基于方形对称邻域的局部离群点检测方法 *

揭财明, 刘慧君, 朱庆生
(重庆大学 计算机学院, 重庆 400044)

摘　要: 针对 NDOD(outlier detection algorithm based on neighborhood and density)算法在判断具有不同密度分布的聚类间过渡区域对象时存在的不足,以及为了降低算法时间复杂度,提出一种基于方形对称邻域的局部离群点检测方法。该算法改用方形邻域,吸收基于网格的思想,通过扩张方形邻域快速排除聚类点及避免"维灾";通过引入记忆思想,使得邻域查询次数及范围成倍地减小;同时新定义的离群度度量方法有利于提高检测精度。实验测试表明,该算法检测离群点的速度及精度均优于 NDOD 等算法。

关键词: 数据挖掘; 离群检测; 方形对称邻域; 局部离群度
中图分类号: TP311.13; TP391　　**文献标志码**: A　　**文章编号**: 1001-3695(2012)02-0472-03
doi:10.3969/j.issn.1001-3695.2012.02.018

Square symmetric neighborhood based local outlier detection algorithm

JIE Cai-ming, LIU Hui-jun, ZHU Qing-sheng
(*College of Computer Science*, *Chongqing University*, *Chongqing* 400044, *China*)

Abstract: NDOD may result in wrong estimation when objects are in the location where the density distributions in multiple clusters are significantly different. To void this problem and reduce the computational complexity, this paper proposed a new density based algorithm named SSNOD(square symmetric neighborhood based local outlier detection algorithm). By utilizing the grid-based idea, the algorithm partitioned dataset with square neighborhood and expaned neighborhood rapidly, it could get rid of non-outliers quickly and overcome "dimension curse". By absorbing memory idea, the times of neighborhood query and range were significantly decreased. Besides, computation accuracy could be improved within the novel metrics. Experimental result shows SSNOD is not only efficient in the computation but also more effective than NDOD in detection accuracy.

Key words: data mining; outlier detection; square symmetric neighborhood; local outlier degree

0　引言

离群点检测是数据挖掘技术的重要研究领域之一,用于发现数据集中与其他数据明显不一致的数据对象。近年来,离群点问题的研究在数据库、数据挖掘、机器学习和统计学领域受到广泛关注,有着非常广阔的应用前景,如欺诈检测、入侵检测、故障检测和异常自然气候的发现等[1,2]。

离群点检测算法大致可以分为基于分布、基于偏差、基于聚类、基于距离和基于密度五类[3,4]。其中基于密度的比较具有代表性的算法有 LOF[5]、DBSCAN[6]、NBC[7]、利用对称关系对离群点排序[8]、ODBSN[9]、NDOD[10]等。前四种算法对数据集中的每个对象进行邻域查询时都必须遍历一次数据集,导致时间复杂度较高,限制其在高维海量数据集中的应用;ODBSN 算法中对于离群点离群度的计算方法易导致有意义离群点的丢失及部分离群点离群度不可度量,该缺陷在 NDOD 算法中得到了良好的改进;但是对于 ODBSN 算法在判断具有不同密度分布的聚类间过渡区域对象时容易导致误判,在 NDOD 算法中并没有得到有效的改进。

本文通过分析 NDOD 算法,改进其不足,并且引入记忆思想,成倍减少邻域查询次数及范围,从而降低算法的时间复杂度。

1　NDOD 算法存在的不足

1.1　NDOD 算法的主要思想

NDOD 算法分两个阶段进行,包括初选和精选。其主要思想是:a)吸收基于网格算法的思想,用方形邻域分割数据空间,然后进行方形邻域的扩张,快速排除聚类点,即建立候选异常点集合;b)对候选异常点进行邻域(圆形邻域)查询,以计算其离群度;c)离群度大于某个阈值的候选异常点为异常点。

1.2　NDOD 算法的不足

在基于邻域和密度的异常点检测算法 NDOD 中,对象 p 的异常程度定义为:对象 p 的影响空间中对象的局部密度平均值与对象 p 的局部密度的比值。对于该算法涉及的相关概念,通过图 1 来解释说明。

1)假定有数据集 D,图 1 中 p、q、r、m 和 n 均为 D 中的对象,k-邻域中 k 取值为 3,R_P 表示圆 p 的半径,以求对象点 p 的异常程度为例:

a)点 p 的 k 邻域。$K_{\mathrm{KNB}}(p)=\{q,m,n\}$。

b)点 p 的反 2 次 k 邻域。若点 q 是点 p 的 k 邻域中的对象,且点 q 的 k 邻域仍然包含 p,即 $R2K_{\mathrm{KNB}}(p)=\{q\in K_{\mathrm{KNB}}(p)\mid p\in K_{\mathrm{KNB}}(q)\}$。图 1 中 $R2K_{\mathrm{KNB}}(p)=\{q\}$。

收稿日期: 2011-08-05; **修回日期**: 2011-09-15　　**基金项目**: 国家自然科学基金资助项目(61073058)

作者简介: 揭财明(1985-),男,江西宁都人,硕士研究生,主要研究方向为数据挖掘(c_m_jie@163.com); 刘慧君,男,讲师,博士,主要研究方向为 Web 挖掘、电子商务;朱庆生(1956-),男,安徽当涂人,教授,博导,博士,主要研究方向为数据挖掘、图像处理、数字农业.

c)影响空间。$N_{\mathrm{NIS}_k}(p)=K_{\mathrm{KNB}}(p)\cup R2K_{\mathrm{KNB}}(p)=\{q,m,n,q\}$。点 p 的局部密度为

$$L_{LD_k}(p)=\frac{1}{R_p}$$

点 p 的局部异常因子为

$$N_{\mathrm{NLOF}_k}(p)=\frac{\sum_{q\in N_{\mathrm{NIS}_k}(p)}\frac{L_{LD_k}(q)}{L_{LD_k}(p)}}{|N_{\mathrm{NIS}_k}(p)|}=\frac{R_p}{4}\times\left(\frac{2}{R_q}+\frac{1}{R_m}+\frac{1}{R_n}\right)$$

2)如图 2 所示,该图具有不同密度分布的两个聚类(C_1 较密集,C_2 较为稀疏),p 为 C_2 中的对象且靠近密集聚类 C_1,与对象 q 相比,p 对象的离群程度较低。

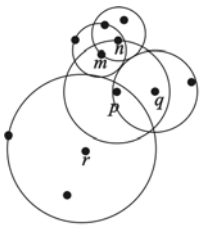

图1　离群度计算

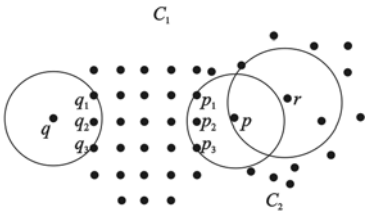

图2　对象p与q离群度比较

以下利用 NDOD 算法比较对象 p、q 的离群度。如图 2 所示,若 k 取值为 3, 对象 p、q 到聚类 C_1 的距离相等,故有 $R_p=R_q$,$K_{\mathrm{KNB}}(p)=\{p_1,p_2,p_3\}$ 而 $K_{\mathrm{KNB}}(q)=\{q_1,q_2,q_3\}$;对于反 2 次 k 邻域,$R2K_{\mathrm{KNB}}(p)=R2K_{\mathrm{KNB}}(q)=\varnothing$;$p$ 与 q 的 k 邻域同分布,$L_{LD_k}(p_1)=L_{LD_k}(q_1)$,$L_{LD_k}(p_2)=L_{LD_k}(q_2)$,$L_{LD_k}(p_3)=L_{LD_k}(q_3)$;进而可以得到 $N_{\mathrm{NLOF}_k}(p)=N_{\mathrm{NLOF}_k}(q)$,即对象 p 与 q 的异常程度相同,与实际情况对象 p 异常程度低于对象 q 不相符。

对于图 2 所存在的异常点误判的问题,是因为对象 p 处于具有不同密度分布的聚类 C_1 与 C_2 间过渡区域且 p 属于 C_2 但是更接近 C_1,计算异常程度时更多地考虑了 C_1 对它的影响而未考虑 C_2 对它的影响。该问题可以通过对反向 k 邻域的重新定义得到解决。

2　改进算法——SSNOD

以下介绍基于方形对称邻域的局部离群点检测算法 SSNOD 的基本概念以及算法描述。由于 SSNOD 算法与数据集的维数无关,适用于任意维数的数据集。为了描述方便及易于理解,下面以二维数据集为例进行介绍。

2.1　基本概念

给定二维数据集 D,其大小为 n,p、q 均为 D 中的对象。

定义 1　方形邻域。以 $p(x_0,y_0)$ 为中心、e 为边长的方形区域内的对象集合,称为点 p 的方形邻域,记做 $S_e(p)$,即

$$S_e(p)=\{q(x,y)\in D\,|\,|x-x_0|\leqslant e/2 \text{ 且 } |y-y_0|\leqslant e/2\}$$

定义 2　邻近方形邻域。与方形邻域 S 有相邻点的八个方形邻域,称为邻近方形邻域,记做 $N_e(S)$。

定义 3　密集邻域及稀疏邻域。若点 p 的方形邻域中对象数(不含点 p)不少于 k(事先设定值),则称该方形邻域为密集邻域,密集邻域中所有对象点为聚类点;否则,该称方形邻域为稀疏邻域,其内所有对象点为候选离群点。

定义 4　k 距离及 k 方形邻域。以 $p(x_0,y_0)$ 为中心,包含距离点 p 最近的 k 个对象时方形区域的边长,称为点 p 的 k 距离,记做 $d_k(p)$,该邻域称为点 p 的 k 方形邻域,记做 $N_k(p)$。

定义 5　反 k 方形邻域。设 $q\in D$ 且 $p\in N_k(q)$,则称对象 q 为 p 反 k 方形邻域中的对象,记做 $R_k(p)$,即 $R_k(p)=\{q\in D\,|\,p\in N_k(q)\}$。

定义 6　基于方形邻域的影响空间。k 方形邻域与反 k 方形邻域的并集称为基于方形邻域的影响空间,记做 $IS_k(p)$,即 $IS_k(p)=N_k(p)\cup R_k(p)$。

定义 7　局部密度。点 p 的局部密度定义为点 p 的 k 距离的倒数,记做 $LD_k(p)$,即

$$LD_k(p)=\frac{1}{d_k(p)}$$

定义 8　局部离群度。点 p 的局部离群度为点 p 基于方形邻域的影响空间内对象的局部密度的平均值与 p 的局部密度的比值,记做 $LOD_k(p)$,即

$$LOD_k(p)=\frac{\sum_{q\in IS_k(p)}LOD_k(q)/LOD_k(p)}{|IS_k(p)|}$$

从定义可知,若点 p 影响空间内对象局部密度平均值越大,而 p 的局部密度越小,则 $LOD_k(p)$ 取值将越大;若大于某个阈值(大于 1 的值),可以判定点 p 为离群点。

对于 NDOD 算法在判断具有不同密度分布的聚类间过渡区域对象时存在的不足(如图 2 中点 p、q 异常度的误判),在引入局部离群度后得到了良好的改进。以下通过图 3 中点 p、q 离群度的比较加以分析说明。

SSNOD 算法伪代码如下:

```
Algorithm SSNOD ( D , e , k )
//粗选:建立候选离群点集及候选邻域
  while( ! D. isAllVisited( ) )
    p = selectUnvisitedPoint( D );
    SP = getSquareNeighbor( p , e );
    NextSP = SP;
    while( ! NextSP. isAllVisited( ) or NextSP. isEmpty( ) )
      if( |NextSP| > k ) then
        NextSP. set( ClusterLabel ) ;
else
  NextSP. set( CandidateLabel ); CandidateOutlierSet. add( NextSP );
//建立候选邻域并为其内所有点创建索引标记
  CandidateNeighbor.
    setNeighborIndex( p, NextSP ) ;
  end if
  NextSP = getNearSquareNeighbor( NextSP );
  end while
end while
//精选:计算候选离群点的局部离群度
for each object p in CandidateOutlierSet DO
  if( unvisited( p ) ) then
/* 搜索邻域组成时,优先考虑具有相同索引标记的点集以减小搜索范围 */
  SN = getKSN( p ); //获取 k 方形邻域
  RSN = getRSN( p );//获取反 k 方形邻域
unvisited( p ) = FALSE;
//T 为大于 1 的阈值
if( LOD( p ) > T ) then
    OutlierSet. add( p );
end if
end if
end for
```

类似于图 2、3 中 k 值仍取 3,点 p、q 与 C_1 等距;点 p 的 k 方形邻域 $N_k(p)=\{p_1,p_2,p_3\}$,点 q 的 k 方形邻域 $N_k(q)=\{q_1,q_2,q_3\}$;对于反 k 方形邻域而言,$R_k(p)=\{r\}$,$R_k(q)=\varnothing$;

故 $IS_k(p)=N_k(p)\cup R_k(p)=\{p_1,p_2,p_3,r\}$，而 $IS_k(q)=N_k(q)\cup R_k(q)=\{q_1,q_2,q_3\}$；由于点 p、q 方形邻域内点的密度分布完全相同，于是有 $LD_k(p_1)=LD_k(q_1)$，$LD_k(p_2)=LD_k(q_2)$，$LD_k(p_3)=LD_k(q_3)$；对于点 p、q、r 的局部密度而言，由于 $d_k(p)=d_k(q)<d_k(r)$，于是有 $LD_k(p)=LD_k(q)>LD_k(r)$，因此：

$$LOD_k(p)=(LD_k(p_1)+LD_k(p_2)+LD_k(p_3)+LD_k(r))/4LD_k(p)$$

$$LOD_k(q)=(LD_k(q_1)+LD_k(q_2)+LD_k(q_3))/3LD_k(q)$$

显然有 $LOD_k(p)<LOD_k(q)$，即点 p 的局部离群度低于点 q，符合实际情况。

2.2　算法介绍

基于方形对称邻域的离群点检测算法 SSNOD 包括粗选和精选两个阶段，相比 NDOD 算法，SSNOD 作了以下改进：

a)粗选和精选均在方形邻域中进行(NDOD 算法粗选时采用方形邻域、精选时采用圆形邻域)，从而减少了邻域切换的开销。

b)粗选时，若检测到稀疏邻域，则在将其内所有点标记为候选离群点的同时形成候选邻域。候选邻域的概念正是引入记忆思想的结果。所谓记忆思想是指粗选阶段的稀疏邻域极可能成为精选阶段的 k-邻域，记录这些稀疏邻域将极大地减少候选离群点离群度计算时的邻域查询范围和次数。

c)引入更为合理和更具意义的反 k 邻域的定义，使得离群度计算更精确。

3　实验测试

实验平台：操作系统 Linux Server、Intel P4 CPU 2.80 GHz、内存 2 GB、编程工具 MATLAB 7.1。

本章将从实验结果及运行效率等方面分析 SSNOD 算法的优势。

3.1　实验结果

实验在如图 4 所示的合成数据集上进行；图 5 为算法 SSNOD 粗选阶段划分的方形邻域。

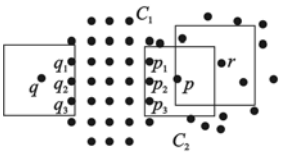

图3　基于方形对称邻域的对象p与q离群度比较

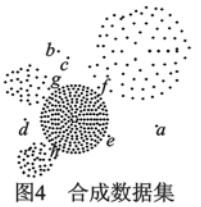

图4　合成数据集

NDOD 算法计算结果：e 的离群度低于点 f 和 g，若阈值 T 取 1.2，则误将靠近密集聚类的稀疏聚类点 f、g 判断为离群点。SSNOD 算法计算结果：点 f 和 g 离群度低于点 e，能正确地找出所有离群点，并不会导致误判。

3.2　运行效率

为验证 SSNOD 算法的性能优势，本实验采用随机数发生器产生服从正态分布的具有不同密度的聚类数据集，数据集大小从 5 000 到 50 000 不等，并随机添加 200 个离群点，与 LOF、NDOD 算法进行时间复杂度比较。实验结果转换为图 6 所示的线性图，其中算法名称中的数字表示 k 的取值，百分比为数据集大小等于 50 000 时的离群点查准率。

从图6 可以看出，SSNOD 算法的运行时间与 k 的取值密切相关，该算法的运行效率明显优于 LOF 算法，并且也优于 NDOD 算法；随着 k 值逐渐增大，算法运行时间逐渐减少；为兼顾离群点的查找精度，k 值不能无限增大，从实验结果可知，该数据集 $e=30$、$k=40$ 时，实验效果比较理想。

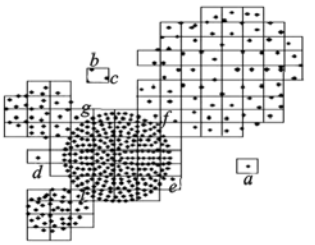

图5　算法SSNOD粗选阶段方形邻域划分

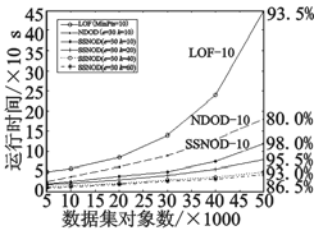

图6　算法运行效率比较

4　结束语

本文通过分析 NDOD 算法，改进其不足之处，提出基于方形对称邻域的局部离群点检测算法。该算法将已有经典算法的邻域改为方形邻域，融入基于网格算法的思想，通过扩张方形邻域快速排除聚类点及避免“维灾”；引入记忆思想及新的离群度度量方法，成倍地减少邻域查询次数及范围。实验结果表明，SSNOD 算法无论是离群点检测精度还是检测速度均取得了良好的效果。今后的研究内容是：建立自适应数学模型，能根据数据集自动确定最优的方形邻域边长 e 及参数 k；将 SSNOD 算法应用于如欺诈检测等实际应用领域中。

参考文献：

[1] 张卫旭，尉宇. 基于密度的局部离群点检测算法[J]. 计算机与数字工程，2010，**38**(10)：11-14.

[2] 胡彩平，秦小麟. 一种基于密度的局部离群点检测算法 DLOF[J]. 计算机研究与发展，2010，**47**(12)：2110-2116.

[3] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques[M]. 2nd ed. San Francisco: Morgan Kaufmann, 2006.

[4] MARATEB H R, ROJAS-MARTINEZ M, MANSOURIAN M, *et al*. Outlier detection in high-density surface electromyographic signals[C]//Proc of the 32nd Annual International Conference. 2010: 4850-4853.

[5] BREUNIG M M, KRIEGEL H, NG R T, *et al*. LOF: identifying density-based local outliers[C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000: 93-104.

[6] ESTER M, KRIEGEL H P, SANDER J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proc of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1996: 226-231.

[7] ZHOU Shui-geng, ZHAO Yue, GUAN Ji-hong, *et al*. A neighborhood-based clustering algorithm[C]//Proc of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2005: 361-371.

[8] JIN Wen, TUNG A K H, HAN Jia-wei, *et al*. Ranking outliers using symmetric neighborhood relationship[C]//Proc of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2006: 93-104.

[9] 黄添强，秦小麟，叶飞跃. 基于方形邻域的离群点查找新方法[J]. 控制与决策，2006，**21**(5)：541-545.

[10] 陶运信，皮德常. 基于邻域和密度的异常点检测算法[J]. 吉林大学学报：信息科学版，2008，**26**(4)：398-403.