doi:10.3969/j.issn.1672-4348.2017.06.017

一种基于近邻关系的新型离群评估算法

张洁玲

(福建江夏学院 电子信息与科学学院, 福建 福州 350108)

摘要: 针对传统离群点检测算法的局限性进行研究,利用数据对象之间的相邻关系,提出了一种基于密度和距离相结合的离群检测算法,该算法解决了基于距离的离群检测算法不能准确识别局部离群点的问题,有效避免由于稀疏和密集簇过于邻近的而出现离群点误判的情况。通过在人工模拟数据及真实数据集上的实验测试证明改进算法的可行性,该算法能更有效地检测出数据集中的离群对象。

关键词: CDD 算法; k-近邻; 离群评估

中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1672-4348(2017)06-0591-06

A new outlier evaluation algorithm based on the nearest neighbor relationship

Zhang Jieling

(School of Electrical and Information Science, Fujian Jiangxia University, Fuzhou 350108, China)

Abstract: Aiming at the limitations of traditional outliers detection algorithm, a new outlier detection algorithm based on the combination of density and distance was proposed according to the neighboring relationships of the data. The new algorithm solves the problem that the distance-based algorithm cannot identify local outliers, and effectively avoids wrong detection of outliers when the sparse clusters and dense clusters are too close. Experiments on artificial and real datasets prove that the improved algorithm is feasible and it can detect the outliers in the datasets more effectively.

Keywords: CDD algorithm; k-nearest neighbor; outlier evaluation

离群点数据被认为是与其他观测值有较大差别、怀疑由不同机制产生的异常观测值,这些异常数据可能来源于不同的类、自然变异,以及数据测量或收集误差。现实生活中,由于异常事例常常隐藏着有价值和出乎意料的知识,挖掘异常事例及离群数据往往比常规情况更加令人关注。离群点的检测作为目前数据挖掘技术中重要的研究领域,广泛应用于包括信用卡欺诈发现、网络安全入侵检测、生态系统失调预测、犯罪行为发现及预防医疗检查等众多行业领域研究[1]。此外,离群点检测也常用于检测数据集中的异常样本,剔除"脏数据"以提高如聚类和分类计算的数据分析质量。

目前异常检测数据挖掘主要包含基于模型、密度、聚类和距离等技术。基于模型(distribution-based)的检测技术通过估计概率分布的参数来创建数据分布模型,不能很好地与模型相拟合的对象则被判别为异常数据。该技术不适用于数据的统计分布事先未知或没有训练数据可用的情况。基于密度(density-based)的检测技术通过计算每个数据对象的密度评估值,将低密度区域中的数据对象检测为离群点,如 LOF 算法[2]、MDEF[3]、

收稿日期: 2017-08-25
基金项目: 福建省自然科学基金资助项目(2017J01513);福建江夏学院青年科研人才培育基金项目(JXZ2014010)
通讯作者: 张洁玲(1981-),女,福建福州人,讲师,硕士,研究方向:数据库与数据挖掘。

COF[4]和 NLOF[5]等算法,这些算法可适用于具有不同密度区域的数据集,但对初始参数的选择非常敏感。基于聚类(clustering-based)的检测技术通过执行聚类操作,将远离其他簇的小簇标识为离群对象,或使用目标函数来评估对象属于簇的程度,根据离群点评估值隔离异常数据。基于距离即邻近性度量(distance-based)的技术,不需要事先了解数据分布模式,将远离大部分其他对象的对象判定为异常数据,最常见的方法是 Ramaswamy 等人[6]提出的 k-近邻(k-th nearest neighber)离群挖掘算法,当某个数据对象与其近邻数据对象的距离很大时,则认为该数据对象位于稀疏区域并成为离群对象。该算法通过构造 KNN 图,依据每个数据点到 k^{th}最近邻数据点的距离进行排序,将排序列表中距离数值较大的若干数据点视为离群点。

在众多离群挖掘算法中,基于距离和密度的离群点挖掘算法是最有代表性和最有效的挖掘方法。

1 问题的提出

使用 KNN 算法对图 1 所示数据集进行离群点检测,当参数 k 取值为 7 时,数据点 A 和 B 将获得相同的离群评估值,A、B 数据点的 k 近邻距离均为 p,则被视作相同等级的离群点;而当 k 取值为 8 时,该方法误将 A 视作比 B 更强的离群点。从直观角度上不难看出,B 比 A 更趋向于成为离群点。若离群检测算法仅从距离角度考虑离群评估,存在着明显的缺陷。

文献[7]提出的算法将数据点与其 k 个最近邻数据点的距离之和作为判别离群对象的评估值以实现 KNN 算法的改进。由于数据点 B 远离于其他数据点,其与 k 近邻距离之和即离群评估值大于数据点 A 的离群评估值,此改进算法虽然可以解决 KNN 算法对于图 1 所示数据集存在的离群检测问题,然而对于图 2 所示的密度差异较大的簇集,无论采用 k-近邻距离或是求距离之和的方法,都只能检测出全局离群点 D,而无法识别密集簇附近的局部离群点 C,且很可能误将稀疏簇中的大部分数据判定为离群对象。因此,基于距离的离群检测算法适用于密度相近的簇集,对于稀疏各异的簇集,检测错误率将大幅提高。

Breaunig 等人[2]提出的 LOF(local outlier factor)算法通过引入局部离群因子来评估对象相对于邻域的孤立程度,对分布密度相差较大的簇集有较好的检测结果。该算法不从全局的角度考虑数据对象的离群特性,而是计算各数据对象与其最近的 k 个邻居的局部密度比,以获取离群系数。局部密度较大的数据对象具有较小的 LOF 评估值,LOF 评估值较大的数据对象则被判别为离群点。如图 2 所示的数据集中,由于离群点 C 和 D 都与其邻近的数据点有较大的密度差异,若参数选择得当,使用 LOF 算法可以有效识别这两个离群点。然而,对于数据点 G 和 C 来说,由于他们的最近邻密度对象均属于密集簇(近邻数据点的密度相近),且数据对象 C 比 G 更靠近密集簇,根据 LOF 算法,数据对象 G 将获得比 C 更高的离群评估值,即算法认为 G 将比 C 更有可能成为离群点,很明显地,G 只是与密集簇相近的一个稀疏簇中的正常数据对象。与基于距离的离群检测算法对比,LOF 算法虽然能够有效检测出局部离群点,但当两个密度各异的簇集相互靠近时,该算法很容易将稀疏簇与密集簇交界的数据对象误判为离群点。

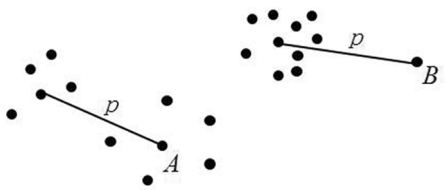

图 1 具有相同离群评估值的两个数据对象 A 和 B(k=7)

Fig.1 Data A and B with equal outlier evaluation values

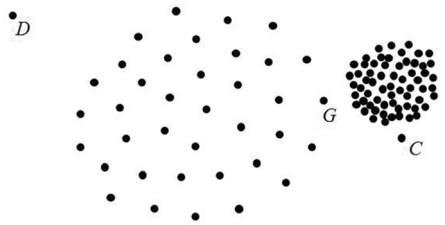

图 2 密度差异较大的两个簇集

Fig.2 Two clusters with great differences in density

综上所述,本文提出一种新的数据对象离群评估算法,使得在密度差异大的簇集中,也能实现正确的离群评估。

2　基于密度和距离相结合的新型离群检测算法 CDD 算法

数据对象成为离群点的概率大小不仅与周围邻居数据点的密度相关，还与其偏离其它数据点的距离大小相关，数据对象周围邻居密度越小、偏离邻居数据点的距离越大，则成为离群点的可能性就越大。综合考虑这两方面因素，提出一种基于密度和距离相结合的新型的离群检测算法 CDD（combination of density based and distance based outlier detection algorithm），过程描述如下：

首先基于 k-近邻构造有向近邻图，其中，对邻域的选择沿用传统经典 k-最近邻域理论[8]。将各数据点作为图的顶点，指向与自己最相近的 k 个邻居数据点，如图 3 所示，然后结合该数据点与近邻数据点的距离以及该数据点与周围数据的密切关系来综合衡量其孤立的程度，获取离群评估值，具体描述如下：

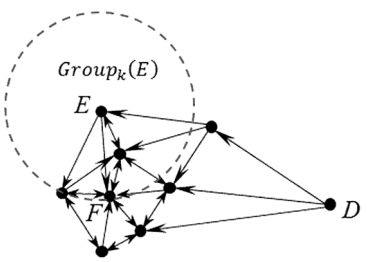

图 3　数据对象指向最邻近的 k 个邻居（k=3）

Fig.3　Data pointing to k nearest neighbors（k=3）

定义 1　对于给定对象的数据集 $X=\{x_1,x_2,\cdots,x_n\}$，$\|x_i-x_j\|$ 为两个数据对象 x_i 和 x_j 的距离，$x_i,x_j\in X$。令 $k\_\mathrm{Dist}(x_i)$ 为数据对象 x_i 与其第 k 个最近数据对象的距离，$\mathrm{Group}_k(x_i)$ 表示数据对象 x_i 的 k 个最近邻数据对象的集合，即：

$$\mathrm{Group}_k(x_i)=\{x_p\in X-\{x_i\}\mid \|x_i-x_p\|\leqslant k\_Dist(x_i)\}.$$

当 k=3 时，图 3 显示出数据集 X 中各数据对象与其 k-最近邻居的关系。其中，每个顶点 x_i 代表一个数据矢量，每条边都指向该对象最近的 k 个邻居，即 $\mathrm{Group}_k(x_i)$ 中包含的 k 个数据对象。

定义 2　令 $\mathrm{Neigh}_k(x_i)$ 为数据集 X 中所有指向 $x_i(x_i\in X)$ 的矢量对象的数量。对于 $\forall x_i\in X$，若 $x_i\in\mathrm{Group}_k(x_p)(p\in[1,n])$，则 $\mathrm{Neigh}_k(x_i)=\mathrm{Neigh}_k(x_i)+1$。

$\mathrm{Neigh}_k(x_i)$ 从全局角度考虑数据对象 x_i 的邻域关系，从很大程度上评估出数据点 x_i 与邻近对象关系的密切程度。其数值越大，x_i 数据对象成为其它数据对象的 k-近邻的情况越多，则表示数据对象 x_i 存在于高密度区域的概率越大，其成为离群点的可能性则越低；反之，对于某数据对象来说，若没有或者极少其他数据对象的 k-近邻指向它，那么该数据对象很可能是离群点。如图 3 所示的数据集中，对于周围数据点密集的 F 点来说，其 $\mathrm{Neigh}_k(F)$ 值为 6，而 H 点远离大部分数据点使其不属于任何对象的 k-近邻集合，$\mathrm{Neigh}_k(H)$ 值为 0。

定义 3　$|\mathrm{Group}_k(x_i)|$ 表示包含数据对象 x_i 的 k-近邻的集合的大小，那么数据点 x_i 与其 k 个最近邻数据对象的平均距离为：

$$\overline{\mathrm{Dist}(x_i,k)}=\frac{\sum_{x_j\in\mathrm{Group}_k(x_i)}\|x_i-x_j\|}{|\mathrm{Group}_k(x_i)|}.$$

定义 4　对于给定对象的数据集 $X=\{x_1,x_2,\cdots,x_n\}$，则每个数据对象的离群评估得分可表示为：

$$\text{Outlier-Score}_k(x_i)=(\mathrm{Neigh}_k(x_i)+1)^{-1}\times\overline{\mathrm{Dist}(x_i,k)}$$

$\overline{\mathrm{Dist}(x_i,k)}$ 实现对数据 x_i 与 k 近邻平均距离的评估，$(\mathrm{Neigh}_k(x_i)+1)^{-1}$ 则将所有数据点的指向邻居的密集程度标准化到[0,1]的范围，该数值越大，则相应的数据对象成为孤立点的可能性也就越大。图 3 中，数据对象 H 处于稀疏区域且无任何邻居指向它，则 $(\mathrm{Neigh}_k(H)+1)^{-1}=1$。数据对象周围密度越稀疏、与 k 近邻数据的距离越大，则该数据的离群评估得分 $\text{Outlier-Score}_k(x_i)$ 越高，本算法将离群评估值较大的前 n 个数据点判断为离群点。

3　实验分析

实验测试环境为 Intel E6300 2.80GHz CPU，4G 内存，Windows 7 操作系统，使用 VC++6.0 编程软件分别对模拟数据和真实数据进行计算，真实数据来源于 UCI 机器学习数据库中的 Lymphography 和 Wisconsin breast Cancer 数据集，实验对比新算法 CDD、KNN 算法和 LOF 算法在数据布局不同的情况下对离群数据点的挖掘准确率。

3.1 模拟数据的离群检测对比

为了较好地展现新算法的优点,实验生成如图 4 所示的一组模拟数据,在 3 个不同密度的簇集(方点)中加入若干离群数据(圆点),采用 KNN、LOF 和 CDD 3 种算法实现离群数据的挖掘计算,最近邻个数 k 取值范围为 2~8,数据挖掘的结果如图 5-图 7 所示,分别显示了在不同的 k 近邻取值情况下 3 种算法获得的离群评估值。图 5-图 7 按离群评估值从高到低显示出前 8 个数据对象,其中,横坐标表示 k 的取值,纵坐标表示相应的离群评估值。

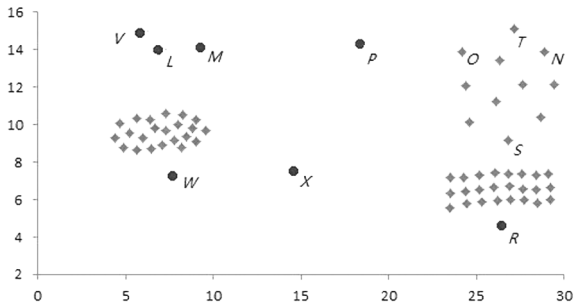

图 4 仿真数据集

Fig.4 Simulation dataset

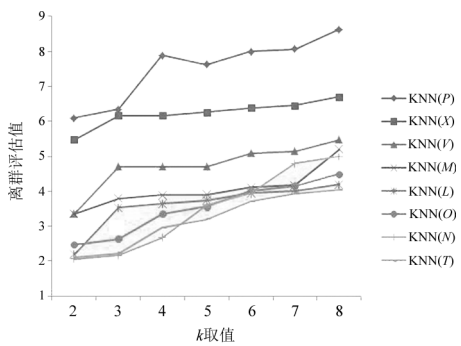

图 5 KNN 算法离群计算结果

Fig.5 Outlier detection results of KNN algorithm

从图 5 可以看出,离群点 W 和 R 并未被 KNN 算法检测出来,由于稀疏簇中的正常数据点(如 O、T 和 N)的离群评估值均大于距离密集簇较近的离群点 W 和 R,KNN 算法并不能很好对密度差异较大的簇集进行离群点的识别,实验结果表明稀疏簇中的所有数据的离群值都高于离群点 W 和 R,仅考虑距离因素的 KNN 算法,错检率较高。图 6 中,由于数据点 R 和 S 的邻近簇相同,而数据点 R 比 S 更靠近邻近簇,LOF 算法则作出了错误

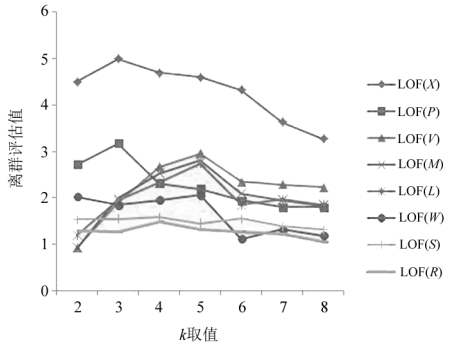

图 6 LOF 算法的离群计算结果

Fig.6 Outlier detection results of LOF algorithm

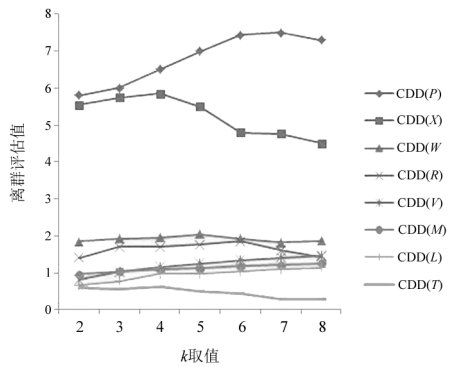

图 7 CDD 算法离群计算结果

Fig.7 Outlier detection results of CDD algorithm

的判断,认为正常数据点 S 的离群程度高于独立数据点 R。当疏密差距较大的两个簇集相近时,稀疏簇边界上的数据点通常会被 LOF 算法误判为离群数据,而新算法 CDD 则解决了由于互相靠近的簇集密度不均衡所带来的边缘数据点离群值错误评估的问题,图 7 的检测结果显示,由 CDD 算法获得的前 7 位数据对象已准确涵盖本实验所有离群点,且它们的离群评估值也客观地体现出了各数据点的离群程度。

另一方面,k-近邻算法、LOF 算法对参数 k 的取值具有较大的依赖性,为了获取高质量的离群检测结果,需要多次调整参数 k 的取值,并根据现实经验来判断最佳的离群检测结果。如图 6 所示,随着 k 值的变化,LOF 算法下的各数据点的离群评估值也有较大的波动,k 值的选定在一定程度上影响了离群检测结果的准确性和稳定性,而 CDD 算法得出的各数据点的离群评估值的大小排序则相对稳定(见图 7),离群值走势不随 k 值的变化剧烈波动,说明 CDD 算法有效弱化了参数

k 对离群检测结果的影响，提高离了群点检测的稳定性和准确率。

3.2　真实数据集离群检测对比

3.2.1　Lymphography 数据集

Lymphography 数据集包含了来源于前南斯拉夫肿瘤学研究所的 148 条淋巴系造影术数据，每个数据包含 18 个分类属性。数据被划分为 4 个簇，其中两个簇分别包含了 81 和 61 个数据对象，而另外两个小簇分别包含 2 个和 4 个数据对象，本实验将占比较小的两个小簇中的 6 个数据对象作为歧异值进行离群检测，使用 KNN、LOF 和 CDD 3 种算法分别对该数据集实现数据挖掘并进行离群检测结果对比。各算法检测出的离群点中，正确的离群点所占比率越高，则表示该算法性能越好[9]。

表 1 列出了当最近邻居个数 k 取值为 5～10 时，3 种算法获取的离群评估平均值排名前 n 位的离群数据的检测率对比，其中，检测率指检测到正确的离群数据量与离群数据总量的比值。从表 1 可以看出，CDD 算法检测获取的前 6 个离群数据对象中已包含 5 个正确的离群对象，检测获取的前 8 个数据对象即涵盖了出所有离群对象，而 KNN 算法和 LOF 算法获取的前 8 个离群对象中，离群检测率仅为 66.7% 和 88.3%；LOF 算法须提取前 10 个离群数据才能找出所有离群点，KNN 算法离群检测率更低，需要提取前 12 个离群数据。由此可见，在 Lymphography 数据集上，CDD 算法在离群检测率方面明显优于其他两种算法。

3.2.2　Wisconsin breast Cancer 数据集

Wisconsin breastCancer 乳腺癌数据集包含 569 组数据，每个数据对象包含 32 个分类属性，其中良性肿瘤细胞特征有 357 例，恶性肿瘤细胞特征有 212 例，本实验将选取 357 个数据对象良性肿瘤细胞和部分恶性肿瘤细胞（32 个数据对象），共 389 个数据对象作为检测样本，使用 3 种算法分别对其进行离群数据对象（恶性肿瘤特征数据）的检测，算法选取 k 值为 10～15，测试对比结果如表 2 所示。

表 1　3 种算法在 Lymphography 数据集中的运行结果对比

Tab.1　Comparison of results of three algorithms in Lymphography dataset

检测获取的前 n 个离群数据	KNN 算法		LOF 算法		CDD 算法	
	正确个数	检测率/%	正确个数	检测率/%	正确个数	检测率/%
6	4	66.7	5	83.3	5	83.3
8	4	66.7	5	83.3	6	100
10	5	83.3	6	100	6	100
12	6	100	6	100	6	100

表 2　3 种算法在 Wisconsin breast Cancer 数据集中的运行结果对比

Tab.1　Comparison of results of three algorithms in Wisconsin breast cancer dataset

检测获取的前 n 个离群数据	KNN 算法		LOF 算法		CDD 算法	
	正确个数	检测率/%	正确个数	检测率/%	正确个数	检测率/%
43	26	81.3	24	75.0	30	93.8
46	27	90.0	24	81.3	32	100
52	29	90.6	26	81.3	32	100
57	32	100	31	96.9	32	100
63	32	100	32	100	32	100

分析实验结果可以看出，KNN 算法和 LOF 算法分别需要检测出前 57 个和 63 个离群评估值较大的数据对象，才能找到所有恶性肿瘤离群数据。算法 CDD 检测出的前 46 个数据对象中已经包含所有 32 个离群数据，而 KNN 和 LOF 算法检测到的前 46 个数据中，仅分别包含 27 个和 24 个正确

离群数据。因此,对于 Wisconsin breast Cancer 数据集,CDD 算法的离群检测有效性还是优于 KNN 和 LOF 算法。

4 结语

本文提出的改进算法 CDD 通过构造有向邻近图,从簇中各数据对象的近邻密度疏密以及分布距离大小两方面进行综合考量,有效量化各数据对象的离群强弱程度并获得离群对象队列。通过在不同的数据集上的实验对比,结果表明新型算法的离群点检测准确度明显高于 k-近邻算法及 LOF 算法,且有效弱化了离群算法对初始参数 k 取值的依赖性。

在此改进基础上,针对不同数据集的特征,进一步研究算法以降低时间复杂度、提高算法运行的效率,是今后要继续研究的方向。

参考文献:

[1] He Zengyou, Xu Xiaofei, Deng Shengchun. Discovering cluster-based local outliers[J]. Pattern Recognition Letters,2003, 24(9):1641-1650.

[2] Breunig M, Kriegel H P, Ng R, et a1. LOF:Identilying densityhased local outliers[C]//. Proc of the ACM SIGMOD International Conference on Management of Data,2000:93-104.

[3] Papadimitrious S, Kitawaga H, Gibbons P B, et al. LOCI: Fast outlier detection using the local correlation integral[C]//. Proc of the 19th International Conference on Data Engineering,2002:315-326.

[4] Tang J, Chen Z, Fu A, et al. Enhancing effectiveness of outlier detections for low-density pattern[J]. Proceedings of the 6th PAKDD,2002,2236:535-548.

[5] 王敬华,赵新想,张国燕,等.NLOF:一种新的基于密度的局部离群点检测算法[J].计算机科学,2013,40(8):181-185.

[6] Ramaswamy S, Rastogi R, Kyuseok S. Efficient algorithms for mining outliers from large data sets[C]//. Proc of the ACM SIGMOD International Conference on Management of Data,2000:427-438.

[7] Angiulli F, Pizzuti C. Outlier mining in large high-dimensional data sets[J]. IEEE Trans. Knowledge and Data Eng,2005, 2(17):203-215.

[8] 范小刚,朱庆生,万家强.基于 K-近邻树的离群检测算法[J].计算机应用研究,2015,32(3):669-673.

[9] 周世波,徐维祥.一种基于偏离的局部离群点检测算法[J].仪器仪表学报,2014,35(10): 2293-2297.

(特约编辑:黄家瑜)