文章编号:1671 - 6833(2012)03 - 0117 - 04

BRINK:基于局部质变因子的聚类边界检测算法

邱保志[1],杨 洋[1],杜效伟[2]

(1.郑州大学 信息工程学院,河南 郑州 450001;2.漯河职业技术学院,河南 漯河 462000)

摘 要:为有效检测聚类的边界,提出了一种基于局部质变因子的聚类边界检测算法(BRINK).该算法使用加权欧式距离技术解决现有聚类边界检测算法不能适用于高维数据的问题,根据局部质变因子在聚类的边界具有稍大于 1 的特征来识别聚类的边界.实验结果表明,该算法能有效的检测出聚类的边界.

关键词:边界检测;局部质变因子;聚类

中图分类号:TP391 **文献标志码**:A doi:10.3969/j.issn.1671 - 6833.2012.03.030

0 引言

聚类的边界检测是数据挖掘新兴的研究领域之一,聚类的边界点是位于高密聚类边沿,它们通常具有 2 个或 2 个以上聚类的特征,其归属并不明确[1],有效的提取聚类边界不但可以提高聚类的精度,还可以研究聚类边缘的特性,因此聚类的边界点具有很重要的研究价值和广泛的应用价值.目前对聚类边界的研究才刚刚起步,聚类边界检测的算法还不是很多.

CHEN Yi-xia 等基于聚类边界点的反向 k 近邻值小于聚类内部点的反向 k 近邻值这一事实提出了聚类边界检测算法 Border[1],对不含噪声的数据集,Border 能有效地识别聚类的边界,然而对含有噪声的数据集和多密度数据集来说 Border 都不能正确识别聚类的边界.BRIM[2]边界点检测算法利用边界点的正向半邻域内分布着较多的点,负向半邻域内分布着较少的点这一特征标记边界点,解决了 BORDER 不能有效识别噪声点和聚类边界点的问题,但该算法参数选择困难且不能用于高维数据.Band[3]算法根据聚类的边界点具有一个较大的变异系数这一原理识别边界,它能够有效地识别含有噪声的多密度聚类的边界,但不能用于高维数据聚类边界的识别.

为了能有效地检测聚类的边界,笔者利用局部质变因子特性提取边界和去除噪声,利用加权的欧式距离使算法适用于高维数据,提出一种基于局部质变因子的聚类边界检测算法.

1 BRINK 算法

1.1 相关概念

在高维空间中,基于欧式距离的方法衡量数据间的相似度会导致"差距趋零"现象的发生,笔者利用加权的欧式距离[4]来解决这一问题.

定义 1 (维度的权重)D 是数据集,$p \in D$,$N_\varepsilon^{A_i}(p)$表示对象 p 在属性 A_i 上的 ε 邻域,t 为权值,$|N_\varepsilon^{A_i}(p)|$ 表示对象 p 在属性 A_i 上的邻居数,点 p 的各维度权重定义如下

$$w_i = \begin{cases} t & |N_\varepsilon^{A_i}| \geqslant K; \\ 1 & |N_\varepsilon^{A_i}| < K. \end{cases} \tag{1}$$

式中:$t \geqslant 1$ 且 t 为整数.

如果 $|N_\varepsilon^{A_i}(p)|$ 大于给定的近邻阈值 K,就认为对象 p 在属性 A_i 的 ε 邻域是密集,给予其赋予较大的权重,否则赋予其较小的权重.

定义 2 对象 p,q 的加权的欧式距离定义为

$$dist_p(p,q) = \sqrt{\sum_{i=1}^{d} w_i (\Omega_{A_i}(p) - \Omega_{A_i}(q))^2}, \tag{2}$$

其中 w_i 为对象 p 在第 i 个维度上的权重,$\mathrm{dist}_p(p,q)$ 表示对象 p 对于对象 q 的加权欧式距离.

定义 3 对任意的自然数 K,p 的 K - 距离

收稿日期:2011 - 10 - 13;修订日期:2012 - 01 - 05

基金项目:河南省重点科技攻关资助项目(112102310073);河南省教育厅自然科学研究计划项目(2009A520028)

作者简介:邱保志(1964 -),男,郑州大学教授,博士,硕士生导师,从事数据挖掘的研究,E-mail:iebzqiu@zzu.edu.cn.

(K-$dist(p)$)为 p 和某个对象 o 之间的距离,这里的 o 满足:

(1)至少存在 K 个对象 $o' \in D \setminus \{p\}$,使得 $d(p,o') \leqslant (p,o)$

(2)至多存在 $K-1$ 个对象 $o' \in D \setminus \{p\}$,使得 $d(p,o') < d(p,o)$.

定义4 对象 p 的 K 距离邻域为包含所有与 p 的距离不超过 K-$dist(p)$ 的对象,即

$$N_{K-dist(p)}(p) = \{q \in D \setminus \{p\} \mid d(p,q) \leqslant K\text{-}dist(p)\}. \quad (3)$$

为了方便,对象 p 的 K 距离邻域简写为 $N_K(p)$.

定义5 给定自然数 K,对象 p 相对于对象 o 的可达距离为

$$reach-dist(p,o) = \max\{K\text{-}dist(o), d(p,o)\}. \quad (4)$$

定义6 用 MinPts 表示 p 的邻域中最小的对象个数,那么对象 p 的局部可达密度(记为 lrd)为对象 p 与它的 MinPts - 距离邻域的平均可达距离的倒数:

$$lrd_{MinPts}(p) = 1/\left[\frac{\sum_{o \in N_{MinPts}(p)} reach\text{-}dist_{MinPts}(p,o)}{\mid N_{MinPts}(p) \mid}\right]. \quad (5)$$

定义7 点 p 的局部质变因子(LOF)[5] 定义为

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{\mid N_{MinPts}(p) \mid}. \quad (6)$$

依据局部异常因子的定义,局部异常因子具有如下特性:在簇内的对象的 LOF 值约等于1,在簇边缘的对象的 LOF 值略大于1,而离簇的距离越远,对象的 LOF 的值越大,并且 LOF 的值与该对象附近的其他对象的分布密度有关[6].

定义8 边界点:数据集中任意对象 p 的局部质变因子 $LOF_{MinPts}(p)$ 满足:

$$\alpha < LOF_{MinPts}(p) < \beta, \quad (7)$$

则称点 p 为边界点. 根据定义7的描述,因为边界对象的局部质变因子具有稍大于1的特性,所以 α 取1,β 取1.05较为合适. 这里 α,β 不作为参数.

1.2 BRINK 算法描述

算法的主要思想:首先扫描整个数据集,计算出数据集中的每个对象在每一维上的权重,其次根据加权的欧式距离计算出每个对象在数据集中的 K 近邻和每个对象在其邻域内的可达距离,然后根据对象的可达距离计算出每个对象的局部可达密度. 最后根据局部可达密度得出每个对象的局部质变因子,并依据每个对象的质变程度标记聚类的边界,算法描述如下.

输入:近邻阈值 K,权值 t;

输出:聚类的边界对象;

步骤1:权重的计算. 扫描整个数据集,计算每个对象在每一维属性上的权重,如果该对象在某一维上具有的邻居数大于近邻阈值 K,就赋予其权值 t,否则就赋予其权值1.

步骤2:K 近邻的计算. 根据步骤1得出的数据集中每个对象在每一维上的权值和公式(2)得出每个对象与其他对象加权的欧式距离,进而得出每个对象在数据集中的 K 近邻.

步骤3:局部可达密度的计算. 首先根据公式(4)计算出数据集中每个对象在其邻域范围内的可达距离,然后利用公式(5)计算出每个对象的局部可达密度.

步骤4:局部质变因子的计算. 根据步骤3得出的每个对象的局部可达密度,利用公式(6)计算出数据集中每个对象的局部质变因子.

步骤5:边界的输出. 把质变因子的值在1到1.05的对象输出.

2 实验结果及分析

实验环境:CPU 为 Intel(R) dual - core 2.60GHz,内存为1.99G,操作系统为 Windows XP professional,算法编写环境为 VC + +6.0.

2.1 实验结果

笔者以一个含有噪声的均匀分布二维的数据集和一个含有噪声的二维多密度数据集验证算法在低维空间中检测边界的能力和去除噪声的能力;使用两个真实数据集来验证发现高维聚类边界的能力.

图1(a)给出的是含有噪声的,不同形状的均匀数据集,含有9 993个数据对象;图1(b)是 Border 算法的边界检测结果($k=25$,$n=1\ 200$);图1(c)是 Band 算法的边界检测的结果($k=20$,$w=1.11$,$BPT=0.26$);图1(d)是本算法(BRINK)的运行结果,使用的参数:近邻阈值 $K=100$,权重 $t=1$.

从图1可以看出,在含有噪声的均匀数据集上,Border 算法不能够区分边界点与噪声点,BRINK 与 Band 两种算法都能够很好的区分边界

点与噪声点，正确识别聚类的边界.

图 2(a)数据集含有 5 034 个数据对象，含有不同形状的非均匀聚类且含有噪声. 图 2(b)是 Border 算法的结果(k = 120，n = 1 200)；图 2(c)是 BRIM 算法的结果(Eps = 40，δ = 60)；图 2(d)是 BRINK 算法的运行结果，使用的参数：近邻阈值 K = 20，权重 t = 1.

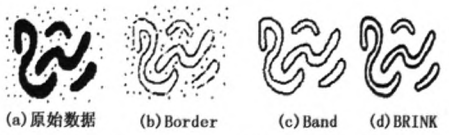

图 1　三种算法的边界检测结果比较

Fig. 1　Three kinds of algorithm for boundary detection results

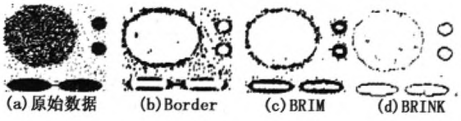

图 2　三种算法的边界检测结果比较

Fig. 2　Three kinds of algorithm for boundary detection results

从图 2 可以看出，Border 算法在含有噪声的非均匀数据集上不能正确的区分聚类的边界点与噪声点，BRIM 算法虽然能够去除一部分噪声，但是吸收了靠近聚类边缘的噪声点. BRINK 算法能够识别聚类的边界，但由于本数据集的大圆中部分地方的密度过于稀疏，以至于大圆内部的有些点被误认为是聚类的边界.

真实数据集“biomed”(*http://lib. stat. cmu. edu/datasets/*)包含 207 个数据对象，每个对象 4 个属性. 该数据集分为两类：病毒感染者(75 人)和正常人(134 人，其中有 30 个病毒携带者). 这里 30 个病毒的携带者就是所要找的聚类边界. 表 1 是 BRINK 算法在“biomed”上运行的结果，使用的参数 K = 20，t = 4. 表 1，2 中用准确率和召回率两个指标来验证 BRINK 算法的有效性，这里令 A = 实验结果中检索到是边界对象，B = 实验结果中检索到不是边界对象，C = 实验结果中未检测到的边界对象，则准确率 = $A/(A+B)$，召回率 = $A/(A+C)$.

从表 1 中可以看出，实验结果得出的 36 人中既包含了 30 个真实的边界对象(病毒携带者)，又包含了 6 个正常人，这一检测结果对疾病防控效果没有负面影响.

表 1　真实数据集“biomed”边界检测结果

Tab. 1　Boundary detection results fordata set “Biomed”

数据集	真实边界对象	实验结果	准确率	召回率
biomed	30	36	83.3%	100%

Breast Cancer (http://archive. ics. uci. edu/ml/)数据集包含 699 个数据对象，每个对象有 10 个属性，它含有两个聚类：恶性肿瘤患者(241 人)和良性肿瘤患者(458 人. 其中 37 个可能发展成为恶性肿瘤的患者)，从医学意义上看这 37 人就是聚类的边界. 表 2 是 BRINK 算法在“Breast Cancer”上运行的结果，所使用的参数 K = 20，t = 5.

表 2　真实数据集“Breast Cancer”边界检测结果

Tab. 2　Boundary detection results for “Breast Cancer”

数据集	真实边界对象	实验结果	准确率	召回率
Breast cancer	37	29	78.3%	78.3%

从表 2 可以看出，实验结果所得的 29 人全部包含在真实的边界对象 37 人当中，所以 BRINK 算法能够检测出高维聚类空间的边界.

以上 4 个实验结果表明，BRINK 算法不但对含有噪声的均匀密度和非均匀密度的数据集有较好的效果，而且能用于高维数据的聚类边界检测.

2.2　算法的时间复杂度分析

在本算法中步骤 1 的时间复杂度为 $O(kn^2)$，步骤 2 的时间复杂度为 $O(n)$，步骤 3 的时间复杂度为 $O(n)$，所以本算法的时间复杂度为 $O(kn^2)$，如果使用索引树结构，算法的时间复杂度可以降为 $O(knlogn)$. 从图 3 可以看出本算法(BRINK)在同规模的数据集上运行时间不如 BRIM，但优于 BORDER.

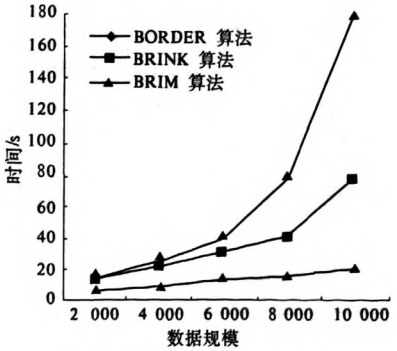

图 3　三种算法运行时间对比

Fig. 3　Running time of three algorithms compared

2.3 参数讨论

BRINK算法有两个参数,即近邻阈值K与权值的参数t,一般来说K值的大小会影响边界检测的结果与算法的执行效率,最近邻数一般不宜过大或过小,过大会影响算法的执行效率,过小局部质变因子就没有意义.对于小规模数据集近邻阈值K的取值一般在10到30较为合适;对于大规模数据集K的取值一般在10到110较为合适.权值参数t的值会影响数据集对象间的差异,权值过小,在中高维数据空间中对象间的差异会不明显.经过大量实验表明,对于低维数据t一般取1较为合适,对于高维数据t一般取2到6较为合适.

3 结论

笔者提出了一种基于局部质变因子的聚类边界检测算法BRINK,该算法既能用于带有噪声的均匀密度和非均匀低维数据集中聚类边界识别,又能适用于高维数据集中聚类边界的识别,解决了现有聚类边界算法不能识别高维数据聚类边界的问题.

参考文献:

[1] CHEN Yi-xia, HSU W, LEE M L, et al. BORDER: Efficient Computation of Boundary Points[J]. IEEE transaction on knowledge and data engineering, 2006, 18(3): 289-303.

[2] QIU Bao-zhi, YUE Feng, SHEN Jun-yi, et al. BRIM: AnEfficient Boundary Points Detecting Algorithm [C]//Proc. Of Advances in Knowledge Discovery andDataMining. Heidelberg: Springer, 2007: 761-768.

[3] 薛丽香,邱保志. 基于变异系数的边界点检测算法[J]. 模式识别与人工智能,2009,22(5):799-802.

[4] 黄王非,陈黎飞,姜青山,等. 基于子空间维度加权的密度聚类算法[J]. 计算工程,2010,36(9):65-67.

[5] 杨风召,朱扬勇, IncLOF:动态环境下局部异常的增量挖掘算法[J]. 计算机研究与发展,2004,41(3):477-484.

BRINK: An Algorithm of Boundary Points of Clusters Detecton Based On Local Qualitative Factors

QIU Bao-zhi[1], YANG Yang[1], DU Xiao-wei[2]

(1. School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Luohe Vocational and Technical College, Luohe 462000, China)

Abstract: In order to detect boundary points of clusters efficiently, we present an algorithm of boundary points detection based on local qualitative factors (BRINK). This algorithm uses weighted euclidean distance to solve high dimensional data problem which most of the existing clusters detecting algorithms can not deal with. According to the feature of local qualitative factors, the individual finds that it is lightly larger than 1 in boundary points of clusters. we can detect the boundary points with the former two processes. As shown by the experimental results, BRINK can detect boundary points in noisy high-dimensional datasets containing clusters of arbitrary shapes, sizes and different densities.

Key words: boundary detection; local qualitative factor; cluster