DOI:10.7652/xjtuxb201408012

应用相似度测量的图离群点检测方法

李涛[1,2], 肖南峰[2]
(1.华南农业大学现代教育技术中心, 510640, 广州; 2.华南理工大学计算机科学与工程学院, 510640, 广州)

摘要: 针对传统离群点检测方法精确度不高的问题,提出了一种同时基于全局和局部视野综合考虑的离群点检测方法,并将其成功应用于事务图数据集的离群点检测。该方法利用极大公共频繁子图来测量任意两个事务图之间的相似度,提出利用基于公共近邻的裁剪方法对相似矩阵进行裁剪,通过计算数据结点的往返距离得出各个结点的离群值评分,弥补了传统基于稳态分布随机游走的离群点检测方法的缺陷。实验结果表明:该方法在事务图数据离群点检测方面的性能明显优于基于subdue的方法,精确度和错误报警率以及召回率提高了约10%。

关键词: 数据挖掘;离群点检测;相似度;图
中图分类号: TP391 **文献标志码**: A **文章编号**: 0253-987X(2014)08-0067-06

A Graph Anomalies Detection Method Based on Graph Similarity

LI Tao[1,2], XIAO Nanfeng[2]
(1. Modern Education and Technology Center, South China Agricultural University, Guangzhou 510640, China;
2. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China)

Abstract: An anomalies detection method based on considering both the global and the local perspectives is proposed to solve the low accuracy problem in traditional anomalies detection, and is successfully applied to outlier detection in transaction graph data sets. The method evaluates the similarity between any two graphs based on maximum common frequent subgraph, and then cuts out the similarity matrix based on common neighbor. The round-trip distance for a data node is calculated and is used as its anomalies score, that makes up the defect of traditional outlier detection methods based on the steady-state distribution and random walk. Experiments in real datasets show that the performance of the proposed method is better than the performance of the method based on subdue. The precision, the recall rate and the false alarm rate are improved by about 10%.

Keywords: data mining; anomalies detection; similarity; graph

离群点检测又称异常模式检测,是找出数据中某些数据对象的行为不同于预期对象的过程。离群点分为3类:全局离群点、情景离群点和集体离群点[1-2]。图数据的离群点通常归类为集体离群点,集体离群点检测与其他类型离群点检测的关键区别是结构通常不是明确定义的,必须作为离群点检测过程的一部分来发现。目前在离群点挖掘领域研究较多的是对矢量数据的离群点挖掘,对于图数据的离群点挖掘研究进行得不够充分。近期主要研究包括:文献[3]提出了不确定离群点检测方法并可以应用于图数据;文献[4]利用半监督算法从在线社交网络中探测离群点;文献[5]利用矢量数据对象的相似

收稿日期: 2013-12-10。 作者简介: 李涛(1978—),男,博士生;肖南峰(通信作者),男,教授,博士生导师。 基金项目: 国家自然科学基金资助项目(61171141),广东省产学研省部合作专项资金资助项目(2012B091100448)。
网络出版时间: 2014-06-30 网络出版地址: http://www.cnki.net/kcms/detail/61.1069.T.20140630.1837.002.html

图探测离群点;文献[6]在谱图中检测离群点;文献[7]在带权的大图中检测离群结点的框架;文献[8]在子空间属性图中检测离群点;文献[9]扩展谱图聚类在边带权的大图中进行离群子结构检测;文献[10]在社会网络和信息网络中挖掘进化社区离群点;文献[11]提出了一种条件离群点检测算法;文献[12]在图流中检测离群点;文献[13]基于 KNN 图进行时空离群点检测。本文与上述文献的区别在于本文研究对象是事务图集中的图数据,也就是一个图数据集中包含许多规模相对较小的图,要检测的离群点既不是图的单个节点,也不是单条边,而是一个完整的图数据对象。本文同时基于全局和局部的视野来综合考虑单个图像相对数据集全局和局部的差异与不同,从中发现特征异于正常图的离群点图,而不是单独从局部或者全局比较两个图之间的差异。

1 检测方法

定义 1 事务图数据。一个图 P 的结点集合表示为 $V(P)$,边的集合表示为 $E(P)$。对于两个图 P 和 P',如果 P'中存在子图与 P 同构,可表示为 $P\subseteq P'$。P'被称为 P 的超图,P 被称为 P'的子图。在图挖掘中,一个图模式是一个图 P 自身。给定一个图数据的集合 $D=\{G_1,G_2,\cdots,G_n\}$,对于任何图模式 p,p 的支持数据库表示为 D_p,定义为 $D_p=\{G_i\mid p\subseteq G_i,1\leqslant i\leqslant n\}$,一个图模式 p 是频繁的当且仅当对于一个给定的支持度阀值 σ 存在 $|D_p|/|D|\geqslant\sigma$。

本文方法的思想是基于子结构相似性,所以结构异于正常图的图数据对象都可以归结为图离群点,算法流程概述如下。

步骤 1 对图数据集 D 进行频繁子图挖掘,抽取能表示图数据对象的特征矢量 $\boldsymbol{F}$。

步骤 2 挖掘数据集中所有图数据对象两两之间的极大公共频繁子图,以极大公共频繁子图作为两个图数据对象之间的相似度测量标准,计算任意两个图数据对象之间的相似度。

步骤 3 生成一个带权的数据图 G 来表示原始的图数据集,边的权值就是结点之间的相似度。再以邻接矩阵 $\boldsymbol{A}$ 来表示图 G,这个邻接矩阵称为数据图 G 的相似矩阵。

步骤 4 设置权值阀值 T,对相似矩阵 $\boldsymbol{A}$ 执行裁剪,删除一些权值过低的边。然后计算数据图 G 中结点相互之间的公共近邻数,更新图 G 对应的相似矩阵 $\boldsymbol{A}$ 中的元素为公共近邻的基数。

步骤 5 根据图 G 对应的相似矩阵计算结点之间的往返距离,然后计算每个结点的离群值评分。

步骤 6 最后对离群值评分进行排序,顶点的 N 个结点被判断为离群点,它们对应原始事务图数据集中的若干图数据对象。

1.1 图数据的特征值抽取

现有的基于子结构相似性的离群点挖掘技术侧重于局部结构,会导致较高的误报率。本文提出的方法是基于图之间的最大子结构相似性,因此具有更强的适应性。因为一个图数据对象可能是若干个频繁子图组成的,所以抽取频繁子图来代表一个图是常用的方法。采用频繁子图作为特征值,利用频繁子图挖掘算法在预设的支持度阀值条件下生成频繁子图,一旦生成了频繁子图,就给数据集中的每个图构建一个基于频繁子图的特征向量 $\boldsymbol{F}$,这样图数据对象就可以通过一组二元特征矢量来表示。构建特征矢量的方法如下:假设 p_i 是一个图中的频繁子图模式,P 是所有频繁子图的集合,那么每个图 G_i 的特征矢量可以被编码为一个 $|P|$ 维矢量

$$\boldsymbol{x}_{ip}=\begin{cases}1, & \text{if } p\subseteq G_i\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

1.2 图数据的相似度测量

因为属于离群点的图异于正常图,不像正常图那样拥有很多频繁子图成分,所以可以采用极大公共频繁子图识别离群图。我们采用极大公共频繁子图来测量两个图对象之间的相似度。

定义 2 极大公共频繁子图。假设 $G_1,G_2\in D$ 是两个图,P 是事务图集 D 中发现的频繁子图集合。设 $P_{G_1\cap G_2}$ 是 G_1、G_2 同时包含的公共频繁子图集合,极大公共频繁子图 $S_{G_1\cap G_2}\subseteq P_{G_1\cap G_2}$ 满足如下条件:如果存在频繁子图 $p_1\in S_{G_1\cap G_2}$,那么不存在 $p_1\subseteq p_2$ 的频繁子图 $p_2\in S_{G_1\cap G_2}$。

本文算法的关键思想是比较两个图,抽取近似的最大公共频繁子图。采用极大公共频繁子图来测量图之间的相似度比其他基于启发式方法要准确得多。本文算法也适用于存在多重标号的图。抽取极大公共频繁子图算法的思想是:只要频繁子图 S 能够覆盖尽可能最大数量的规模大小是 1 的频繁子图(在图数据库中,规模大小是 1 的频繁子图集合表示为 L_1),那么就可以认为 S 是任意两个给定图之间的极大公共频繁子图。

算法步骤如下:

步骤 1 首先对两个图 G_1 和 G_2 的特征矢量 $\boldsymbol{F}$ 进行交叉运算,可以发现 G_1 和 G_2 都拥有公共频繁子图的集合 C;

步骤 2　选择规模最大的子图 $p\in C$,集合 S 初始状态时只包含 p,此时 p 变成极大公共频繁子图的第一个组成部分;

步骤 3　边覆盖集(set of covering edges, SOCE)设为 E,初始化为包含所有属于 p 的规模为 1 的频繁子图;

步骤 4　从 C 中选择能够覆盖 L-E 中最大数量频繁子图边集的下一个公共频繁子图作为下一个极大公共频繁子图组成部分加入 S 中;

步骤 5　重复步骤 4 直到 L_1 被完全覆盖或者 $|C|=0$,此时 S 中的频繁子图集合就是两个图 G_1 和 G_2 的极大公共频繁子图。

取得极大公共频繁子图后,采用一个启发式算法来计算两个图之间的相似度。

定义 3　基于极大公共频繁子图的相似度。设 $S=\{p_1,p_2,\cdots,p_k\}$ 是组成两个图 G_1 和 G_2 的极大公共频繁子图的 k 个频繁子图组分集合,G_1 和 G_2 之间的相似度 S 通过下式计算

$$S(G_i,G_j)=\begin{cases}0, & \text{if } G_1=G_2\\ \left(\sum_{i=1}^{k}N(p_i)-|C|\right)^2/N(G_i)N(G_j), & \text{otherwise}\end{cases}\tag{2}$$

式中:$N(G_i)$ 表示图 G_i 中结点和边的数量之和;$N(p_i)$ 是第 i 个极大公共频繁子图中结点和边的数量总和;$|C|$ 是频繁子图集合 S 中所有子图的公共结点的数量。

1.3　建立相似矩阵

目前常用的基于距离或者密度的离群点检测算法对于探测小规模离群点聚类效果不佳,显示较高的误报率,采用基于图的方法可以解决这个问题。本文方法的思想是把原始的图数据集中的对象表示为一个带权的无向图 G,称 G 为数据图,其中每个结点表示一个图数据对象,每条边都会被分派一个权值,这个权值等于结点之间的相似度。通过 $n\times n$ 的邻接相似矩阵 $\boldsymbol{A}$ 来表示对应的图 G,其中 n 是图数据对象的个数。相似矩阵中每个元素都表示任意两个图数据对象 G_i 和 G_j 之间的相似度。如果 G_i 和 G_j 之间的相似度大于 0,那么在相似矩阵 $\boldsymbol{A}$ 的 (i,j) 位置上就存在一个非零元素,值为式(2)计算得来的 $S(G_1,G_2)$。通过这种方式计算得来的数据图 G 必然是一个完全图,避免了环的存在,而其对应的邻接相似矩阵 $\boldsymbol{A}$ 除对角线元素为 0 外,其余位置必定都非 0。

1.4　基于公共近邻的相似度裁剪

相似矩阵 $\boldsymbol{A}$ 可能过于稠密,这样会导致包括离群点和正常结点在内的很多数据对象都拥有巨大数量的邻居结点,这样很难把离群点从正常结点中区分开来,因此有必要过滤一些不必要的相似度值,也就是删除一些不必要的边。本文设置一个阈值 T,凡是权值低于阈值 T 的边都会被删除,从而限制邻居结点只是与当前结点具有高度相似值的结点集合。删除低相似度边后,不是与邻居结点具有高度相似值的结点会被孤立为离群点。

阈值 T 太大或者太小都会出现问题,在本文中对阈值 T 的设置考虑数据集的相似度的分布规律,假设 X 是相似度值的集合,μ 是平均值,σ 是 X 的标准偏差。经过抽取少量数据进行训练,会发现阈值 T 在区间 $[\mu-\sigma,\mu)$ 之间会产生较高的预测准确率,也就是说阈值 T 应该被选择不超过均值,但是最小也不能低于均值的一个标准偏差。这种方法的优势是阈值 T 的选择仅仅依赖于数据集本身,而目前其他基于距离和密度的算法都需要用户指定某些参数。

基于公共近邻的相似度裁剪算法步骤如下:

步骤 1　检测相似矩阵的每个元素,如果元素 $A(X,Y)>T$,就设置 $A(X,Y)=1$,否则设置 $A(X,Y)=0$,这个过程实际上是把相似矩阵 $\boldsymbol{A}$ 中的元素离散化;

步骤 2　逐行对相似矩阵中的每对结点计算公共近邻,例如第一行的元素集合是{0,0,1,1,1,0},第二行是{0,0,0,1,1,0},两者取交运算得{0,0,0,1,1,0},取非 0 元素的集合{1,1},这就是两个结点的公共近邻的集合,公共近邻的个数是 2,此时设置 $A(X,Y)=A(Y,X)=2$,然后设置对角线元素 $A(X,Y)=0$;

步骤 3　如果计算发现两个结点之间没有公共近邻,那么设置 $A(X,Y)=A(Y,X)=0$。

上面的裁剪方法如果删除了太多的低权值边,会破坏图 G 的连通性,导致下面对数据图 G 进行离群值评分计算可能无法正常进行。为维持连通性,在裁剪之前首先生成数据图 G 的最小生成树。如果在裁剪数据图 G 后发现它不再连通,就把最小生成树对应位置上的被删除的边重新加到数据图 G 中,从而维持数据图 G 的连通性,但是因为删除了一些边,图 G 在一定程度上变得稀疏,裁剪后的相似矩阵 $\boldsymbol{A}'$ 也变成稀疏矩阵。

1.5　离群值评分计算

传统的基于随机游走的离群值评分存在缺陷。图 G 上的随机游走是通过一个有限 n 态时间齐次

马尔科夫链描述的结点序列。在时间 t 游走到结点 i，然后在 $t+1$ 时间选择结点 j 的概率是由图的边权值决定的，$p_{i,j}=P(s(t+1)=j\mid s(t)=i)=w_{ij}/d_{ii}$，其中 $d_{ii}=\sum_{j\in \mathrm{adj}(i)} w_{ij}$，adj($i$)是结点 i 的邻居，设 $\boldsymbol{W}$ 是条目为 p_{ij} 的过渡矩阵，$\boldsymbol{A}$ 是图 G 邻接相似矩阵，$\boldsymbol{D}$ 是条目为 d_{ii} 的对角矩阵，存在 $\boldsymbol{W}=\boldsymbol{D}^{-1}\boldsymbol{A}$。$\pi_i(t)$ 是时间 t 到达结点 i 的概率，$\pi(t)=[\pi_1(t),\pi_2(t),\cdots,\pi_n(t)]^{\mathrm{T}}$ 是时间 t 的概率状态分布，转移到时间 $t+1$ 的概率是 $\pi(t+1)=\boldsymbol{P}^{\mathrm{T}}\pi(t)$，存在 $\pi(t)=(\boldsymbol{P}^{\mathrm{T}})^t\pi(0)$，其中 $\pi(0)$ 是初始状态分布，对于所有 $t>0$，如果 $\pi(t)=\pi(0)$，分布 $\pi(t)$ 就是稳态分布。如果对一个图 G 进行随机游走，那么最少访问的结点就最可能是一个全局的离群点。

定义 4 离群值评分 给定一个随机游走图 G 和它的过渡矩阵 $\boldsymbol{W}$，稳态分布 π_i 就是结点 i 的离群值评分，其中存在 $\forall i, 1\leqslant i\leqslant n$。

但是，这种方式的离群值评分存在一个缺陷，就是稳态分布只根据全局的随机游走来识别离群点，所有的结点都被无差异地对待，因此存在局部差异的结点可能无法被识别为离群点，这些结点的离群值评分可能很高但是却无法被检测到。为了避免这个缺陷，本文引入往返距离的概念，它既可以考虑全局问题，也可以考虑到局部的差异。对于随机游走存在到达距离 $h(i,j)$ 和往返距离 $c(i,j)$ 两个测量标准。到达时间 $m(i,j)$ 是算法从结点 i 开始，到达结点 j 所需要经历的步长之和，这里步长以过渡矩阵 $\boldsymbol{W}$ 中的元素表示的过渡概率 p_{ij} 表示为

$$h(i,j)=\begin{cases}0, & i=j\\ 1+\sum_{k\in \mathrm{adj}(i)} p_{ik}h(k,j), & \text{其他情况}\end{cases} \tag{3}$$

往返距离 $c(i,j)$ 是算法从结点 i 开始随机游走，到达结点 j，然后返回结点 i 所经历的距离

$$c(i,j)=h(i,j)+h(j,i) \tag{4}$$

显然可以看出，到达距离 $h(i,j)$ 是不对称的，但是往返距离 $c(i,j)$ 却是对称的。往返距离 $c(i,j)$ 存在如下性质：①$c(i,j)\geqslant 0$；②$c(i,j)=0$ 当且仅当 $i=j$ 时；③$c(i,j)=c(j,i)$；④$c(i,j)\leqslant c(i,k)+c(k,j)$。往返距离 $c(i,j)$ 可以通过计算图 G 的拉普拉斯矩阵的伪逆矩阵求得，伪逆矩阵是对不是满矢或者不是方阵的矩阵的逆的推广，图的拉普拉斯矩阵和它的伪逆矩阵都是对称和半正定的。设 $\boldsymbol{L}=\boldsymbol{D}'-\boldsymbol{A}$ 表示图 G 的拉普拉斯矩阵，$\boldsymbol{L}^+$ 表示对应的伪逆矩阵，那么往返距离 $c(i,j)$ 可以表示为

$$c(i,j)=V_G(l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+) \tag{5}$$

式中：$V_G=\sum_{i=1}^{n} d_{ii}$ 是图的容积；l_{ij}^+ 是伪逆矩阵 $\boldsymbol{L}^+$ 在 (i,j) 位置的元素。事实上往返距离是图拉普拉斯矩阵的特征向量空间中的欧几里德距离。下面将利用往返距离来检测离群点。根据图的有关理论，存在如下结论。

结论 1 给定一个聚类 C 和一个在 C 之外的结点 s，s 与聚类 C 边界上的结点 t 连通。如果向聚类 C 中增加一些结点或者边让 C 变得更稠密，那么 s 和 t 之间的往返距离会增大。

根据结论 1，聚类越稠密，在聚类之外的结点和聚类之内的结点之间的往返距离越大，就越容易检测出聚类之外的离群点，所以可以采用往返距离来有效探测局部离群点。因为在同一个聚类内部的结点的往返距离比较短，而不同聚类之间结点的往返距离比较长，所以任意两个结点之间的往返距离可以同时抓取这两个结点之间的距离以及它们的局部邻居结点的密度。因此，采用这种方法可以同时探测局部和全局范围的离群点。最后的问题是如何计算结点的离群值评分。本文采用的方法是计算任意结点 i 到图中其他所有结点的往返距离的平均值，这个平均值就作为结点 i 的离群值评分。算法步骤如下：

步骤 1 从原始图的数据矩阵构建一个完全的相似矩阵 $\boldsymbol{A}$，然后利用公共近邻来裁剪相似矩阵 $\boldsymbol{A}$，获得裁剪后的相似矩阵 $\boldsymbol{A}'$；

步骤 2 计算相似矩阵 $\boldsymbol{A}'$ 表示的图拉普拉斯矩阵 $\boldsymbol{L}$ 和它对应的伪逆矩阵 $\boldsymbol{L}^+$；

步骤 3 根据伪逆矩阵 $\boldsymbol{L}^+$ 利用式(5)计算任意两个结点之间的往返距离 $c(i,j)$；

步骤 4 计算每个结点 i 到其他所有结点往返距离的平均值 $\sum_{j=1}^{k} c(i,j)/k$，作为结点 i 的离群值评分 Score(i)；

步骤 5 对所有结点的离群值评分 Score(i)排序，采取最大似然估计计算最顶端的 N 个数据结点并识别为离群点，这里的数据结点对应原始图数据集中的单个图数据对象。

2 实验分析

本文采用数据挖掘中的几个常用测量标准来评估实验：正确肯定(T_{P})，即实际的响应中被正确预测为响应的个数；正确否定(T_{N})，即实际的非响应中被

正确预测为非响应的个数;错误肯定(F_P),即实际的非响应中被错误预测为响应的个数;错误否定(F_N),即实际的响应中被错误预测为非响应的个数。

精确度设为　$P=T_P/T_P+F_P$　(6)

错误报警率设为　$F=F_P/F_P+T_N$　(7)

召回率设为　$R=T_P/T_P+F_N$　(8)

本文采用 3 个不同真实数据集来评估算法。第 1 个数据集 PTC,是化学毒物预测数据集,按照实验动物的类别该数据集又包含 4 类数据:雄性老鼠(MM)、雌性老鼠(FM)、雄性兔子(MR)、雌性兔子(FR)。第 2 个数据集 DTP Aids,是艾滋病化学成分预测数据集。第 3 个数据集 DTP Cancer,是化学成分对癌症的治疗功能预测数据集。为加快算法运行,我们从中随机选择一些图作为测试对象。数据集的属性见表 1。图离群点也是从这些数据集中随机选择。采用 gSpsn 频繁子图挖掘算法[14]来生成频繁子图,每个数据集采用的最小支持度见表 1。

表 1　数据集属性

数据集	实例数	离群点数	最小支持度
PTC_MM	220	17	0.05
PTC_MR	215	22	0.05
PTC_FM	219	16	0.05
PTC_FR	240	18	0.05
DTP Aids	900	45	0.10
DTP Cancer	7 500	380	0.10

表 2 给出了本文算法与文献[15]基于 subdue 的图离群点检测算法的平均精确度 P 和平均错误报警率 F 的比较。

表 2　实验结果

数据集	P		F	
	本文算法	文献[15]算法	本文算法	文献[15]算法
PTC_MM	0.330	0.270	0.059	0.060
PTC_MR	0.310	0.180	0.079	0.080
PTC_FM	0.480	0.180	0.060	0.069
PTC_FR	0.180	0.180	0.082	0.082
DTP Aids	0.340	0.120	0.039	0.045
DTP Cancer	0.450	0.120	0.041	0.045

一般来说精确度越高越好,错误报警率越低越好。从表 2 可以看到,本文算法比文献[15]算法表现出更高的精确度。只有在 PTC_FR 数据集是个例外,两者的精确度相同,错误报警率也相同,这可能是 PTC_FR 数据集子结构相似度差异大使文献[15]算法表现出了比较好的性能。实验结果证实了基于极大公共频繁子图抽取特征值的有效性。考虑到真实世界数据集具有一定的复杂性,本文算法具有比较高的满意度。

采用规模较大的两个数据集 DTP Aids 和 DTP Cancer 测试算法的运行时间和伸缩性,运行时间以对数刻度来表示。图 1 给出了实验结果,可以看到本文算法运行时间超过了文献[15]算法。本文算法涉及到矩阵运算花费较多时间,计算公共近邻裁剪也需要较多时间,算法的时间复杂度达到$O(n^3)$,n 是结点的个数。虽然本文算法时间复杂度较高,但是还是能在可容忍的限度内。

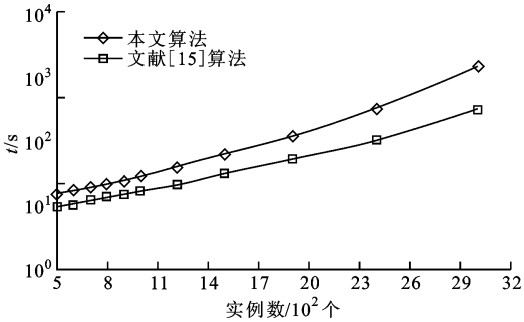

图 1　DTP Aids 数据集实例数和运行时间的关系图

接着测试算法的稳定性。采用表 1 中的数据集,测试当数据集中离群点的比例改变时算法的精确度。由图 2 可见,本文算法精确度比较稳定,而文献[15]算法精确度有所下降。图 3 是关于错误报警率 F 的测试结果,由图可见,当离群点比例增大时,本文算法表现出相对低的错误报警率,但是文献[15]算法的错误报警率增长得比较快,几乎是本文算法的 3 倍多。

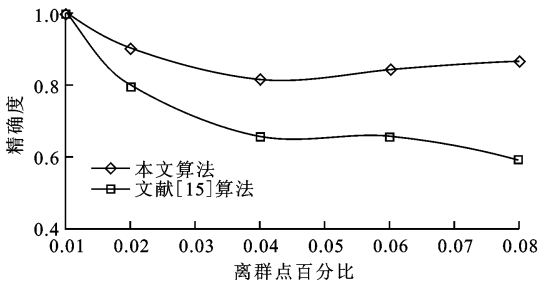

图 2　DTP Aids 数据集改变离群点百分比时的精确度

最后比较本文算法采用传统随机游走的离群值评分和采用往返距离评分的效果差异。采用随机游走的版本记为版本 A,采用往返距离的版本记为版本 B,也是前面实验测试的版本。用表 1 描述的数据集测试召回率,结果见图 4。由图 4 可见,在所有的数据集中,版本 A 召回率超过了版本 B,这说明

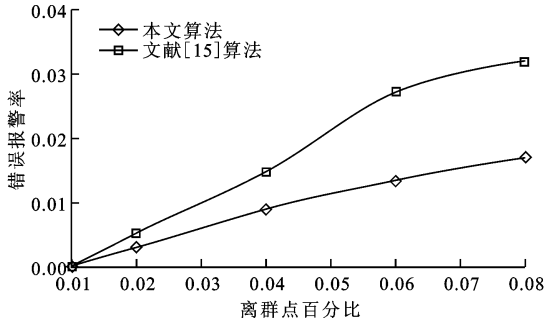

图 3 DTP Aids 数据集改变离群点百分比时的错误报警率

采用往返距离的离群点检测算法综合性能超过了采用随机游走的算法。因为往返距离既可以从全局视野考虑,也能从局部视野考虑,特别是可以识别分布在稠密聚类周边的结点,弥补了随机游走的弱点,表现出更高的精确度和完整性。

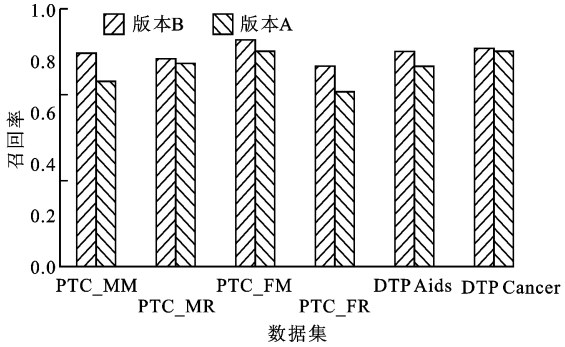

图 4 两个算法版本的召回率比较

3 结 语

本文研究了无监督的基于相似度测量的图数据离群点检测方法,介绍了基于相似度测量的图离群点检测框架。在真实数据集上的测试结果证实了本文方法在图数据离群点检测方面的性能优于之前的基于 subdue 的方法,但是在算法的时间复杂度方面还存在不足。未来的研究方向包括研究一种近似方法,可以改善目前的算法在时间消耗上的不足,以及发展一种基于磁盘的图离群点检测算法来支持超大图挖掘。

参考文献:

[1] HAN Jiawei. Data mining: concepts and technology [M]. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[2] 林海. 离群检测及离群释义空间查找算法研究 [D]. 重庆: 重庆大学, 2012.

[3] 于浩, 王斌, 肖刚, 等. 基于距离的不确定离群点检测 [J]. 计算机研究与发展, 2010, 47(3): 474-484.
YU Hao, WANG Bin, XIAO Gang, et al. Distance-based outlier detection on uncertain data [J]. Journal of Computer Research and Development, 2010, 47(3): 474-484.

[4] HASSANZADEH R, NAYAK R. A semi-supervised graph-based algorithm for detecting outliers in online-social-networks [C]// Proceedings of the 28th Annual ACM Symposium on Applied Computing. New York, USA: ACM, 2013: 577-582.

[5] MOONESINGHE H, TAN P N. Outrank: a graph-based outlier detection framework using random walk [J]. International Journal on Artificial Intelligence Tools, 2008, 17(1): 19-36.

[6] EGOZI A, KELLER Y, GUTERMAN H. A probabilistic approach to spectral graph matching [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 18-27.

[7] AKOGLU L, MCGLOHON M, FALOUTSOS C. Oddball: spotting anomalies in weighted graphs [J]. Lecture Notes in Computer Science, 2010, 6119: 410-421.

[8] MULLER E, SANCHEZ P I, MULLE Y, et al. Ranking outlier nodes in subspaces of attributed graphs [C]// Proceedings of the 2013 IEEE 29th International Conference on Data Engineering Workshops. Piscataway, NJ, USA: IEEE, 2013: 216-222.

[9] GUO Yanrong, WU Guorong, JIANG Jianguo, et al. Robust anatomical correspondence detection by hierarchical sparse graph matching [J]. IEEE Transactions on Medical Imaging, 2013, 32(2): 268-277.

[10] GUPTA M, GAO Jing, SUN Yizhou, et al. Integrating community matching and outlier detection for mining evolutionary community outliers [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2012: 859-867.

[11] 杨茂林. 离群检测算法研究 [D]. 武汉: 华中科技大学, 2012.

[12] AGGARWAL C C, ZHAO Yuchen, YU P S. Outlier detection in graph streams [C]// Proceedings of the 2011 IEEE 27th International Conference on Data Engineering. Piscataway, NJ, USA: IEEE, 2011: 399-409.

(下转第 79 页)