

基于相对比重的扩展隔离森林算法

刘俊成, 董 东

(河北师范大学 计算机与网络空间安全学院, 河北 石家庄 050024)

摘 要: 由于局部离群点被密度相似的正常点掩盖, 不易被隔离, 使得扩展的隔离森林算法 (EIF) 对这类离群点的识别效果不理想。针对此问题, 提出基于相对比重的扩展隔离森林算法 (Relative Proportion-Extended Isolation Forest, RP-EIF)。该算法仍然基于随机斜度和随机截距划分超平面, 生成隔离森林, 但根据预测样本落入的叶子节点与其父节点的相对比重计算离群分数排名, 而不使用基于路径长度的排名。把全局排名替换为考虑邻域数据分布局部排名增强了算法对局部离群点的敏感性, 同时还减少了算法的时间复杂度。在离群点检测数据库 (ODDS) 的 5 个公开数据集上验证 RP-EIF 算法的有效性和算法效率, 并与 EIF 算法、GIF 算法、iForest 算法、COPOD 算法、LOF 算法进行了对比。实验表明: RP-EIF 算法在 5 个 ODDS 公开数据集上的准确率高出 EIF 算法 1 至 4 个百分点, 高出其他 5 个算法 2 至 38 个百分点。而且在 5 个数据集上的时间消耗 RP-EIF 算法要比 EIF 算法减少约 30%。

关键词: 大数据挖掘; 离群点检测; 局部离群点; 扩展的隔离森林算法; 相对比重

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2023)06-0016-06

doi:10.3969/j.issn.1673-629X.2023.06.003

Extended Isolation Forest Algorithm Based on Relative Proportion

LIU Jun-cheng, DONG Dong

(School of Computer and Cyber Security, Hebei Normal University, Shijiazhuang 050024, China)

Abstract: Since local outliers are covered by normal points with similar density, they are not easy to be isolated, so the extended isolation algorithm (EIF) is not effective in identifying such outliers. To solve this problem, an Relative Proportion-Extended Isolation Forest algorithm is proposed. The algorithm still divides the hyperplane based on random slopes and random intercepts to generate isolation forests, but ranks the outlier score based on the relative proportions of the leaf nodes that the predicted samples are fallen into with their parent nodes, rather than the path length-based ranking. Replacing the global ranking with local ranking considering the neighborhood data distribution enhances the algorithm's sensitivity to local outliers and reduces the algorithm's time complexity. The effectiveness and algorithm efficiency of the RP-EIF algorithm are tried on 5 public datasets in the Outlier Detection Databases (ODDS). Compared with EIF algorithm, GIF algorithm, iForest algorithm, COPOD algorithm, LOF algorithm, the accuracy of the RP-EIF algorithm on the 5 ODDS public datasets is 1 to 4 percentage points higher than the EIF algorithm, and 2 to 38 percentage points higher than the other 5 algorithms. Moreover, the time consumption of the RP-EIF algorithm on the 5 datasets is about 30% less than that of the EIF algorithm.

Key words: big data mining; outlier detection; local outliers; extended isolated forest algorithm; relative proportion

0 引 言

随着大数据时代的到来, 离群点检测成为国内外研究的热点^[1]。近年来, 离群点检测算法分为四种: 基于统计、基于距离、基于密度以及基于聚类^[2]。以箱线图 (BOXPLOT)^[3] 和基于连接函数的离群点检测 (COPula-based Outlier Detection, COPOD) 算法^[4] 为代表的基于统计的方法, 多数根据数据的分布去判断离群点, 模型一旦建立便可快速地完成检测, 该类方法适

用于定量实值数据集。但因假设分布, 在缺乏关于分布的先验知识的情况下可应用性不高。K 近邻 (K-Nearest Neighbor, KNN) 算法^[5], 根据预先指定的距离度量, 计算出样本点之间的距离后排序, 取最上层的样本作为离群点。其不依赖于假设分布来拟合数据, 与基于统计方法相比更适用于现实中分布多样的数据集。但是算法的复杂度较高。LOF (Local Outlier Factor) 算法^[6] 使用局部离群因子判别离群点, 对局部

收稿日期: 2022-08-17

修回日期: 2022-12-20

基金项目: 教育部教育考试院“十四五”规划支撑专项课题 (NEEA2021064)

作者简介: 刘俊成 (1997-), 男, 硕士研究生, 研究方向为数据挖掘及其应用; 通信作者: 董 东 (1971-), 男, 副教授, 硕士, 研究方向为大数据分析等。

离群点十分有效。基于相对密度的离群值检测 (Relative Density-based Outlier Source, RDOS) 算法^[7] 引入相对密度来测量对象的局部离群值。离群点是聚类的副产品, 对聚类算法 DBSCAN^[8]、CHAMELEON^[9]、CLARANS^[10] 加以修改都可用于离群点检测。这些方法大多通过考虑样本点与簇之间的关系检测离群点。该类方法为无监督方法, 从集群中学习后, 可以插入额外的新点, 这使其能适应增量模式, 因此更适用于数据流中的离群点检测。但该类方法大多需要事先指定簇的数量 K , 且检测结果对于 K 值较敏感^[11]。

隔离森林 (Isolation Forest, iForest) 算法^[12] 不依靠距离或密度作为相似度量, 而是通过随机采样的方式构建隔离树, 并利用离群点与正常点在隔离树中深度不同这一特性去判别离群点。该算法的时间与空间复杂度很低, 但对局部离群点的敏感性不高。针对该问题, 结合 LOF 算法与 iForest 算法形成了一种两阶段离群点检测算法^[13]。首先, 利用 LOF 算法进行离群点检测, 后使用 iForest 算法在其结果集中进行筛选。这样提高了局部离群点检测的准确率, 但时间开销太大^[14]。基于隔离森林的快速离群点检测 (Fast Isolation Forest, FIF) 算法^[15], 根据根节点的数据分布筛选样本子集避免无关隔离树的产生, 以及使用黄金切割法进行节点划分, 在保证准确率几乎不变的情况下极大提高了 iForest 算法的效率。

扩展的隔离森林 (Extended Isolation Forest, EIF) 算法^[16] 使用随机法向量与随机截距确定分割超平面, 解决了 iForest 算法的轴平行问题。但 EIF 算法时间开销较大, 且对局部离群点不敏感, 易产生局部离群点被密度相似的簇掩盖等问题。基于随机子空间的隔离森林 (Extended Isolation Forest based on Random Subspace, RS-EIF) 算法^[17], 结合子空间的思想来创建隔离森林, 相较于 EIF 算法减少约 60% 的时间开销。广义隔离森林 (Generalized Isolation Forest, GIF) 算法^[18], 首先将采样数据全部投射到单位法向量上, 然后在投影的最大最小值间选择切割点, 避免生成无效的空节点, 提高了算法的效率。以上两种算法虽然减少了 EIF 算法的时间消耗, 但仍未解决局部离群点的掩盖问题。

该文提出基于相对比重的扩展隔离森林 (Relative Proportion-Extended Isolation Forest, RP-EIF) 算法, 不再根据样本点在隔离树中的路径长度去判断离群点, 而是根据样本点所在叶节点上的数据量与其直接父节点的数据量比重去判断。这种基于相对比重的局部排名方式, 优化了 EIF 算法在局部离群点检测上的不足, 同时节省了算法的计算开销, 增加了其工程应用价值。

1 局部离群点

EIF 算法根据样本点在隔离森林中的平均路径长度去判别离群点。这对全局离群点的检测十分有效, 原因是基于路径长度的全局排名方式无法考虑到拥有特殊分布的局部离群点, 导致局部离群点被掩盖。在图 1 中, a_1 、 a_2 为局部离群点, 簇 C_2 的正常样本点密度与局部离群点 a_1 、 a_2 的密度相似。这就导致 C_2 中的一些边缘样本点可能会与 a_1 、 a_2 在隔离树中拥有相同甚至是更短的平均路径长度, 导致 EIF 算法产生局部离群点掩盖问题, 降低其对局部离群点的敏感性。

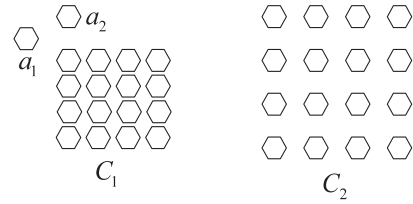


图1 局部离群点

2 PR-EIF 算法

2.1 相对比重

相对比重 (Relative Proportion) 为隔离树中叶节点的样本量与父节点的样本量之比。样本点 x 的离群分数 $p_i(x)$ 定义为:

$$p_i(x) = \frac{\text{size}(T_i(x))}{\text{size}(T'_i(x))} \quad (1)$$

其中, T_i 为隔离森林中第 i 棵隔离树, $T_i(x)$ 为待预测样本点 x 落入 T_i 中的叶节点, $T'_i(x)$ 为 $T_i(x)$ 的父节点, $\text{size}(T'_i(x))$ 表示父节点上的样本量, $\text{size}(T_i(x))$ 表示叶节点上的样本量。 $p_i(x)$ 的值域为 $(0, 1]$ 。

样本 x 在森林上的最终离群得分 $P(X)$ 是该样本在每棵隔离树上离群分数 $p_i(x)$ 的均值:

$$P(X) = \frac{1}{n} \sum_{i=1}^n p_i(x) \quad (2)$$

当对数据集中的每个样本 x 完成离群得分的计算后, 根据分数大小进行升序排序, 最后选取前若干个样本点作为离群点。

2.2 RP-EIF 算法的实现过程

EIF 算法在构建隔离树时叶节点允许的最小样本量为 1。而 RP-EIF 算法使用考虑邻域数据分布的局部排名方式, 故设置叶子节点允许的最小样本量为参数 Minsize (默认值为 5)。当节点上的样本量小于等于 5 时, 该节点便停止生长。由于超平面在 N 维空间中至少与一个维度至多与 N 个维度相交, 故设置参数 Exlevel, 根据需要调整相交的维数。

为了方便计算子节点与父节点的样本量比重, 在隔离树节点中增加 parent 属性引用父节点。隔离树的构建如算法 1 所示。

算法 1: iTree(X , Height, e , MinSize=5, Exlevel)

输入: X —当前的数据集, Height—树的高度限制(默认为 $\log_2(\text{采样量})$), e —当前树的高度, MinSize—节点允许的最小样本数, Exlevel—扩展水平

输出: 二叉隔离树

1. IF $e \geq \text{Height}$ or X 的数据量 $\leq \text{MinSize}$ THEN
2. 返回节点 Node(X , $\vec{n} = \text{null}$, $\vec{p} = \text{null}$, parent,
3. left=null, right=null, node_type=“叶子”)
4. ELSE
5. 根据扩展水平 Exlevel, 选择随机法向量 \vec{n}
6. 在样本范围内选择随机截距 \vec{p}
7. $X_l \leftarrow$ 满足 $(x - \vec{p}) \cdot \vec{n} > 0$ 的点
8. $X_r \leftarrow$ 满足 $(x - \vec{p}) \cdot \vec{n} \leq 0$ 的点
9. 返回节点 Node(X , \vec{n} , \vec{p} , parent,
10. left \leftarrow iTree(X_l , Height, $e+1$, MinSize, Exlevel),
11. right \leftarrow iTree(X_r , Height, $e+1$, MinSize, Exlevel),
12. node_type=“内部节点”)
13. END

隔离森林由 n 棵隔离树组成, 具体步骤如算法 2 所示。

算法 2: rpForest(X , n , φ , Exlevel, MinSize=5)

输入: X —数据集, n —隔离树的数量, φ —随机采样的样本个数, Exlevel—扩展水平, MinSize—节点允许的最小样本数

输出: 隔离森林 rpForest

1. 隔离森林集合 Forest 初始化为空
2. 设置每棵树的高度限制 Height 为 $\log_2(\varphi)$ 的上取整
3. FOR $i = 1$ TO n DO
4. $X_{\text{sub}} \leftarrow$ 在 X 中随机抽取 φ 个样本点
5. Forest \leftarrow iTree(X_{sub} , Height, e , MinSize=5, Exlevel) \cup
6. Forest
7. END

离群分数的计算如算法 3 所示。

算法 3: nodeSize(x , T)

输入: x —一个样本, T —一棵隔离树

输出: 样本 x 落入叶节点的直接父节点与该节点的样本量比重

1. IF T 为叶节点 THEN
2. 返回该节点上数据量与其直接父节点的数据量的比重
3. ELSE
4. $\vec{n} \leftarrow$ 用于该节点划分的法向量
5. $\vec{p} \leftarrow$ 用于该节点划分的截距
6. IF $(x - \vec{p}) \cdot \vec{n} \leq 0$ THEN
7. 在节点的右子树递归 nodeSize(x , T . right)
8. ELSE
9. 在节点的左子树递归 nodeSize(x , T . left)
10. END

11. END

通过算法 3 得到样本点 x 落在叶节点的数据量及其父节点的数据量比重, 再根据公式(2)计算出最终离群分数, 如算法 4 所示。

算法 4: computeScore(test, iTree)

输入: test—待预测的样本集, iTree—一棵隔离树

输出: 离群分数集合 S

1. 初始化一个离群分数的空集合 S
2. FOR $i = 0$ TO 集合 test 的数据量 DO
3. FOR $j = 0$ TO 隔离树的数量 n DO
4. Score = Score +
($\frac{\text{样本 test}[i] \text{ 在树 iTree}[j] \text{ 上叶节点的数据量}}{\text{该叶节点父节点的数据量}}$)
5. END
6. $S[i] = (\text{Score} / n)$
7. 返回离群分数集合 S
8. END

3 实验结果与分析

3.1 实验环境

实验均在 CPU 为 AMD Ryzen 5 3500U, 2.10 GHz, 运行内存为 16 GB, 操作系统为 Windows 10 的 PC 机上进行。

采用 PR-EIF 算法、EIF 算法、GIF 算法、iForest 算法 4 种基于树型结构的建议参数^[12]: 创建 100 棵隔离树、随机采样数为 256。

COPOD 算法无需设置任何参数^[4]。最近邻的数量达到 10 时 LOF 的标准差开始稳定, 同时考虑到时间消耗, 因此将 LOF 算法最近邻数设置为 $10^{[6]}$ 。

3.2 实验数据集

为了更好地验证 RP-EIF 算法在不同数据量不同维度数据集上的性能, 本次实验所选的 5 个数据集包括从低维度到高维度、低样本量到高样本量。数据集的具体属性如表 1 所示。

表 1 实验所用数据集

数据集	样本数	维度	离群比例/%
Breastcancer	367	30	2.73
Forest Cover	286 048	10	0.90
Ionosphere	351	33	36.0
Mammography	11 183	6	2.32
Satellite	6 435	36	32.0

Breastcancer 为诊断乳腺癌数据集, 选取其中 10 个恶性诊断作为离群点, 所有的良性诊断作为正常点。Forest Cover 为描述森林覆盖的多分类数据集, 在做离群点检测时, 第 2 类被认为是正常类, 第 4 类为离群类, 离群比例为 0.9%。Ionosphere 为电离层数据集, 有一个属性的值全为零, 该属性被丢弃, 其中坏类被视

为离群类,好类被视为正常类。Mammography 为乳腺 X 光数据集共有 11 183 个实例,其中 260 个钙化实例作为离群点。Satellite 为 Landsat 卫星数据集,2、3、4 类作为离群类,其余所有类作为正常类。

3.3 评价指标

一般而言,离群点检测可以看作样本类别不平衡的二分类问题,因此实验使用受试者工作特征曲线(Receiver Operating Characteristic, ROC)以及 ROC 曲线下面积(the Area Under the ROC, AUC)来评价算法的性能。

二分类问题结果分为 4 类:真正类(True Positive, TP)、真负类(True Negative, TN)、假正类(False Positive, FP)、假负类(False Negative, FN)。ROC 曲线的 X 轴为假正比例(False Positive Rate, FPR),即预测的正类中实际负实例占所有负实例的比例。Y 轴为真正比例(True Positive Rate, TPR),即预测的正类中实际正实例占所有正实例的比例。假正比例与真正比例的计算见公式(3)、(4)。AUC 代表 ROC 曲线下面积的大小,值域为[0,1],越接近 1 算法性能越优秀。

$$FPR = \frac{FP}{FP + TN} \quad (3)$$

$$TPR = \frac{TP}{TP + FN} \quad (4)$$

3.4 局部离群点敏感性的验证

对 RP-EIF 算法进行局部离群点敏感性验证。为便于可视化,选取鸢尾花数据集的 Sepal.Width、Petal.Width 两维度作为二维数据集进行实验。首先对该数据集的局部离群点进行人工标注,标注局部离群点 14 个,共 150 个样本点。标注后的数据集如图 2 所示。可以看到数据集由两个簇组成,人工标注的局部离群点分布在两个簇的周围。

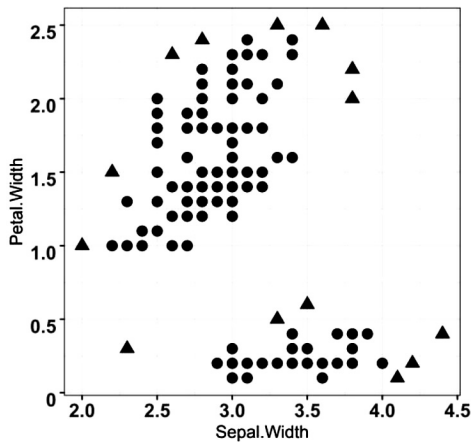
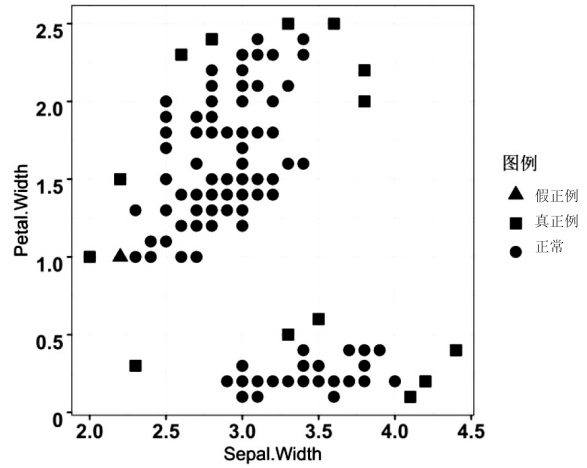


图2 实验所用二维数据集

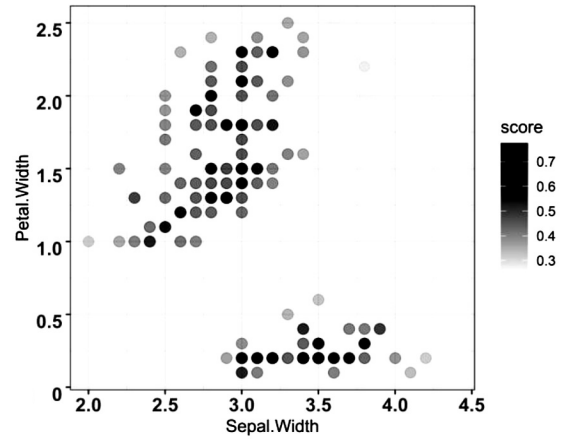
使用 RP-EIF 算法在该数据集上进行离群点检测,根据离群分数升序排序选取前 15 个样本点作为离群点。结果如图 3(a)所示。其中真正例 14 个,假正

例 1 个。经计算可知,算法的局部离群点识别率为 100% (真正例数量与离群点总数之比),识别准确率为 93.3% (真正例数量与检测出的离群点数之比)。

同时,为了更好地展示算法离群分数的分布情况,这里绘制离群分数分布热图,如图 3(b)所示。越靠近簇的中心密集区域离群分数越高(注意该算法分数越低代表样本点为离群点的可能性越大),而越靠近簇边缘的稀疏区域离群分数越低。这说明 RE-EIF 算法在进行离群分数计算时,考虑到了数据的分布,对局部离群点具有较高的敏感性。



(a) RP-EIF 算法检测结果



(b) RP-EIF 算法离群分数分布

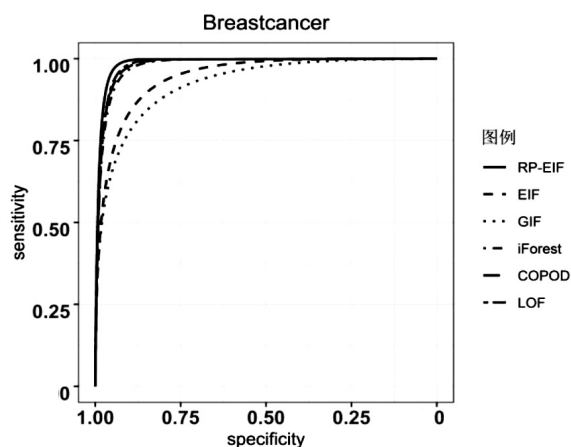
图3 RP-EIF 算法的检测结果与离群点的分布

3.5 各算法 AUC 及运行时间对比

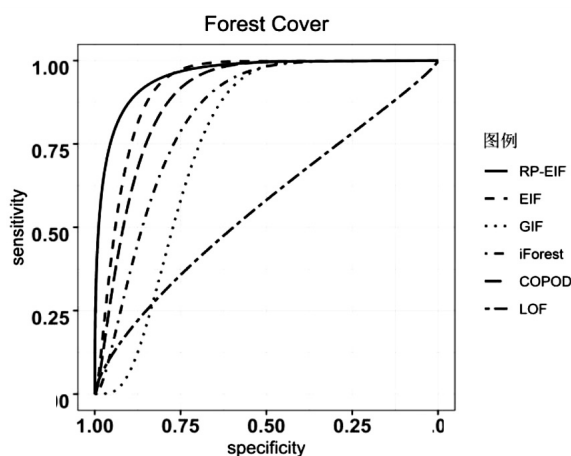
在 5 个 ODDS 数据集上验证算法的准确率与算法效率。并与 5 种离群点检测算法(EIF、GIF、iForest、COPOD、LOF)进行比较分析。

6 种算法的 ROC 曲线如图 4 所示。RP-EIF 算法在 5 个 ODDS 数据集上的 ROC 曲线均优于其他 5 种算法。其中在 Forest Cover、Ionosphere、Mammography 数据集上的 ROC 曲线明显更靠近左上方。在图 4(a)中,可以发现 6 种算法在 Breastcancer 数据集上均可很好地识别离群点。

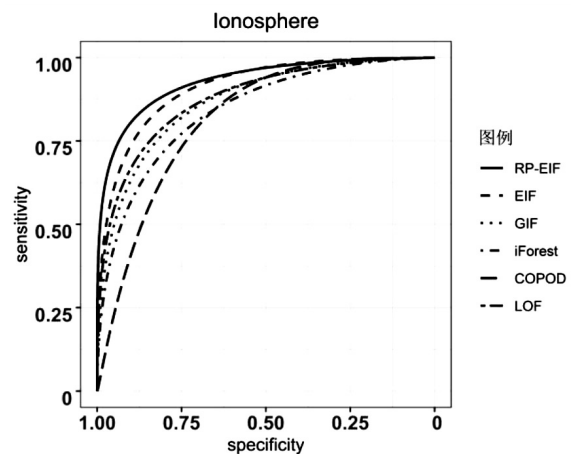
由表 2 可知,RP-EIF 算法相较于 EIF 算法,准确度从 0.960 提高到 0.986,提高约 3 百分点,这是由于 RP-EIF 算法在离群分数计算阶段使用相对比重的局部排名方式,考虑了数据点与其邻域点的分布关系,使算法对局部离群点更加敏感。在图 4(e)中,5 种算法在 Satellite 数据集上的离群点识别效果略差于其他 4 个数据集。但是 RP-EIF 算法在该数据集上准确率高出 EIF 算法、GIF 算法、iForest 算法、COPOD 算法、LOF 算法 2 ~ 13 百分点。



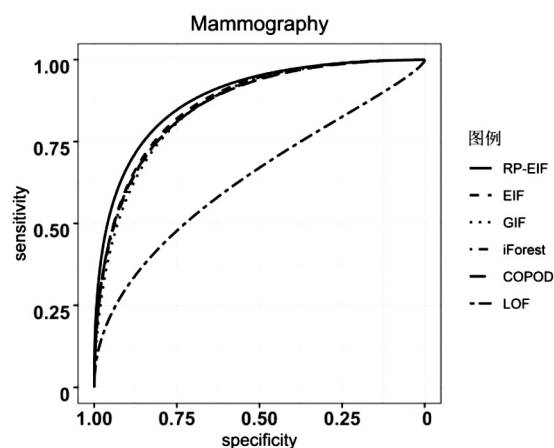
(a) Breastcancer 数据集 ROC 曲线



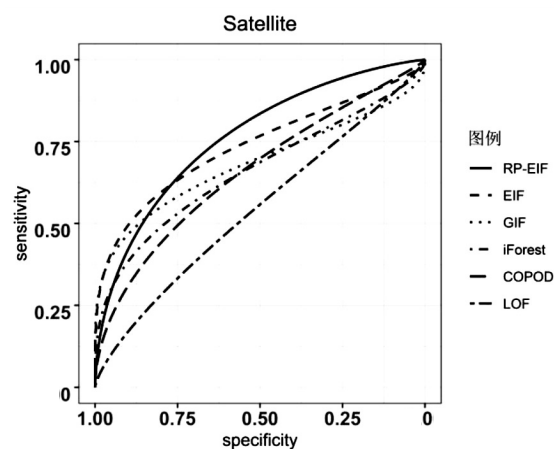
(b) Forest Cover 数据集 ROC 曲线



(c) Ionosphere 数据集 ROC 曲线



(d) Mammography 数据集 ROC 曲线



(e) Satellite 数据集 ROC 曲线

图 4 不同算法在不同数据集上的 ROC 曲线

各算法的运行时间如表 3 所示。运行时间取 5 次运行结果的平均值作为最后结果。由表 3 可知,RP-EIF 算法的运行时间比 EIF 算法快约 30%。原因是 RP-EIF 算法在构建隔离树的过程中,当节点上的样本数小于或等于 5 时,该节点就停止生长。使 RP-EIF 算法的森林模型比 EIF 算法收敛更快。同时,在离群分数计算阶段,RP-EIF 算法无需根据条件调整返回的路径长度,这也使算法的效率提高。GIF 算法的运行时间略少于 RP-EIF 算法,这是由于 GIF 算法避免了隔离树的空节点问题,使算法的效率提高,但是其准确率却低于 RP-EIF 算法。而 iForest 算法的时间消耗少于 RP-EIF 算法,是由于 RP-EIF 算法需要进行大量的高维向量运算。但运行时间的略长,带来的却是准确率的显著提高,在 5 个 ODDS 数据集上 RP-EIF 算法的准确率高出 iForest 算法 5 ~ 14 百分点。基于统计的 COPOD 算法不需要计算相似性度量,模型一旦建立,便可快速完成检测,因此其时间消耗在 6 个算法中最少。但其在高维或数据量大的数据集上检测精度不佳。基于密度的 LOF 算法,在数据量较少时运行时间比 RP-EIF 算法略快。这是由于基于树型结构的算法,无论数据的多少,都需要将每个待预测样本点 x 遍

历森林中的每棵隔离树,而 LOF 算法在数据量较少时可以很快得到局部离群因子,所以在数据量较少时 LOF 算法时间消耗较少。但 LOF 算法在数据集 Mammography、Satellite 上的时间消耗是 RP-EIF 算法的 6~15 倍。在大型数据集 Forest Cover 上由于 LOF 算法的复杂度过高无法完成有效检测。

表2 各算法 AUC

数据集	RP-EIF	EIF	GIF	iForest	COPOD	LOF
Breastcancer	0.986	0.960	0.935	0.981	0.982	0.985
Forest Cover	0.961	0.920	0.835	0.821	0.891	NA
Ionosphere	0.931	0.913	0.871	0.836	0.818	0.910
Mammography	0.884	0.865	0.860	0.863	0.874	0.646
Satellite	0.780	0.760	0.704	0.709	0.658	0.547

表3 各算法运行时间 s

数据集	RP-EIF	EIF	GIF	iForest	COPOD	LOF
Breastcancer	1.2	1.7	1.1	0.4	0.02	0.3
Forest Cover	776.7	1 180.2	721.3	695.4	1.9	NA
Ionosphere	1.4	2.0	1.2	0.5	0.03	0.4
Mammography	22.0	32.0	18.3	5.4	0.04	347.7
Satellite	18.0	26.5	14.3	12.2	0.3	110.3

4 结束语

隔离森林算法在大数据上识别离群点的表现出色。基于相对比重的概念,提出了基于相对比重的扩展隔离森林算法。优化了算法在离群分数计算阶段的排名方式,增强了算法对于局部离群点的敏感性,提高了算法的准确率与效率。在 5 个 ODDS 数据集使用 RP-EIF 算法进行离群点识别,并与 5 种离群点检测算法(EIF 算法、GIF 算法、iForest 算法、COPOD 算法、LOF 算法)进行了比较,验证了 RP-EIF 算法在准确率与算法效率两方面的有效性。在之后的工作中,计划将算法应用于实际的大数据上,并进一步探索在深度森林上的表现。

参考文献:

- [1] WANG H, BAH M J, HAMMAD M. Progress in outlier detection techniques: a survey [J]. IEEE Access, 2019, 7: 107964-108000.
- [2] WANG X C, WANG X L, WILKES M. Developments in unsupervised outlier detection research [M]. Xi'an: Xi'an Jiaotong University Press, 2021: 13-36.
- [3] VANDERVIJVEREN E, HUBERT M. An adjusted boxplot for skewed distributions [J]. Computational Statistics & Data Analysis, 2004, 52(12): 5186-5201.
- [4] LI Z, ZHAO Y, BOTTA N, et al. COPOD: copula-based outlier detection [C]//Proceedings of the 2020 IEEE international conference on data mining. Sorrento: IEEE, 2020: 1118-1123.
- [5] COVER T, HART P. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 2003, 13(1): 21-27.
- [6] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers [C]//ACM SIGMOD international conference on management of data. Dallas: ACM, 2000: 93-104.
- [7] TANG B, HE H B. A local density-based approach for outlier detection [J]. Neurocomputing, 2017, 241(7): 171-180.
- [8] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise [C]//Proc of the 2nd international conference on knowledge discovery & data mining. Portland Oregon: AAAI, 1996: 226-231.
- [9] KARYPIS G, HAN E H, KUMAR V. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling [J]. Computer, 1999, 32(8): 68-75.
- [10] NG R T, HAN J W. Efficient and effective clustering methods for spatial data mining [C]//Proc of the 20th international conference on very large data bases. San Francisco: Morgan Kaufmann Publishers Inc, 1994: 144-155.
- [11] 周玉, 朱文豪, 房倩, 等. 基于聚类的离群点检测方法研究综述 [J]. 计算机工程与应用, 2021, 57(12): 37-45.
- [12] FEI T L, KAI M T, ZHOU Z H. Isolation forest [C]//IEEE international conference on data mining. Washington: IEEE, 2008.
- [13] XIAO Y, LU A T, HAN J. Filtering and refinement: a two-stage approach for efficient and effective anomaly detection [C]//International conference on data mining. Florida: IEEE, 2009.
- [14] JIN W, TUNG A, HAN J. Mining top-n local outliers in large databases [C]//Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2001: 293-298.
- [15] 冯嘉琛, 蔡江辉, 杨海峰. 一种改进隔离森林的快速离群点检测算法 [J]. 小型微型计算机系统, 2019, 40(11): 2418-2423.
- [16] HARIRI S, KIND M C, BRUNNER R J. Extended isolation forest with randomly oriented hyperplanes [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(4): 1479-1489.
- [17] 谢雨, 蒋瑜, 龙超奇. 基于随机子空间的扩展隔离森林算法 [J]. 计算机应用, 2021, 41(6): 1679-1685.
- [18] LESOUPLE J, BAUDOUIN C, SPIGAI M, et al. Generalized isolation forest for anomaly detection [J]. Pattern Recognition Letters, 2021, 149: 109-119.