基于相对距离的反 k 近邻树离群点检测

杨晓玲[1,2],冯 山[1],袁 钟[2]
(1. 四川师范大学数学科学学院,四川成都 610066;2. 西南交通大学信息科学与技术学院,四川成都 611756)

摘 要: 针对分布复杂且离群类型多样的数据集进行离群检测困难的问题,提出基于相对距离的反 k 近邻树离群检测方法 RKNMOD(Reversed K-Nearest Neighborhood).首先,将经典欧氏距离、对象局部密度和对象邻域结合,定义了对象的相对距离,能同时有效检出全局和局部离群点.其次,以最小生成树结构为基础,采取最大边切割法以快速分割离群点和离群簇.最后,人工合成数据集和UCI数据集试验均表明,新算法的检测准确率更高,为分布异常且离群类型多样的数据集的离群检测提供了一条有效的新途径.

关键词: 离群点;离群簇;反 k 近邻;最小生成树;相对距离度量
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2020)05-0937-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.05.014

Outlier Detection Based on Reversed *K*-Nearest Neighborhood MST of Relative Distance Measure

YANG Xiao-ling[1,2], FENG Shan[1], YUAN Zhong[2]
(1. *School of Mathematical Science, Sichuan Normal University, Chengdu, Sichuan* 610066, *China*;
2. *School of Information Science and Technology, Southwest Jiaotong University, Chengdu, Sichuan* 611756, *China*)

Abstract: For outlier detection difficulty of data sets with complex distribution and various types of outliers, a new outlier detection method based on reversed k-nearest neighborhood MST of relative distance measure is proposed. Firstly, relative distance of object is defined with the combination of classical distance, local density and neighborhood of object, which can be used to detect global outliers and local outliers both. Secondly, on basis of minimum spanning tree structure, by tactics of maximum-edge-cutting, outliers and outlier clusters can be obtained. Finally, experiments of synthetic and UCI data sets show that the new algorithm is much more correct and effective. It is a new effective way for detecting outliers of data sets with abnormal distribution and diversity outlier types.

Key words: outlier; outlier cluster; reversed nearest neighborhood; MST (minimum spanning tree); relative distance metric

1 引言

离群点检测在欺诈行为、医疗卫生、公共安全及入侵监测等领域具有十分重要的作用[1,2].随着数据采集的硬件技术和数据组织的软件技术不断发展,从事机器学习和数据挖掘等领域研究的学者们分别从不同角度对离群点检测问题进行了研究.例如,围绕距离和密度离群点检测,文献[3]对相关的离群点检测方法进行了综述.针对无监督离群点检测,文献[4]用19种算法同时对不同应用领域的10个数据集进行了全面的检测评估,发现基于距离的离群检测效果更好,且稳定.从计算机科学角度,文献[5]将不同离群检测法进行了联系以讨论离群点检测问题.在分类上,根据数据集与被考察对象的主体差异性,可分为全局离群和局部离群两类.根据离群对象规模,可分为单点和多点离群(簇).按照离群点检测思想,经典离群检测主要分为基于统计[6,7]、基于邻近性[8~12]和基于聚类[13]等方法.

统计离群检测一般分为两步:(1)假设数据集服从某一分布;(2)根据分布假设,将不遵守分布模型的对象判为离群对象.例如,Yamanishi 等人[14]利用有参高

收稿日期:2019-07-04;修回日期:2019-09-09;责任编辑:马兰英
基金项目:国家自然科学基金(No.61673285,No.61976182,No.61572406);四川省青年科技基金(No.2017JQ0046);四川省国际科技创新合作重点项目(No.2019YFH0097)

斯混合模型,将对象和模型的相异程度值作为对象的离群分数,分数大于指定阈值时被视为离群点. 针对高维数据集,文献[7]提出了快速 HBOS(Histogram-based Outlier Score)算法,对数据特征相互独立的数据集检测效果较好. 一般来讲,统计检测适于高维数据集,但要对数据集的分布规律进行估计. 由于绝大多数数据集的分布并不能事先知道,也难以用单一分布特征刻画,分布假设条件往往不易满足,故统计检测法存在明显局限.

实际上,许多经典聚类算法可直接或稍加修改后用于离群检测,也可通过引入某种特殊机制减少离群点对聚类结果的不良影响. 比如,文献[15]提出了一种基于马氏距离的聚类离群点检测方法用于入侵检测领域. 一般来讲,聚类有效性高度依赖于所用聚类方法[1],对不规则分布数据集,其检测效果往往不佳,使得聚类离群检测不一定最优. 对大型数据集,聚类的时空开销往往很大.

基于邻近性的离群点检测包括距离和密度两种检测模式. 无论哪种模式,用户都要指定对象的合理邻域阈值,并对对象集中各对象考察其合理邻域中的其他对象. 如果对象集中大多数对象远离某个对象,该对象就被视为对象集中的一个离群点.

距离离群检测最早由 Knorr 等人[16]提出. 为克服距离离群程度信息的不足,Ramaswamy 等人[17]提出了基于 k 近邻的离群度量方法,但它并不包含对象的所有 k 个最近邻的全部信息,不能很好地反映对象邻域的紧密或稀疏情况. 为了更好地反映邻域内的对象分布情况,提高计算精度,文献[18]用 k 近邻对象的平均距离作为离群分数,但它对每个对象都计算离群分数,时空开销较大. 为此,文献[19]基于抽样来考查被测对象与其近邻的关系,并计算其离群分数. 另外,对 k 近邻算法参数 k,其选择会影响算法性能和结果. 文献[20]提出了不需设置参数 k 的自然最近邻概念,其节点邻居由算法自适应计算形成.

密度离群检测一般通过局部离群因子 LOF(Local Outlier Factor)度量数据点的离群程度[21],LOF 值越大,对象离群的可能性越大. 对局部离群检测,文献[22,23]对 LOF 又做了一些改进. 对分布异常数据集,文献[24]提出了基于反向 k 近邻的局部离群度度量方法,它不仅考虑了数据点的 k 邻域关系,还考虑了数据点的反向 k 邻域,以避免复杂数据分布下的 LOF 算法误判. 文献[25]研究了距离法用于高维数据集的离群点检测,指出了反 k 近邻中的对象对离群点检测的有用性. 文献[4]指出,相比聚类离群检测,基于近邻性的离群检测性能较好. 在基于邻近性的离群点检测算法中,KNN(K-Nearest Neighbor)算法[8]效果较好,对涉及局部离群点的检测,LOF 算法[21]效果较好. 为此,选择 KNN 和 LOF 与本文的 RKNMOD 进行比较.

实际上,对分布复杂且离群类型多样的数据集,单一检测法往往是有局限的. 新方法既要能克服现有检测方法缺陷,又要能保留其优势. 近年来,人们已经开始进行方法融合的检测研究. 比如,融合密度和聚类方法的优势,文献[26,27]提出了基于 k 近邻树的离群点检测. 融合距离和聚类的优势,文献[12]提出了一种新的检测方法. 目前,以方法融合思想研究离群点检测的文献不多,但生活中有大量这样的数据集离群检测需要用到融合思想展开研究. 例如,在电子欺诈中,存在相似性很高的对象、相似性不高但明显偏离其余对象的对象以及分布异常的对象等情形. 最小生成树法[27]能够适应数据集多样分布和可变密度的有效检测,但它没有用到数据集的分布特征,对数据集分布异常的检测效果较差. 数据集分布异常时,用反 k 近邻密度影响因子法[24]也难以检出整体离群簇.

为此,本文提出了融合经典距离、局部密度和邻域关系的相对距离度量,它既适于距离法擅长的全局离群检测,也适于密度法擅长的局部离群检测. 新方法首先要构建一棵基于相对距离度量的最小生成树,然后对其进行子树切割分析,以进行离群检测. 显然,最小生成树子树切割同时包含了离群点和离群簇检测. 相比文[27]算法,新算法能更有效地检测分布异常数据集中的离群点和离群簇.

2 基本概念

定义 1(离群点) 设 $U=\{x_1,x_2,\cdots,x_n\}$ 为论域,各对象由 m 个属性描述,f 是对象离群度量函数,U 中全体对象的离群映射结果为 Y,U 中大多数对象的离群映射结果为 Y_1. 那么,则集合 $O=\{x\in U|f(x)\in Y-Y_1\}$ 中对象为离群点.

定义 2[12](最小生成树) 对无向连通网络 $G=(V,E)(|V|=n)$ 及其生成树 $G_i'=(V_i',E_i')$,如果 $\sum_{j=1}^{n-1}e_j'|e_j'\in E_i'$ 最小,称 G_i' 为 G 的最小生成树或 MST,记为 $TG_{\min}$.

定义 3[24](对象的 k 近邻) 对非空有限对象集 $U=\{x_1,x_2,\cdots,x_n\}$,$\forall x\in U$,U 中与 x 最靠近的 k 个对象的集合(x 除外)称为 U 中对象 x 的 k 近邻或 k 邻域,记为 $KNN_k(x)$.

定义 4[24](对象的反 k 近邻) 对非空有限对象集 $U=\{x_1,x_2,\cdots,x_n\}$,$k\in\mathbb{N}^+$,k 近邻包含对象 x 的对象集合称为 x 的反 k 近邻,记为 $RNN_k(x)$,即:

$$RNN_k(x)=\{y\in U|y\neq x,x\in KNN_k(y)\}$$

3　基于反 k 近邻树的离群点检测

结合经典欧氏距离和密度的 k 近邻度量[27]用对象及其最相似的 k 个对象间的近邻关系进行离群检测,忽略了反 k 近邻中对象间的相似性或近邻关系,导致在处理分布异常数据集时检测效果不好. 实际上,以欧氏距离和反 k 近邻为基础,并融合对象密度信息,也能反映数据对象间的距离关系. 为此,基于融合思想,一种新的相对距离度量也可以适用于分布异常的数据集离群检测.

3.1　相对距离度量的构建

$KNN_k(x)$是 U 中与 x 最相似对象的集合,x 的邻域信息与 $KNN_k(x)$密切相关. 若 $y \in KNN_k(x)$,也可认为 y 与 x 相似. 易知,x 的反 k 近邻也隐含了 x 的邻域信息. 融合了对象相似性和反向相似性的 $RNN_k(x) \cup KNN_k(x)$可以更全面地反映 x 的邻域信息,获得更多与 x 相似相关的对象信息. 若 $IS_k(x)$表示 x 的 $KNN_k(x)$与 $RNN_k(x)$的并,有:$IS_k(x) = KNN_k(x) \cup RNN_k(x)$.

定义 5(对象的 k 近邻密度)　对 U 中对象 x 和整型常量 k,假设 $KNN_k(x)$中离 x 最远的对象与 x 的距离为 $k_{dist}(x)$,对象 x 的 k 近邻密度可定义为:$d_k(x) = 1/k_{dist}(x)$.

用 $k_{dist}(x)$度量 x 的邻域分布密度时,如果 x 离群,$k_{dist}(x)$一般较大,$d_k(x)$较小. 因 $KNN_k(x)$中对象与 x 相似,$KNN_k(x)$中对象的离群性与 x 的相似的可能性大. $\forall y \in U, y \neq x, y \in RNN_k(x)$,$x$ 离群时 y 离群的可能性大. 若同时用 $d_k(x)$和 $IS_k(x)$的平均密度 $\sum_{y \in IS_k(x)} d_k(x)/|IS_k(x)|$(即 x 邻域平均密度)考察,可弥补简单欧氏距离检测的不足.

定义 6(对象的邻域密度影响因子)　对 U 中 x 和整型 k,$d_k(x)$是 x 的 k 密度,$\sum_{y \in IS_k(x)} d_k(x)/|IS_k(x)|$是平均密度,则 x 的邻域密度影响因子可定义为:

$$DOF(x) = d_k(x)/\left(\sum_{y \in IS_k(x)} d_k(y)/|IS_k(x)|\right)$$

$DOF(x)$融合了 x 的 k 近邻密度及其邻域平均密度,$d_k(x)$可反映 x 的 k 近邻对象分布疏密程度. x 的 k 近邻密集时 $d_k(x)$较大,x 的邻域平均密度与 $d_k(x)$较为接近,$DOF(x)$接近 1. x 显著离群时,x 的邻域平均密度相比 k 近邻密度 $d_k(x)$会增大,$DOF(x)$显著减小. x 邻近离群数据对象时,其离群可能性较大. 此时,x 的邻域平均密度相比 k 近邻密度 $d_k(x)$可能略微减小,$DOF(x)$一般略微增大. 对密度分布不均匀区域,x 的邻域平均密度低于 $d_k(x)$时 $DOF(x)$变高,反之变低. 故随着分布的不同,$DOF(x)$值可以更客观地反映对象间的距离. 可将欧氏距离与密度影响因子融合,以便更科学、合理地度量对象间的距离.

定义 7(对象的相对距离)　对论域 U,对象 x 与对象 y 的相对距离可定义为:

$$RD(x,y) = \max(DOF(x), DOF(y))\, d(x,y)$$

易知,$RD(x,y)$可增强密度分布不均匀时对象间的差异性. x 和 y 位置不同时,密度影响因子也不同,密度分布的不同可在 $RD(x,y)$中反映. 对 $RD(x,y)$度量,选择密度影响较大对象的密度影响因子并结合其欧氏距离,对离群对象 x 而言,其邻域内的对象的密度影响因子将变大,邻域内的对象间的相对距离也会增大,它提升了离群点邻域对象的离群可能性,符合“物以类聚”的分类思想.

3.2　基于相对距离的反 k 近邻树离群点检测算法

结合欧氏距离与密度,以 $KNN_k(x)$为基础,文献[27]提出了最小生成树离群点检测. 与 $KNN_k(x)$不同,$RNN_k(x)$采用邻域包含 x 的对象相似性度量以弥补邻域边界对象遗漏[27],用 $RNN_k(x)$为基础的相对距离度量构造反 k 近邻 MST[28],通过割边策略[27]进行子树划分以获得离群点或簇. 当 MST 子树中的结点数≤k 时,子树被视为离群点(单点子树)或离群簇(多点子树). 对 MST 检测而言,分割子树中的结点数小到规定占比即可结束离群检测. 因此,RKNMOD 算法,见算法 1,处理流程如图 1 所示.

算法 1　基于相对距离度量的反 k 近邻树离群点检测(RKNMOD)

输入:数据集 U,$|U| = n$,近邻对象数阈值 k 和离群点数阈值 $\nu\%$
输出:离群点集合 OS_k
(1)对 U 进行规范化处理,$OS_k = \varphi$
(2) for $i = 1$ to n
　　计算 $KNN_k(x_i)$
　　计算 x_i 到 $KNN_k(x_i)$中其他数据对象 x_j 的距离 $d(x_i, x_j)$
　　计算 $RNN_k(x_i)$
　　计算 $IS_k(x_i)$
　　计算 $DOF(x_i)$
　end for
(3)for $i = 1$ to n
　　for $j = 1$ to n
　　计算 $RD(x_i, x_j)$
　　end for
　end for
(4)根据 RD 矩阵构建最小生成树.
(5) while $|OS_k| > \nu\%$
　　切割森林中最大边 l_{max},形成子树 T_1 和 T_2.
　　if $|T_1| < k$ then
　　$OS_k = OS_k \cup T_1$
　　end if
　　if $|T_2| < k$ then
　　$OS_k = OS_k \cup T_2$
　　end if
　end while

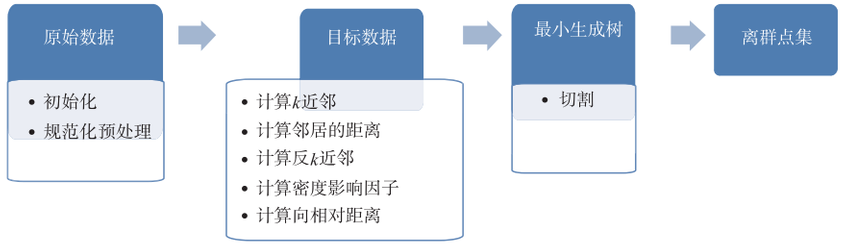

图1 算法1的处理流程

RKNMOD 算法的时间复杂度取决于 x 邻域内的对象的相对距离计算、MST 建立和割边法查找离群点的过程. 假设论域对象数为 n,x 的邻域对象数为 k,因邻域内的对象间的相对距离由反 k 邻域寻找、欧氏距离和相对距离计算组成,其时间复杂度为 $O(kn^2)$. 用 Kruskal 算法邻接表进而构建 MST 的时间复杂度为 $O(kn\log(kn))$. 而对 n 条边排序并以割边法检测离群点的时间复杂度为 $O(n\log n)$. 综上时间复杂度为 $O(kn^2)+O(kn\log(kn))+O(n\log n)$. 与文献[27]方法相比,RKNMOD 连通图的边少,其实际计算距离和构建 MST 的时间复杂度也相应降低了.

3.3 算法的演算实例

例 1 设 $U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$,对象属性结构 $A=\{a,b,c\}$(表 1),其三维分布如图 2 所示. 离群检测结束阈值 10%. RKNMOD 的演算过程如下:

表 1 论域 U

U	a	b	c
x_1	I	4	0.7
x_2	II	7	0.4
x_3	I	1	0.6
x_4	II	2	0.3
x_5	II	8	0.5
x_6	III	10	0.8

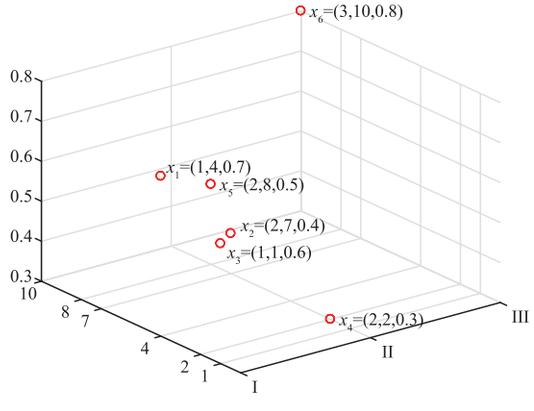

图2 论域U的数据对象分布图

假设采用最小-最大规范化,$k=3$,U 的 k 近邻、反 k 近邻及 IS_k 的演算结果如表 2 所示. 论域 U 的相对距离如表 3 所示. U 的反 k 近邻 MST 结构如图 3 所示.

对图 3,按最大边切割后的结果如图 4 所示. 其中 x_6 是离群点. 因检出离群点占比 16.7% 已超过占比阈值 10%,离群判定算法停止.

表 2 U 的 k 近邻、反 k 近邻和并集 IS_k

U	KNN_k	RNN_k	IS_k
x_1	$\{x_3,x_5,x_2\}$	$\{x_3,x_5,x_6\}$	$\{x_3,x_5,x_2,x_6\}$
x_2	$\{x_5,x_4,x_1\}$	$\{x_1,x_3,x_4,x_6,x_5\}$	$\{x_1,x_3,x_4,x_5,x_6\}$
x_3	$\{x_1,x_4,x_2\}$	$\{x_1,x_4\}$	$\{x_1,x_2,x_4\}$
x_4	$\{x_2,x_5,x_3\}$	$\{x_2,x_5,x_3\}$	$\{x_3,x_5,x_2\}$
x_5	$\{x_2,x_4,x_1\}$	$\{x_2,x_4,x_1,x_6\}$	$\{x_1,x_2,x_4,x_6\}$
x_6	$\{x_1,x_2,x_5\}$	φ	$\{x_1,x_5,x_2\}$

表 3 U 中数据对象的相对距离

U	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	2.516	0.313	0	0.969	0.981
x_2	2.516	0	2.628	1.225	0.475	2.745
x_3	0.313	2.628	0	0.898	0	0
x_4	0	1.225	0.898	0	0.647	0
x_5	0.969	0.475	0	0.647	0	0.988
x_6	0.981	2.745	0	0	0.988	0

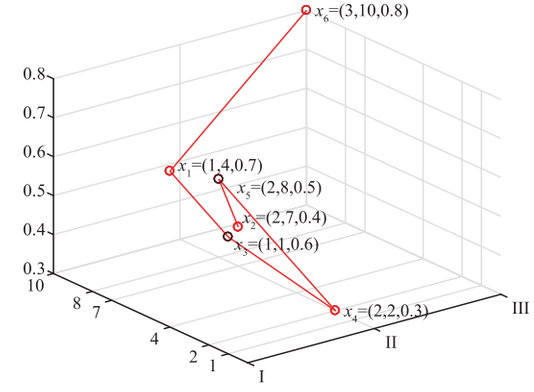

图3 k=3时U的反k近邻最小生成树结构

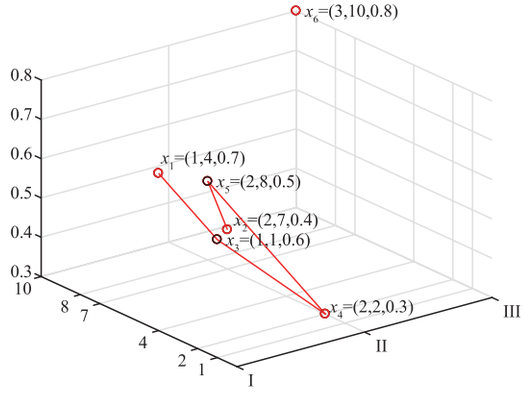

图4　对图3按最大边切割所得的森林结构

4　实验结果和分析

为了验证 RKNMOD 的检测精度,本节实验用 LOF[21]、KNN[8]、INFLO(InFLuenced Outlierness)[24]和 KNMOD(Outlier Detection Based on K-nearest Neighborhood MST)[27]与 RKNMOD 进行比对. 其中,KNN 在基于邻近性的离群点检测中效果较好,LOF 对涉及局部离群点的检测效果较好,INFLO 对分布异常的数据集进行检测的效果较好,KNMOD 适于密度分布不同且离群类型多样的数据集. 为验证 RKNMOD 算法的适用性和有效性,实验中分别采用人工合成和 UCI 数据集进行验证. 以检出离群点数与真实离群点数的占比评价,占比越高,精度越好. 以 O_m/m 表示,m 是检测出离群点数,O_m 是真实离群点数.

实验硬件环境是 Intel(R) Core(TM) i5 - 3337CPU @ 1.80GHz、4.00GB 内存;软件环境是 Windows 7 操作系统,MATLAB R2016b 编程语言.

4.1　人工数据集实验分析

为了验证 RKNMOD 算法能够对分布形状多样和分布异常的数据集进行有效离群检测. 本文首先构造了三个人工数据集进行实验.

(1)Data1 数据集

Data1 有 474 个对象,其二维平面分布如图 5 所示. Data1 中有 6 个点离群点、一个正态分布数据对象簇和两个均匀分布数据对象簇. 实际上,能检出的离群簇的大小与 k 相关,它影响离群簇检测结果的有效性. 观察图 5 易知,两个均匀分布簇各有 9 个对象,要完全检出其离群点和离群簇,要取 $k=10$. 如果只检测离群单点,可取 $k=6$. 因此,对整个数据集的离群检测可分别取 $k=6$ 和 $k=10$ 构建 Data1 的相对距离 MST(图 6 和图 7). 对图 6 和图 7 依次分别进行 3 次最大边切割后的森林结构分别如图 8 和图 9 所示.

可见,基于相对距离的反 k 近邻树离群检测可同时检测出离群点和离群簇. 对 Data1,$k=6$ 可有效检测出

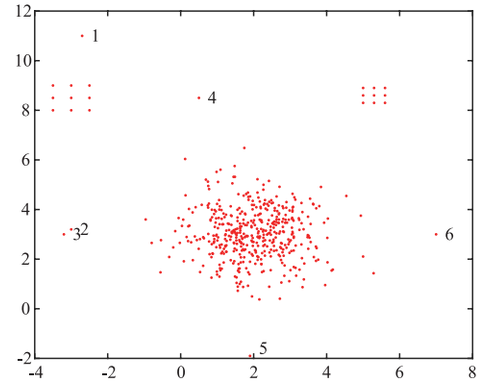

图5　Data1数据集的数据对象分布图

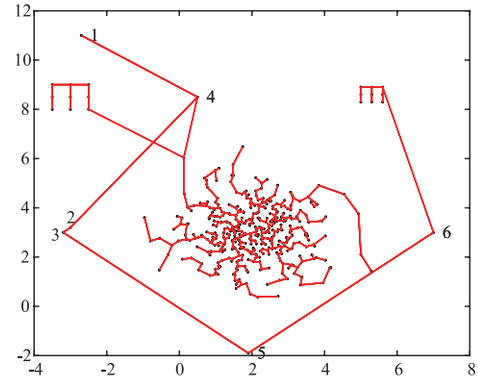

图6　k=6时Data1的相对距离最小生成树结构图

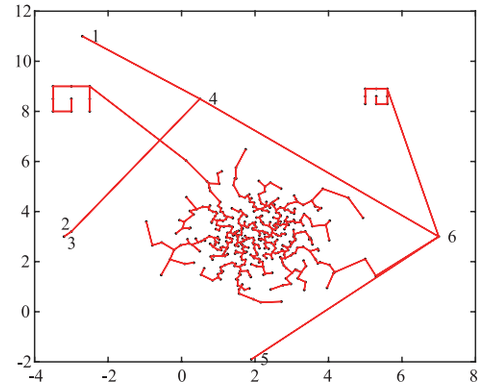

图7　k=10时Data1的相对距离最小生成树结构图

所有 6 个离群点. $k=10$ 时能检测出所有 6 个离群点和两个小均匀分布离群簇,表明 RKNMOD 适应能力很强. 经过对 5 种算法的反复对比实验,取各算法准确率最高的优选参数进行检测,对比结果如表 4 所示.

表 4　对 Data1 的离群检测准确率结果对比表

LOF	KNN	INFLO	KNMOD	RKNMOD
24/24 6/18	24/24 6/23	24/25 6/19	24/24 6/6	24/24 6/6

表 4 第 1 行是检出 6 个离群点和 2 个均匀分布簇的准确率,第 2 行是只检出 6 个离群点的准确率. 可见,

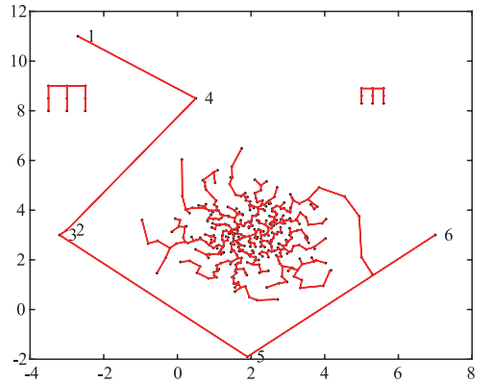

图8 k=6时Data1的相对距离最小生成树3次切割后的森林结构

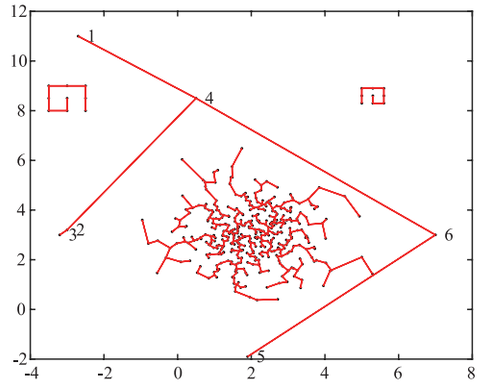

图9 k=6时Data1的相对距离最小生成树3次切割后的森林结构

只检测出 6 个离群点时,LOF、KNN 和 INFLO 的准确率较 RKNMOD 和 KNMOD 要低得多. 同时检出 6 个离群点和 2 个均匀分布簇时,RKNMOD 不仅保持了最好的准确率,还能检测出 KNMOD 算法所擅长的离群点和离群簇.

(2) Data2 数据集

为了检验全局和局部离群检测效果,Data2 由 78 个对象构成(图 10). 其中,对象 1、2、3、4、5 和 6 最有可能离群,1、3 和 5 可能是全局离群,2 和 4 可能是局部离群. 数据集中另有 4 个不同密度的数据对象簇. 此时,在全局和局部离群检测以试验结果最优参数 $k=3$ 来构建 MST(图 11)并进行 6 次切割后,整个森林结构如图 12 所示.

可见,RKNMOD 算法在 Data2 中能直接检出 6 个离群点,对既有全局离群又有局部离群的数据集能够进行有效离群检测. 各对比算法的反复实验检测的最优结果对比如表 5 所示.

表 5 各对比算法离群检测准确率对比表

LOF	KNN	INFLO	KNMOD	RKNMOD
– – –	6/22	6/7	6/6	6/6

由表 5 可知,KNN 和 LOF 的检测没有 INFLO、KNMOD 和 RKNMOD 高效. 对 Data2,后三个算法都检测出了全部 6 个离群点. 可见 RKNMOD 算法能同时检测出局部和全局离群点,且效果较好.

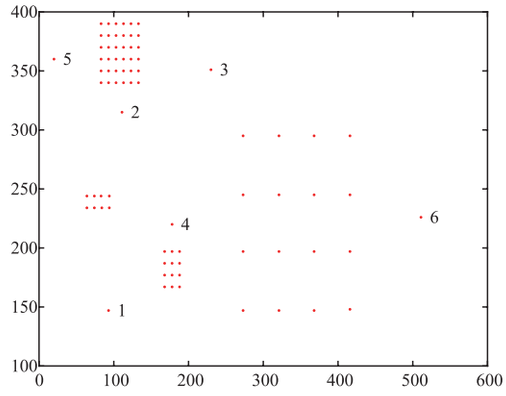

图10 Data2数据集中的数据对象分布图

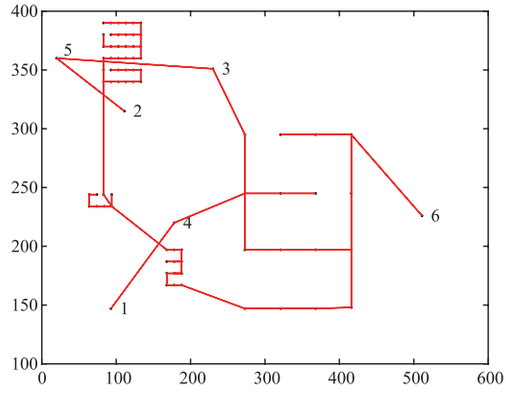

图11 k=3时Data2的相对距离最小生成树结构图

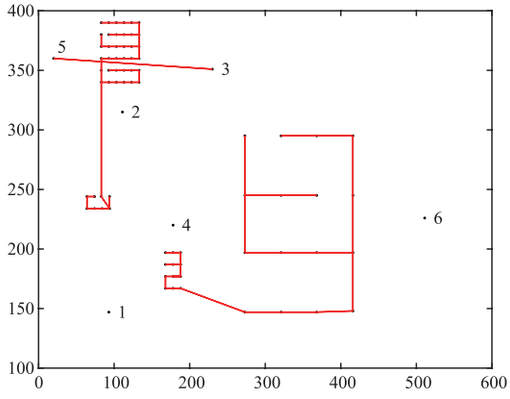

图12 k=3时Data2的MST切割最大6条边后的森林结构

(3) Data3 数据集

为了对分布异常数据集进行离群检测验证,Data3 给出了 45 个数据对象(图 13),其中,对象 1 和 2 与均匀密集簇 C_1 距离相同,对象 2 在 C_1 附近的另一均匀稀疏簇 C_2 中. 从相对距离变化幅度看,相比 2,1 的离群度

应更大[13]. 对 3 和 4,它们与 C_2 的距离较远,相对于 1 和 2,其离群程度应更大. 因此,4 个对象最可能的离群大小关系应该是:4→3→1→2[13],即 $f(4)>f(3)>f(1)>f(2)$.

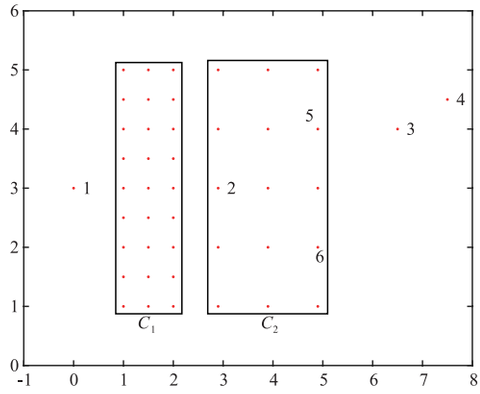

图13　Data3数据集中的数据对象分布图

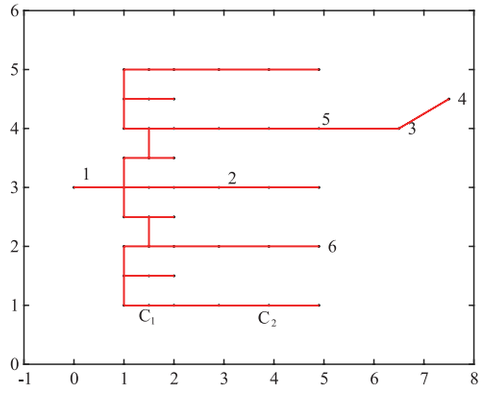

图14　k=4时Data3的相对距离最小生成树结构图

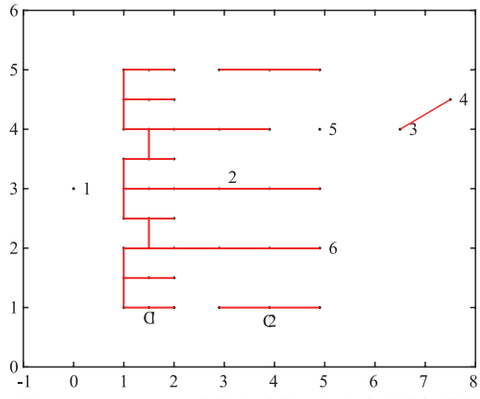

图15　k=4时Data3的离群点检测切割的森林结构展示

对 Data3,以试验结果最优参数 $k=4$ 构建 MST(图 14),RKNMOD 的离群检测结果如图 15 所示. 可见,在稀疏簇周围,点的离群度大于稀疏簇和密集簇,它符合人们的直觉认知. 多次试验结果表明,$k=4$ 时,Data3 的对象邻域内数据对象间存在较强的特征相似性. 5 是离群点 3 和 4 的近邻,其离群性会受到 3 和 4 影响,离群性应大于对象 6. Data3 的实验结果对比如表 6 所示.

表 6　Data3 中数据对象的离群度大小检测结果对比

LOF	KNN	INFLO	KNMOD	RKNMOD
4>3>2>1	4>3>1>2	1>4>2>3	1>2>3=4	3=4>1>2

表 6 是各对比算法检测结果中 4 个对象的离群度大小关系. 易知,KNN 和 RKNMOD 的检测结果更符合直觉认知,而 LOF、INFLO 和 KNMOD 的检测存在不符合直觉认知的误判. 可见,由于 RKNMOD 结合了 k 近邻和反 k 近邻信息建立相对距离度量,它既有 KNN 所具有的优势,又能增加离群点邻域内对象的离群程度.

4.2　UCI 数据集实验分析

为进一步验证算法的有效性,我们以 UCI 数据库[29]中的 Wine、Wisconsin Breast Cancer(WBC)、Lymphography(Lym)和 Iris 等数据集进行进一步的实验比对. 实验中,对各数据集作了如下处理:

(1)如文献[27]一样,对 Wine 和 Iris 数据集,只从其某个簇中随机选若干对象构成离群簇,并与数据集中其他对象一起构成新的实验数据集. 相对于其他簇,所选离群数据对象可以被认为是该数据集中由不同机制所产生的离群簇.

(2)对 WBC 数据集,为了构成不平衡分布数据集,可仿照文献[30]中做法,从 WBC 中移去一些属于"malignant"类的数据对象. 最终,WBC 中包括了 483 个对象,其中 39 个属于"malignant"类,444 个属于"benign"类. 假定少数类"malignant"看作稀有类. 4 个数据集的基本特征信息如表 7 所示.

表 7　所选 UCI 实验数据集的基本特征信息

特征 / 数据集	对象个数	属性个数	离群点数
Lym	148	18	6
Wine	134	13	15
WBC	483	9	39
Iris	115	4	15

假定要检出给定数据集所有离群点. 参数 k 取值过大时,对象邻域内会出现误判信息,取值过小时,又可能使对象的邻域信息获取不足. 故一般来讲,k 从 2 到数据对象数的一半中选取,并在所选 UCI 数据集中计算对应的准确率(图 16).

由图 16 可知,k 值小于 50 的范围内各算法的离群检测准确率达到最高. 文献[7]的研究表明,KNN 的 k 值范围为[10,50]. 因此,本文实验中,RKNMOD 和 KNN 的 k 均从 10~50 中取值,其他对比算法从 1 到对象个数范围内取值,在此基础上,从实验结果中选择最

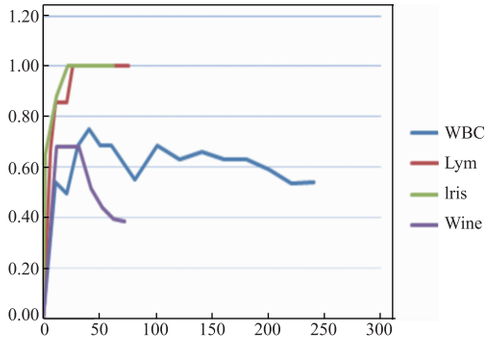

图16 所选UCI数据集的不同k值对应的检测准确率

佳结果及其 k 值进行展示. 多次反复实验的各算法最优效果检测参数如表 8 所示,其对应的离群检测准确率结果对比如表 9 所示.

表 8 各对比算法的实验参数 k 的取值信息

数据集 \ 算法	LOF	KNN	INFLO	KNMOD	RKNMOD
Lym	20	15	19	140	15
Wine	80	30	80	5	19
WBC	263	15	410	150	41
Iris	100	30	45	50	40

表 9 各算法的检测准确率结果对比

数据集 \ 算法	LOF	KNN	INFLO	KNMOD	RKNMOD
Lym	0.75	0.55	0.67	0.4	1
Wine	0.19	0.83	0.71	0.83	0.88
WBC	0.66	0.72	0.63	0.55	0.74
Iris	1	1	0.88	1	1

因 Iris 的稀疏离群簇离其他对象距离较远,各算法检测效果均不错. 因 Lym 是全局离群,适于局部离群点(簇)检测的 KNMOD 的检测效果并不理想. 由于 Wine 的簇内数据对象高度相似,只关注局部统计量的 LOF 和 INFLO 的检测准确率低. 综上,相对于各对比算法,融合反 k 近邻信息的相对距离算法 RKNMOD 显著提高了离群点邻域内对象的离群性,能同时兼顾局部离群、全局离群和离群簇等多种类型的离群检测以及分布复杂的数据集的离群检测.

5 结论

结合经典距离、对象局部密度、对象邻域关系,提出了一类新的相对距离度量,结合 MST 分割法进行离群点和离群簇检测. 理论分析和对比实验结果表明,新方法适应能力强,对分布形状多样和密度分布异常的数据集,其检测更有效.

下一步的工作可从以下方面展开:(1)k 值的自适应提取与优化;(2)新度量如何应用在高维数据集;(3)如何提升算法的时间效率.

参考文献

[1] Xue Z X, Shang Y L, Feng A F. Semi-supervised outlier detection based on fuzzy rough C-means clustering[J]. Mathematics and Computers in Simulation, 2010, 80(9): 1911 – 1921.

[2] Han J W, Kamber M, Pei J. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann, 2011. 543 – 584.

[3] 薛安荣,姚林,鞠时光,陈伟鹤,马汉达. 离群点挖掘方法综述[J]. 计算机科学,2008,35:11 – 18.
Xue A R, Yao L, Ju S G, Chen W H, Ma H D. Survey of outlier mining[J]. Computer Science, 2008, 35: 11 – 18. (in Chinese)

[4] Markus G, Seiichi U, Dongxiao Z. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data[J]. PLOS ONE, 2016, 11(4): 1 – 31.

[5] Aggarwal, Charu C. Outlier Analysis[M]. Berlin, Germany: Second Edition, Springer, 2017.

[6] Wu D F. A regression sequences based method for high dimensional outlier detection [J]. Journal of Discrete Mathematical Sciences and Cryptography, 2017, 20(4): 931 – 943.

[7] Goldstein M, Dengel A. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm[A]. Poster and Demo Track of the 35th German Conference on Artificial Intelligence (KI-2012)[C]. Germany: 2012. 59 – 63.

[8] Ramaswamy S. Rastogi R. Shim K. Efficient algorithms for mining outliers from large data sets[A]. ACM Sigmod International Conference on Management of Data[C]. New York, USA: ACM, 2000. 427 – 438.

[9] Jiang F, Sui Y F, Cao C G. A rough set approach to outlier detection[J]. International Journal of General Systems, 2008, 37(5): 519 – 536.

[10] Yuan Z, Zhang X Y, Feng S. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures[J]. Expert Systems with Applications, 2018, 112(12): 243 – 257.

[11] Chen Y M, Miao D Q, Zhang H Y. Neighborhood outlier detection[J]. Expert Systems with Applications, 2010, 37(12): 8745 – 8749.

[12] Wang X, Wang X L, Ma Y, et al. A fast MST-inspired kNN-based outlier detection method[J]. Information Systems, 2015, 48(3): 89 – 112.

[13] Tang X Q, Zhu P. Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space[J].

IEEE Transactions on Fuzzy Systems,2013,21(5):814 – 824.

[14] Yamanishi K,Takeuchi J I,Williams G,et al. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms[J]. Data Mining and Knowledge Discovery,2004,8(3):275 – 300.

[15] Wu N,Zhang J. Factor-analysis based anomaly detection and clustering[J]. Decision Support Systems,2006,42(1):375 – 389.

[16] Knorr E M,Ng R T. Algorithms for mining distance-based outliers in large datasets[A]. Proceedings of the 24th VLDB Conference[C]. New York,USA:VLDB,1998. 392 – 403.

[17] Ramaswamy S,Rastogi R,Kyuseok S. Efficient algorithms for mining outliers from large data sets[A]. Proceedings of the 2000 ACM SIGMOD international conference on Management of data[C]. New York,USA:ACM,2000. 427 – 438.

[18] Angiulli F,Pizzuti C. Fast outlier detection in high-dimensional spaces[A]. Proceeding of 6th European Conference on Principles of Data Mining and Knowledge Discovery[C]. Finland (Helsinki):ACM ,2002. 15 – 26.

[19] Wu M,Jermaine C. Outlier Detection by sampling with accuracy guarantees[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York,USA:ACM,2006. 767 – 772.

[20] 朱庆生,唐汇,冯骥. 一种基于自然最近邻的离群检测算法[J]. 计算机科学,2014,41(3):276 – 278.
Zhu Q S,Tang H,Feng J. Outlier detection algorithm based on natural nearest neighbor[J]. Computer Science,2014,41(3):276 – 278. (in Chinese)

[21] Breunig M M,Kriegel H P,Ng R T,Sander J. LOF:Identifying density-based local outliers[A]. ACM Sigmod International Conference on Management of Data[C]. New York,USA:ACM,2000. 93 – 104.

[22] 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.
Hu C P,Qin X L. A density-based local outlier detecting algorithm[J]. Journal of Computer Research and Development,2010,47(12):2110 – 2116. (in Chinese)

[23] 王敬华,赵新想,张国燕. 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.
Wang J H,Zhao X X,Zhang G Y. NLOF:A new density-based local outlier detecting algorithm[J]. Computer Science. 2013,40(8):181 – 185. (in Chinese)

[24] Jin W,Anthony K H,Han J W,Wang W. Ranking outliers using symmetric neighborhood relationship[A]. Pacific-Asia Conference on Knowledge Discovery and Data Mining[C]. Berlin,GER:Springer,2006. 577 – 593.

[25] Radovanovic M,Nanopoulos A,Ivanovic M. Reverse nearest neighbors in unsupervised distance-based outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(5):1369 – 1382.

[26] 朱利,邱媛媛,于帅,原盛. 一种基于快速 k-近邻的最小生成树离群点检测方法[J]. 计算机学报,2017,40(12):2857 – 2870.
Zhu L,Qiu Y Y,Yu S,Yuan S. A fast kNN-based MST outlier detection method[J]. Chinese Journal of Computers,2017,40(12):2857 – 2870. (in Chinese)

[27] Zhu Q S,Fan X G,Feng J. Outlier detection based on K-Neighborhood MST[A]. IEEE International Conference on Information Reuse and Integration[C]. Las Vegas,USA:IEEE,2015. 718 – 724.

[28] Alsuwaiyel M H. Algorithms Design Techniques and Analysis[M]. Beijing:Publishing House of Electronics Industry,2013. 239 – 246.

[29] Bay S D. The UCI KDDN repository[DB/OL]. http://kdd. Ics. Uci. edu. 2011-10-15.

[30] Harkin S,He H X,Williams G J,et al. Outlier detection using replicator neural networks[A]. Proc of the 4th Int Conf on Data Warehousing and Knowledge Discovery[C]. Aix-en-Provence:Springer-Verlag,2002. 170 – 180.

作者简介

杨晓玲　女,1993 年 1 月出生,四川乐山人,2016 年和 2019 年分别在西华师范大学和四川师范大学获得理学学士和理学硕士学位. 现为西南交通大学计算机科学与技术专业的博士生,从事粗糙集、离群点检测和数据挖掘有关的研究.
E-mail:yangxlt1993@163.com

冯　山(通讯作者)　男,1967 年 7 月出生,重庆丰都人,2003 年在中国科学院获得博士学位. 现为四川师范大学硕士导师,从事智能开发平台和算法分析与设计有关的研究.
E-mail:fengshanrq@sohu.com