文章编号: 1001-0920(2019)07-1433-08 DOI: 10.13195/j.kzyjc.2017.1649

基于邻域链的数据异常点检测

梁绍一, 韩德强†

(1. 西安交通大学 电信学院, 西安 710049;
2. 中国电子科技集团公司 航天信息应用技术重点实验室, 石家庄 050081)

摘 要: 异常点检测(outlier detection)领域的大量研究都集中于一类“基于密度的”方法,这类方法能够克服许多传统异常点检测方法的缺陷,但仍大多使用基于几何距离的方式进行数据点局部密度的估计,导致在某些情况下反直观结果的出现. 针对该问题,用一种基于邻域链的方法取代传统方法进行局部密度的估计,设计新的异常点检测方法. 实验结果表明,对比经典的基于密度的异常点检测方法LOF(Local outlier factor)以及几种基于LOF的改进方法,所提出的方法能够更加准确地区分正常和异常数据点,避免反直观结果的出现.

关键词: 数据挖掘; 异常点检测; 局部密度; 局部异常因子; 欧氏距离; 邻域链

中图分类号: TP181 **文献标志码:** A

Outlier detection based on neighborhood chain

LIANG Shao-yi, HAN De-qiang†

(1. College of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. CETC Key Laboratory of Aerospace Information Applications, China Electronics Technology Group Corporation, Shijiazhuang 050081, China)

Abstract: Many research works in the area of outlier detection are focused on the so called "density-based" methods. Such kind of methods can counter-act many drawbacks of the traditional outlier detection methods. However, most existing density-based methods use geometric-distance-based approaches to estimate the data point's local density, which leads to incorrect results in certain cases. To resolve the problem, the traditional local density estimation method is substituted by a neighborhood-chain-based method, and a new outlier detection method is proposed. Compared to the local outlier factor (LOF) and several of related modifications, the proposed one can find the outliers more accurately.

Keywords: data mining; outlier detection; local density; local outlier factor; Euclidean distance; neighborhood chain

0 引 言

对于数据挖掘领域的许多应用(例如欺诈检测、电子商务中的犯罪活动检测、机器故障检测等)而言,寻找数据中的“异常点”往往比寻找数据中正常和大量出现的模式更有意义[1]. 近年来,出现许多异常点检测方法,这些方法通常被分为4类[2-3]:基于分布(Distribution-based)的方法[4-5]、基于聚类(Clustering-based)的方法[6-7]、基于距离(Distance-based)的方法[8-9]以及基于密度(Density-based)的方法[10]. 其中基于距离和基于密度的方法最为常见.

在早期的基于距离的方法中,如果某个数据点与其所属数据集中大部分数据点距离都较远,则该数据点将被认为是一个异常点. 这一类方法的代表为文献[8-9]等. 在后期的发展中[11-14],许多基于距离的方法已不再通过一个固定的距离阈值来判别异常点,而是通过计算某个点到其k近邻点的距离(k-nn distance)来确定该点属于异常点的程度(outlierness). 基于距离的异常点检测方法对异常点的定义非常直观,也易于实现,因而应用广泛. 然而,当数据集中的点所形成的类团具有不同的密度时,这一类方法无法保证异常点的准确判别.

相比于基于距离的异常点检测方法,基于密度的方法在处理具有不同密度分布的数据时有着明显的优势. 这一类方法将异常点定义为那些与其邻域点的局部密度具有很大差异的数据点. 局部异常因子[10](Local outlier factor, LOF)方法就是一种经

收稿日期: 2017-12-05; **修回日期:** 2018-02-25.

基金项目: 国家自然科学基金项目(61573275, 61671370); 国家973计划项目(2013CB329405); 陕西省科技计划项目(2013KJXX-46); 中央高校基本科研业务费专项资金项目(xjj2016066); 中国博士后科学基金项目(2016M592790); 中国电子科技集团公司航天信息应用技术重点实验室高校合作课题项目(KX172600034).

†通讯作者. E-mail: deqhan@mail.xjtu.edu.cn.

典的基于密度的异常点检查方法. 在LOF中,每个被测试的点将使用该点到它的各k邻域点的可达距离计算所谓的局部可达密度(Local reachability density, LRD). 同时,被测点的k邻域中的每个点也将根据它们的邻域点计算各自的局部可达密度. 被测点的局部可达密度与其各个邻域点的局部可达密度相差越大,被测点就越有可能是一个异常点. 基于LOF方法的基本思想已经出现许多改进和变种方法. 例如,Schubert等[15]将LOF方法中的可达距离替换为k-nn距离,进而得到一种更简单高效的数据点局部密度估计;Jin等[16]提出的INFLO(Influenced outlierness)方法同时使用被测点的邻域点以及反邻域点(Reverse nearest neighbors)来估计被测点的局部密度,以更好地处理数据集中含有密度差异较大类团的情况;Kriegel等[17]提出的LoOP(Local outlier probabilities)方法基于均方距离构造了一种更为鲁棒的局部密度估计方式;Zhang等[18]使用被测点到其近邻点距离的均值与各近邻点之间距离均值的比值来定义被测点属于异常点的程度;杨茂林等[19]使用所谓的"相异度"定义两数据点间的可达距离并使用剪枝的方法基于相异度矩阵寻找异常点. 有关上述方法更为详细的介绍和对比参见文献[20-21].

在基于密度的异常点检测方法中,被测点的局部密度估计对最终结果起决定性的作用. 上述各方法以不同方式估计局部密度,各有侧重,但它们有一个共同点,即都是基于数据点之间的几何距离(如欧氏距离)构造的. 本质上,这些方法都是通过被测点周围的邻域点到被测点之间某种距离度量的平均值的倒数(或其他自变量与变量具有反向变化趋势的函数)来量化被测点局部密度的. 被测点周围的邻域点到被测点的平均距离越小,换言之,被测点周围一定距离范围内邻域点数目越多,则被测点局部密度越大. 在这种度量方式下,被测点的局部密度仅由其邻域点与被测点的距离决定,而忽略了各邻域点及被测点之间的相互关系(比如邻域点在被测点周围分布的形状等). 这种信息的忽视,导致在某些情况下无法准确地判别异常点.

针对上述问题,本文提出一种基于LOF的改进方法. 该方法使用之前提出的一种数据点间基于邻域链的相似性度量[22]取代LOF中估计局部密度时使用的数据点间基于几何距离得到的所谓"可达距离". 不同于几何距离,数据点间基于邻域链的相似性度量(Closeness measure based on neighborhood chain, CMNC)在量化两个数据点间的相似性时,不仅考虑了两个数据点在特征空间中位置的远近,也同时考虑了这两点与其邻域点的相互关系. 因此,即使正常点与异常点周围一定范围内的邻域数目相同,使用基于CMNC构造的局部密度估计方法,也能够利用被测点周围邻域不同的分布方式,区分正常与异常数据点. 实验结果表明,对比LOF方法及现有的几种基于LOF的改进方法,本文提出的新方法能够更准确地判别异常数据点.

1 局部异常因子异常点检测方法

本节将介绍局部异常因子(LOF)方法的基本思想及其缺陷.

1.1 LOF方法的基本思想

首先介绍LOF中的一些基本定义.

k-距离(k_dist): 数据点o的k-距离$k_\mathrm{dist}(o)$是一个能够满足下面两个条件的距离: 1) 至少有k个点$\hat{o}$,使得$d(o,\hat{o}) \leqslant k_\mathrm{dist}(o)$; 2) 最多有$k-1$个点$\hat{o}$,使得$d(o,\hat{o}) < k_\mathrm{dist}(o)$.

k-可达距离($\mathrm{reach\_dist}_k$) : k-可达距离带有方向信息. 点o到点p的k-可达距离即为点o 的k-距离与点o到点p的欧氏距离中较大的那一个,其定义为

$$\mathrm{reach\_dist}_k(p \leftarrow o) = \max\{k\_\mathrm{dist}(o), d(p,o)\}. \quad (1)$$

局部密度(LRD): 一个点p的局部密度定义为

$$\mathrm{LRD}(p) = 1\Big/\frac{\sum\limits_{o\in \mathrm{kNN}(p)} \mathrm{reach\_dist}_k(p\leftarrow o)}{|\mathrm{kNN}(p)|}. \quad (2)$$

其中: $|\mathrm{kNN}(p)|$表示点p的k-距离邻域点数目,一般而言,$|\mathrm{kNN}(p)| \geqslant k$.

利用上述定义,LOF方法对数据集中的每个点计算一个局部离群因子LOF. LOF表示某个点周围的k-距离邻域点(即到某个点的距离小于该点的k-距离的所有点)的局部密度(LRD)与该点局部密度的比值的均值. 假设p是数据集中的一个点,p的局部离群因子LOF定义如下:

$$\mathrm{LOF}_k(p) = \frac{1}{|\mathrm{kNN}(p)|}\sum_{o\in \mathrm{kNN}(p)}\frac{\mathrm{LRD}_k(o)}{\mathrm{LRD}_k(p)}. \quad (3)$$

由LOF方法的定义可以看出,某个点的LOF值越大,表示该点的局部密度越小于其邻域点的局部密度,该点就越有可能是一个异常点.

1.2 LOF方法的缺陷

在LOF中,被测点的密度是由该点的k-距离邻域点,以及这些邻域点各自的k-距离邻域点决定的,因而被称作"局部"密度. 相比于传统异常点检测方法,LOF能够较好地适用于数据点所形成的类团密

度不一致的数据集. 但基于LOF中的局部密度定义, 某些情况下, 该方法仍然不能正确地判别异常点.

如图1所示, 数据点周围的圆圈半径表示由LOF方法得到的该点的LOF值. 圆圈越大, 该点越可能是异常点.

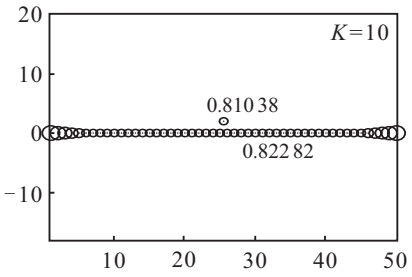


图 1　LOF方法无法正确地判别异常点

从图1中可以看到, 异常数据点的LOF值与正常点的LOF值相差无几, 而一些距离较近的正常数据点的LOF值反而会大于异常点的LOF值, 并且一些正常数据点(例如处于两端边缘的数据点)的LOF值会异常大. 同时发现, 当将LOF方法中的参数k设置得更大时, 该现象更加明显. 这也说明LOF方法对邻域大小k的选择较为敏感.

这种问题是由LOF中局部密度估计方法的缺陷导致的. 在LOF中, 某个点的局部密度实际上是由该点周围一定距离半径内邻域点的数目决定的, 而忽略了被测点与这些邻域点之间的相互关系. 该问题在较为复杂的数据集中体现得也比较明显. 如图2所示, 在doublemoon数据集中, 人为地在数据集末尾加入一个异常点, 即图中实线箭头所指. 使用LOF算法得到的该离群点的LOF值(即点周围圆圈半径大小)并没有比其他正常点显著地大. 而图2中很多边缘点(虚线星形箭头所指)的LOF值都与真正的异常点非常接近, 甚至超过了真正异常点的LOF值. 此时无法通过数据点的LOF值找到真正的异常点.

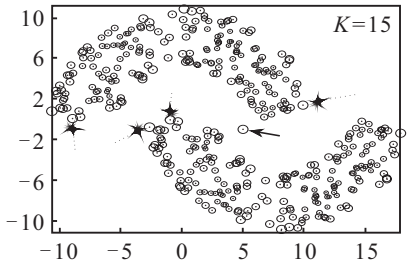


图 2　LOF方法的问题在较为复杂数据集中的体现

doublemoon数据集中各点的LOF值如图3所示. 图3中箭头所指的真正异常点(即数据集末尾人为添加的点)的LOF值并非最大, 此时无法通过数据点的LOF值找到真正的异常点.

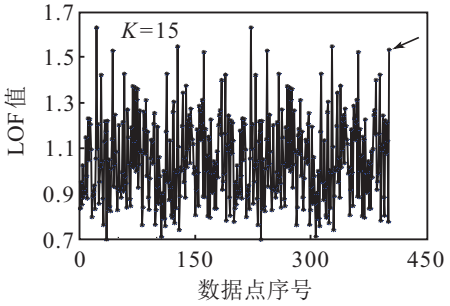


图 3　doublemoon数据集中各点的LOF值

1.3　导致LOF方法缺陷的本质原因及改进方法

在LOF算法中, 点的局部密度是基于该点邻域内其他点到该点的k可达距离进行计算的, 且可达距离是基于几何距离来定义的. 然而, 两个点之间的几何距离(如欧氏距离)仅能够反映出这两点在其特征空间中的近似程度, 且这种近似程度仅与这两点有关而与它们周围的其他点无关. 在很多情况下, 两点间的几何距离较近并不意味着它们同属于高密度区域, 而反之亦然.

单纯以基于几何距离的方式进行数据点局部密度的估计实际上反映的是被测点周围一定距离半径内其他点数目的多少, 而忽略了被测点与其周围邻域点的相互关系, 这导致了上述问题的产生. 为了更加准确地对数据点的局部密度进行估计, 并利用局部密度的差异更好地判别异常点, 本文使用之前提出的一种基于邻域链的数据点间相似性度量(CMNC)[22]来替换LOF方法中数据点间的可达距离. 在CMNC度量下, 两个点的“远近”取决于这两点是否是“紧密连通”的, 这一特点将对解决上述LOF方法的缺陷提供帮助.

2　基于邻域链的数据点间相似性度量

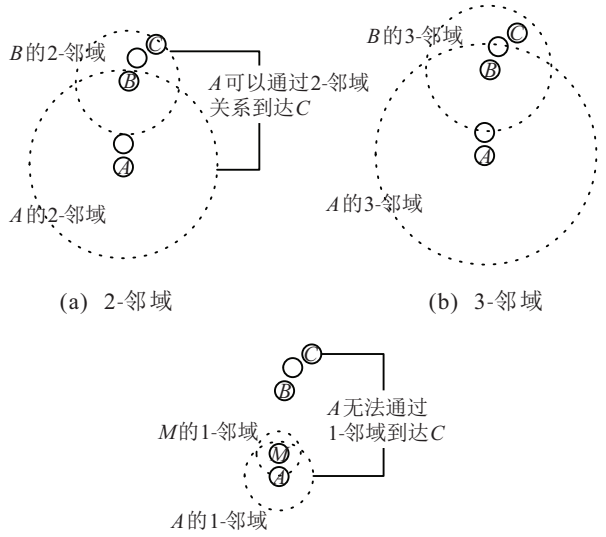
基于邻域链的数据点间相似性度量(CMNC)[22]是一种数据点间成对相似性度量. CMNC通过量化在两个点之间建立一条邻域链的困难程度来度量这两点间的相似性.

2.1　数据点间邻域链的建立

数据点间的邻域链由一系列的点组成, 其中包含一个起点和一个终点. 在链中的每一个点(除去起点)都是其前一个点的k-近邻成员. 下面用一个例子来说明邻域链的建立.

在图4中, A, M, B, C来自于同一个数据集. 显然在图4(a)中, B在A的2-近邻内, 同时C也在B的2-近邻内. 因此, 通过一个中间点B, 一条基于2-邻域关系的从A到C的链条就形成了. 如图4(b)所示, B同样也在A的3-近邻(或更大的近邻范围)内, 同时

C显然也属于B的3-近邻(或者对应的更大的近邻范围). 这意味着也可以通过3-邻域关系建立从A到C的链条,但显然,基于2-邻域建立的邻域链所花费的“代价”更低. 实际上,在图4中,2-邻域是建立从A到C邻域链所需的最小邻域范围(图4(c)展示了基于1-邻域无法建立从A到C的邻域链). 因此,建立A到C的邻域链的最小邻域数为2.



(a) 2-邻域 (b) 3-邻域

(c) 1-邻域

图4 建立点A到点C邻域链

下面给出邻域链的定义. 假设$\Omega \subseteq R^n$是一个数据集且$A, C \in \Omega$. 令k为一个正整数,其使得一串在Ω中的点$\{A, M_1, M_2, \cdots, M_q, C\}$满足

$$\begin{cases} M_1 \in \text{Neighbors}(A, k); \\ M_i \in \text{Neighbors}(M_{i-1}, k),\ 1 < i \leqslant q; \\ C \in \text{Neighbors}(M_q, k). \end{cases} \tag{4}$$

其中$\text{Neighbors}(\cdot, k)$表示包含某个数据点及其$k$邻域点的集合. 如果这样的正整数$k$存在,则从$A$到$C$的邻域链可以建立,且定义

$$R(A, C) = \min(k) \tag{5}$$

为建立从A到C邻域链的最小邻域数.

2.2 通过邻域链度量两点间的相似性

通过两个量来综合度量两点间(假设为A和C)的相似性,分别是“邻域可达性代价”($\text{NRC}(A, C)$)及“邻域可达性跨度”($\text{NRS}(A, C)$).

邻域可达性代价被定义为:分别以A和C为起点,建立到达对方的邻域链所需的最小邻域数中较大值的函数,即

$$\text{NRC}(A, C) = f(\max(R(A, C), R(C, A))), \tag{6}$$

其中$f(\cdot)$一般取指数函数.

邻域可达性跨度定义为:分别以A和C为起点,建立到达对方的邻域链中相邻两点间距离的最大值,即

$$\text{NRS}(A, C) = \max\{S(A, C), S(C, A)\}. \tag{7}$$

其中:$S(A, C)$表示从A到C的邻域链中,相邻两点间距离的最大值.

通过以上两个量化值定义两点间基于邻域链的相似性(CMNC),即

$$\text{CMNC}(A, C) = \frac{1}{\text{NRC}(A, C) \cdot \text{NRS}(A, C)}. \tag{8}$$

2.3 CMNC度量与欧氏距离的对比

直观来看,欧氏距离度量的是两个数据点特征空间上的距离,这两个点的远近仅与它们在特征空间中的位置有关,而CMNC度量(CMNC并不是严格意义上的距离,其不满足三角不等式)考察的是两个数据点之间是否是通过其他数据点“紧密相连”的. 在CMNC度量下,如果两个数据点是紧密相连的,则它们是相“近”的,而如果两个数据点之间存在密度间隔,则它们相距较“远”.

在聚类问题中,人类通过直观观察认为两个点属于同一类,并不一定是因为这两点间的欧氏距离较短. 例如图5所示,在(实心与空心点)两个类团中,A与B之间的欧氏距离远小于A与C之间的距离,但很明显,A与C更可能属于同一类团. 因为根据流形假设[23],A与其局部邻域内的点(在这里可以认为是最近邻点)应属于同一类,而该最近邻点又与其局部邻域内的点属于同一类. 这种关系可以(沿着图中实心点)传导到C,即A与C应属于同一类. 在该聚类问题下,A与C相比A与B应该更“近”,或相似性更高. 在CMNC度量下,紧密相连的A和C将具有很高的相似性,而A与B由于被隔断,将被赋予很低的相似性. CMNC能够表达出两点间连通性(或者说两点在同一流形上)这一特点是其与欧氏距离的主要区别.

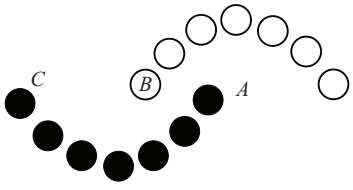


图5 聚类中CMNC与欧氏距离对比示例

3 基于邻域链的C-LOF方法

在1.2节中讨论了LOF方法的缺陷. 这些缺陷是由LOF方法中基于几何距离定义的“可达距离”导致的. 因此,本节使用基于邻域链的相似性度量CMNC

替换LOF方法中可达距离的定义. 由CMNC的定义可以看出,在该度量下,对于存在密度间隔的两个点,即使它们之间的几何距离很近,它们的相似性也较低. 而对于紧密连通的两点,即使它们之间的几何距离较远,在CMNC度量下,它们的相似性也较高. 这一特点使得CMNC较适合用来解决前文所述的LOF方法的问题.

3.1 方法定义

基于邻域链的LOF方法(C-LOF)定义为

$$C\text{-LOF}_k(p)=\frac{1}{|C\text{kNN}(p)|}\sum_{o\in C\text{kNN}(p)}\frac{\text{CLRD}_k(o)}{\text{CLRD}_k(p)}. \tag{9}$$

其中: $|C\text{kNN}(p)|$表示点p的在CMNC度量下的k邻域数, $\text{CLRD}_k(p)$表示p的局部密度,其定义为

$$\text{CLRD}_k(p)=1\Big/\frac{\sum_{v\in C\text{kNN}(p)}\text{CMNC}(v,p)}{|C\text{kNN}(p)|}. \tag{10}$$

与LOF方法相同,某个数据点的C-LOF值越大,表明该点越有可能是一个异常点.

下面以伪代码的形式给出C-LOF方法的具体计算步骤。

算法输入: 数据集D,数据集样本数n,邻域大小K,异常点个数r.

算法输出: r个异常点.

算法开始:

For $i=1:n$

计算数据集中所有点到数据点p_i的CMNC度量值.

找到与p_i在CMNC度量下最近的K个点,计算它们与p_i的平均CMNC值.

依据式(1)计算p_i的局部密度$\text{CLRD}_k(p_i)$.

For $j=1:K$

依据式(10)计算p_i的K邻域中每个点o_j的局部密度$\text{CLRD}_k(o_j)$.

End

依据式(9)计算p_i的C-LOF值.

对数据集中所有点按照其C-LOF值由大到小排序,并选出最前面的r个点输出.

算法结束.

3.2 运算复杂度分析

对于数据集中的每一个点,为了计算其C-LOF值,需要寻找其K邻域点以及这些邻域点各自的K邻域点. 因此,对于一个点,得到其C-LOF值需要进行$K+1$次邻域查找,且每次邻域查找需要计算n次CMNC度量. 这是C-LOF方法最主要的运算复杂度来源,相比这部分运算,其他运算(例如对各点C-LOF值排序)的耗时是可以忽略的. 在文献[22]中,详细讨论了CMNC度量的运算复杂度,并得到结论,即在最坏情况下,计算两点间的CMNC度量运算复杂度为$O(n(n-1)^2)$. 因此,在最坏情况下,计算C-LOF值的运算复杂度为$O((K+1)n^2(n-1)^2)$.

4 实验及分析

4.1 基于人造数据集的实验

在本节中,首先使用与图1中完全相同的例子来验证C-LOF方法在传统LOF方法失效时是否能够正确地判别异常点. 实验结果如图6所示,图6中圆圈半径的大小表示某个数据点的C-LOF值. 从图6中可以看到,相比于传统LOF方法,C-LOF能够明显地区分出正常点与异常点,且能够克服端点处数据出现异常的问题. 在与图2完全相同的例子中,C-LOF方法得到的结果如图7和图8所示,可以看到,使用C-LOF方法时,数据集中的异常点与其他正常点的C-LOF值相差非常大,可以被轻易地判别出来(即图中箭头所指数据点).

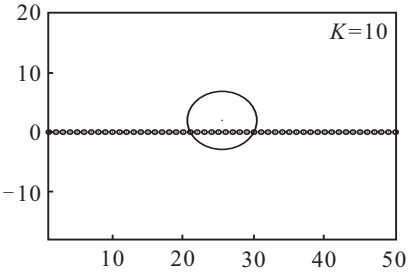


图6　C-LOF方法区分正常与异常点

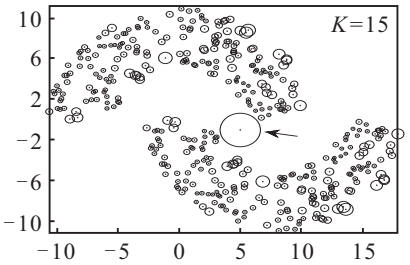


图7　C-LOF方法在“doublemoon”数据集中的结果

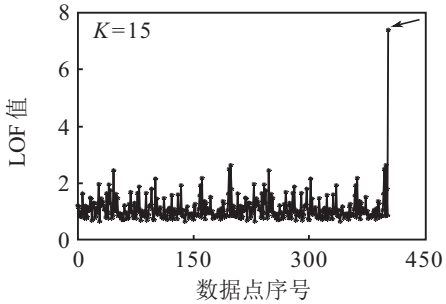


图8　“doublemoon”数据集中各数据点的C-LOF值

4.2 基于真实数据集的实验

下面将在实际数据集上对LOF方法、C-LOF方法以及两种较新的基于LOF的改进方法LoOP[17]和LDOF[18]的性能进行量化评估.

本节使用的数据均来自于UCI数据集. UCI数据集被广泛用于评估分类方法的实验,但由于异常点一般被认为是稀少且分散的,直接将UCI数据集用于异常点检测方法的评估并不合适. 因此,参考文献[2,20]中的做法,首先对所使用的数据集进行降采样处理. 降采样处理将对所使用数据集的某一类进行随机采样,保留该类一定数目的数据点(即采样点)并删除该类其他数据点. 在采样率较低时,随机采样点相比于数据集中的其他点可以看作是分散的异常点. 由于选择的采样类别与采样率不同,基于同一个原始数据集可以生成多个含有不同数目不同位置异常点的“变种”数据集. 对于异常点检测方法而言,这些“变种”数据集可以视为各不相同的数据集. 在实验中基于6个原始数据集(iris, wine, liver, sonar, ionosphere, wpbc)生成共24个变种数据集,表1列出了这些数据集及它们的生成方法.

表1 实验所用数据集及生成方法

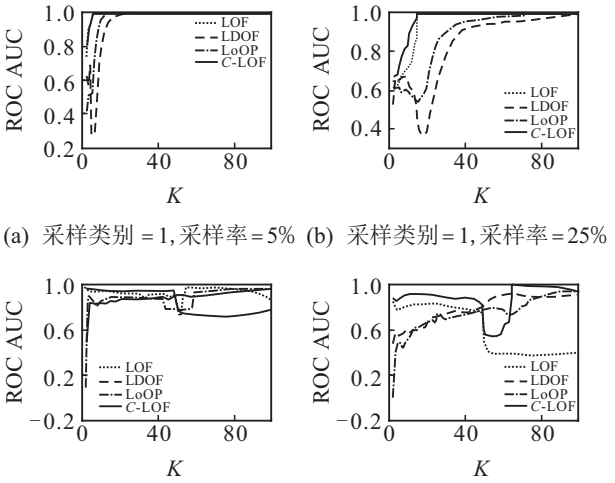
数据集	采样类别	采样率/%	异常点数
iris_1	1	5	3
iris_2	1	25	13
iris_3	3	5	3
iris_4	3	25	13
wine_1	1	5	3
wine_2	1	15	15
wine_3	3	5	2
wine_4	3	25	12
liver_1	1	5	7
liver_2	1	25	36
liver_3	2	5	10
liver_4	2	25	50
sonar_1	1	5	5
sonar_2	1	25	25
sonar_3	2	5	6
sonar_4	2	25	28
ionoshpere_1	1	5	12
ionoshpere_2	1	25	56
ionoshpere_3	2	5	6
ionoshpere_4	2	25	32
wpbc_1	1	5	8
wpbc_2	1	25	38
wpbc_3	2	5	3
wpbc_4	2	25	12

在实验中,选取了一种在评估异常点检测方法时较为常用[2,20-21]的ROC(Receiver operating characteristics)曲线AUC(Area under curve)值评估方法. ROC平面的纵坐标为“真阳率”(TPR),即真正的异常点出现在所确定异常点集合中的概率. ROC平面的横坐标为“假阳率”(FPR),即正常点出现在所确定的异常点集合中的概率.

实验中被评估的各异常点检测方法的输出都是数据集中每一个点的“异常度”分数,且该分数越高,表示该点越可能是一个异常点. 将某个方法的输出按照“异常度”分数由高到低进行排序,并设置一个阈值(top-n)确定前n个点为数据集中的异常点. 当阈值以及该阈值下的异常点确定后,就可以在ROC平面画出一个点.

当遍历阈值参数top-n所有可能的取值后,就能够得到ROC平面中的一条曲线,显然,依据ROC平面横纵坐标的意义,该曲线下面积(AUC)越大,代表被评估方法的性能越好. 因此,固定一个邻域大小k就能够得到一条ROC曲线,并计算出该曲线对应的一个ROC AUC值. 当选取不同的邻域大小k时,就能得到不同的ROC曲线,即可得到由对应的ROC AUC值连成的曲线. 显然,在不同的k下,ROC AUC曲线值越接近1,代表该方法性能越好.

图9～图14展示了在各个测试数据集中4种异常点检测方法在不同k值下的ROC AUC曲线. 可以看出,在大部分k值下,C-LOF方法得到的ROC AUC值是最高的.

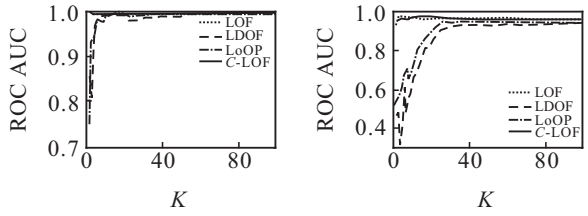


(a) 采样类别=1,采样率=5% (b) 采样类别=1,采样率=25%

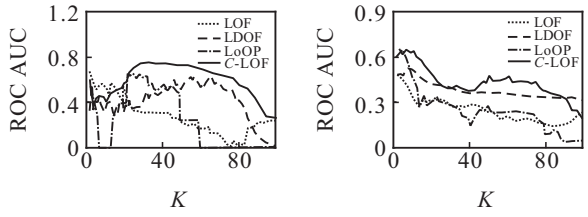
(c) 采样类别=3,采样率=5% (d) 采样类别=3,采样率=25%

图9 各方法在选取不同k值时的ROC AUC值(基于iris的4个变种数据集)

通过实验可以看出,在所选择的人造及真实数据集上,C-LOF方法能够正确地找出传统LOF方法无法判别的异常点,并且依据ROC AUC量化指标可知其在绝大部分实验数据集上表现出了更好的性能以及对所选择邻域大小更低的敏感度.

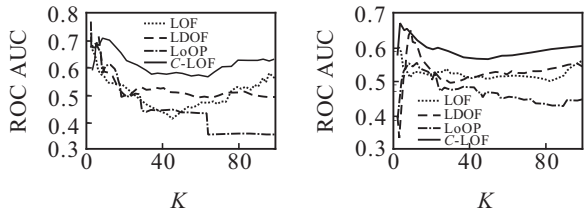


(a) 采样类别=1, 采样率=5%　(b) 采样类别=1, 采样率=25%

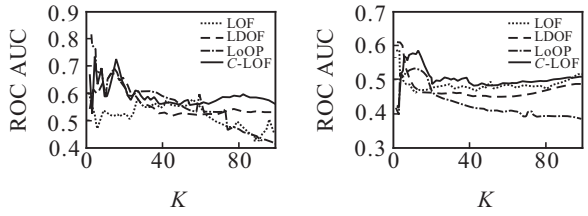


(c) 采样类别=3, 采样率=5%　(d) 采样类别=3, 采样率=25%

图 10　各方法在选取不同k值时的ROC AUC值
(基于wine的4个变种数据集)

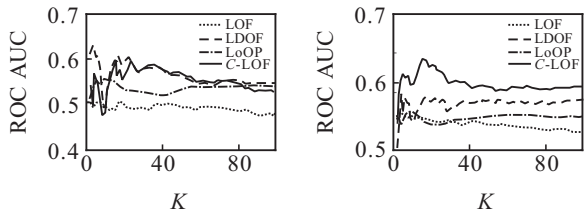


(a) 采样类别=1, 采样率=5%　(b) 采样类别=1, 采样率=25%

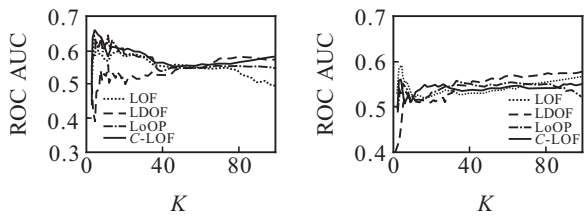


(c) 采样类别=2, 采样率=5%　(d) 采样类别=2, 采样率=25%

图 11　各方法在选取不同k值时的ROC AUC值
(基于sonar的4个变种数据集)

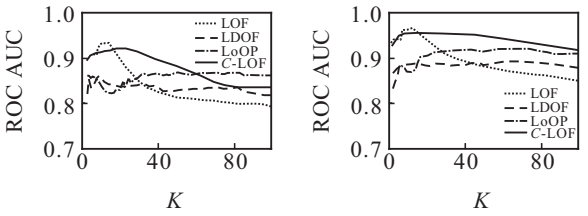


(a) 采样类别=1, 采样率=5%　(b) 采样类别=1, 采样率=25%

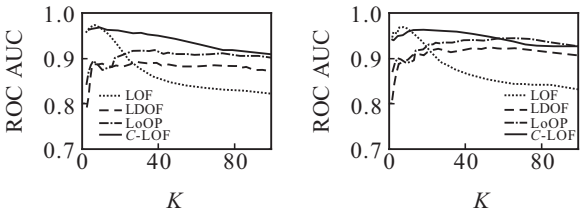


(c) 采样类别=2, 采样率=5%　(d) 采样类别=2, 采样率=25%

图 12　各方法在选取不同k值时的ROC AUC值
(基于liver的4个变种数据集)

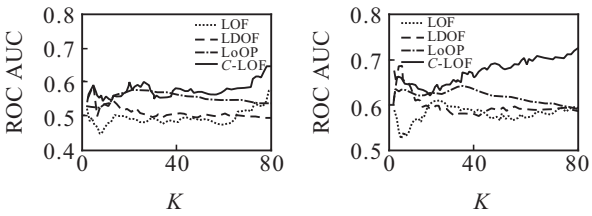


(a) 采样类别=1, 采样率=5%　(b) 采样类别=1, 采样率=25%

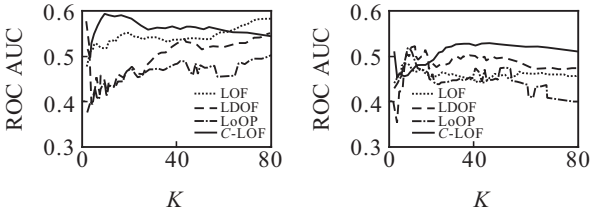


(c) 采样类别=2, 采样率=5%　(d) 采样类别=2, 采样率=25%

图 13　各方法在选取不同k值时的ROC AUC值
(基于ionosphere的4个变种数据集)



(a) 采样类别=1, 采样率=5%　(b) 采样类别=1, 采样率=25%



(c) 采样类别=2, 采样率=5%　(d) 采样类别=2, 采样率=25%

图 14　各方法在选取不同k值时的ROC AUC值
(基于wpbc的4个变种数据集)

5　结　论

本文提出了一种基于邻域链的改进LOF方法, 即C-LOF. 该方法使用基于邻域链的方式重新定义了传统LOF方法中的可达距离, 解决了传统LOF方法在某些情况下无法正确找到异常点, 以及LOF值对所选邻域大小较为敏感等问题. 在实验中, 采用了数个人造及真实数据集对C-LOF方法及其他几种对比方法进行了测试, 并使用ROC AUC值对各被测方法进行了量化评估. 实验结果表明, 本文提出的C-LOF方法能够有效地找出传统方法无法发现的异常点, 且在被测数据集上有更好的量化指标.

由于引入了邻域关系, C-LOF在计算每两点间的可达距离时都可能需要遍历数据集中的其他点. 在未来的工作中, 拟通过引入某些快速kNN算法来减少寻找邻域的时间复杂度.

参考文献(References)

[1] Bolton R J, Hand D J. Statistical fraud detection: A review (with discussion)[J]. Statistical Science, 2002, 17(3): 235–255.

[2] Tang B, He H B. A local density based approach for outlier detection[J]. Neurocomputing, 2017, 241(2): 171-180.

[3] Jin W, Tung A K, Han J. Mining top-n local outliers in large databases[C]. Proc of the 7th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2001: 293-298.

[4] Barnett V, Lewis T. Outliers in statistical data[M]. Wiley: & Sons, 1994: 335-338.

[5] Hwkins D M. Identification of outliers[M]. New York: Springer, 1980: 613-615.

[6] Zhang T, Ramakrishnan R, Livny M. BIRCH: A new data clustering algrithm and its applications[J]. Data Mining and Knowledge Discovery, 1997, 1(2): 141-182.

[7] Brito M, Chavez E, Quiroz A, et al. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection[J]. Statistics & Probability Letters, 1997, 35(1): 33-42.

[8] Knorr E M, Ng R T. A unified notion of outliers: Properties and computation[C]. Proc of the 3rd ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 1997: 219-222.

[9] Knorr E M, Ng R T. Algorithms for mining distance based outliers in large datasets[C]. Proc of the 24th Int Conf on Very Large Data Bases. New York: Morgan Kaufmann Publishers Inc, 1998: 392-403.

[10] Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying density-based local outliers[C]. Proc of ACM Sigmod Record. Madison: ACM, 2000: 93-104.

[11] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets[C]. Proc of the ACM Int Conf on Management of Data (SIGMOD). Dallas: ACM, 2000: 427-438.

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

[13] 刘一民, 文俊杰, 王岚君. 基于空-时近邻与似然比检验的传感器网络异常点检测[J]. 清华大学学报: 自然科学版, 2017, 57(11): 1196-1201.
(Liu Y M, Wen J J, Wang L J. Outlier detection based on spatio-temporal nearest neighbors and a likelihood ratio test for sensor networks[J]. J of Tsinghua University: Science and Technology, 2017, 57(11): 1196-1201.)

[14] 杨金伟, 王丽珍, 陈红梅, 等. 基于距离的不确定数据异常点检测研究[J]. 山东大学学报: 工学版, 2011, 41(4): 34-37.
(Yang J W, Wang L Z, Chen H M, et al. Distance-based outlier detection over uncertain data[J]. J of Shandong University of Technology: Engineering Science, 2011, 41(4): 34-37.)

[15] Schubert E, Zimek A, Kriegel H P. Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection[J]. Data Min Knowl Discov, 2014, 28(1): 190-237.

[16] Jin W, Tung A K H, Han J, et al. Ranking outliers using symmetric neighborhood relationship[C]. Proc of the 10th Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD). Singapore: Springer, 2006: 577-593.

[17] Kriegel H P, Kroger P, Schubert E, et ak. LoOP: Local outlier probabilities[C]. Proc of the 18th ACM Conf on Information and Knowledge Management (CIKM). Hong Kong: ACM, 2009: 1649-1652.

[18] Zhang K, Hutter M, Jin H. A new local distance based outlier detection approach for scattered real-world data[C]. Proc of the 13th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 813-822.

[19] 杨茂林, 卢炎生. 基于剪枝的海量数据离群点挖掘[J]. 计算机科学, 2012, 39(10): 152-156.
(Yang M L, Lu Y S. Outlier mining in mass data based on pruning algorithm[J]. Computer Science, 2012, 39(10): 152-156.)

[20] Campos G O, Zimek A, Sander J, et al. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study[J]. Data Min Knowl Disc, 2016, 30(4): 891-927.

[21] Schubert E, Wojdanowski R, Zimek A. On evaluation of outlier rankings and outlier scores[C]. Proc of the 2012 SIAM Int Conf on Data Mining. Anaheim: SIAM, 2012: 1047-1058.

[22] Liang S Y, Han D Q, Zhang L, et al. A novel clustering oriented closeness measure based on neighborhood chain[C]. Proc of the 2017 Int Joint Conf on Nerual Networks (IJCNN). Anchorage, IEEE: 2017: 997-1004.

[23] Benjio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

作者简介

梁绍一(1987–), 男, 博士生, 从事聚类分析的研究, E-mail: liangsymail@163.com;

韩德强(1980–), 男, 教授, 博士生导师, 从事模式分类、信息融合等研究, E-mail: deqhan@mail.xjtu.edu.cn.

(责任编辑: 闫 妍)