtain highly accuracy results in different situations. From this point of view, HetPathMine is an efficient method for classification.

On the other hand, HetPathMine can obtain more accuracy results in different dataset compared with the state-of-the-art algorithm, GNetMine. We can see from the result, in each dataset, HetPathMine keeps a high accuracy and has a nearly 5% improvement in accuracy compared with GNetMine.

Overall, HetPathMine performs the best on all datasets with different seed fractions. It is more stable than LLGC and wvRN and more effectiveness than GNetMine. The result showed in Figure 3 also demonstrates that by mining heterogeneous network, one can get more meaningful result than the homogeneous ones.

4.4 Case Study on Meta-path Selection

In this section, we study the learned weights for each meta path by HetPathMine. The rank for the three selected meta paths used in our test is showed in Table.3.



(a) Efficiency in DataSet-1 (b) Efficiency in DataSet-2

(c) Efficiency in DataSet-3 (d) Efficiency in DataSet-4

Fig. 3. Performances under different seed fractions (0.1, 0.2, 0.3, 0.4 and 0.5) in each class are tested. Seed fraction determines the percentage of labeled information. If the seed fraction is 0.1, and the number of object is 1000, then there are 100 objects labeled in the dataset. For each given seed fraction, we average the results over 10 different selections. The classification result is showed in Figure. In our experiment, we set $\mu = 0.01$ in the meta path selection model, and set $\lambda = 2$ for the transductive classification model.

For the ease of illustrate, the weights for each meta path is mapped to the range $[0, 1]$. The meta path "$author-paper-conference-paper-author$" always has a highest weight compared with the other two meta paths. This is consistent with the human intuition: different conference have different interesting topic. Such as ICML is mainly for machine learning and ECIR is mainly for information retrieval. Authors intend to submit their papers to the conference which has their interesting topics. The meta path "$author - paper - author$" also has a high weight in HetPathMine too. This means that two authors are very likely to have similar research interest if they have co-author relationship. The meta path "$author-paper-term-paper-author$" has the lowest weight in HetPathMine. This is consistent with real world scenarios: it is not rare that two papers from different areas can have the same term. For example, the words "algorithm" and "method" appear in many papers from different areas.

Table 3. Meta Paths Weight Comparison

Rank	Meta Path	Weight
1	Author-Paper-Author	$0.5 \sim 0.6$
2	Author-Paper-Conference-Paper-Author	$0.3 \sim 0.4$
3	Author-Paper-Term-Paper-Author	$0.1 \sim 0.2$

5 Conclusion

In this paper, we study the transductive classification problem in heterogeneous information network and propose a novel algorithm, HetPathMine. Different from the transductive classification method of homogeneous network, HetPathMine utilize the relation path information of the network by introduce the concept meta path. On the other hand, different from GNetMine, HetPathMine have distinct advantages in managing the classification problem that each type of object has different classification criteria. The experimental result demonstrates the effectiveness of HetPathMine: (1) HetPathMine can achieve good accuracy in comparison with the existing methods, and (2) the weight obtained by HetPathMine for each meta path is consistent with human intuition or real-world situations. The experimental result also showed that by mining heterogeneous information network, more meaningful result could be obtained.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61373051; the National Science and Technology Pillar Program (Grant No. 2013BAH07F05), Jilin Province Science and Technology Development Program (Grant No. 20111020); Project of Science and Technology Innovation Platform of Computing and Software Science (985 engineering), and the Key Laboratory for Symbolic Computation and Knowledge Engineering, Ministry of Education, China.

References

1. Sun, Y., Han, J.: Mining heterogeneous information networks: principles and methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery 3(2), 1–159 (2012)
2. Gao, J., Liang, F.E.: On community outliers and their efficient detection in information networks. In: KDD 2010, pp. 813–822. ACM (2010)
3. Taskar, B., Abbeel, P., Koller, D.: Discriminative probabilistic models for relational data. In: UAI 2002, pp. 485–492. Morgan Kaufmann Publishers Inc. (2002)
4. Castells, M.: The rise of the network society: The information age: Economy, society, and culture, vol. 1 (2011), `Wiley.com`
5. Even, S.: Graph algorithms. Cambridge University Press (2011)
6. Zhou, D., Bousquet, O.E.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, vol. 16(16), pp. 321–328 (2004)

7. Wu, M., Schölkopf, B.: Transductive classification via local learning regularization. In: AISTATS 2007, pp. 628–635 (2007)
8. Sun, Y., Han, J.E.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: EDBT 2009, pp. 565–576. ACM (2009)
9. Macskassy, S.A., Provost, F.: A simple relational classifier. Technical report, DTIC Document (2003)
10. Ji, M., Sun, Y., Danilevsky, M., Han, J., Gao, J.: Graph regularized transductive classification on heterogeneous information networks. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part I. LNCS, vol. 6321, pp. 570–586. Springer, Heidelberg (2010)
11. Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T.: Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: VLDB 2011 (2011)
12. Getoor, L., Taskar, B.: Introduction to statistical relational learning. The MIT Press (2007)
13. La Fond, T., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: WWW 2010, pp. 601–610. ACM (2010)
14. Sun, Y., Norick, B., Han, J., Yan, X., Yu, P.S., Yu, X.: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In: KDD 2012, pp. 1348–1356. ACM (2012)
15. Cai, D., Shao, Z., He, X., Yan, X., Han, J.: Mining hidden community in heterogeneous social networks. In: LinkKDD 2005, pp. 58–65. ACM (2005)
16. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis, vol. 821. Wiley (2012)
17. Mintz, M.E.: Distant supervision for relation extraction without labeled data. In: ACL 2009, pp. 1003–1011. Association for Computational Linguistics (2009)
18. Nguyen, T.V.T., Moschitti, A., Riccardi, G.: Convolution kernels on constituent, dependency and sequential structures for relation extraction. In: EMNLP 2009, pp. 1378–1387. Association for Computational Linguistics (2009)
19. Sun, Y.E.: Co-author relationship prediction in heterogeneous bibliographic networks. In: ASONAM 2011, pp. 121–128. IEEE (2011)
20. Macskassy, S.A., Provost, F.: Classification in networked data: A toolkit and a univariate case study. The Journal of Machine Learning Research 8, 935–983 (2007)