双类型异质网中基于排序和聚类的离群点检测方法

彭 涛[1,2],杨妮亚[1],徐原博[1],王冰冰[1],刘 露[1]
(1. 吉林大学计算机科学与技术学院,吉林长春 130012;
2. 符号计算与知识工程教育部重点实验室(吉林大学),吉林长春 130012)

摘 要: 挖掘隐藏在网络中不同于正常数据对象的离群点是数据挖掘的重要任务之一. 目前,针对双类型异质信息网络离群点检测的研究工作相对较少,原本适用于同质网络的离群点检测方法将很难适用于双类型异质网络. 为此,提出了异质信息网络中基于排序和聚类的离群点检测方法(RKBOutlier). 从异质信息网络中抽取两种类型的对象以及链接两种对象的语义信息,将待检测的数据作为属性对象,将另一类型数据作为目标对象,对目标对象进行聚类来检测属性对象在各个聚类中的分布情况,数据分布异常的对象即为离群点. 将排序和聚类相结合来显著提高聚类的准确度. 实验结果表明,RKBOutlier 可以在双类型异质信息网络中有效地检测出离群点.

关键词: 离群点检测; 排序; 聚类; 目标对象; 属性对象
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)02-0281-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.02.004

An Outlier Detection Method Based on Ranking and Clustering in Bi-typed Heterogeneous Network

PENG Tao[1,2], YANG Ni-ya[1], XU Yuan-bo[1], WANG Bing-bing[1], LIU Lu[1]
(1. *College of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
2. *Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun, Jilin* 130012, *China*)

Abstract: Mining the outliers that are different from normal data objects in the network is one of the important tasks in data mining. At present, the research aiming at outlier detection in bi-typed heterogeneous information network is relatively small. The methods which are applicable to homogeneous network can not be applied to bi-typed heterogeneous networks. Therefore, we propose a Rank-Kmeans Based Outlier detection method, called RKBOutlier, in heterogeneous information network. The two kinds of the objects and the connected semantic information are extracted from the heterogeneous information network. One type of the objects is regarded as the attribute objects, another type of the objects is regarded as the target objects. We perform cluster partitioning on target objects to detect the distribution of the attribute objects in each cluster. The objects which are abnormal at data distribution are considered to be the outliers. Ranking and clustering are combined to significantly improve the accuracy of clustering. The experimental results show that RKBOutlier can effectively detect outliers in bi-typed heterogeneous information network.

Key words: outlier detection; ranking; clustering; target object; attribute object

1 引言

异质信息网络挖掘是数据挖掘的一类重要问题. 识别和分析网络中有趣而稀少的模式具有很重要的现实意义. 离群点检测旨在发现大量数据集合中一部分离群的数据,其分布的规律不同于正常的数据[1]. 离群点不同于噪声,离群数据本身携带着重要的信息,可以用于数据分析. 离群点检测在日常生活中有着广泛的应用,如网络入侵[2]、医疗诊断[3]、金融诈骗[4]等等.

在异质信息网络中,不同类型的节点和链接关系

收稿日期:2016-09-26;修回日期:2016-11-14;责任编辑:蓝红杰
基金项目:国家自然科学基金(No. 60903098);吉林大学研究生创新基金(No. 2016183, No. 2016184)

包含着丰富的语义信息,使得离群点检测过程变得更加复杂.例如,文献信息网络包含会议/期刊、作者、论文等多种类型的节点以及它们之间的多种链接关系.传统的异质信息网络离群点检测是通过分析整个网络中各个对象之间的链接关系进而找出分布异常的数据对象,这使得离群点检测变得十分复杂并且效率较低.受到以上想法的启发,本文提出一种双类型异质信息网络离群点检测方法.该方法主要解决 3 个问题.第 1 个问题是如何在简化异质信息网络的同时不丢失对象之间的语义信息.第 2 个问题是如何表示双类型网络中的数据对象.第 3 个问题是如何准确地找出异质网络中分布异常的数据对象.以异质信息网络经典的数据集 DBLP 为例,数据集中存在一些作者,他们所发表论文的领域不专一.发生这种情况,极可能是他所发表的论文中,存在挂名作者的情况.在本文中,我们将这样的作者视为分布异常的数据对象.针对上述问题,我们在异质信息网络中提取 2 种类型的数据对象以及它们之间的链接关系.在双类型异质信息网络中,将对象分为属性对象和目标对象 2 种类型的对象.属性对象作为目标对象的特征,属性对象与目标对象之间的链接权值作为特征值,构造邻接矩阵.这样我们在检测异质网络离群点的过程中,既不丢失语义信息也提高了整体的效率.由于排序算法对聚类有一定程度的促进作用,本文提出基于排序的聚类方法 Rank_Kmeans,在目标对象簇的初始化划分中引入排序算法,在提高算法准确率的同时,也提高了效率.在检测离群点的过程中,我们定义了双类型异质信息网络中离群属性的概念,检测属性对象在目标对象聚类领域的分布情况,给出判断属性对象是否离群的标准,找出分布异常的属性对象,即为离群点.

本文的主要贡献如下:(1)我们提出了一种双类型离群点检测方法 RKBOutlier,通过提取异质网络中 2 种类型的数据对象并分析属性对象的数据分布情况来检测离群点.(2)提出排序和聚类相结合的方法并应用到双类型网络离群点检测中来提高离群点检测的效率.(3)将属性对象作为目标对象的特征表示,对目标对象进行聚类,通过分析属性对象的数据分布来检测离群点.(4)我们定义了双类型异质信息网络离群因子的概念,来判断属性对象的离群程度.(5)通过在不同数据集中进行实验,结果表明,我们提出的双类型异质信息网络离群点检测方法可以有效地进行离群点检测.

2 相关工作

离群点检测是数据挖掘中一个关键的问题,在同质信息网络中,研究者们做了很多深入的研究. Gunisetti[5] 使用统计方法比较每一个位置和它的邻居找出异常的数据. Aktolga 等人在文献[6]中将离群点检测技术用在信息检索中. Zimek 等人[7] 研究二次抽样技术在离群点探测器之间诱导多样性.文献[8]中做出了新的改进,使用数据扰动作为新技术在个体离群点探测器以及等级积累方法结合个人的离群值排名构造一个离群点检测上诱导多样性.文献[9]给出了一个使用位置敏感哈希(激光冲徊化)技术基于距离的孤立点检测的近似算法.江峰等人[10] 将粗糙集中边界的概念与 Knorr 等所提出的雨季距离的离群点检测方法结合在一起,提出一种新的离群点检测方法 BDOD.随后,江峰等人[11] 又提出了 RSOD 算法,该算法利用粗糙集,理论中的知识熵和属性等概念构建三种类型的序列,通过分析序列中元素的变化情况来检测离群点.

由于网络中充斥着许多不同类型的对象和链接关系,异质信息网络中的离群点检测问题逐渐引起了研究者们的关注. Ayushi 等人[12] 针对异质信息网络离群点检测问题,提出长方体异常值检测图. Manish 等人[13] 提出一个集成优化框架,通过快照和紧密耦合方式进化的离群值识别构建 outlier-aware 社区匹配找出分布异常的数据.除此以外,很多研究者们还可以在聚类的基础上做离群点检测的研究.例如 Basu 等人[14] 提出了 K-means 聚类方法,该方法是经典的基于距离的聚类方法,两个对象的距离越近,在一个簇中的可能性越大.在文献[15]中,Van 等人分析全脑覆盖的静息状态数据组,提出了基于体素的标准化切割图聚类方法. Sun 等人[16] 提出了 NetClus 聚类方法.将异质网络中基于排序的聚类方法应用到星型网络中. Zhuang 等人[17] 提出了在异质网中根据离群性排序子网中节点的离群点检测方法 BMSim,首先在各个子网中排序网络中的节点查找出离群点,进而分析整个网络中的离群点. Gupta 等人[18] 提出了 CDOutliers 算法,该方法基于联合非负矩阵分解进行离群点检测. Qi 等人[19] 在输出聚类的社会媒体网络中设计了一个和网络内容和链接结构相一致的概率测量模型,并且在真实的社交网络中测试展示了优势.

研究者们在同质网络和多类型异质网络上做了很多的离群点检测方面的研究,但关于双类型异质网络离群点检测方法的研究还很少.在很多情况下,与多类型网络相比较来说,双类型网络在模型表示和计算的过程中相对简单.针对上述情况,本文针对双类型异质信息网络提出了 RKBOutlier 离群点检测方法.

3 问题定义

在本节中,先给出几个相关的定义,然后给出双类型异质信息网络中离群点检测的形式化定义.

定义 1(异质信息网络)[20] 给定一个有向图 $G=$

$\langle V,E\rangle$. 其中 V 是节点的集合,E 是边的集合. 存在一个节点类型映射:$V\rightarrow A$ 和一个边类型映射:$E\rightarrow R$. 如果网络中节点类型数 $|A|>1$ 或者边类型数 $|R|>1$,这个信息网络被称为异质信息网络. 反之为同质信息网络.

定义 2(双类型异质信息网络)[20]　给定一个异质信息网络 $G=\{(X\cup Y),W\}$. X 和 Y 代表两种类型的对象集合,W 代表对象之间链接关系的集合,其中,$\boldsymbol{W}_{XY}(i,j)=p_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$. 对于任意 $\boldsymbol{x}_i\in X$,存在 $\boldsymbol{x}_i=(p_{i1},p_{i2},\cdots,p_{in})$.

以 DBLP 数据集为例,两种节点类型分别为会议和作者,将会议作为目标对象. 因此,一个会议可以表示为向量 $\boldsymbol{x}_i=(p_{i1},p_{i2},\cdots,p_{in})$,其中 p_{ij} 表示在会议 i 上作者 j 发表的论文数. 如图 1 所示,会议 x_1 表示为 $x_1=(2,2,0,0)$,会议 x_2 表示为 $x_2=(0,1,2,0)$,会议 x_3 表示为 $x_3=(0,0,3,1)$. 将每个作者在这个会议上发表论文的数作为特征值,利用余弦相似度量计算两个会议的相似程度. 则 x_1 和 x_2 的相似程度为:

$$sim(x_1,x_2)=\frac{2\times0+2\times1+0\times2+0\times0}{\sqrt{2^2+0^2}+\sqrt{2^2+1^2}+\sqrt{2^2+0^2}+\sqrt{0^2+0^2}}=0.321.$$

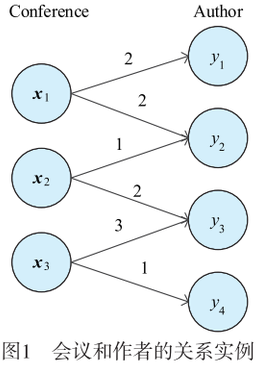

图1　会议和作者的关系实例

定义 3(目标对象的排名分数)[20]　双类型信息网络 $G=\{(X\cup Y),W\}$,对于 $x_i\in X$,则 x_i 的排名分数分布为:

$$r_X(x)=\frac{\sum_{j=1}^{n}\boldsymbol{W}_{XY}(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{W}_{XY}(i,j)}\tag{1}$$

问题 1(双类型异质网络中的离群属性)　给定一个双类型异质信息网络 $G=\{(X\cup Y),W\}$. 其中,目标对象集合表示为 $X=(x_1,x_2,\cdots,x_m)$,属性对象集合表示为 $Y=(y_1,y_2,\cdots,y_n)$. 对于任意目标对象 $x_i\in X$,x_i 关于属性对象 y_j 的值记为 p_{ij}. 因此存在 $x_i=(p_{i1},p_{i2},\cdots,p_{in})$,本文通过分析计算第 j 个属性对象 y_j 在目标对象聚类领域的分布,找出分布异常的属性对象.

例如在文献信息网络中,我们提取出会议集合 X,作者集合 Y. 在会议 x_i 上作者 y_j 发表的论文数量为 $p_{ij}(i=1,2,\cdots,m;j=1,2,\cdots,n)$,构造目标对象和属性对象的邻接矩阵. 会议集合 X 按照领域划分为 K 个簇,检测出在各个领域中分布不唯一的作者.

4　基于排序和聚类的异质信息网络离群点检测方法

在本节中,我们提出了 RKBOutlier 方法,用于在双类型异质信息网络中检测领域异常的属性对象.

4.1　Rank_Kmeans 聚类过程

为了提高聚类的准确度,我们在对目标对象聚类前对其进行排序. 将排名分数按照从小到大的顺序排列,形成了一段呈上升趋势的折线,可以称其为排名分数折线.

给定排名分数折线上相邻两个对象 i 和 j,其中 $i,j\in X$ 且 $r_X(i)\leqslant r_X(j)$,对象 i 和对象 j 的排名分数之差为对象 i 和 j 的斜率,记为 s_i. 在排名分数折线上,如果存在相邻对象 i 和 j,使得 $s_i\neq s_j$,则对象 i 是该排名分数折线上的一个折点. 依次抽取排名分数折线上所有折点,记为 $T=\{t_0,t_1,\cdots,t_u,\cdots t_n\}$,其中 $0\leqslant u\leqslant n,0\leqslant n\leqslant m-1$ 且 $\forall t_u\in X$. 我们将目标对象聚类成 K 个簇. 那么:

(1)当 $n\geqslant K-1$ 时,选择集合 T 中对象 $t_u(1\leqslant u\leqslant n)$ 所对应的斜率 s_{t_u} 最大的 $K-1$ 个折点作为分割点,将对象分成 K 个簇.

(2)当 $0<n<K-1$ 时,除了集合 T 中的所有 N 个对象均作为分割点外,在集合 $X-T$ 的 $m-n$ 个对象中选择 $K-1-n$ 个 $s_i(i\in X)$ 最大的对象作为分割点.

(3)当 $n=0$ 时,此时排名分数折线上没有折点,即所有对象的排名分数相同. 这时,只需要将对象平均分成 K 个簇.

遵循以上步骤将目标对象初始化分成 K 个簇,选取聚类内部中间排名的对象作为聚类簇的中心 $o_i(i=1,2,\cdots,K)$.

目标对象初始化聚类后,余弦相似性作为度量标准调整聚类. 在双类型信息网络中,属性对象作为目标对象的特征表示. 其中,每个属性对象与目标对象的链接权值作为目标对象的特征值.

定义 4(目标对象间的余弦相似度)[21]　给定一个双类型信息网络 $G=\{(X\cup Y),W\}$,对于任意 $x\in X_k$,其中 $1\leqslant k\leqslant K$. o_k 是聚类 X_k 的中心对象,则 x 与 o_k 的余弦相似度的计算方法如下:

$$sim(\boldsymbol{o}_k,\boldsymbol{x})=\frac{\boldsymbol{o}_k,\boldsymbol{x}}{\|\boldsymbol{o}_k\|\|\boldsymbol{x}\|}\tag{2}$$

通过对每个聚类中目标对象与中心对象的余弦相似度的计算,以属性对象为特征,属性对象与目标对象之间的链接权值作为特征值对聚类中的对象进一步划

分. 如果目标对象与所在聚类的中心对象之间的余弦相似度小于给定的阈值 α,那么,我们将这个对象调整到与其相似度最大的聚类中. 根据对象与聚类中心的相似程度来调整每个簇内对象的分布. 反复用每个簇内其他对象代替簇的中心点,使得误差平方和 E[22] 最小,进而得到 K 个聚类的最佳中心点. 误差平方和 E 的计算公式如下:

$$E = \sum_{k=1}^{K} \sum_{x \in X_k} sim(\boldsymbol{x}, \boldsymbol{o}_k)^2 \tag{3}$$

4.2 基于 Rank_Kmeans 的离群点检测方法

在双类型异质信息网络 $G = \{(X \cup Y), W\}$ 中,X 为目标对象,Y 为属性对象. 在 4.1 节中,我们通过计算余弦相似度得到目标对象的 K 个聚类. 在本节中,我们将找出在目标对象聚类领域分布异常的属性对象.

对于任意属性对象 $y_j \in Y(1 \leqslant j \leqslant n)$,对应于目标对象 $x_i \in X(1 \leqslant i \leqslant m)$,存在唯一的向量表示 $\boldsymbol{y}_j = (p_{1j}, p_{2j}, \cdots, p_{mj})$. 以 DBLP 数据集为例,会议作为目标对象,作者作为属性对象. 利用会议之间的相似性得到会议的 K 个聚类后. 一般来说,每个作者的研究领域是专一的. 那么,分布正常的作者发表的论文会集中在一个聚类上. 在本文中,我们的目的是找出在目标聚类领域分布异常的属性对象. 为了分析属性对象在每个目标对象聚类领域的分布情况,属性对象 y_j 的归一化向量 $\boldsymbol{y}_J$ 的计算方法如下:

$$\boldsymbol{y}_J = (p_{1j}, p_{2j}, \cdots, p_{mj}),\ j = 1, 2, \cdots, n \tag{4}$$

其中,p_{ij} 的计算公式如下:

$$p_{ij} = \frac{p_{ij}}{\sum_{i=1}^{i=m} (p_{ij})} \tag{5}$$

数据分布方差的大小反应数据分布的集中程度,对于归一化后的任意 $y_j \in Y$,期望 $\bar{y}_J$ 和方差 S_{y_j} 的计算方法如下:

$$\bar{y}_J = \frac{\sum_{i=1}^{i=m} (p_{ij})}{m} \tag{6}$$

$$S_{y_j} = \sum_{i=1}^{i=m} (p_{ij} - \bar{y}_J)^2 \tag{7}$$

在本文中,我们将离群因子的概念引入到双类型异质信息网络中,离群因子反映了每个属性对象偏离整体数据的程度. 对于每个属性对象 y_j,离群因子 ε_{y_j} 的计算方法如下:

$$\varepsilon_{y_j} = \frac{S_{y_j}}{\sum_{j=1}^{j=n} S_{y_j}} \tag{8}$$

因为,分布异常的属性对象在目标对象聚类领域得到的向量的方差会比较大. 所以,若属性对象的离群因子 ε_{y_j} 大于给定阈值 φ,则该属性对象被视为离群点,反之则是正常点. 我们把异质信息网络中提取出来的两种类型对象以矩阵的形式体现,矩阵中的每一个值代表目标对象和属性对象之间的链接权重. 以属性对象为特征,为了提高目标对象划分的精度,排序与聚类相结合,得到目标对象的领域划分. 依次分析计算每一个属性对象得到对应的离群因子,结合给出的离群点判别标准,若属性对象在目标对象聚类领域的分布不唯一,则该属性对象被视为离群点. 基于排序和聚类的离群点检测算法描述如算法 1 所示.

算法 1 离群点检测算法 RKBOutlier

输入:双类型网络 $G = \{(X \cup Y), W\}$,目标对象 $X = (x_1, x_2, \cdots, x_m)$,属性对象 $Y = (y_1, y_2, \cdots, y_n)$,$X$ 与 Y 之间的链接关系 $\boldsymbol{W}_{XY}(i,j) = p_{ij}(i = 1, 2, \cdots, m; j = 1, 2, \cdots, n)$,余弦相似度阈值 α,离群因子阈值 φ

输出:离群点集合 *RKBOutliers*

1. 计算 $\boldsymbol{r}_X(x_i) = \frac{\sum_{j=1}^{n} p_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}}$. //$\boldsymbol{r}_X(x_i)$ 是在期刊 x_i 上发表的文章占总体文章的比值. $\boldsymbol{x}_i = (p_{i1}, p_{i2}, \cdots, p_{in})(i = 1, 2, \cdots, m)$.
2. 以升序的方式对 $\boldsymbol{r}_X(x_i)$ 进行排序.
3. 得到初始的聚类 $X_k(i = 1, 2, \cdots, m)$.
4. for each $x_i \in X_k$
5. for $h = 1$ to K
6. 计算 o_h 值//o_h 是第 h 个簇的中心
7. 计算 $sim(\boldsymbol{o}_h, \boldsymbol{x}_i)$
8. end for
9. if $sim(\boldsymbol{o}_k, \boldsymbol{x}_i) < \alpha$ then
10. 找到最大的 $sim(\boldsymbol{o}_h, \boldsymbol{x}_i)$ 值并且记录对应的簇 f
11. end if
12. end for
13. $X_f \leftarrow X_f \cup x_i$
14. for each $x_i \in X$
15. $p_{ij} = \frac{p_{ij}}{\sum_{i=1}^{m} p_{ij}}$
16. $y_j = (p_{1j}, p_{2j}, \cdots, p_{mj})$
17. end for
18. for each $y_j \in Y$
19. $\bar{y}_J = \frac{\sum_{i=1}^{m} p_{ij}}{m}$
20. $S_{y_i} = \sum_{i=1}^{m} (p_{ij} - \bar{y}_J)^2$
21. $\varepsilon_j = \frac{S_{y_j}}{\sum_{j=1}^{n} S_{y_i}}$
22. if $\varepsilon_j < \varphi$ then
23. $RKBOutliers \leftarrow RKBOutliers \cup y_j$
24. end if
25. end for

5　实验与结果

在这一节中,我们提出了基于排序和聚类的异质信息网络离群点检测算法,并在 2 个数据集上进行测试.

5.1　度量标准

我们将采用两个常用的度量标准(准确率[23]和召回率[24])来评估基于排序和聚类的离群点检测算法性能. 我们假定 I 是检测出的离群点集合,J 是存在于各个聚类中所有离群点的集合. 因此,准确率和召回率的计算方法如下:

$$precision = \frac{|I \cap J|}{|I|} \times 100\% \tag{9}$$

$$recall = \frac{|I \cap J|}{|J|} \times 100\% \tag{10}$$

因为这两个指标不成正相关,我们采用 *precision* 和 *recall* 的平均值 *F-Measure*[25]来作为评估标准,定义 *F-Measure* 值如下:

$$F\text{-}Measure = \frac{(\beta^2 + 1)\, precision \times recall}{\beta^2 \times precision + recall} \tag{11}$$

其中 β 是一个反应准确率和召回率相对重要程度的权值,若 β 大于 1,则召回率的重要性大于准确率,反之亦然. 在本文中将 β 值设为常量 1.

5.2　数据集

(1) DBLP[26]数据集是异质网络中常用的实验数据,数据集是以引文的形式存储在 xml 文件中,其中包括会议/期刊、作者、发表时间等等. 我们读取 20000 条数据记录,提取期刊、作者的信息作为实验数据. 将出现在实验中的期刊人工标记它所属的领域. 以读取 20000 条数据为例,出现 22 个期刊,人工标记期刊所属的领域.

(2) AMiner[27]数据集是社会网络挖掘常用的数据集,其中包括作者、会议、主题等信息. 我们主要使用 a-miner-topic-data-pubs. xml 文件中的数据. 对 AMiner 数据集进行了预处理,提取出 6 类目标对象,即涉及到 6 个领域的论文和与之对应的作者.

5.3　实验结果

在本节中,我们通过 5 个实验对本文提出的方法进行评估. 在第 1 个实验中,我们首先对聚类进行评估. 聚类的结果由两个参数决定,分别是聚类的数量 γ 和实验读取数据集的大小 μ. 表 1 列出了随着 γ 和 μ 的取值的变化相应的 *precision* 值.

表 1　不同参数下 rank-Kmeans 算法的准确率

precision	$\mu=10000$	$\mu=12500$	$\mu=15000$	$\mu=17500$	$\mu=20000$	$\mu=22500$	$\mu=25000$
$\gamma=4$	76.92%	56.25%	70.00%	63.64%	68.18%	69.23%	57.14%
$\gamma=5$	84.62%	68.75%	60.00%	72.73%	85.71%	72.73%	61.54%
$\gamma=6$	68.75%	70.00%	84.62%	81.82%	95.45%	85.71%	76.92%
$\gamma=7$	84.61%	71.25%	75.00%	81.25%	87.27%	84.62%	86.92%
$\gamma=8$	72.73%	76.79	80.45%	76.92%	75.74%	80.25%	82.81%

我们使用 rank-Kmeans 方法对读取到的 DBLP 数据集中的目标对象进行聚类. 将聚类结果与目标对象实际所属的聚类做对比,得出 rank-Kmeans 方法的 *precision* 值. 由表 1 可以分析出,当取值为 6,读取的数据集取值为 20000 时,*precision* 值可以达到 95.45%,从而可以证明 rank-Kmeans 方法的有效性. 因为,目标结果的聚类结果对离群对象的检测具有一定的促进作用. 因此,在离群点检测过程中,我们将 μ 设置为 20000,γ 设置为 6.

在第 2 个实验中,我们分别在 DBLP 数据集和 A-Miner 数据集上评估余弦相似度阈值 α 和离群因子阈值 φ 对 RKBOutlier 算法的影响. 如图 2 所示,由于在两个数据集中提取的目标对象和属性对象的数量不同,在 DBLP 数据集中,当 α 取值为 0.8 时,对应的 *F-Measure* 值达到峰值. 在 AMiner 数据集中,当 α 取值为 0.85 时,对应的 *F-Measure* 值达到峰值. 由于峰值右侧对参数 α 要求的严格程度要高于峰值左侧对参数 α 的要求,因此,在峰值的左侧,曲线的走势比较平缓. 而在峰值的右侧,α 的取值接近 1 的时候,*F-Measure* 值急剧下降. 如图 3 所示,在 DBLP 数据集和 AMiner 数据集上,当 φ 取值为 1 时,对应的 *F-Measure* 值达到峰值,并且参数 φ 在 1 的两侧取值时,*F-Measure* 值会急剧下降.

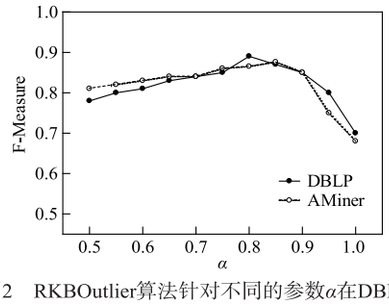

图2　RKBOutlier算法针对不同的参数α在DBLP和AMiner数据集上值的比较

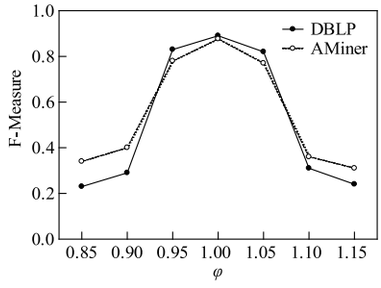

图3 RKBOutlier算法针对不同的参数φ在DBLP和AMiner数据集上值的比较

在第3个实验中,我们对算法受参数γ和μ的影响情况进行分析,即测试算法伸缩性.分别在DBLP和AMiner数据集上进行实验.取数据集$\mu=20000$,余弦相似度的阈值α在DBLP和AMiner数据集上分别取0.8和0.85,离群因子φ为1.实验结果如图4所示.RKBOutlier算法受参数γ影响不大,而且两种数据集上的*F-Measure*值接近于拟合,说明RKBOutlier算法在不同数据集上的泛化性良好.接下来,我们分析参数μ对RKBOutlier算法*F-Measure*值的影响.如图5所示,*F-Measure*值随着μ的变化波动较小.因此,RKBOutlier算法的伸缩性良好.当μ取20000时,*F-Measure*值达到最大.从图4和图5可以看出,rank-Kmeans聚类结果的*precision*值与RKBOutlier算法*F-Measure*值是成正向相关的.当选取了合适的γ值和μ值,rank-Kmeans聚类的*precision*值会对RKBOutlier算法的*F-Measure*值产生一定的促进作用.对于更大规模的数据,我们首先需要对聚类进行准确度实验,得到合适的参数,再进一步做更大数据环境下的离群点检测研究.

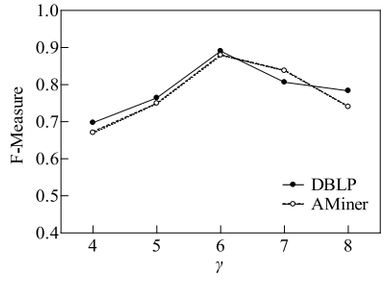

图4 RKBOutlier算法针对不同的参数γ在DBLP和AMiner数据集上值的比较

在第4个实验中,我们分别用五种算法(BM-Sim[17]、CDOutliers[18]、基于K-means[14]、基于Normalized Cut[15]和基于NetClus[16])和我们提出的RKBOutlier算法在两个数据集上进行测试.图6分别表示在DBLP数据集和AMiner数据集上,本文方法和对比离群点检测算法的*F-Measure*值.结果表明RKBOutlier算法明显优于其他五种算法.RKBOutlier算法将目标对象、属性对象以及其链接的语义信息提取出来.通过对目标对象

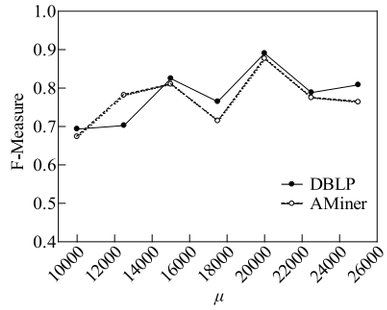

图5 RKBOutlier算法针对不同的参数μ在DBLP和AMiner数据集上的*F-Measures*值

聚类来找出离群的属性对象.在离群点检测的过程中,若属性对象分布在目标对象聚类的数量越多,属性对象的离群因子越大,当大于给定阈值时,该属性对象即被视为离群点.因此,我们能有效识别出二分网络中存在的离群点.而且在聚类前进行排序,减少了迭代的次数,提高了准确率.

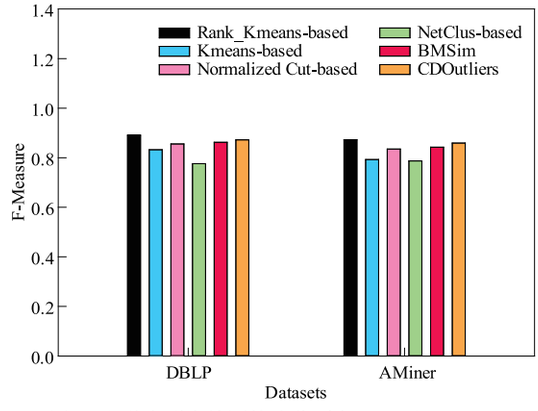

图6 六种离群点检测算法分别在DBLP和AMiner数据集上*F-Measures*值的比较

在第5个实验中,我们对算法的效率进行测试.实验结果如图7所示,基于Rank_Kmeans离群点检测算法的运行时间明显少于基于K-means离群点检测算法的运行时间.RKBOutlier算法的时间开销主要花费在排序、聚类和离群点检测上.在实验中,目标对象的数量为n,聚类的个数为k.因此,排序部分的时间复杂度为$O(n^2)$;聚类部分计算的时间复杂度O(nk);离群点检测部分的时间复杂度为O(n).随着数据量的增加,算法消耗的时间呈线性增长趋势而非指数增长趋势,聚类前引入排序过程.给出了目标对象的初始聚类,加速了聚类的过程,很大程度上减少了迭代次数,不仅提高了聚类的准确率,而且提高了效率.

6 结论

本文针对双类型异质信息网络提出了一种基于排序和聚类的离群点检测方法.从异质信息网络中抽取

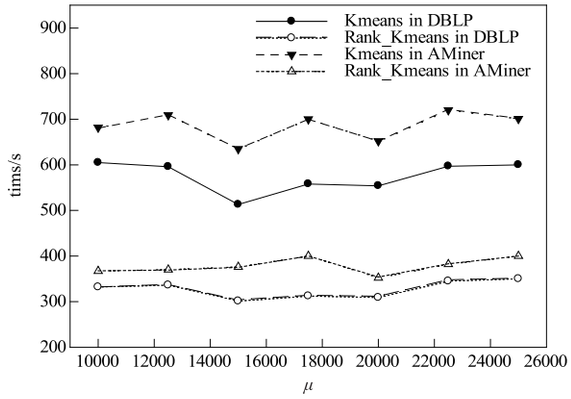

图7　基于K-means和基于Rank_Kmeans离群点方法在DBLP和AMiner数据集上执行时间的比较

目标对象和属性对象,排序与聚类相结合,提高了算法的准确率和效率. 我们引入了双类型异质信息网络离群因子的概念,从离群的属性对象这个新的角度检测异质信息网络中的离群点,找出在目标对象聚类领域分布不唯一的属性对象. 与 BMSim 算法、CDOutliers 算法、基于 K-means、基于 Normalized Cut 和基于 NetClus 离群点检测算法进行对比,并且在 DBLP 和 AMiner 数据集上得到了有效的验证,显示出 RKBOutlier 算法的有效性.

参考文献

[1] Aggarwal C C, Yu P S. Outlier detection for high dimensional data[J]. Acm Sigmod Record, 2001, 30(2): 37 – 46.

[2] Koc L, Mazzuchi T A, Sarkani S. A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier[J]. Expert Systems with Applications, 2012, 39: 13492 – 13500.

[3] Kaganov A, Lakhany A, Chow P. FPGA acceleration of multifactor cdo pricing[J]. ACM Transactions on Reconfigurable Technology and Systems (TRETS), 2011, 4(2): 20 – 25.

[4] Kuklisova-Murgasova M, Quaghebeur G, Rutherford M A, et al. Reconstruction of fetal brain MRI with intensity matching and complete outlier removal[J]. Medical Image Analysis, 2012, 16(8): 1550 – 1564.

[5] Gunisetti L. Outlier detection and visualization of large datasets[A]. Proceedings of the International Conference on Emerging Trends in Technology[C]. New York: ACM, 2011. 522 – 524.

[6] Aktolga E, Ros I, Assogba Y. Detecting outlier sections in us congressional legislation[A]. Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011[C]. New York: ACM, 2011. 235 – 244.

[7] Zimek A, Gaudet M, Campello R J G B, et al. Subsampling for efficient and effective unsupervised outlier detection ensembles[A]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2013. 428 – 436.

[8] Zimek A, Campello R J G B, Sander J. Data perturbation for outlier detection ensembles[A]. Proceedings of the 26th International Conference on Scientific and Statistical Database Management[C]. New York: ACM, 2014. 13: 1 – 12.

[9] Pillutla M R, Raval N, Bansal P, et al. LSH based outlier detection and its application in distributed setting[A]. Proceedings of the 20th ACM International Conference on Information and Knowledge Management[C]. New York: ACM, 2011. 2289 – 2292.

[10] 江峰,杜军威,毕跃飞等. 基于边界和距离的离群点检测[J]. 电子学报,2010,38(3):700 – 705.
Jiang Feng, Du Junwei, Mu Yuefei, et al. Outlier detection based on boundary and distance[J]. Acta Electronica Sinica, 2010, 38(3): 700 – 705. (in Chinese)

[11] 江峰,杜军威,葛艳等. 基于粗糙集理论的序列离群点检测[J]. 电子学报,2011,39(2):345 – 350.
Jiang Feng, Du Junwei, Ge Yan, et al. Sequence outlier detection based on rough set theory[J]. Acta Electronica Sinica, 2011, 39(2): 345 – 350. (in Chinese)

[12] Dalmia A, Gupta M, Varma V. Query-based graph cuboid outlier detection[A]. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015[C]. New York: ACM, 2015. 705 – 712.

[13] Manish G, Gao J, Sun Y Z, et al. Integrating community matching and outlier detection for mining evolutionary community outliers[A]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2012. 859 – 867.

[14] Basu S, Banerjee A, Mooney R J. Semi-supervised clustering by seeding[A]. 18th International Conference on Machine Learning[C]. San Francisco, CA: Morgan Kaufmann, 2002. 27 – 34.

[15] Van d H M, Mandl R, Hulshoff P H. Normalized cut group clustering of resting-state FMRI data[J]. Plos One, 2008, 3(4): e2001.

[16] Sun Y Z, Yu Y, Han J W. Ranking-based clustering of heterogeneous information networks with star network schema[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2009. 797 – 806.

[17] Zhuang H, Zhang J, Brova G, et al. Mining query-based subnetwork outliers in heterogeneous information networks[A]. IEEE International Conference on Data Mining[C]. Piscataway, NJ: IEEE, 2014. 1127 – 1132.
[18] Gupta M, Gao J, Han J. Community distribution outlier detection in heterogeneous information networks [A]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases[C]. Berlin: Springer, 2013. 557 – 573.
[19] Qi G J, Aggarwal C C, Huang T S. On clustering heterogeneous social media objects with outlier links[A]. Proceedings of the 5th ACM International Conference on Web Search and Data Mining [C]. New York: ACM, 2012. 553 – 562.
[20] Sun Y Z, Han J W, Zhao P X, et al. RankClus: Integrating clustering with ranking for heterogeneous information network analysis[A]. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology[C]. New York: ACM, 2009. 565 – 576.
[21] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques[A]. KDD Workshop on Text Mining[C]. Piscataway, NJ: IEEE, 2000. 400 (1): 525 – 526.
[22] Han J W, Kamber M, Pei J. Data Mining Concepts and Techniques[M]. Third Edition, San Francisco, CA: Morgan Kaufmann, 2012. 102 – 120.
[23] Zhang K, Hutter M, Jin H. A new local distance-based outlier detection approach for scattered real-world data [A]. Advances in Knowledge Discovery and Data Mining[C]. Berlin: Springer, 2009. 813 – 822.
[24] Tzeng J Y, Byerley W, Devlin B, et al. Outlier detection and false discovery rates for whole-genome DNA matching[J]. Journal of the American Statistical Association, 2003. 98(461): 236 – 246.
[25] Croft W B, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice[M]. Reading: Addison-Wesley, 2010. 23 – 37.
[26] Ley M. The DBLP computer science bibliography: Evolution, research issues, perspectives[A]. String Proceedings and Information Retrieval[C]. Berlin: Springer, 2002. 1 – 10.
[27] Tang J, Zhang J, Yao L, et al. Arnetminer: Extraction and mining of academic social networks[A]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM, 2008. 990 – 998.

作者简介

彭 涛 男,1977 年生,博士、教授. 主要研究方向为数据挖掘、信息检索、机器学习.

杨妮亚 女,1992 年生,硕士. 主要研究方向为数据挖掘、机器学习.

徐原博 男,1990 年生,博士. 主要研究方向为数据挖掘、机器学习.

王冰冰 女,1990 年生,硕士. 主要研究方向为数据挖掘、搜索引擎.

刘 露(通信作者) 女,1989 年生,博士. 主要研究方向为数据挖掘、异质信息网络挖掘、离群点检测.
E-mail: liulu12@ mails. jlu. edu. cn