第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1110-06

XmR控制图的异常点检测算法研究

陈丽芳[1,2],王荣杰[1],刘云庆[1],周 旭[1]

(1.华北理工大学理学院,河北唐山 063000;2.河北省数据科学与应用重点实验室,河北唐山 063000)

摘要:针对隔离森林异常点检测方法计算烦琐、耗时长等不足,提出基于XmR控制图的异常点检测算法.通过计算样本属性的单值均值、移动极差及其均值,绘制X图与mR图的控制界限和中心线,同时在图中绘制样本的单值属性;根据X图中超出界限的点对应的样本序号,与mR图中超出界限的点对应的样本序号加1,取并集,从数据中将其删除,然后将删除异常点后的数据代入CART、随机森林和支持向量机算法中进行实验验证.结果表明该方法与隔离森林方法相比具有更快的速度和更好的精度,为异常点检测提供了一种新的研究思路.

关键词:XmR控制图;异常点检测;控制界限;中心线
中图分类号:O1 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.08.010

引用格式:陈丽芳,王荣杰,刘云庆,等. XmR控制图的异常点检测算法研究[J]. 中国科学技术大学学报,2020,50(8):1110-1115,1186.
CHEN Lifang, WANG Rongjie, LIU Yunqing, et al. Research on outlier detection algorithm of XmR control chart[J]. Journal of University of Science and Technology of China, 2020, 50(8):1110-1115,1186.

Research on outlier detection algorithm of XmR control chart

CHEN Lifang[1,2], WANG Rongjie[1], LIU Yunqing[1], ZHOU Xu[1]
(1. *College of Science, North China University of Technology, Tangshan* 063000, *China*;
2. *Hebei Key Laboratory of Data Science and Application, Tangshan* 063000, *China*)

Abstract: A novel outlier detection algorithm was proposed based on the XmR control chart to address the complicated calculation and its time-consuming method in detecting isolated forest anomalies. By calculating the single-valued mean, its moving range and average of the sample attributes, we can draw the control limits and centerlines of the X and mR charts, and the single-valued attributes of the samples in the chart. According to the points in the X chart that exceeds the limits Sample number, add 1 to the sample number corresponding to the point that exceeds the limit in the mR graph, we take the union and delete it from the data, and then replace them after the deletion of the anomaly point with the CART. We use the random forest and support vector machine algorithm for experimental validations. The results show that this method has a faster speed and better precisions compared with the isolation forest method, which provides a new research idea for outlier detection.

Key words: XmR control chart; outlier detection; control limit; centerline

0 引言

“互联网+”时代,各行各业快速积累产生海量数据,如何有效地挖掘数据背后隐藏的有价值信息是一个重要挑战.实际产生的数据集中往往存在异常数据,异常数据作为一种特殊数据形式,如何快速、高效地对异常数据进行检测和处理,成为目前数据挖掘领域中一个新的研究热点.

本文主要研究数据挖掘中以异常数据为研究对象的异常挖掘.异常数据,是指在数据集中与大部分数据对象不一致,明显偏离数据集中的其他数据且不满足于数据的一般模式或行为的数据对象,“孤立点”、“新颖点”、“偏离点”、“稀有类”、“离群点”、“噪音”、“异常点”等都是指异常数据.异常点检测是指利用算法从数据集中挖掘出异常点并将其删除或单独分析处理,在不同的研究领域异常点检测具有不同程度的价值和意义.在欺诈检测[1]中,通过异常点检测能够发现异常交易,确保财产安全,减少不必要的损失;在网络入侵检测[2]中,通过异常点检测算法可以找出潜在的威胁.

目前关于异常点检测的算法,可以从两个方面进行分析,一是模型的角度,包括基于统计模型的

收稿日期:2020-04-29;**修回日期**:2020-08-06
基金项目:河北省自然科学基金(F2014209086)资助.
作者简介:陈丽芳,女,1973年生,博士/教授.研究方向:机器学习、智能计算、数据挖掘.E-mail:hblg_clf@163.com
通讯作者:周旭,男,硕士/讲师.E-mail:sxzhouxu@126.com

异常点检测算法[3-4]、基于邻近度的异常点检测算法[5-7]、基于子空间的异常点检测算法[8-9]、基于图论的异常点检测算法、基于集成学习的异常点检测算法[10]；二是数据挖掘的角度，分为无监督和有监督，无监督的异常检测算法包括基于聚类分析的方法[11-12]和基于局部离群因子方法[13-16]，有监督的异常检测算法包括基于神经网络的方法、基于支持向量机的方法以及基于决策树[17]的方法. 为了提高机器学习的准确率，学者对数据的异常点检测进行了研究，杨晓晖等[18]通过用随机超平面隔离数据集，采用孤立森林算法和层次化集成学习对数据集进行异常检测，提高了复杂数据和多维数据中异常检测的准确性和稳定性. 李春生等[19]通过引入“属性隶属度”的概念，提出基于改进距离和异常点检测算法，解决了由于数据分布不均匀导致的检测准确率较低的问题. Hussin 等[20]提出利用余弦函数进行循环函数关系模型的平均圆误差的新方法，还使用了行删除方法检测影响最大测量值的观察结果，从而将它们识别为异常值. 周志华等[21]用隔离森林进行异常点检测，首先随机选择子样本集，然后随机选择维度进行划分，最后将树中具有很短的平均高度所对应的样本判定为异常点. 该方法每次切割数据空间都是随机选取一个维度，建完树后仍然有大量的维度信息没有被使用，导致选取的属性不够全面，导致算法可靠性降低；另外，若构建的树较多，会增加样本平均高度的计算量，耗费更多的时间. 沈琰辉[5]提出基于邻域离散度的异常点检测算法，解决了很多异常点检测算法计算量大的问题. 冀汶莉[22]提出 RDU-SMOTE-RF 异常数据识别方法，提高了异常数据的识别准确率.

关于异常点检测的研究虽然取得一定的进展，但是现有算法均存在不足，基于统计学的异常点检测算法，主要是对给定的数据建立统计模型，该算法对模型假设是否成立依赖性较大；基于聚类的异常点检测算法，簇的个数会影响产生的离群点集，且通过聚类算法产生的簇的质量对产生的离群点的质量影响也非常大；基于距离的异常点检测算法，比如 k 近邻算法对参数的选取较敏感，且不能考虑密度的变化；基于密度的异常点检测算法，比如局部离群因子算法对参数的选取也比较敏感.

本文提出 XmR 控制图的异常点检测算法，该算法不涉及参数的选取、计算简单且耗时相比其他算法短，弥补了其他异常点检测算法的不足，可以有效挖掘出数据中的异常点.

1　理论知识

1.1　XmR 控制图

统计过程控制（statistical process control，SPC），有广义和狭义之分，狭义 SPC 是指控制图，一种对生产过程的关键质量特性值进行测定、记录、评估、并监测过程是否处于控制状态的图形方法. 最早的控制图是休姆哈特提出的 P 图，后来此类控制图都被称为 SPC 控制图. SPC 控制图可分为两类[23]：一是用于连续型数据的控制图——计量值控制图，包括均值-极差控制图、均值-标准差控制图、中位数-极差控制图以及单值-移动极差控制图；二是用于离散型数据的控制图——计数型控制图，包括不合格品数控制图，不合格品率控制图、缺陷数控制图和单位缺陷数控制图.

SPC 控制图有很多作用，在质量诊断方面，可以度量过程的稳定性，即过程是否处于统计控制状态；在质量控制方面，可以确定何时对过程加以调整，何时需使过程保持相应的稳定状态；在质量改进方面，可以确认某过程是否得到了改进.

XmR 控制图是 SPC 控制图中的一种，又称单值-移动极差图. 该图由三条平行于横轴的直线构成，中心线（CL），控制上限（UCL）以及控制下限（LCL），控制上限和控制下限表示可以接受的变异范围. XmR 控制图包括 X 图和 mR 图. X 图由样本各属性的所有单值，中心线（均值）、控制上限、控制下限组成；mR 图是由单值移动极差值 mR，中心线（mR 的均值）、控制上限和控制下限组成.

应用 XmR 控制图的目的是及时发现过程中出现的异常，判断异常的原则之一就是出现了“小概率事件”. 为此，判断的准则有两类：第一类是数据超出控制界限，在稳定状态下，数据越出界限的概率为 0.27%；第二类是数据虽在控制界限内，但排列的形状有缺陷，即控制图中的扫描点落在 UCL 与 LCL 之外或扫描点在 UCL 和 LCL 之间的排列不随机，则表明数据异常.

鉴于此考虑到异常点检测与质量控制问题有相似之处，可以借鉴控制图思想进行数据构图，从而找出异常点，进而对数据进行清洗，以利于后期数据分类建模，最终达到提升分类器性能的目的.

1.2　CART 决策树

CART 决策树是由 Breiman 提出的一种决策树构建算法，其核心思想是：通过相应的判断条件，从原始数据集中找到一个最优属性进行二分和细化，不断重复上面的操作，直到满足对象的自动分类为止. 相比其他决策树，CART 决策树计算速度较快. CART 算法，即分类回归树算法，是决策树构建的一种算法，CART 决策树的生长过程如图 1 所示，CART 的构建过程采用二分循环分割方法，每次划分都把当前样本集划分为两个子样本集，使决策树的节点均有两个分支，由此构成二叉树. 如果分支属性有多于两个的取值，在分裂时会对属性值进行组合，选择最佳的两个组合分支. 在 CART 决策树中，有根节点、非叶子节点和叶子节点，根节点是最先开始划分的属性，叶子节点是已经完成分类且不能再分类的包含最终结果的节点，有唯一的类别值（回归值）可计算，非叶子节点是在该节点处分类没有结束，尚未获得最终的值. 分类回归树是较成熟的决策树构建方法，该算法既可以用于分类，也可以用于回归，分类树主要用于变量是离散值的情况，回归树主要用于变量是连续值的情况，该算法结构清晰，易于理解且准确度高.

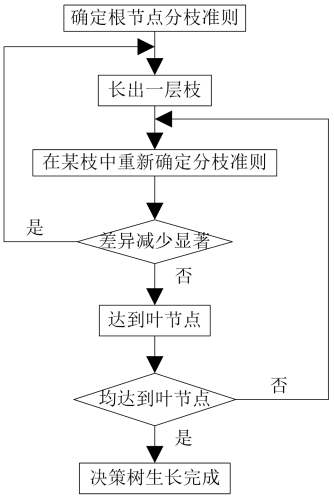

图 1 决策树生长过程

Fig. 1 Decision tree growth process

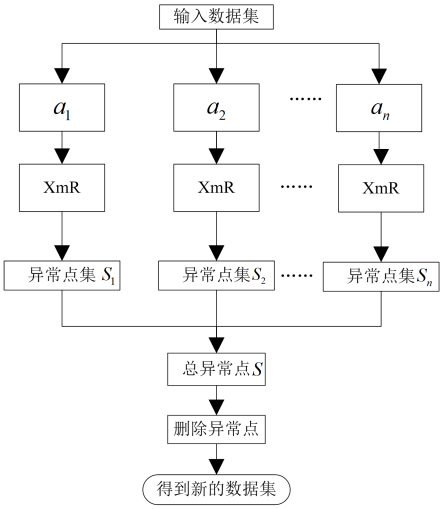

图 2 算法流程图

Fig. 2 Algorithm flowchart

2 算法流程与实例计算

2.1 算法步骤

假设 $\{X_1, X_2, X_3, \ldots, X_n\}$ 是一组测量数据集,样本属性的个数是 n,则该控制图的算法步骤如下:

Step1 计算每个属性所有数据的移动极差 mR,即相邻差的绝对值;

$$\mathrm{mR} = |X_{i+1} - X_i|, 1 \leqslant i \leqslant n-1 \quad (1)$$

Step2 计算每个属性所有数据的平均值 $\overline{X}$;

Step3 计算每个属性所有数据的移动极差的平均值 $\overline{\mathrm{mR}}$;

Step4 采用公式(2)~(7)计算控制界限;

$$\text{X 图:中心线 CL} = \overline{X}; \quad (2)$$

$$\text{上控制限 UCL} = \overline{X} + 2.66 \times \overline{\mathrm{mR}} \quad (3)$$

$$\text{下控制限 LCL} = \overline{X} - 2.66 \times \overline{\mathrm{mR}} \quad (4)$$

$$\text{mR 图:中心线 CL} = \overline{\mathrm{mR}} \quad (5)$$

$$\text{上控制限 UCL} = 3.27 \times \overline{\mathrm{mR}} \quad (6)$$

$$\text{下控制限 LCL} = 0 \quad (7)$$

Step5 绘制 X 图与 mR 图;

Step6 找出每个属性的异常点,即每个属性 X 图中超出控制上限与控制下限的点的横坐标表示的样本是异常点,mR 图中超出控制上限与控制下限的点的横坐标表示的样本序号加 1 对应的点是异常点,总的异常点是所有异常点的并集;

Step7 将筛选出的异常点删除,进行后续分类计算.

以上算法步骤流程如图 2 所示,该算法的优势是计算步骤少且简单易行.

2.2 实例计算

从 New-Thyroid1 数据集(KEEL 数据集的一种)中,随机选取 15 个样本(见表 1),采用本文提出的 XmR 控制图算法进行异常点检测,每个样本都有 5 个属性:a_1, a_2, a_3, a_4, a_5.

第一步 根据公式(1)计算样本中每个属性的 mR 值,见表 2.

第二步 根据公式(5)~(7)分别计算每个属性 mR 图中的中心线,控制上限和控制下限,见表 3.

第三步 根据公式(2)~(4)分别计算每个属性 X 图中的中心线,控制上限和控制下限,见表 4.

第四步 根据表 4 及表 3 中的值分别绘制每个属性的 X 图与 mR 图,如图 3~7 所示.

第五步 X 图的异常点是超出界限所对应的横坐标表示的样本序号,从图 5 的 X 图中得出 2 号样本的 a_3 属性异常,从图 6 的 X 图中得出 13 号样本和 14 号样本的 a_4 属性异常,从图 7 的 X 图中得出 14 号样本的 a_5 属性异常. mR 图的异常点是超出界限所对应的横坐标的样本序号加 1,从图 6 的 mR 图中得出 13 号样本和 15 号样本的 a_4 属性异常,从图 7 的 mR 图中得出 15 号样本的 a_5 属性异常,而总的异常点是 X 图中的异常点与 mR 图中的异常点的并集(见表 5),然后将异常点从数据集中剔除得到新的数据集.

表 1 new-thyroid1 中的随机样本

Tab. 1 Random samples in new-thyroid1

样本	a_1	a_2	a_3	a_4	a_5	类别
X_1	105	7.3	1.5	1.5	−0.1	2
X_2	67	23.3	7.4	1.8	−0.6	1
X_3	111	8.4	1.5	0.8	1.2	2
X_4	89	14.3	4.1	0.5	0.2	1
X_5	105	9.5	1.8	1.6	3.6	2
X_6	110	20.3	3.7	0.6	0.2	1
X_7	84	21.5	2.7	1.1	−0.6	1
X_8	113	11.1	1.7	0.8	2.3	2
X_9	97	7.8	1.3	1.2	0.9	2
X_{10}	106	13.4	3	1.1	0	1
X_{11}	104	6.3	2	1.2	4	2
X_{12}	112	5.9	1.7	2	1.3	2
X_{13}	120	1.9	0.7	18.5	24	2
X_{14}	118	3.6	1.5	11.6	48.8	2
X_{15}	106	9.4	1.7	0.9	3.1	2

表 2　样本属性的 mR 值

Tab. 2　mR value of sample attribute

样本	a_1 值	a_1 的 mR 值	a_2 值	a_2 的 mR 值	a_3 值	a_3 的 mR 值	a_4 值	a_4 的 mR 值	a_5 值	a_5 的 mR 值	类别
X_1	105		7.3		1.5		1.5		−0.1		2
X_2	67	38	23.3	16	7.4	5.9	1.8	0.3	−0.6	0.5	1
X_3	111	44	8.4	14.9	1.5	5.9	0.8	1	1.2	1.8	2
X_4	89	22	14.3	5.9	4.1	2.6	0.5	0.3	0.2	1	1
X_5	105	16	9.5	4.8	1.8	2.3	1.6	1.1	3.6	3.4	2
X_6	110	5	20.3	10.8	3.7	1.9	0.6	1	0.2	3.4	1
X_7	84	26	21.5	1.2	2.7	1	1.1	0.5	−0.6	0.8	1
X_8	113	29	11.1	10.4	1.7	1	0.8	0.3	2.3	2.9	2
X_9	97	16	7.8	3.3	1.3	0.4	1.2	0.4	0.9	1.4	2
X_{10}	106	9	13.4	5.6	3	1.7	1.1	0.1	0	0.9	1
X_{11}	104	2	6.3	7.1	2	1	1.2	0.1	4	4	2
X_{12}	112	8	5.9	0.4	1.7	0.3	2	0.8	1.3	2.7	2
X_{13}	120	8	1.9	4	0.7	1	18.5	16.5	24	22.7	2
X_{14}	118	2	3.6	1.7	1.5	0.8	11.6	6.9	48.8	24.8	2
X_{15}	106	12	9.4	5.8	1.7	0.2	0.9	10.7	3.1	45.7	2

表 3　mR 图中的相关值

Tab. 3　Related values in the mR graph

	a_1	a_2	a_3	a_4	a_5
中心线	16.93	6.56	1.86	2.86	8.29
上控制限	55.36	21.45	6.08	9.35	27.11
下控制限	0	0	0	0	0

表 4　X 图中的相关值

Tab. 4　Correlation values in X graph

	a_1	a_2	a_3	a_4	a_5
中心线	103.13	10.93	2.42	3.01	5.89
上控制限	148.16	28.38	7.37	10.62	27.94
下控制限	58.10	−6.52	−2.53	−4.6	−16.16

表 5　异常点

Tab. 5　Outlier

样本	a_1	a_2	a_3	a_4	a_5	类别
2	67	23.3	7.4	1.8	−0.6	1
13	120	1.9	0.7	18.5	24	2
14	118	3.6	1.5	11.6	48.8	2
15	106	12	9.4	5.8	1.7	2

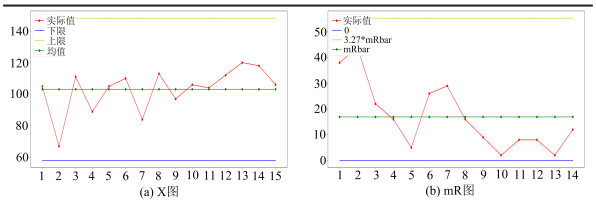

图 3　属性 a_1 的 X 图与 mR 图

Fig. 3　X graph and mR graph of attribute a_1

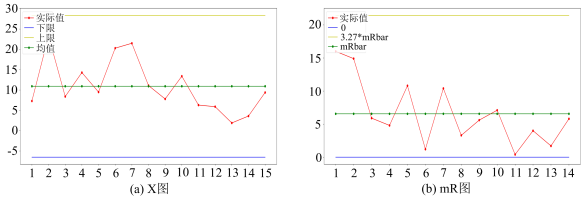

图 4　属性 a_2 的 X 图与 mR 图

Fig. 4　X graph and mR graph of attribute a_2

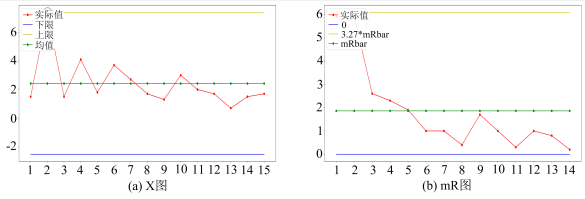

图 5　属性 a_3 的 X 图与 mR 图

Fig. 5　X graph and mR graph of attribute a_3

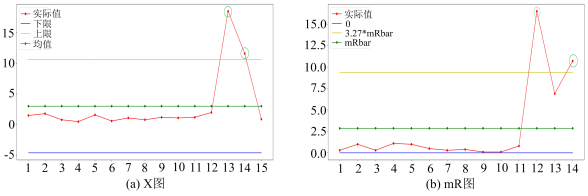

图 6　属性 a_4 的 X 图与 mR 图

Fig. 6　X graph and mR graph of attribute a_4

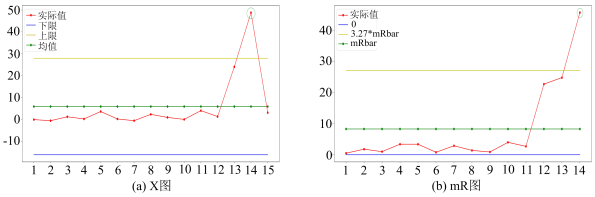

图 7　属性 a_5 的 X 图与 mR 图

Fig. 7　X graph and mR graph of attribute a_5

3　仿真实现

在 KEEL 数据集和 UCI 数据集上进行实验，分别采用 XmR 异常点检测算法和隔离森林算法进行仿真，并对结果进行对比分析.

3.1　数据准备阶段

本文运行数据集的硬件环境是 Intel(R) Core (TM) i7-8550U CPU @ 1.80GHz 1.99 GHz；RAM：8GB. 软件环境是 Python 3.7.3，Anaconda 4.8.3，Spyder 4.1.3；操作系统是 Windows 家庭中文版，系

统类型为 64 位操作系统,基于 x64 的处理器. 基于以上环境,用本文提出的 XmR 异常点检测算法对 KEEL 数据集和 UCI 数据集中的数据进行测试,这些数据集中包括二分类数据和多分类数据,这里只对二分类数据进行测试,数据集的基本信息见表 6.

表 6 数据集基本信息

Tab. 6 Basic data set information

数据集	样本数量	属性个数	类别数	正类	负类	不平衡比
glass0	214	9	2	70	144	2.06
haberman	306	3	2	81	225	2.78
vehicle2	846	18	2	218	628	2.88
vowel0	988	13	2	90	898	9.98
yeast1	1484	8	2	429	1055	2.46
yeast3	1484	8	2	163	1321	8.10
magic04	19020	10	2	6688	12332	1.84

3.2 评价标准

数据的异常点检测问题属于二分类问题,本文使用 G-means 值,F-measure 值以及 AUC 值衡量算法的优越性.

$$G\text{-mean}=\sqrt{\text{TPR}\times\text{TNR}} \tag{8}$$

$$F\text{-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{9}$$

式中,精确率 Precision 和召回率 Recall 计算公式为

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

ROC 曲线可以用于评价分类器的好坏,这条曲线关注两个指标:真正率(TPR)和假正率(FPR). 在 ROC 曲线中,横坐标是 FPR,纵坐标是 TPR.

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}},\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{12}$$

式中,TP 表示少数类正确分类数,FP 表示多数类错误分类数,TN 表示多数类正确分类数,FN 表示少数类错误分类数.

3.3 结果分析

为了说明本文算法的优越性,采用 3 种分类器:CART 决策树、随机森林以及支持向量机,并与隔离森林方法进行了比较,几种分类器在不同数据集上的对比结果如表 7~10 所示.

表 7 时间对比分析表

Tab. 7 Time comparison analysis table

分类器异常点检测方法	CART			随机森林			支持向量机		
	XmR	隔离森林	无	XmR	隔离森林	无	XmR	隔离森林	无
glass0	2.76	21.35	0.40	100.71	104.73	73.92	3.72	22.54	0.98
haberman	0.83	23.69	0.48	92.84	106.77	88.72	3.90	22.83	1.37
vehicle2	8.60	22.97	0.59	116.61	136.48	110.03	12.80	33.21	4.05
vowel0	4.47	28.91	0.66	121.64	141.67	116.55	4.71	28.98	1.09
yeast1	9.72	23.01	0.62	146.30	2248.78	192.05	9.88	17.23	0.80
Yeast3	9.64	17.22	0.51	117.85	125.62	105.15	9.65	17.18	0.58
magic04	339.14	22.19	2.06	759.44	399.91	381.20	346.62	28.85	2290.59

表 8 G-means 均值-标准差表

Tab. 8 G-means mean-standard deviation table

分类器异常点检测方法	CART			随机森林			支持向量机		
	XmR	隔离森林	无	XmR	隔离森林	无	XmR	隔离森林	无
glass0	0.78±0.07	0.76±0.07	0.77±0.08	0.85±0.06	0.84±0.06	0.84±0.07	0.40±0.27	0.41±0.28	0.33±0.29
haberman	0.53±0.07	0.52±0.10	0.51±0.09	0.48±0.09	0.47±0.09	0.47±0.11	0.11±0.16	0.11±0.16	0.14±0.15
vehicle2	0.94±0.02	0.94±0.02	0.94±0.02	0.98±0.02	0.97±0.01	0.98±0.01	0.90±0.10	0.87±0.15	0.89±0.13
vowel0	0.96±0.04	0.95±0.04	0.95±0.04	0.97±0.03	0.97±0.03	0.96±0.03	0.91±0.05	0.90±0.04	0.90±0.05
yeast1	0.63±0.03	0.62±0.03	0.63±0.04	0.66±0.04	0.65±0.04	0.65±0.04	0.52±0.03	0.51±0.04	0.51±0.04
yeast3	0.81±0.05	0.80±0.05	0.81±0.06	0.83±0.05	0.82±0.05	0.83±0.05	0.75±0.05	0.74±0.05	0.74±0.05
magic04	0.80±0	0.79±0	0.79±0.01	0.85±0	0.85±0	0.85±0	0.51±0.18	0.47±0.18	0.46±0.19

表 9　*F*-measure 均值-标准差表

Tab. 9　*F*-measure mean-standard deviation table

分类器异常点检测方法	CART			随机森林			支持向量机		
	XmR	隔离森林	无	XmR	隔离森林	无	XmR	隔离森林	无
glass0	0.70±0.09	0.76±0.09	0.77±0.10	0.79±0.07	0.80±0.08	0.79±0.08	0.36±0.28	0.35±0.27	0.30±0.28
haberman	0.36±0.08	0.35±0.10	0.34±0.02	0.32±0.10	0.31±0.10	0.32±0.11	0.16±0.17	0.15±0.17	0.15±0.15
vehicle2	0.91±0.03	0.92±0.03	0.91±0.03	0.98±0.02	0.97±0.01	0.97±0.02	0.85±0.11	0.83±0.17	0.85±0.15
vowel0	0.92±0.06	0.91±0.06	0.90±0.05	0.96±0.03	0.97±0.03	0.96±0.04	0.85±0.06	0.85±0.06	0.84±0.07
yeast1	0.50±0.04	0.50±0.04	0.51±0.04	0.56±0.05	0.55±0.05	0.55±0.05	0.40±0.04	0.39±0.05	0.39±0.04
yeast3	0.67±0.06	0.67±0.06	0.67±0.07	0.75±0.06	0.74±0.07	0.75±0.06	0.68±0.06	0.67±0.06	0.66±0.07
magic04	0.86±0	0.85±0	0.86±0	0.92±0	0.91±0	0.91±0	0.67±0.26	0.68±0.25	0.64±0.30

表 10　AUC 均值-标准差表

Tab. 10　AUC mean-standard deviation table

分类器异常点检测方法	CART			随机森林			支持向量机		
	XmR	隔离森林	无	XmR	隔离森林	无	XmR	隔离森林	无
glass0	0.79±0.06	0.77±0.06	0.77±0.07	0.86±0.06	0.85±0.06	0.85±0.06	0.61±0.09	0.62±0.10	0.59±0.10
haberman	0.57±0.06	0.56±0.07	0.56±0.07	0.56±0.05	0.56±0.06	0.57±0.06	0.51±0.03	0.51±0.03	0.52±0.02
vehicle2	0.93±0.02	0.94±0.02	0.94±0.02	0.98±0.02	0.97±0.01	0.97±0.01	0.91±0.08	0.89±0.11	0.90±0.09
vowel0	0.96±0.04	0.95±0.04	0.95±0.03	0.96±0.03	0.97±0.03	0.96±0.03	0.92±0.04	0.91±0.04	0.90±0.05
yeast1	0.65±0.03	0.64±0.03	0.65±0.03	0.69±0.03	0.69±0.03	0.68±0.03	0.62±0.02	0.61±0.02	0.61±0.02
yeast3	0.83±0.04	0.82±0.04	0.82±0.05	0.84±0.04	0.83±0.04	0.84±0.04	0.78±0.04	0.77±0.04	0.77±0.04
magic04	0.79±0	0.80±0	0.79±0.01	0.86±0	0.85±0	0.85±0	0.63±0.07	0.61±0.07	0.61±0.07

从表 7 可以看出，对比 CART、随机森林以及支持向量机这 3 种分类器的结果，XmR 异常点检测算法对于大部分数据集用时明显减少，提高了计算效率，但对于个别数据集（如数据集 magic04）效果不理想，这说明该算法还有一定的提升空间.

G-means 是一种整体上衡量数据集分类性能的指标，该值越大，算法的性能越好，从表 8 可以看出，对于大部分数据集，XmR 异常点检测算法的 *G*-means 值是高于隔离森林的 *G*-means 值，但是当 XmR 异常点检测算法与隔离森林结合时，对数据集 glass0 的分类效果不是很理想. 总体来看，XmR 异常点检测算法优于隔离森林.

F-measure 值越大，表示算法的性能越好，从表 9 可以看出，对 CART 决策树来说，XmR 异常点检测算法对数据集 haberman、vowel0 和 magic04 的分类效果比隔离森林算法效果好；对随机森林分类器来说，XmR 异常点检测算法对于大部分数据集效果理想，但是对于个别数据集，比如 glass0 和 vowel0 效果不是很理想，对于支持向量机来说，XmR 异常点检测算法的 *F*-measure 值对于大部分数据集是高于隔离森林的 *F*-measure 值. 简而言之，XmR 异常点检测算法与不同的分类器结合会有不同的分类效果，因此本文提出的算法具有一定的可行性.

从表 10 可以看出，对于任意一种分类器，XmR 异常点检测算法对数据集 yeast3 的分类准确率最高，对数据集 haberman 的效果不是很理想，说明本文提出的算法具有一定的优越性，但是还有待提高.

由表 7～10 可知，这些数据在经过 XmR 异常点检测后，*G*-means 均值，*F*-measure 均值，AUC 均值相比隔离森林要高，但是稳定性有待提升.

仿真实验结果整体来看，XmR 异常点检测算法可以用来检测出异常点且计算比较简单，用时较短，相比隔离森林方法，时间效率和准确率方面均有明显优势.

4　结论

本文提出的基于 XmR 控制图的异常点检测算法，通过计算样本属性的单值均值、移动极差及属性均值并绘制 *X* 图和 mR 图，能够快速检测异常点，不需要大量的计算，改善了传统异常点检测方法中由于计算繁琐所带来的耗时较长、参数选择敏感等缺陷. 仿真结果表明，该方法所清洗的数据，在 CART、随机森林以及支持向量机传统智能算法中均具有较好的分类性能. XmR 控制图的异常点检测算法是对异常点检测研究的一种有益尝试和补充，具有一定的应用价值，后续研究中，课题组将把该算法进一步应用于网络入侵检测、公共卫生检测、银行欺诈检测等不同领域，经过异常检测数据预处理后进行挖掘，更加有利于获取数据中的隐含价值，提升数据挖掘效率.

参考文献（References）

[1] 王东. 数据挖掘在检测农业补贴中欺诈行为的应用——基于异常检测与神经网络模型[J]. 平顶山学院学报，2015，30(05)：75-78.

[2] 王立英. 异常点检测算法及在网络入侵检测中的应用研究[D]. 济南：山东师范大学，2020.

[3] 王康，周治平. 高斯核密度估计方法检测健康数据异常值[J]. 计算机科学与探索，2019，13(12)：2094-2102.

（下转第 1186 页）