



离群点检测算法的评价指标

宁进^{1,2}, 陈雷霆^{1,2,3}, 罗子娟⁴, 周川^{1,2*}, 曾慧茹^{1,2}

(1. 电子科技大学 计算机科学与工程学院, 成都 611731; 2. 数字媒体技术四川省重点实验室(电子科技大学), 成都 611731;

3. 电子科技大学 广东电子信息工程研究院, 广东 东莞 523808;

4. 中国电子科技集团公司第二十八研究所 信息系统工程重点实验室, 南京 210007)

(* 通信作者电子邮箱 zhouchuan@uestc.edu.cn)

摘要:随着离群点检测技术的深入研究和广泛应用,越来越多的优秀算法被提出来,然而,现有的离群点检测技术的评价仍然沿用传统分类算法的测量指标,存在着评价指标单一、适应性差的问题。针对这些问题,提出了一类高真正率指标(HT_AUC)和二类低假正率指标(LF_AUC)。首先,整理常用的离群点检测评价指标,分析其优缺点和适用场景;然后,在已有的曲线下面积(AUC)方法的基础上,分别针对高真正率(TPR)要求和低假正率(FPR)要求,提出了一类高真正率指标和二类低假正率指标,为离群点检测算法的效果评价和量化集成提供了更合适的指标。在真实数据集上的实验结果表明,与传统评价指标的相比,所提出的方法更能满足一类高真正率和二类低假正率要求。

关键词:离群点检测;评价指标;曲线下面积;真正率;假正率

中图分类号:TP181 **文献标志码:**A

Evaluation metrics of outlier detection algorithms

NING Jin^{1,2}, CHEN Leiting^{1,2,3}, LUO Zijuan⁴, ZHOU Chuan^{1,2*}, ZENG Huiru^{1,2}

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan 611731, China;

2. Digital Media Technology Key Laboratory of Sichuan Province (University of Electronic Science and Technology of China), Chengdu Sichuan 611731, China;

3. Institute of Electronic and Information Engineering in Guangdong, University of Electronic Science and Technology of China, Dongguan Guangdong 523808, China;

4. Information System Engineering Laboratory, The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing Jiangsu 210007, China)

Abstract: With the in-depth research and extensive application of outlier detection technology, more and more excellent algorithms have been proposed. However, the existing outlier detection algorithms still use the evaluation metrics of traditional classification, which leads to the problems of singleness and poor adaptability of evaluation metrics. To solve these problems, the first type of High True positive rate-Area Under Curve (HT_AUC) and the second type of Low False positive rate-Area Under Curve (LF_AUC) were proposed. First, the commonly used outlier detection evaluation metrics were analyzed to illustrate their advantages and disadvantages as well as applicable scenarios. Then, based on the existing Area Under Curve (AUC) method, the HT_AUC and the LF_AUC were proposed aiming at the high True Positive Rate (TPR) demand and low False Positive Rate (FPR) demand respectively, so as to provide more suitable metrics for performance evaluation as well as quantization and integration of outlier detection algorithms. Experimental results on real-world datasets show that the proposed method is able to better satisfy the demands of the first type of high true rate and the second type of low false positive rate than the traditional evaluation metrics.

Key words: outlier detection; evaluation metric; Area Under Curve (AUC); True Positive Rate (TPR); False Positive Rate (FPR)

0 引言

离群点,也可称为异常点,是数据集中与大多数点不一致,或是由不同机制产生的数据^[1]。例如在海上安防系统中^[2]中,入侵船只被看作是异常点,需要拦截。雷达数据处理

中^[3],噪声被看作是离群点,需要过滤以防止干扰建模。

近年来,离群点检测算法依然是数据挖掘的热点方向。各种基于统计、基于邻近性、基于分类、基于聚类、基于集成的方法等^[4-5]层出不穷,以取得更好的离群点检测效果。离群点检测算法的输出通常为离群得分,得分越高,越可能是离群

收稿日期:2020-02-13;修回日期:2020-04-22;录用日期:2020-04-28。

基金项目:四川省科技计划项目(2019YJ0177, 2019YJ0176, 2019YFQ0005)。

作者简介:宁进(1991—),女,四川成都人,博士研究生,主要研究方向:离群点检测、数据挖掘; 陈雷霆(1966—),男,四川成都人,教授,博士生导师,主要研究方向:图像处理、计算机图形、虚拟现实; 罗子娟(1985—),女,江西泰和人,高级工程师,硕士,主要研究方向:遥感影像目标识别、影像变化检测; 周川(1977—),男,四川成都人,讲师,博士,主要研究方向:计算机动画、机器视觉、医学图像分析、数据挖掘; 曾慧茹(1994—),女,江西赣州人,博士研究生,主要研究方向:深度学习、医学信息学。



点。基于统计的方法对正常数据建模,用与正常模式的偏离程度来表示离群得分。基于邻近性的方法用与邻居差异程度来表示离群得分。基于分类的方法用与分界线的偏离程度来衡量离群得分。基于聚类的方法视离群点为聚类的副产物,用与正常簇的偏离程度来衡量离群得分。基于集成的方法通过集成多个结果得到最终的离群得分。

由于离群点本身的少量、多变,以及难以预知、难以建模的特点,离群点检测算法常采用无监督方法。再加上缺少离群点的标签,使得离群点检测的评价变得困难。离群点检测一般使用外部度量来进行评价,这种度量需要已有的真实标签来进行。现有的离群点检测算法评价指标主要分为三类,如图 1。第一种是阈值法,在离群得分的基础上,利用所设置的阈值来划分预测的离群点集。将预测的离群点集与真实的离群点标签作对比,用检测率、精确度等统计值来评价算法效果。第二种是曲线法,将阈值法的全参数下的指标绘制连续的曲线,曲线越“凸”,表示算法效果越好。第三种是整合法,用曲线下的面积来衡量算法效果,值越大,表示算法的效果越好。

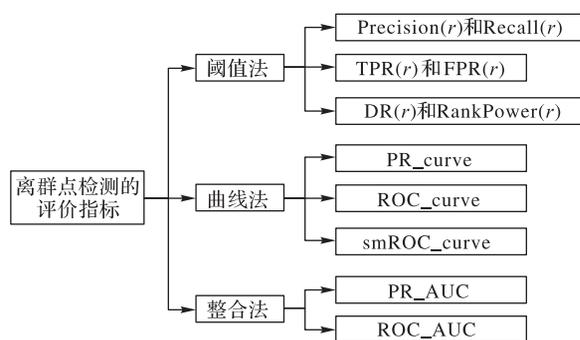


图 1 离群点检测算法评价指标

Fig. 1 Evaluation metrics of outlier detection algorithm

近年来,一些改进的方法也被提出来了。例如 Zhang 等^[6]提出了一种带标准化的精确度的均值,以包含离群度排位信息;但是,这种方法在没有调整的时候会产生错误^[7]。Klement 等^[8]针对受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线丢失离群得分信息的问题,提出了一种平滑的 ROC 曲线,通过对 ROC 曲线加入平滑分量以保留离群得分信息,对评价算法的差异更具有一致性。此外,Marques 等^[9]提出了一种不需要真实标签的内部评价方式,这种方式基于离群得分的相对评价,但是计算复杂度太高。

尽管已有很多适合的评价指标,但很多离群点检测文献仍然存在评价方法选择不当、使用不当的问题,使得所得出的结论站不住脚。例如,如果错将正常点标记 1,异常点标记 0,得出的评价指标虚高。再例如,使用阈值法时,阈值设置不合理,得出的指标结果偏差也大。此外,离群点检测算法的评价要求常常分为两类:一类要求高真正率,例如在疾病检测中,要求检测到所有患病者,即使存在将正常人归为患病类;二类要求低假正率。例如在垃圾邮件检测中,要求不能把有用邮件误归为垃圾邮件,即使漏检部分真正的垃圾邮件。总之,由于离群点检测算法的特殊性,目前,仍然缺乏针对离群点检测问题的专门的系统评价方法研究。

本文首先对离群点检测算法的已有评价指标做了一个详细的整理,为研究者评价所提出的算法提供评价指标的说明和参考;然后针对已有指标不能区分一类和二类要求的问题,提出了一类高真正率评价指标 (High True positive rate-Area Under Curve, HT_AUC) 和二类低假正率评价指标 (Low False positive rate-Area Under Curve, LF_AUC),通过计算证明和在真实数据集上与已有方法的对比实验,说明了本方法的适

用性。

1 常用的离群点检测算法评价指标

设 N 个点的离散数据集 D 中, O 表示真实的离群点集 (令集合大小 $|O|=n$), NO 表示正常点集 (令集合大小 $|NO|=m$)。离群点检测算法大多返回离群得分 (Outlier Score, OS)^[1], 可以是距离、密度、概率等。离群得分越高,越可能是离群点。 $OS(p)$ 表示点 p 的离群得分。 $rank(p)$ 表示点 p 的离群得分在 OS 中的排位,离群得分越高, $rank$ 值越小,位次越高。离群点的标签应是正类 (这里用“1”表示);正常点的标签应是负类 (这里用“0”表示)。

1.1 阈值法

步骤 1 设定阈值。

首先通过设定阈值将离群得分转化成算法判定的 \hat{O} 集和 \widehat{NO} 集。一种方法是关于离群得分的阈值,根据经验设置阈值 α (可以是固定值,可以是统计值等^[10])。

$$\hat{O} = \{ p | OS(p) \geq \alpha \}$$

另一种是 TOP r ($1 \leq r \leq N$, 评价的时候只要真实的标签可用,那么 r 就可以设为 n), 表示将离群得分排在前 r 的点判为离群点。

$$\hat{O} = \{ p | rank(p) \leq r \}$$

步骤 2 计算评价指标对。

离群点检测算法所采用的评价指标对主要有 3 组,分别是精确度 (Precision) 和召回率 (Recall)^[11-13], 真正率 (True Positive Rate, TPR) 和假正率 (False Positive Rate, FPR)^[14-15], 检测率 (Detection Rate, DR) 和排位力 (Rank power, Rp)^[16-17]。计算方法如表 1。其中: TP 表示将离群点标记为离群点的量; FP 表示将正常点标记为离群点的量; FN 表示将离群点标记为正常点的量; TN 表示将离群点标记为离群点的量。

表 1 阈值法的评价指标计算方法

Tab. 1 Evaluation metrics calculation method of threshold method		
指标	第一种表达式	第二种表达式
Precision	$ \hat{O} \cap O / \hat{O} $	$TP / (TP + FP)$
Recall/TPR/ DR	$ \hat{O} \cap O / O $	$TP / (TP + FN)$
FPR	$ \hat{O} \cap NO / NO $	$FP / (TN + FP)$
Rp	$\frac{k(k+1)}{2} \sum_{i=1}^k rank(i)$; $k = \hat{O} \cap O $	

$Recall=TPR=DR$, 也称为检测准确率,表示预测出的真实离群点数量占所有的真实离群点数量的比,值越高,表示算法效果越好。但这个单一的指标存在着漏洞,即: $|\hat{O}|$ 越大,检测准确率越高。当算法预测所有数据为离群点,即 $|\hat{O}| = |D|$ 时,检测准确率为 1。所以,只有这一个指标还不足以说明算法的效果。 $Precision$ 表示预测出的真实离群点数量占预测的离群点数量的比,值越高,表示算法效果越好。 FPR 表示预测错误的离群点 (真实的正常点预测为离群点) 占正常点数量的比,值越低,表示算法效果越好。 Rp 反映了预测的真实离群点在 $rank$ 中的排位情况,值越高,表示算法的效果越好。所有的离群点排位在 $rank$ 前列时, $Rp=1$ 。 $Precision$ 、 FPR 和 Rp 作为检测准确率的补充增强,可以弥补检测准确率的漏洞;此外 Rp 还利用了 $rank$ 信息,对算法要求更高。

阈值法简单有效,可以直接评价离群点检测算法实验结果的优劣。但是有如下 3 个缺陷:

1) 参数依赖。例如, α 值太高 (或者 r 太小), 漏标多, 评价



值会偏低;α 值太低(或者 r 太大),错标多,评价值会偏高。

2)参数设置困难。大部分论文在使用这种方法评价算法时,会设置 r=|O|,这需要提前知道数据集有多少真实的离群点。然而在实际应用中,很难提前获取真实离群点的量。

3)丢失了 rank 和 score 信息。不能表示算法结果的整体好坏。此外,即使是 Rp 利用了部分 rank 信息,仍然区分不了如下情况。例如表 2:取 r=4 的时候,检测准确率 DR1=DR2=0.5, Rp1=Rp2=0.6,这种情况下,算法 1 和算法 2 的评价结果相同,无法区分好坏。

表 2 Rank Power 的例子

Tab. 2 Examples of Rank Power

算法	rank
算法 1	0 1 1 0 1 0 0 0
算法 2	1 0 0 1 1 0 0 0

4)对于 Precision 和 Recall,在参数相同的情况下,一些好的算法常常要么高 Precision 低 Recall,要么低 Precision 高 Recall。

1.2 曲线法

为了摆脱参数依赖,整合 rank 信息,以更精确地评价各个算法的优劣。通过从 1 到 N 变化参数 r,得到对应的 N 组 Precision 和 Recall。依次连接每对 (Recall(r), Precision(r)) 点绘制 Precision-Recall (PR) 曲线^[13-18](如图 2(a))。同样,通过从 1 到 N 变化参数 r,依次得到对应的 TPR(r) 和 FPR(r),FPR 作横坐标,TPR 作纵坐标,绘制 ROC 曲线^[19-21](如图 2(b))。由于 ROC 曲线比 PR 曲线更直观,且具有单调性,所以一般情况下,多使用 ROC 曲线。ROC 曲线越“凸”,表示算法的效果越好。smROC^[8]在 ROC 曲线的基础上增加了离群得分信息,使得修改的 ROC 曲线更加平滑(如图 2(c))。

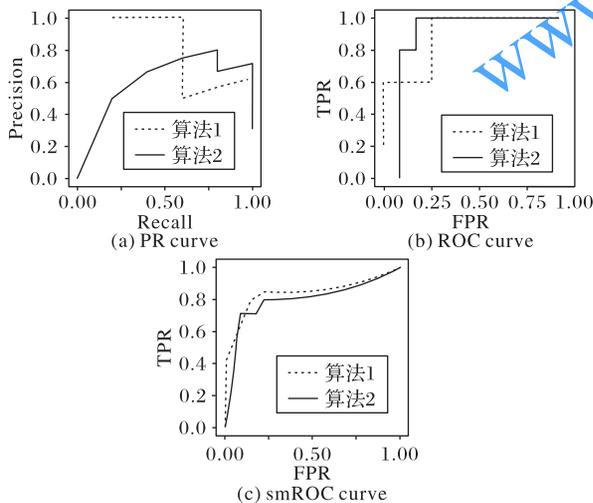


图 2 PR curve、ROC curve 和 smROC curve 的示例

Fig. 2 Examples of PR curve, ROC curve and smROC curve

ROC 曲线应用在离群点检测算法的结果评价上,具有直观、简便、精确的优点,且不受离群点检测数据集类别的有偏性的影响,在一定程度上是很成功的,具有广泛的应用;但仍然有如下缺陷:

1)不够清楚。很多时候,一个算法不会完全地比另一个算法“凸”,例如图 2(b),或者更加错综复杂,算法的优劣需要进一步分情况讨论。

2)不能扩展。大部分离群点检测算法都有除阈值以外的其他参数依赖,例如,基于邻近性的算法依赖参数 k(邻域的大小),基于一类支持向量机算法依赖核函数的选择,用 ROC 曲

线只能展示特定参数下算法的差异。

1.3 整合法

为了验证算法与非阈值参数的关系,通常需要整合曲线,直接用一个数值来体现算法综合能力。使得该数值既有阈值的简单直观性,并保留曲线法的精确性。已经知道,曲线法评价好的算法比坏的算法更“凸”,于是可以用一种曲线的整合形式,即曲线下的面积(Area Under Curve, AUC)来评价算法。数值越高,表示算法效果越好。

PR_AUC^[22-23]是 PR 曲线下的面积,可以由离群点的平均精确度计算。

证明 在 PR 曲线中,随着 r 的增加,当第 r 个数据点真实标签为 1 时, Precision 变为 Precision(r), Recall 增加 1/n, 对应变化面积为 $\frac{1}{n} \left(\frac{P(\text{rank}(i)) + P(\text{rank}(i-1))}{2} \right)$ 。当第 r 个数据点真实标签为 0 时, Recall 不变, Precision 减少, 曲线垂直下降, 变化面积为 0, 所以 PR_AUC 可以计算如下:

$$PR_AUC = \frac{1}{n} \left(\frac{P(\text{rank}(i)) + P(\text{rank}(i-1))}{2} \right) = \sum_{i=0}^{n-1} \frac{1}{n} \frac{P(\text{rank}(i))}{2} + \sum_{i=0}^{n-1} \frac{1}{n} \frac{P(\text{rank}(i-1))}{2} = \sum_{i=0}^{n-1} \frac{1}{n} \frac{P(\text{rank}(i))}{2} + \sum_{i+1=0}^{n-1} \frac{1}{n} \frac{P(\text{rank}(i))}{2} = \sum_{i=0}^{n-1} \frac{1}{n} \frac{P(\text{rank}(i))}{2} + \sum_{i=0}^{n-1} \frac{1}{n} \frac{P(\text{rank}(i))}{2} - \frac{1}{n} \frac{P(\text{rank}(0))}{2} = \frac{1}{n} \sum_{i=0}^{n-1} P(\text{rank}(i))$$

ROC_AUC^[24-25]是 ROC 曲线下的面积,也可由数据集中离群点-正常点对的均值来计算。

证明 离散情况下,在 ROC 曲线中随着 r 增加,当第 r 个数据点真实标签为 1 时,TPR 增加 1/n, FPR 不变, 对应 ROC 曲线垂直上升, 变化面积为 0。当第 r 个数据点真实标签为 0 时, TPR 不变, FPR 增加 1/m, 对应变化面积为 $\frac{1}{m} TPR(\text{rank}(i))$, 所以 ROC_AUC 可以计算如下:

$$ROC_AUC = \sum_{i \in N_0} \frac{1}{m} TPR(\text{rank}(i)) = \frac{1}{m} \sum_{i \in N_0} \frac{TP(\text{rank}(i))}{n} = \frac{1}{m * n} \sum_{i \in N_0} \sum_{j \in O} I(\text{rank}(i), \text{rank}(j))$$

$$其中 I(\text{rank}(i), \text{rank}(j)) = \begin{cases} 1, & \text{rank}(i) > \text{rank}(j) \\ 0.5, & \text{rank}(i) = \text{rank}(j) \\ 0, & \text{rank}(i) < \text{rank}(j) \end{cases}$$

用曲线法评价离群点检测算法效果时,不受数据集中离群点比例的影响。但整合为 ROC_AUC 后,只要求曲线像左上角“凸”,很难保证算法同时有高真正率和低假正率,丢失了曲线的细节信息,不能同时满足一类和二类要求。例如表 3 中, 算法 1 的 $ROC_AUC_1 = \frac{4+4+4}{3 \times 5} = 0.8$, 算法 2 的

$$ROC_AUC_2 = \frac{5+5+2}{3 \times 5} = 0.8$$

算法 1 和算法 2 的 ROC_AUC 值相同,但实际差别很大。算法 1 在 r=4 时,就能检测出所有离群点,而算法 2 在 r=6 的时候才能检测出所有离群点;算法 r=2 时,能检测出 2 个离群点,且未将正常点误判为离群点,而算法 1 无论 r 等于多少,都存在将正常点误判为离群点。

在实际应用中,对于算法 1 和算法 2 有着不同的适用场景。算法 1 适合要求高检测准确率的场景,即要求所有离群点的 rank 靠前,例如疾病检测;算法 2 适合要求低错误率的场景,即要求所有正常点的 rank 靠后,例如垃圾邮件检测。



代价敏感(Meta Cost)方法^[1]通过引入代价因子,作为TPR和FPR的权衡。代价因子 $c(Y, n)$ 表示将正常点预测为离群点的代价, $c(N, y)$ 表示将离群点预测为正常点的代价。通过修改 I 函数,每项不等式右乘 $\frac{c(Y, n)}{c(N, y)}$,为不同类型的错误分类设置不同的代价。当设置 $c(Y, n) > c(N, y)$ 时,表示正常点预测为离群点的代价更高,最终的meta_AUC比ROC_AUC更小;当设置 $c(Y, n) < c(N, y)$ 时,表示离群点预测为正常点的代价更高,最终的meta_AUC比ROC_AUC更大。这种方法通过设置两个代价因子来权衡参数依赖,需要依靠经验设置。代价因子的可解释性较弱,不便于使用。

表3 ROC_AUC的例子
Tab. 3 Examples of ROC_AUC

算法	rank
算法1	0 1 1 1 0 0 0 0
算法2	1 1 0 0 0 1 0 0

综上所述,阈值法适合在应用决策时使用,曲线法适合算法效果的精确展示,整合法适合在参数控制时使用。已有的离群点检测评价方式常常采用以上指标的综合方案^[8],以便优势互补,充分验证算法的效果。

2 方法

2.1 高真正率和低假正率指标

定义1 一类高真正率要求:要求TPR接近1,对应ROC曲线向顶部“凸”。

定义2 二类低假正率要求:要求FPR接近0,对应ROC曲线向左部“凸”。

例如在疾病检测中,将患病(标签为“1”)错标记为正常(“0”),会导致该患者得不到治疗。如果是传染病,漏检还会发生进一步传染,产生严重的后果。因此,疾病检测系统要求一类高真正率,检测到所有患病者,即使存在将正常人归为患病类,可以进一步检测排除“疑似类”。在垃圾邮件检测中,将重要邮件(标签为“0”)误判为垃圾邮件(标签为“1”),会给收件人带来难以估量的影响。因此垃圾邮件检测系统要求二类要求低假正率,要求不能把重要邮件误判为垃圾邮件,即使漏检部分真正的垃圾邮件。

为了同时解决已有整合法的信息丢失和参数依赖的问题,适应一类高真正率和二类低假正率要求,本文提出了HT_AUC和LF_AUC。

$$HT_AUC = \frac{1}{m*n} \sum_{i \in H} \sum_{j \in O} I(\text{rank}(i), \text{rank}(j)) + \frac{1}{m} * [\alpha * m] \quad (1)$$

其中: $\alpha \in [0, 1]$ 是控制变量,表示求算法的ROC曲线在 $FPR > \alpha$ 时具有高TPR, H 表示 NO 中rank值在后 $[(1 - \alpha) * m]$ 的点的集合。式(1)中第一个加项表示ROC曲线后 $1 - \alpha$ 部分曲线的面积,第二个加项表示忽略ROC曲线前 α 部分曲线的面积,适应一类要求。

$$LF_AUC = \frac{1}{m*n} \sum_{i \in L} \sum_{j \in NO} I(\text{Rank}(i), \text{Rank}(j)) + \frac{1}{n} * [(1 - \alpha) * n] \quad (2)$$

其中: $\alpha \in [0, 1]$ 是控制变量,表示求算法的ROC曲线在 $TPR < \alpha$ 时具有低FPR, L 表示 O 中rank值在前 $[\alpha * n]$ 的点的集合。式(2)中第一个加项表示ROC曲线下面 α 部分曲线的面积,第二个加项表示忽略ROC曲线后 $1 - \alpha$ 部分曲线的面积,适应二类要求。

例如,表3中算法1和算法2,取 $\alpha = 0.2$,可以计算出:

$$HT_AUC_1 = \frac{3 + 3 + 3 + 3}{3 \times 5} + 0.2 = 1$$

$$HT_AUC_2 = \frac{2 + 2 + 3 + 3}{3 \times 5} + 0.2 \approx 0.867$$

$$LF_AUC_1 = \frac{4 + 4}{3 \times 5} + \frac{1}{3} = \frac{13}{15}$$

$$LF_AUC_2 = \frac{5 + 5}{3 \times 5} + \frac{1}{3} = 1$$

$HT_AUC_1 > HT_AUC_2$,说明算法1更能适应一类要求, $LF_AUC_2 > LF_AUC_1$,说明算法2更能适应二类要求。

2.2 证明

1)当 $\alpha = 0$ 时,

$$HT_AUC = ROC_AUC; LF_AUC = 1$$

2)当 $\alpha = 1$ 时,

$$HT_AUC = 1; LF_AUC = ROC_AUC$$

3)当 $0 < \alpha < 1$ 时,

$$HT_AUC = \sum_{i \in H} \frac{1}{m} TPR((\text{rank}(i)) + 1 * \frac{1}{m} * [\alpha * m]) =$$

$$\frac{1}{m} \sum_{i \in H} \frac{TP(\text{rank}(i))}{n} + \frac{1}{m} * [\alpha * m] =$$

$$\frac{1}{m * n} \sum_{i \in H} \sum_{j \in O} I(\text{rank}(i), \text{rank}(j)) + \frac{1}{m} * [\alpha * m]$$

$$LF_AUC = \sum_{i \in L} \frac{1}{n} (1 - FPR(\text{Rank}(i))) + 1 * \frac{1}{n} * [(1 - \alpha) * n] =$$

$$\sum_{i \in L} \frac{1}{n} (1 - FPR(\text{Rank}(i))) + \frac{1}{n} * [(1 - \alpha) * n] =$$

$$\sum_{i \in L} \frac{1}{n} \left(1 - \frac{FP(\text{Rank}(i))}{m} \right) + \frac{1}{n} * [(1 - \alpha) * n] =$$

$$\frac{1}{m * n} \sum_{i \in L} \sum_{j \in NO} I(\text{Rank}(i), \text{Rank}(j)) + \frac{1}{n} * [(1 - \alpha) * n]$$

本方法通过调整参数 α 控制一类高真正率或者二类低假正率要求的程度。对于HT_AUC,表示在容忍 $FPR = \alpha$ 的情况下整合TPR越高越好, α 越小越接近ROC_AUC;对于LF_AUC,表示在满足 $TPR = \alpha$ 的情况下整合FPR越低越好, α 越大越接近ROC_AUC。相较于Meta Cost中的代价因子,本文方法的参数可解释性更强,更容易设置,参数依赖性更低。

3 实验结果与分析

3.1 实验准备

数据集取自UCI的30个真实数据集^[26],表4展示了这些真实数据集的特征。将数量稀少的类或者特选类中的数据点作为离群点,剩余的数据点作为正常点。

1)一类和二类要求。为了验证本文评价方法的有效性,本文首先细化一类要求:要求在 $FPR=40\%$ 时,离群点检测算法的TPR越高,算法效果越好。这种要求表示在同等等容错下,检测准确率越高的算法越能满足高真正率要求。然后细化二类要求:要求在 $TPR=80\%$ 时,离群点检测算法的FPR越低,算法效果越好。这种要求表示在同等等检测率下,检测错误率越低的算法越能满足低假正率要求。实验平台为3.4 GHz CPU, 8 GB RAM, Windows10系统, PyCharm社区版,采用Python编程。

2)离群点检测方法。使用下列4种经典的离群点检测算法^[18,27]作为评价指标的对比算法:局部异常因子(Local Outlier Factor, LOF)、K最近邻(K Nearest Neighbor, KNN)、孤立森林(Isolation Forest, IF)、不稳定因子(INStability factor, INS)。这4种不同类型的算法在每个数据集上的检测结果有不同程度的差异,本实验的目的即比较出更能区分这些算法



在不同要求下效果优劣的评价指标。

3)对比方法。将本文提出的HT_AUC和LF_AUC方法与已有的PR_AUC,ROC_AUC以及meta_AUC(代价比分别设为1.25和0.8)作对比。一类要求的评价方法对比策略:以每个算法在FPR=40%时的TPR值作为基准指标,按从大到小对算法排序,再对比HT_AUC与其他3个方法的评价排序,与基准指标越接近(排序的欧式距离越小)的评价方法越好;同理,二类要求的评价方法对比策略:以每个算法在TPR=80%时的FPR值作为基准指标,按从小到大对算法排序,再对比HT_AUC与其他3个算法的评价排序,与基准指标越接近(排序的欧式距离越小)的评价方法越好。

表 4 真实数据集的描述

Tab. 4 Description of real-world datasets

数据集 标签	数据集名	维数	类数	离群类	总数据 量	离群点 数量
1	Wine-red	11	6	<5	1 599	63
2	Wine-white	11	7	<5	4 898	183
3	Wine	13	3	3	138	8
4	Messidor	19	2	1	1 151	611
5	Hepatitis	19	2	1	155	32
6	Ionosphere	33	2	b	351	126
7	BreastTissue	9	6	car	106	21
8	Banknote	4	2	1	1 372	610
9	Faults	27	2	1	1 941	673
10	Iris-versicolor	4	3	versicolor	150	50
11	Iris-setosa	4	3	setosa	150	50
12	Iris-virginica	4	3	virginica	150	50
13	Sonar	59	2	M	208	20
14	Lung-cancer	56	3	2	32	13
15	Wholesale	6	2	2	440	142
16	Seeds	7	3	2	210	70
17	ILPD	10	2	2	579	165
18	Page-blocks	10	5	>1	5 473	560
19	Musk	166	2	0	476	207
20	Phishing	9	3	1	1 353	548
21	Abalone	8	29	<5和>21	4 177	96
22	Breast-cancer	9	2	4	683	239
23	Turkiye	27	13	1	5 820	303
24	Spambase	57	2	1	4 601	1 813
25	HTRU_2	8	2	1	17 898	1 639
26	Magic04	10	2	h	19 020	6 688
27	Mammographic	4	2	1	830	403
28	Transfusion	4	2	1	748	178
29	Pop_failures	18	2	0	540	46
30	Biodeg	41	2	RB	1 055	356

3.2 结果及分析

图3记录了HT_AUC与对比方法在30个真实数据集上的实验结果。可以看出,meta_AUC在大部分数据集上具有最高的差异度,也就是与基准指标的差异最大,这是由于代价因子的影响。PR_AUC和ROC_AUC的方法大部分时候与基准指标差异不大,能基本满足一类高真正率要求。HT_AUC在大部分情况下结果和ROC_AUC一致,部分数据集上能展示出更好的效果。因此,可以得出结论,HT_AUC比其他指标更能满足一类高真正率要求。

图4记录了LF_AUC与对比方法在30个真实数据集上的实验结果。同样的,meta_AUC在大部分数据集上与基准指标的差异较大。PR_AUC和ROC_AUC的方法大部分时候与基准指标差异不大,能基本满足二类低假正率要求。LF_AUC

在大部分情况下结果和ROC_AUC一致,其余数据集上能展示出更好的效果。因此,也可以得出结论,LF_AUC比其他指标更能满足二类低真正率要求。

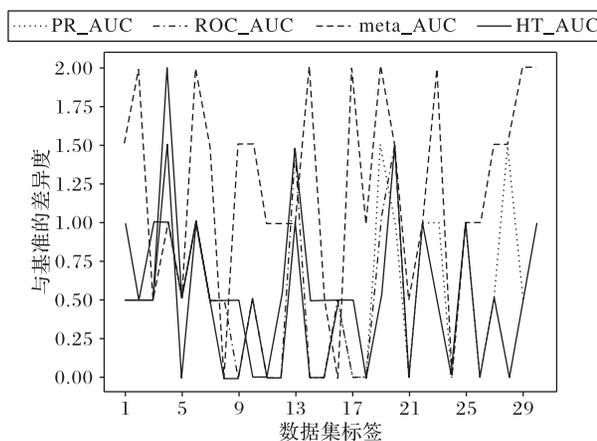


图3 HT_AUC与传统评价方法的结果对比

Fig. 3 Result comparison of the proposed HT_AUC and traditional methods

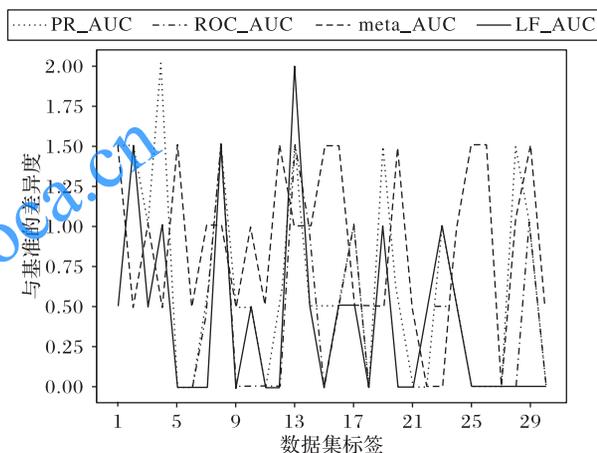


图4 LF_AUC与传统评价方法的结果对比

Fig. 4 Result comparison of the proposed LF_AUC and traditional methods

整体来看,所提出HT_AUC和LF_AUC指标相较于其他方法,与基准指标的差异最小,更能满足一类高真正率要求和二类低假正率要求。该方法可作为具有特别要求系统的评价指标,例如要求一类高真正率的疾病检测可使用HT_AUC指标,要求二类低假正率的垃圾邮件检测可使用LF_AUC指标。

4 结语

本文对离群点检测领域内常见的评价方法作了归纳整理,并提出了满足一类高真正率要求的HT_AUC指标和满足二类低假正率要求的LF_AUC指标。已有离群点检测评价方式建议采用两类以上的评价指标,以便优势互补,充分验证算法的效果。其中,阈值法适合工业选择时使用,曲线法适合算法效果的精确展示,整合法适合在参数控制时使用。实验结果表明,如果应用对算法的真正率和假正率有特殊要求,采用所提出的HT_AUC和LF_AUC指标,能更好地评价所使用的算法。本文所涉及的数据对象主要是离群数据集,未来将继续对序列离群点检测算法的评价方法进行研究。

参考文献 (References)

[1] AGGARWAL C C. Outlier Analysis [M]. 2nd ed. Cham: Springer, 2017: 286-286.



- [2] 王珂, 惠新成, 张遥. 近岸海上安保快艇拦截任务分配模型[J]. 指挥信息系统与技术, 2018, 9(1): 33-38. (WANG K, HUI X C, ZHANG Y. Interception task assignment model for coast guard boats in offshore area [J]. Command Information System and Technology, 2018, 9(1): 33-38.)
- [3] 郑浩, 王伟, 萨出拉. 基于误差检测的杂波点迹过滤技术[J]. 指挥信息系统与技术, 2019, 10(4): 72-76. (ZHENG H, WANG W, SA C L. Clutter plot filtering technology based on error detecting [J]. Command Information System and Technology, 2019, 10(4): 72-76.)
- [4] ZHI H, BAH M J, HAMMAD M. Progress in outlier detection techniques: a survey[J]. IEEE Access, 2019, 7: 107964-108000.
- [5] XU X, LIU W, YAO M. Recent progress of anomaly detection[J]. Complexity, 2019, 2019: No. 2686378.
- [6] ZHANG E, ZHANG Y. Average precision[M]// LIU L, ÖZSU M T. Encyclopedia of Database Systems. Boston: Springer, 2009: 192-193.
- [7] CAMPOS G O, ZIMEK A, SANDER J, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study[J]. Data Mining and Knowledge Discovery, 2016, 30(4): 891-927.
- [8] KLEMENT W, FLACH P, JAPKOWICZ N, et al. Smooth Receiver Operating Characteristics (smROC) curves[C]// Proceedings of the 2011 Joint European Conference on Machine Learning and Knowledge Discovery in Databases, LNCS 6912. Berlin: Springer, 2011: 193-208.
- [9] MARQUES H O, CAMPELLO R J G B, ZIMEK A, et al. On the internal evaluation of unsupervised outlier detection [C]// Proceedings of the 27th International Conference on Scientific and Statistical Database Management. New York: ACM, 2015: No. 7.
- [10] HAN J, KAMBER M, PEI J. Data Mining: Concepts and Techniques [M]. 3rd ed. Burlington, MA: Morgan Kaufmann Publishers, 2011: 351-376.
- [11] FAN H, ZAIANE O R, FOSS A, et al. Resolution-based outlier factor: detecting the top- n most outlying data points in engineering data [J]. Knowledge and Information Systems, 2009, 19(1): 31-51.
- [12] FALCÃO F, ZOPPI T, SILVA C B V, et al. Quantitative comparison of unsupervised anomaly detection algorithms for intrusion detection [C]// Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. New York: ACM, 2019: 318-327.
- [13] FENG F, LIU X, YONG B, et al. Anomaly detection in ad-hoc networks based on deep learning model: a plug and play device [J]. Ad Hoc Networks, 2019, 84: 82-89.
- [14] KOUFAKOU A, GEORGIPOULOS M. A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes [J]. Data Mining and Knowledge Discovery, 2010, 20(2): 259-289.
- [15] TAMA B A, RHEE K H. An in-depth experimental study of anomaly detection using gradient boosted machine [J]. Neural Computing and Applications, 2019, 31(4): 955-965.
- [16] BHATTACHARYA G, GHOSH K, CHOWDHURY A S. Outlier detection using neighborhood rank difference [J]. Pattern Recognition Letters, 2015, 60/61: 24-31.
- [17] HUANG J, ZHU Q, YANG J, et al. A non-parameter outlier detection algorithm based on Natural Neighbor [J]. Knowledge-Based Systems, 2016, 92: 71-77.
- [18] NARITA K, KITAGAWA H. Outlier detection for transaction databases using association rules [C]// Proceedings of the 9th International Conference on Web-Age Information Management. Piscataway: IEEE, 2008: 373-380.
- [19] HANLEY J A, MCNEIL B J. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve [J]. Radiology, 1982, 143(1): 29-36.
- [20] ZHANG J, LI Z, NAI K, et al. DELR: a double-level ensemble learning method for unsupervised anomaly detection [J]. Knowledge-Based Systems, 2019, 181: No. 104783.
- [21] GAUTAM C, BALAJI R, SUDHARSAN K, et al. Localized multiple kernel learning for anomaly detection: one-class classification [J]. Knowledge-Based Systems, 2019, 165: 241-252.
- [22] KIEU T, YANG B, GUO C, et al. Outlier detection for time series with recurrent autoencoder ensembles [C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 2725-2732.
- [23] CHOUHAN N, KHAN A, KHAN H U R. Network anomaly detection using channel boosted and residual learning based deep convolutional neural network [J]. Applied Soft Computing, 2019, 83: No. 105612.
- [24] CHENG L, WANG Y, MA X. A Neural Probabilistic outlier detection method for categorical data [J]. Neurocomputing, 2019, 365: 325-335.
- [25] DUTTA J K, BANERJEE B. Improved outlier detection using sparse coding-based methods [J]. Pattern Recognition Letters, 2019, 122: 99-105.
- [26] ASUNCION A U, NEWMAN D J. UCI machine learning repository [EB/OL]. [2019-07-11]. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- [27] HA J, SEOK S, LEE J S. Robust outlier detection using the instability factor [J]. Knowledge-Based Systems, 2014, 63: 15-23.

This work is partially supported by the Sichuan Science and Technology Project (2019YJ017, 2019YJ0176, 2019YFQ0005).

NING Jin, born in 1991, Ph. D. candidate. Her research interests include outlier detection, data mining.

CHEN Leitong, born in 1966, Ph. D., professor. His research interests include image processing, computer graphics, virtual reality.

LUO Zijuan, born in 1985, M. S., senior engineer. Her research interests include target recognition in remote sensing images, image change detection.

ZHOU Chuan, born in 1977, Ph. D., lecturer. His research interests include computer animation, machine vision, medical image analysis, data mining.

ZENG Huiru, born in 1994, Ph. D. candidate. Her research interests include deep learning, medical informatics.