基于共享最近邻的离群检测算法*

苏晓珂[1], 郑远攀[1], 万仁霞[2]
(1. 郑州轻工业学院 计算机与通信工程学院, 郑州 450002; 2. 北方民族大学 信息与计算科学学院, 银川 750021)

摘 要: 为识别混合属性数据集中的离群点,提出了一种基于共享最近邻的离群检测算法,通过计算增量聚类结果簇间的共享最近邻相似度,不但能够发现任意形状的簇,还可以检测到变密度数据集中的全局离群点。算法时间复杂度关于数据集的大小和属性个数呈近似线性。在人工数据集和真实数据集上的实验结果显示,提出的算法能有效检测到数据集中的离群点。

关键词: 共享最近邻; 离群检测; 任意形状簇; 混合属性
中图分类号: TP181 **文献标志码**: A **文章编号**: 1001-3695(2012)07-2426-03
doi:10.3969/j.issn.1001-3695.2012.07.006

Outlier detection algorithm based on shared nearest neighbor

SU Xiao-ke[1], ZHENG Yuan-pan[1], WAN Ren-xia[2]
(1. *School of Computer & Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou* 450002, *China*; 2. *College of Information & Computation Science, Beifang University for Nationalities, Yinchuan* 750021, *China*)

Abstract: This paper introduced an outlier detection algorithm based on the shared nearest neighbor clustering in order to detect the outliers with the mixed attributes. The algorithm calculated the shared nearest neighbor similarity measure between result clusters caused by the incremental clustering. It could not only find the arbitrary shape clusters but also identify the global outlier in large and high-dimensional dataset with different density. Presented approach had nearly linear time complexity with the number of attributes and the size of dataset which results in good scalability.

Key words: shared nearest neighbor(SNN); outlier detection; arbitrary shape cluster; mixed attributes

0 引言

离群检测的目标是发现数据集中偏离大部分数据的离群点,因为这些数据的偏离也许并非由随机因素产生,而是产生于完全不同的机制。离群检测适用于故障诊断、入侵检测和欺诈检测等多个领域,因此得到了广泛的研究。研究者提出了许多相应的离群检测算法,大致可分为基于统计、基于聚类、基于密度以及基于距离等方法[1]。这些方法对中、小规模低维数值属性数据集具有很好的性能,但面对大规模混合属性数据集时,在伸缩性和准确性等方面存在着不足。

文献[2]使用基于最近邻的三角区域挖掘离群点,本质是将原始数据集进行特征变换,再在变换后的数据集上采用最近邻分类法,虽然能够有效检测到离群点,但因其时间复杂度较高,不适用于大规模数据集。文献[3]研究了面向大规模静态数据集的离群检测问题,在对数据集网格划分的基础上,介绍了数据超方格层次上的密度近似计算与稠密数据主体过滤策略。由于网格划分的特点,算法仅能处理数值属性数据集,若数据维度较高,进行网格划分将导致很多低密度超方格存在。文献[4]提出了一种面向大规模混合属性数据集的离群检测算法,对聚类形成的簇的标记基于整体偏离程度,能够正确识别远离大多数对象的边界离群点,但容易将被正常对象包围的内部离群点误判为正常,导致检测率降低。文献[5]研究了变密度数据集离群检测问题,由于局部离群因子的计算本质上仍是基于距离的,面对大规模数据集时同样存在困难;同时算法重点在识别离群点上,忽略了大量正常数据的内在联系。文献[6]在增量聚类保留的候选离群簇上作离群检测,适用于大多数数据集,但在处理图 1 中不同密度数据集 dataset 1 时遇到了困难。图 1 中的鱼周围分布了八个离群点,鱼尾部分的数据分布稀疏,密度低于其他部分,若采用适合鱼头部分的近邻半径阈值 cnr,则此值相对鱼尾部分来说过低,也即鱼尾部分的正常数据在 cnr 范围内不存在近邻,易被误判为离群点;若采用适合鱼尾部分的 cnr,则鱼头部分的离群点因其有足够的近邻,容易被判为正常,导致检测率降低。该算法在最优情况下仍有 a、b、c 这三个离群点没有检测到,算法也不能识别图中任意形状的簇。

对于离群检测问题,最终的目标是得到尽可能好的检测效果,且使算法时间复杂度比较低。本文将对象的共享最近邻相似度扩展到增量聚类的结果簇上,通过计算结果簇的共享最近邻簇数,不但能够发现任意形状的簇,还可检测到变密度数据集中的全局离群点。

1 相关定义

共享最近邻(SNN)聚类是 Levent Ertoz 提出的一种整合了多种聚类思想的聚类方法,对象间的相似度由共享最近邻数定

收稿日期: 2011-12-08; **修回日期**: 2012-01-06 **基金项目**: 国家自然科学基金资助项目(61163017);郑州轻工业学院博士科研基金资助项目(2010BSJJ039);河南省科技攻关资助项目(122102210125);河南教育厅自然科学基础研究计划资助项目(12B520051)

作者简介: 苏晓珂(1979-),女,河南巩义人,讲师,博士,CCF 会员(E200013017M),主要研究方向为数据挖掘、智能计算(suxiaoke07@126.com);郑远攀(1983-),男,河南邓州人,讲师,博士,主要研究方向为应急管理和安全工程;万仁霞(1975-),男,江西南昌人,副教授,博士,主要研究方向为数据挖掘、智能计算.

义,如果两个对象与相同对象中的大部分相似,则即使直接的相似性度量不能给出它们也相似。共享最近邻本质上度量了一个对象被关于最近邻的相似对象包围的程度,在高密度和低密度区域内的对象一般都具有相对较高的 SNN 密度,而在从低密度到高密度过渡的区域中的对象(簇边界)将趋向具有较低的 SNN 密度。由于共享最近邻相似度量反映数据空间中对象的局部结构,它对空间维度和密度的变化都相对不太敏感[7]。

设 S 是包含 N 个对象的 m 维数据集,C 为 S 的划分簇集,$C=\{C_1,C_2,\cdots,C_l\}$,$\sum_{i=1}^{l}\text{size}(C_i)=S$,$\text{size}(A)$ 指集合 A 的大小,则 SNN 相似度、密度以及噪声簇的定义如下。

定义 1　SNN 相似度。簇 C_1 和 C_2 之间的 SNN 相似度 $\text{sim}(C_1,C_2)$ 定义为 C_1 和 C_2 的最近 k 邻接列表中相同簇的数目,即

$$\text{sim}(C_1,C_2)=\text{size}(nn[C_1]\cap nn[C_2]) \tag{1}$$

其中:$nn[C_1]$ 表示簇 C_1 的最近 k 邻接列表,即与 C_1 距离最近的 k 个簇组成的集合。

$\langle V,E\rangle$ 表示共享最近邻图,顶点 $V\in C$ 指簇集 C 中的所有元素,$\forall C_i,C_j\in C$,C_i 和 C_j 之间有边连接,当且仅当 $C_i\in nn[C_j]$ 且 $C_j\in nn[C_i]$,边的权重由 C_i 和 C_j 的相似度给出。

定义 2　密度。簇 C_1 的密度 $\text{den}(C_1)$ 定义为 C_1 的最近 k 邻接列表中与 C_1 相似的簇数目,即

$$\text{den}(C_1)=\text{count}(\text{sim}(C_1,C_2)\geqslant\alpha) \tag{2}$$

其中:$C_2\in nn[C_1]$,即 C_2 在 C_1 的最近 k 邻接列表中;α 为判断两个簇是否相似的阈值,即两个簇相似的条件是它们共享了大于等于 α 个最近邻。

核簇 ker 即为高密度簇,骨架 ske 为所有核簇构成的集合。给定核簇密度阈值 β,则骨架 ske 为

$$\text{ske}=\{\text{ker}\mid\text{den}(\text{ker})\geqslant\beta\} \tag{3}$$

给定合并阈值 γ,如果 $\text{sim}(C_1,C_2)\geqslant\gamma$,$C_1,C_2\in\text{ske}$,则将它们合并成为一个簇。

定义 3　噪声簇。噪声簇 C_n 定义为所有不与任何一个核簇相似的非核簇,即

$$C_n=\{C_j\mid\text{sim}(C_j,k_i)\leqslant\alpha\},C_j\notin\text{ske},k_i\in\text{ske} \tag{4}$$

2　算法主要思想

依据“两个簇共享的近邻数越多越相似”原则对文献[6]中增量聚类的结果簇进行聚类,得到由结果簇表征的任意形状新簇。因共享最近邻的不完全聚类性,有些结果簇不能被分配到新簇中,将不能被聚类的结果簇也作为新簇,加入到新簇列表中,对所有新簇使用基于距离的离群检测方法识别离群点。具体步骤如下:

输入:原始数据集 S。

输出:离群簇集合。

a)对 S 进行增量聚类,形成对 S 的划分簇集 $C=\{C_1,C_2,\cdots,C_l\}$,$\sum_{i=1}^{l}\text{size}(C_i)=S$。

b)SNNclustering(C),生成新簇集合 $nC=\{nC_1,nC_2,\cdots,nC_p\}$,$\sum_{i=1}^{p}nC_i=C$。

c)对 nC 使用基于距离的离群检测算法,得到最终的离群簇集合。

其中步骤 b)SNNclustering(C)的过程为:对划分簇集 C 中的元素,计算任意两簇间的距离得到距离矩阵 $D_{ij}=d_{(C_i,C_j)}$,$\forall i,j\in[1,l]$,稀疏化 D,对 $\forall C_i$,寻找 C_i 的 k 近邻集 Ck,$d_{ij}=\begin{cases}1 & C_j\in Ck\\ 0 & C_j\notin Ck\end{cases}$;根据定义 1,对稀疏化的 D 构造相似矩阵 sim,由 sim 构造最近邻图$\langle V,E\rangle$;对 $\forall C_i\in V$,计算 C_i 的密度 $\text{den}(C_i)$,得到核簇集合 $\text{ske}=\{C_i\mid\text{den}(C_i)\geqslant\beta\}$,然后聚类 ske 中的核簇,过程为:对 $\forall C_i,C_j$,如果 $\text{sim}(C_i,C_j)\geqslant\gamma$,则 $C_i=C_i\cup C_j$,形成 nC 集合;未被聚类的簇 C_i 同样加入到 nC 中。步骤 c)采用基于距离的离群点定义,即给定邻域半径,依据簇的邻域中包含对象的多少来判定离群点。

3　实验结果与评估

检测率、假正率是度量离群检测方法性能的两个指标。检测率 DR(detection rate)表示被正确检测的离群点占所有离群点的比例;假正率 FR(false positive rate)表示正常记录被检测为离群点的记录数占整个正常记录数的比例。对于离群检测,理想结果是具有高的检测率和低的假正率。为了评估算法的性能,本文在 VC6.0 环境中实现了算法,并在人工数据集和 UCI machine learning repository[8] 提供的真实数据集上进行了测试。

3.1　参数分析

实验中的簇近邻个数阈值 rk 代表离群检测阶段每个新簇需要考察的近邻数,其他参数分析如下。

a)sk 指共享最近邻聚类阶段需要考察的近邻数。若取值小,则同等情况下共享的近邻数就可能减少,核簇出现的机会少,离群簇的个数有可能增多,最终将导致假正率 FR 增高;若取值大,则共享近邻数有可能增多,导致核簇多,因此能被聚类的结果簇数就多。

b)α 指两个增量聚类的结果簇共享最近邻数阈值。若取值小,满足共享最近邻要求的结果簇多,因此可聚类的结果簇就多,导致离群点出现的可能性减少,检测率相对减少;若取值大,可聚类的结果簇少,假正率 FR 增高。

c)β 指核簇密度阈值,即若一个结果簇的密度大于 β,则为核簇,因此,若 β 取值小,满足核簇要求的结果簇的数量相对增多,检测率相对减少;若 β 取值大,满足核簇要求的结果簇数量相对降低,假正率 FR 增高。

3.2　人工数据集检测结果

首先在人工数据集 dataset 1 上评估本算法的检测性能。增量聚类半径阈值为 0.025 时,离群检测结果如表 1 所示。改变 sk 时,结果如表 1 中的 1~3 行,随着 sk 增大,记录满足最近邻要求的可能性增大,SNN 聚类精度逐渐减少;改变 α 时,结果如 4~6 行,随着 α 增大,未聚类的记录数逐渐增多,当 $\alpha=2$ 时,检测结果最理想,SNN 聚类精度达 99.89%,而检测率为 100%,假正率为 0;改变 β 时,结果如 7、8 行,随着 β 增大,未聚类的记录数增多;改变 rk 时,结果如 9、10 行,随着 rk 增大,对近邻个数的要求增加,导致判定为离群点的新簇数量有可能增加,因此检测率逐渐增加,同时假正率也逐渐增加。在第五种情况下,聚类精度和检测结果都达到最优,说明算法在 dataset

1 数据集上是有效可行的。

表 1 Dataset 1 检测结果

序号	*sk*	α	β	*rk*	SNN 簇数	SNN 精度/%	未聚类记录数	DR/%	FR/%
1	6	3	6	7	7	99.87	148	37.50	0.00
2	7	3	6	7	9	99.68	3	75	0.00
3	8	3	6	7	8	76.93	0	62.5	0.00
4	6	1	6	7	10	77.25	0	100.00	0.00
5	6	2	6	7	10	99.89	2	100.00	0.00
6	6	4	6	7	0	/	945	0.00	0.00
7	6	2	2	7	10	77.25	0	100.00	0.00
8	6	2	5	7	9	77.22	1	100.00	0.00
9	6	2	6	4	10	99.89	2	62.50	0.00
10	6	2	6	5	10	99.89	2	87.50	0.00

3.3 Lymphography 淋巴系造影术数据集检测结果

为测试算法对分类属性数据集的检测性能,本文在淋巴系造影术数据集上进行了测试。此数据集包含 148 条记录,每条记录具有 18 个分类属性。所有记录被分为四类,类 1 含 2 条记录,类 2 含 81 条记录,类 3 含 61 条记录,类 4 含 4 条记录。类 1 与类 4 只占整个数据集的 4.05%,可看做离群数据。增量聚类半径阈值为 0.59 时,算法检测结果如表 2 所示,仅列出有代表性的三种情况。

表 2 Lymphography 检测结果

sk	α	β	*rk*	SNN 簇数	SNN 精度/%	未聚类记录数	DR/%	FR/%
4	1	4	2	1	54.73	0	0.00	0.00
4	2	4	2	1	58.78	6	100.00	0.00
4	3	4	2	0	/	148	100.00	0.70

由表 2 可看出,当 $\alpha=2$ 时,达到理想的检测结果,假正率为 0,检测率为 1。增量聚类后形成 10 个结果簇,精度为 77.03%,在结果簇上进行共享最近邻聚类时,其中的 4 个结果簇被聚为一类,其他 6 个仅包含一条原始记录的结果簇未被聚类,虽然最终的聚类精度仅为 58.78%,但 6 个离群数据能被正确地分离出来。聚类精度低的原因在于此数据集中类别为 2 和 3 的正常记录混杂在一起,导致正常记录无法被正确分类,但类别为 1 和 4 的离群点远离正常记录,可被正确识别。

3.4 Wisconsin breast cancer 数据集检测结果

Breast cancer 数据集有 699 条记录,其中良性的 458 条,恶性的 241 条,每条记录包含 9 个数值属性。直观地判断,恶性与良性记录应有明显区别。因此,选取不同比例的两种记录构造分布不平衡的测试集,其中选取 39 条(8%)恶性记录和 444 条(92%)良性记录,期望能够将比例很小的那部分记录从测试集中检测出来。

增量聚类半径阈值为 0.06 时,改变 *sk* 值,检测结果如表 3 中的 1 ~ 4 行,随着 *sk* 的增大,共享最近邻聚成的簇逐渐增多,未聚类的原始记录数逐渐减少,当 $sk=7$、$\alpha=4$、$\beta=1$、$rk=3$ 时,未聚类记录仅有 1 条;当 α 逐渐增大时,检测结果如表中的 5 ~ 7 行,形成的簇数逐渐减少,检测率和假正率都逐渐增大;当 $sk=3$、$\alpha=2$、$\beta=3$、$rk=11$ 时,未聚类记录数达到 457 条,此时算法成为基于距离的离群检测算法;8 ~ 9 行为改变 β 时的检测结果,随着 β 的增大,未聚类的记录数逐渐增多,检测率和假正率都逐渐降低;改变 *rk* 时的结果如 10 ~ 12 行,随着 *rk* 的增大,检测率和假正率都逐渐增高,因数据集中正常记录占 92%,因此每种取值下的聚类精度不低于 80%。

表 3 Wisconsin breast cancer 检测结果

序号	*sk*	α	β	*rk*	SNN 簇数	SNN 精度/%	未聚类记录数	DR/%	FR/%
1	3	4	1	3	0	/	483	10.26	0.00
2	5	4	1	3	2	93.18	14	12.82	0.00
3	6	4	1	3	3	95.39	6	97.44	5.18
4	7	4	1	3	5	95.85	1	97.44	4.73
5	3	0	3	10	20	97.31	0	79.49	2.03
6	3	1	3	10	7	97.57	30	84.62	2.03
7	3	2	3	11	1	80.77	457	97.44	2.93
8	7	0	7	1	6	95.86	0	97.44	4.73
9	7	0	8	1	0	/	483	15.38	0.00
10	3	1	1	1	20	97.31	0	25.64	0.00
11	3	1	1	4	20	97.31	0	79.49	1.35
12	3	1	1	11	20	97.31	0	100.00	15.32

3.5 A1 数据集检测结果

Kddcup99 数据集中的每条记录包含 7 个分类属性和 34 个数值属性。选择与文献[4]中相同的数据子集 A1,包含 38 841 条正常记录和 1 618 条攻击记录,攻击记录占 4%,其中 DoS 攻击占 98.39%, U2R 攻击占 0.06%, R2L 攻击占 0.37%, Probe 攻击占 1.11%。

改变 *sk* 时,结果如表 4 中的 1 ~ 3 行,数据集中正常记录占绝大多数,因此改变 *sk* 对共享最近邻聚类精度影响不大,精度都在 99.00% 以上。改变 α 时,结果如 4 ~ 7 行,随着 α 增大,未聚类的记录数逐渐增多,检测率和假正率都逐渐增大。虽然第一种情况聚类精度可达到 99.94%,但未聚类记录数过多,为 12 012 条,同时假正率也过高,为 4.29%。第三种情况虽然完全聚类所有原始记录,并且聚类精度也可达到 99.67%,但检测率仅为 92.27%。综合表 4 中实验结果,第九种情况相对较理想,具有较高的聚类精度和较好的离群检测结果,说明本文提出的算法在 A1 数据集上的有效性,能使正常记录很好地聚集在一起,并且离群数据能被正确识别。

表 4 A1 检测结果

sk	α	β	rk	SNN 簇数	SNN 精度/%	未聚类数	FR	DR/%			
								全体	DOS	R2L	Probe
3	2	3	7	1	99.94	12 012	4.29	98.83	99.25	33.33	88.89
5	2	3	7	4	99.67	1	0.02	92.27	93.66	0.00	5.56
6	2	3	7	5	99.67	0	0.02	92.27	93.66	0.00	5.56
6	1	3	7	5	99.67	0	0.02	92.27	93.66	0.00	5.56
6	3	3	7	4	99.66	1 135	0.02	92.27	93.66	0.00	5.56
6	5	3	7	1	99.93	3 765	0.32	98.39	99.18	0.00	66.67
6	6	3	7	0	/	40 459	1.19	98.58	99.18	33.33	72.22
6	5	1	7	2	99.93	2 630	0.32	98.39	99.18	0.00	66.67
6	5	1	3	2	99.93	2 630	0.06	98.39	99.18	0.00	66.67
6	5	1	4	2	99.93	2 630	0.32	98.39	99.18	0.00	66.67

3.6 相关算法对比

将本文提出的算法与文献[9]中的 TOD、CBOD 算法和文献[10]中的 ODBUG 算法、文献[6]中的算法在三种数据集上的实验结果进行对比,比较各种算法的检测率和假正率,从而测试这些算法的离群检测效果。

由图 2 可看出,本文提出的方法比前三种 (下转第 2453 页)

(上接第 2428 页)方法的假正率低,检测率高,在 Lymphography 数据集上的检测效果优于文献[6]中的算法,更适用于检测真实数据集上的离群点。

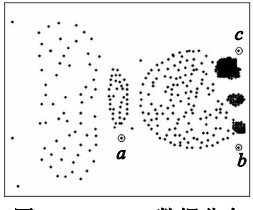

图1 Dataset 1 数据分布

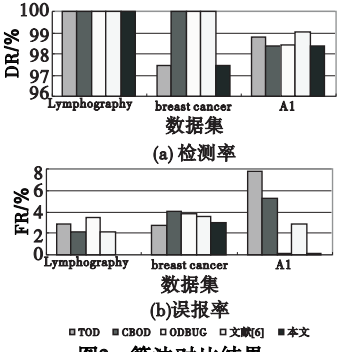

图2 算法对比结果

4 结束语

本文基于 a)数据集中的正常数据占绝大部分,离群点会偏离正常数据;b)正常数据与离群点聚集在不同类中的思想,在对原始数据集增量聚类的基础上,将共享的最近邻簇数作为簇间相似性度量,仅需扫描原始数据集一遍,既可发现任意形状的簇,又能够有效检测到数据集中的离群点,在人工数据集和真实数据集上的实验结果验证了算法的有效性。

参考文献:

[1] PATCHA A, PARK J. An overview of anomaly detection techniques: existing solutions and latest technological trends[J]. Computer Networks, 2007, 51(12): 3448-3470.

[2] TSAI C, LIN C. A triangle area based nearest neighbors approach to intrusion detection[J]. Pattern Recognition, 2010, 43(1): 222-229.

[3] 李存华, 孙志挥. GridOF: 面向大规模数据集的高效离群点检测算法[J]. 计算机研究与发展, 2003, 40(11): 1586-1592.

[4] JIANG Sheng-yi, SONG Xiao-yu. A clustering-based method for unsupervised intrusion detections[J]. Pattern Recognition Letters, 2006, 27(5): 802-810.

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

[6] 苏晓珂, 兰洋. 一种高效混合属性离群检测算法[J]. 小型微型计算机系统, 2010, 31(11): 2282-2286.

[7] LI Xia, JIANG Sheng-yi. A novel fast clustering algorithm[C]//Proc of the Artificial Intelligence and Computational Intelligence. 2009: 284-288.

[8] ASUNCION A, NEWMAN D. UCI machine learning repository[EB/OL]. (2007). http://www.ics.uci.edu/~mlearn/MLRepository.html.

[9] 蒋盛益, 李庆华. 一种两阶段异常检测方法[J]. 小型微型计算机系统, 2005, 26(7): 1237-1240.

[10] 蒋盛益, 姜灵敏. 一种高效异常检测方法[J]. 计算机工程, 2007, 33(7): 166-168.