

基于相关子空间的扩展隔离森林离群检测算法

刘佳, 朱鹏云, 荀亚玲

(太原科技大学 计算机科学与技术学院, 山西 太原 030024)

摘要:扩展隔离森林离群检测作为一种集成离群检测方法,可选取随机斜率的超平面,具有将离群数据与正常数据对象快速分离,时间复杂度较低等优点,但隔离树超平面选取在数据密集区域或含有关键维度数据区域时,严重影响了其离群检测的效果。采用相关子空间思想和方法,提出了一种扩展隔离森林离群检测算法。该算法利用高斯混合模型确定数据对象的相关子空间,从而保证了能够在稀疏数据区域中选取隔离树的切割超平面;隔离树枝分割优先在稀疏数据区域中,选择隔离树超平面的随机截距点,可快速地将离群数据对象从稀疏数据区域中隔离出来,从而避免了在超平面的随机斜率选取时无关属性维度的干扰;将每个数据对象在各隔离树上的平均路径长度归一化后作为离群得分,并选取离群得分最大的若干个数据对象作为离群数据;在UCI数据集上通过实验验证了该算法的有效性,以及抽样数、隔离树个数和近邻数参数对其离群检测效果的影响。

关键词:离群检测;扩展隔离森林;相关子空间;高斯混合模型;稀疏数据区域

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2022)10-0026-08

doi:10.3969/j.issn.1673-629X.2022.10.005

An Extended Isolation Forest Outlier Detection Algorithm Based on Relevant Subspace

LIU Jia, ZHU Peng-yun, XUN Ya-ling

(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: The extended isolation forest outlier detection algorithm, as an ensemble outlier detection method, can select the hyperplane of random slope and has the advantages in separating outliers from normal data and time complexity. But the hyperplane selection of the extended isolation tree in the dense area of the data set or the area with irrelevant dimensions is of great significance to the outlier detection effect. An extended isolation forest outlier detection algorithm is proposed by using the idea and method of relevant subspace. It utilizes Gaussian mixture model to definite the relevant subspace of data objects, which guarantees to select the branching hyperplane of the isolation tree in the sparse data area. During constructing each extended isolation tree, random intercept points of hyperplanes are preferentially selected in the data-sparse region so as to isolate outliers from the data-sparse region quickly. And it can avoid the interference of irrelevant attribute dimensions when selecting the hyperplane's random slope. Then the outlier score of each data object is obtained by normalizing the average path length in each isolation tree, and the selection of several data objects with the largest outlier score is defined as the outliers. Experimental results validate the effectiveness of the algorithm and the effects of parameters, including sub-sample size, the number of isolation tree and nearest neighbors on outlier detection in UCI data sets.

Key words: outlier detection; extended isolation forest; relevant subspace; Gaussian mixture model; sparse data area

0 引言

离群检测是数据挖掘任务的重要内容之一,试图捕获显著偏离多数模式的异常情况^[1],挖掘出潜在的、有意义的知识。离群检测挖掘出的离群数据对象隐藏着非常重要的信息和知识,其背后可能蕴含着更大的研究价值。离群检测已经被广泛应用于入侵检

测^[2-4]、欺诈检测^[5-6]、医疗异常诊断^[7-8]、工业控制系统的异常检测^[9]、无线传感器网络的异常检测^[10]、城市交通流异常检测^[11]、天体光谱数据挖掘^[12]等领域。现有的离群检测算法大致可以分为以下几类:基于统计的方法、基于距离的方法、基于密度的方法、基于聚类的方法和基于集成的方法等。集成离群检测作为一

收稿日期:2021-10-21

修回日期:2022-02-24

基金项目:国家自然科学基金项目(61602335);山西省自然科学基金(201901D211302)

作者简介:刘佳(1997-),女,硕士研究生,研究方向为数据挖掘与并行计算;通信作者:荀亚玲(1980-),女,博士,副教授,研究方向为数据挖掘与并行计算。

种离群检测方法,将多个离群检测算法或模型相结合,形成一个集成框架,提高离群检测性能。

隔离森林(Isolation Forest)^[13]作为一种集成离群检测方法,利用离群数据少且与众不同的特点将离群数据隔离在离树根较近的地方,而将正常数据隔离在离树根较远的一端。隔离森林可以将离群数据与正常数据快速分离,具有较低的线性时间复杂度。在组合多棵树构成森林后,离群检测效果优于多数传统的离群检测算法,但隔离森林在构建隔离树的过程中,分枝仅随机选择一个维度属性进行水平或垂直的切割,且属性可能为无关属性,降低了离群检测效果。扩展隔离森林算法具备了隔离森林离群检测的优点,允许数据切片使用带有随机斜率的超平面,解决了隔离森林中树分枝轴平行的切割使异常分数图中存在偏差等问题^[14],但隔离树超平面选取在数据集的密集区域或含有无关维度的区域,影响离群检测的效果。该文采用相关子空间,提出了一种基于相关子空间的扩展隔离森林离群检测算法,解决了扩展隔离森林离群检测确定超平面随机性强的问题,其主要贡献如下:

- 提出了一种基于相关子空间的分支切割截距点选取方法;
- 给出了一种扩展隔离森林分割平面选择策略;
- 提出了一种基于相关子空间的扩展隔离森林离群检测算法。

1 相关工作

离群检测算法大致分为以下几类:基于统计的方法、基于距离的方法、基于密度的方法、基于聚类的方法和基于集成的方法等。隔离森林是一种独特的集成离群检测方法,它利用一种隔离机制来检测异常,通过递归轴平行细分将每个实例与其余实例隔离开来。隔离森林有较低的线性时间复杂度,避免了基于距离和密度方法进行检测时的计算量大的问题。隔离森林是集成的方法,组合多棵树构成森林后,离群检测效果优于多数传统的离群检测算法。隔离森林离群检测可以用在含有海量数据的数据集上,由于每棵树都是互相独立生成的,因此可部署在大规模分布式系统上来加速运算。

隔离森林作为一类集成离群检测方法,国内外学者对其进行了大量研究,其典型研究工作包括:文献[14]扩展隔离森林离群检测算法可选取随机斜率的超平面,解决了隔离森林中树分枝轴平行的切割使异常分数图中存在偏差等问题。文献[15]提出一种利用隔离概念进行离群检测的方法(isolation using Nearest Neighbor Ensemble, iNNE),采用最近邻算法来代替基于树的方法来隔离数据对象,使用最近邻隔离

超球体来隔离目标空间中的数据,是一种基于隔离思想的高精度集成学习算法,对局部离群数据较敏感,但在数量巨大的数据集中时间开销太大。文献[16]提出了一种基于k-means的隔离森林算法,优化隔离树的结构通过将每个节点分割成 k 个子节点,且在树的每个节点上采用k-means进行划分。该算法可以处理不同应用领域的的数据,但没有考虑随机选择的属性可能为无关属性的问题。文献[17]提到基于树隔离机制的离群检测方法使用的这种快速隔离机制仅局限于距离测量,不能推广到其他常用测量,因此提出了一个LSHiForest通用框架,使用位置敏感哈希(LSH)森林。该框架具有通用性,可以用不同范围的LSH函数实例化,并且快速隔离机制可以扩展到定义LSH函数的任何距离度量、数据类型和数据空间。该文献表明现有的基于树隔离的检测方法是此框架的特例,具有相应的距离度量,且该框架具有较高的时间效率和异常检测效果。文献[18]提出了一种基于模糊孤立森林算法的离群数据检测方法,通过挑选一些有价值的属性对其分别建树组成孤立森林,从多维度出发,对每一维属性的检测结果进行隶属度判断,最后与模糊矩阵进行模糊运算得到最终评价结果。但算法需要从单因素的角度对各等级的模糊子集做隶属度判断,由专家根据评判等级对评价对象进行打分,组成模糊关系矩阵。

隔离森林离群检测算法可以将离群数据与正常数据快速分离,有较低的线性时间复杂度。但隔离森林每次分割随机选取一个维度,高维空间中可能有大量的维度信息没有被使用进行分割,且可能存在无关维度影响树的构建,基于路径长度的全局排名测度对局部异常不敏感,轴平行的细分掩盖了轴平行之间存在的异常。针对隔离森林算法的不足,一些国内外学者进行了改进,但算法选择分割点的方式随机性较强,均没有考虑无关属性维对离群检测的影响。该文提出了一种基于相关子空间的扩展隔离森林离群检测算法,利用多维度随机斜率的超平面,避免了轴平行分割带来的不足;优先在数据分布稀疏的数据集中进行分割,使离群数据快速地从稀疏数据区域中隔离出来;在数据的相关子空间维度上确定超平面,避免了无关维度的干扰,提高了离群检测的准确率和效率。

2 隔离森林与相关子空间

2.1 子空间与隔离森林离群检测

隔离森林离群检测利用了离群数据的两个特征:离群数据占数据集总体规模的比重较小;离群数据相比正常数据的属性值存在明显的差异。在隔离森林中,递归地随机分割数据集,直到所有的样本点都是孤

立的。在该随机分割策略下,离群数据可以快速被隔离,而正常数据需要许多分支切割来隔离。参考文献[13],相关概念定义如下:

定义1(隔离树, Isolation Tree):令 T 是一棵二叉隔离树的节点, T 要么是没有子节点的叶子节点(外部节点),要么是只有两个孩子节点(T_l, T_r)的内部节点。每一次分割都包含属性 q 和分割值 p ,将数据点在属性 q 的值 $q_i < p$ 的数据分到 T_l ,否则将 $q_i \geq p$ 的数据分到 T_r 。

给定 n 个样本数据的数据集 $X = \{x_1, x_2, \dots, x_n\}$,数据集的特征维度为 d 。为其构造一棵隔离树,每次分割需要随机选择一个特征 q 及其分割值 p ,递归地分割数据集 X 构造左子树和右子树,直到满足以下任意一个条件:(1)树达到了限制的高度;(2)节点上只有一个样本;(3)节点上的样本所有特征都相同。

定义2(路径长度):在一棵隔离森林树 Isolation Tree 中,从根节点到叶子节点所经历边的数目称为数据点 x 的路径长度,记为 $h(x)$ 。

由于 Isolation Tree 与二叉查找树的结构等价,外部节点的平均路径长度 $h(x)$ 的估计等价于二叉查找树中查找不成功的平均查找长度。对于给定的数据集 X ,二叉查找树中查找不成功的平均查找长度为:

$$c(n) = 2H(n-1) - (2(n-1)/n) \quad (1)$$

其中, $H(i)$ 为调和数, $H(i) = \ln(i) + 0.577\ 215\ 664\ 9$ (欧拉常数); n 为叶子节点数; $c(n)$ 为给定 n 时 $h(x)$ 的平均值,用来标准化数据点 x 的路径长度 $h(x)$ 。数据点 x 的离群得分计算方法如下:

$$s(x, n) = 2 \frac{-E(h(x))}{c(n)} \quad (2)$$

其中, $E(h(x))$ 为 Isolation Tree 集合中 $h(x)$ 的平均值。当 $E(h(x)) \rightarrow c(n)$ 时, $s \rightarrow 0.5$,即当所有数据均返回的 $s \approx 0.5$ 时,全部样本中没有明显的异常值;当 $E(h(x)) \rightarrow 0$ 时, $s \rightarrow 1$,即当数据返回的 s 非常接近于 1 时,它们是异常值;当 $E(h(x)) \rightarrow n-1$ 时, $s \rightarrow 0$,即当数据返回的 s 远小于 0.5 时,它们有很大的可能为正常值。

隔离森林离群检测每次随机选择一个属性维再随机选择其属性维的一个值作为分割点进行分割,一方面,隔离森林离群检测受到“维灾”的影响,在高维空间中一些无关维度严重地影响离群检测效果;另一方面,选择的分割点若在稠密数据区域,很难将离群数据隔离出来。

定义3(随机斜率 \vec{n}):对于一个 d 维数据集 DS ,选择一个随机斜率进行分支切割,就像在 d 维球单元上均匀地选择一个法向量一样^[14]。设超平面的切割斜率 $\vec{n} = \{n_1, n_2, \dots, n_d\}$ 。 $n_j (j = 1, 2, \dots, d)$ 是从标准

正态分布 $N(0, 1)$ ^[19]中绘制一个随机数。

2.2 高斯混合模型与相关子空间

在隔离森林离群检测中,每次分割在由有意义属性维构成的子空间中,随机选择一个属性,可避免无关属性维对离群检测的干扰。在稀疏数据区域中,随机选择一个属性取值实现分割,可避免稠密数据区域的属性取值分割对离群检测的影响。

在离群检测中,某些相关属性维提供了有价值信息,而有些属性维有价值信息很少甚至没有^[20]。若利用所有属性实现离群检测,就会受到离群信息很少的无关属性干扰。参考文献[20-21],相关子空间是非均匀分布的属性维组成的集合,可有效地体现出“离群数据”的价值信息,其相关定义如下:

定义4:设 DS 是一个 d 维数据集,属性集 $FS = \{A_1, A_2, \dots, A_d\}$,数据对象 x 的最近邻为 $N(x, FS)$ (即局部数据集 LDS),如果 $N(x, FS)$ 在 A_i 属性维上的取值是非均匀分布的,则称 A_i 可以提供有价值的信息,属于相关子空间中的属性维;反之,如果 $N(x, FS)$ 在 A_i 属性维上的取值是均匀分布的,则称 A_i 不能提供有价值的信息,属于不相关子空间中的属性维。

为了通过局部属性维数的稀疏性来度量属性是否服从均匀分布,在文献[21]中,引入了稀疏度的概念。第 i 条数据第 j 维度的数据稀疏度 y_{ij} 为:

$$y_{ij} = \frac{\sum_{r \in p^j(x_i)} (r - c_{ij})^2}{k + 1} \quad (3)$$

其中,数据对象 x_i 的局部数据集 LDS 是由数据对象本身及其 k 近邻构成的, $p^j(x_i)$ 是 x_i 的局部数据集 LDS 在属性维 j 上的集合。 c_{ij} 代表的是 $p^j(x_i)$ 的平均值,

$$c_{ij} = \frac{\sum_{r \in p^j(x_i)} r}{k + 1}。$$

由公式(3)可得知, y_{ij} 代表局部数据属性维的方差, y_{ij} 较大,表明 x_i 的局部数据集在属性维 j 上的值不均匀, x_{ij} 在其局部数据集中的密度较小, x_{ij} 所在的是稀疏子空间(相关子空间);相反, y_{ij} 较小,表明 x_i 的局部数据集在属性维 j 上的值较均匀, x_{ij} 在其局部数据集中的密度较大, x_{ij} 所在的是稠密子空间(不相关子空间)。根据公式(3),可以计算所有数据对象在每个属性维度的稀疏度,并生成整个数据集的稀疏因子矩阵,稀疏度矩阵每个维度的稀疏度 y_{ij} 是由稀疏子空间和稠密子空间混合组成的。通过稀疏度矩阵,可以更方便衡量数据空间的稠密和稀疏区域,去除均匀分布属性维,保留非均匀分布属性维,发现数据空间的相关子空间。

文献[21]中通过预先设定好的阈值对每个维度上稀疏因子进行区分,从而确定各数据对象的相关子

空间和不相关子空间,解决了检测相关子空间时时间复杂度较高的问题,提高了算法的效率,然而通过同一个阈值对所有的维度确定子空间,当各个维度的数据分布有差异时,得到的相关子空间就会有一定的误差。但高斯混合模型利用高斯概率密度函数(正态分布曲线)精确地量化事物,它是一个将事物分解为若干的基于高斯概率密度函数形成的模型。稀疏度矩阵每个维度的稀疏度 y_{ij} 是由稀疏子空间和稠密子空间混合组成的,高斯混合模型的灵活性使其能够很好地适应稀疏度的分布,可以自适应地得到该维度的稀疏子空间和稠密子空间^[22]。因此,利用高斯混合模型识别稀疏度 y_{ij} 所在的子空间,即该维度的相关子空间和不相关子空间。

把第 j 维所有数据对象的稀疏度 y_{ij} 用作高斯混合模型:

$$G(y) = \sum_{r=1}^m p_r N_r(y, u_r, \sigma_r) \quad (4)$$

其中, $N_r(y, \mu, \sigma)$ 是第 r 个高斯分布, p_r 是系数,且 $p_r > 0$, $\sum_{r=1}^m p_r = 1$ 。

第 r 个高斯分布的密度函数为:

$$N_r(y, \mu_r, \sigma_r) = \frac{1}{\sqrt{2\pi}\sigma_r} e^{-\frac{(y-\mu_r)^2}{2\sigma_r^2}} \quad (5)$$

其中,每个部分的密度函数 N_r 含有两个参数,分别为期望 u_r 和标准差 σ_r , u_r 描述正态分布的集中趋势位置, σ_r 描述正态分布中数据分布的离散程度。使用高斯混合模型对每一维度稀疏度进行拟合,把稀疏度分为两个高斯分布,稀疏度较大的分布中的数据构成相关子空间,较小的分布中的数据构成无关子空间,所以该高斯混合模型由两个高斯分布组成,即 $m=2$ 。EM 算法^[23](最大似然估计)是一种改善模型参数估计的迭代算法,可以用来估计模型中的参数,也是 GMM 参数估计最常用的一种方式。使用 EM 算法来估计高斯混合模型各个参数,得到每一维度的每个稀疏度属于两个高斯分布的概率值,在哪个高斯分布的概率值大就属于哪个高斯分布。同时可以得到每一维度的两个高斯分布的均值,均值较大的高斯分布的稀疏度比较大,属于均值较大的高斯分布的数据构成的子空间是稀疏子空间即相关子空间。

文献[22]中使用二进制矩阵 Z 直观表示稀疏因子矩阵的相关性。第 i 个数据对象第 j 维度的值 z_{ij} 的定义如下:

定义 5:对于数据对象 x_i , y_{ij} 是第 i 个数据对象第 j 维度的稀疏度, z_i 是其子空间向量, z_{ij} 是第 i 个数据对象第 j 维度的值,如果 y_{ij} 属于相关子空间,则 $z_{ij}=1$, 如果 y_{ij} 属于不相关子空间,则 $z_{ij}=0$ 。

根据 EM 算法对稀疏度区分的结果和定义 5,可以生成数据对象的子空间向量 z_i 。

3 相关子空间与扩展隔离森林离群检测

3.1 相关子空间与分支切割截距点

扩展隔离森林离群检测具备了隔离森林离群检测的优点,且分割数据使用带有随机斜率的超平面,即在每个分支点上选择一个具有随机“斜率”的分支切点,可以更好地检测隔离森林中树枝轴平行细分掩盖的轴平行之间存在的离群数据。扩展隔离树分枝的每次切割需要两条信息确定超平面:分支切割的随机斜率 \vec{n} 和分支切割的随机截距 \vec{p} 。在扩展隔离森林离群检测构建隔离树的过程中,随机选择分支切割的随机截距点可能在数据集的密集区域,会造成离群数据很难快速地从数据集中隔离出来;从随机选择子空间中生成分支切割的随机斜率受到无关维度的影响。

设 DS 是一个 d 维数据集, x_i 为 DS 中的任意数据对象,依据公式(3)、(4)、(5)和定义 5,可得到 x_i 的子空间向量 z_i 。为了刻画 x_i 所有属性维度的稀疏程度,可定义如下的 x_i 相关维系数 m_i :

$$m_i = \frac{\sum_{j=1}^d z_{ij}}{d} \quad (6)$$

由公式(6)可知, m_i 越大,表明 x_i 在相关子空间中的维度占 DS 全部维度 d 的比值越大,即: x_i 的稀疏维度数量越多, x_i 越可能处于稀疏区域;反之, m_i 越小,表明 x_i 在相关子空间中的维度占 DS 全部维度 d 的比值越小,即: x_i 的稀疏维度数量越少, x_i 越可能处于稠密区域。

利用相关维系数 m_i ,将数据集 DS 分成稠密分布的子集 D_d 和稀疏分布的子集 D_s ,其中: $D_d = \{x_i \mid x_i \in DS, m_i = 0\}$, $D_s = \{x_i \mid x_i \in DS, m_i \neq 0\}$ 。当 $m_i = 0$ 时, x_i 的每个维度都不属于相关子空间属性, x_i 不存在稀疏属性维, x_i 属于稠密区域;当 $m_i \neq 0$ 时,表示 x_i 的维度属于相关子空间属性, x_i 含有稀疏属性, x_i 属于稀疏区域。

由公式(3)可知, x_i 的稀疏度 y_{ij} 是在 x_i 的局部数据集 LDS 上获得的, x_i 的局部数据集 LDS 反映了 x_i 的分布情况,公式(4)和(5)高斯混合模型能很好地区分数据对象属性维的稀疏,因此 x_i 的子空间向量 z_i 能够准确反映局部数据的分布特征, m_i 的大小是由 z_i 的取值决定的,可以较为准确地刻画数据对象分布的稀疏,从而可将 DS 划分为稀疏区域和稠密区域。

在扩展隔离森林离群检测过程中,每次分割确定超平面的随机截距点 \vec{p} 时,若数据分布稀疏的数据集

D_s 中不为空,则在 D_s 中随机选择切割点,否则在稠密分布的数据集 D_d 中随机选择切割点。由于稠密区域的数据对象需要多次分裂才能被完全隔离,而稀疏区域的数据仅需几次分裂就能被完全隔离^[14]。优先从稀疏区域数据子集 D_s 中随机选择切割点,可使离群数据快速地隔离出来,避免了在稠密子空间中分割对隔离离群数据的掩盖。

在定义 3 中,超平面的切割斜率 $\vec{n} = \{n_1, n_2, \dots, n_d\}$,若超平面的切割截距数据对象 $\forall x_i \in D_s$,其子空间向量 $z_i = \{z_{i1}, z_{i2}, \dots, z_{id}\}$,其中:当 $z_{ij} = 0$ 时, $n_j = 0 (j=1, 2, \dots, d)$ 。切割斜率 \vec{n} 只保留在截距点数据对象 x_i 相关子空间维度上的值,将切割斜率 \vec{n} 在截距点数据对象 x_i 不相关子空间维度上的值赋为 0;若超平面的切割截距数据对象 $\forall x_i \in D_d$,其子空间向量 $z_i = \{z_{i1}, z_{i2}, \dots, z_{id}\}$, $z_{ij} (j=1, 2, \dots, d)$ 均为 0,则切割斜率 \vec{n} 只保留在截距点数据对象 x_i 随机一个维度上的值,将切割斜率 \vec{n} 在截距点数据对象 x_i 其余维度上的值均赋为 0。

3.2 扩展隔离森林分割平面选择策略

隔离树在选择分割平面时,若稀疏数据子集 D_s 中不为空,则在 D_s 中随机选择数据对象作为切割截距点 \vec{p} ,在截距点数据对象对应的相关子空间维度上生成超平面的随机斜率 \vec{n} ;否则在稠密分布的数据集 D_d 中,随机选择数据对象作为切割截距点 \vec{p} ,并在截距点数据对象的随机一个维度上生成超平面的随机斜率 \vec{n} 。对于给定 \vec{x} ,数据拆分策略为:

如果 $(\vec{x} - \vec{p}) \cdot \vec{n} < 0$,则将 \vec{x} 划分到左分支,否则它会移动到右分支。

依据该选择策略确定超平面,递归其分割操作,直到当前子树只包含一个节点或达到预设的树高。优先在数据分布稀疏的数据集中,选择分支切割的随机截距点,可快速地使离群数据从稀疏数据区域中隔离出来,并在截距点数据所对应的相关子空间维度上,生成超平面的随机斜率来确定超平面,可避免无关维度的影响,提高扩展隔离森林离群检测的效果。

3.3 扩展隔离森林离群检测算法

综上所述,采用相关子空间,扩展隔离森林离群检测基本思想为:首先,通过 kd 树寻找数据集中每个数据对象的 k 近邻并生成其局部数据集 LDS,依据公式(3)–(5)计算出每个数据对象的稀疏度 y_{ij} 并识别 y_{ij} 所在的子空间,生成数据对象 x_i 的子空间向量 z_i ;然后,随机采样 t 次构造 t 棵不同的隔离树,每次随机采样 ψ 个数据,按照上述分割平面选择策略,构造每棵

隔离树;最后,对整个数据集进行检测,使用 PathLength 函数计算数据对象 x_i 在 gTree 树的路径长度 $h(x)$,并计算 x_i 在各隔离树的平均路径长度,依据公式(2)将平均路径长度归一化后作为数据对象的离群得分,输出离群得分最高的前 M 个数据对象作为离群数据。其算法描述如下:

算法:RSGMM-EIF

输入:数据集 DS,子抽样大小 ψ ,隔离树数量 t ,近邻数 k

输出:离群数据

```

1.  $n = \text{DS}$  的数据个数,  $d = \text{DS}$  的维度
2. for(  $i = 0; i < n; i++$  )
3.  $\text{LDS}_i = \text{getknn}()$ ; //计算第  $i$  个数据对象的局部数据集 LDS
4.  $\text{spl} = \text{getspdegree}()$ ; //依据公式(3)计算第  $i$  个数据对象的稀疏度
5. }
6.  $Z = \text{getsubspace}()$ ; //依据公式(4)和(5)计算相关子空间
7.  $\text{IForest} = \text{iForest}(\text{DS}, t, \psi)$ ; //构造隔离森林
8.  $T = \text{gTree}(\text{DS}, e, l, Z)$ ; //构造 gTree
9. for  $x_i$  in DS
10.  $\text{h\_temp} = 0$ 
11. for  $j$  in range( $t$ )
12.  $\text{depth} = \text{PathLength}(\vec{x}, T, e)$ ;
13.  $\text{h\_temp} = \text{h\_temp} + \text{depth}$ ;
14. }
15.  $\text{Eh} = \text{h\_temp} / t$ ;
16. 依据公式(2),计算数据对象的离群得分;
17. }
18. 选取离群得分最大的  $M$  个数据对象作为离群数据;
19. End RSGMM-EIF

```

在上述 RSGMM-EIF 算法中,函数 iForest(DS, t, ψ)是将 gTree 树组成森林,PathLength(\vec{x}, T, e)是计算数据对象 x_i 在 gTree 树的路径长度,详见文献[14];函数 gTree(DS, e, l, Z)是依据扩展隔离森林分割平面选择策略,构造每棵隔离树,其描述如下:

函数:gTree(DS, e, l, Z)

输入:数据集 DS,当前树的高度 e ,隔离树的高度限制 l ,二进制矩阵 Z

输出:隔离树

//达到限定高度或孩子节点只有一个数据时完成树的构建

```

1. if  $e \geq l$  or  $|X| \leq 1$  then
2. return exNode
3. else
4. 依据公式(6)和(7)计算稀疏分布的数据集  $D_s$ ;
5. if  $\text{len}(D_s) \neq 0$ ,
6. 在  $D_s$  中随机选择数据对象作为切割截距点  $\vec{p}$ ;
7. 在截距点数据对象相关子空间维度上生成超平面的随机斜率  $\vec{n}$ ;

```

```

8. else
9.   在数据集中随机选择数据对象作为切割截距  $\vec{p}$ ;
10.   在截距点数据对象的随机一个维度上生成超平面的随机斜率  $\vec{n}$ ;
11. end if
12.  $X_l = \text{filter}(X, (\vec{x} - \vec{p}) \cdot \vec{n} < 0)$ ;
13.  $X_r = \text{filter}(X, (\vec{x} - \vec{p}) \cdot \vec{n} \geq 0)$ ; #递归构造隔离子树
14. return in Node {Left=iTree( $X_l, e+1, l$ ), Right = iTree( $X_r, e+1, l$ ), Normal $\leftarrow \vec{n}$ , Intercept $\leftarrow \vec{p}$ }
15. end if
16. End gTree

```

算法复杂性分析:参照文献[22],计算数据对象的稀疏度和利用高斯混合模型识别数据对象的相关子空间的时间复杂度是 $O(n \log n) + O(n * k * d) + O(n * d) \approx O(n \log n) + O(n * k * d)$;在构建 gTree 时,每次分割需要生成确定超平面的两个向量并计算不同数据对象的隔离方向,构造隔离森林的时间复杂度是 $O(td\psi \log_2 \psi)$;计算整个数据集离群得分的过程时间复杂度和空间复杂度是 $O(ntd \log_2 \psi)$;RSGMM-EIF 的时间复杂度是 $O(n \log n) + O(n * k * d) + O(td\psi \log_2 \psi) + O(ntd \log_2 \psi)$ 。

4 实验结果及分析

实验环境: Intel(R) Core(TM) i5-1135G7, 16 GB 内存, Windows10 操作系统, 采用 python 语言实现了 RSGMM-EIF 算法及对比算法 IF^[13]、EIF^[14]、iNNE^[15] 和基于密度的局部异常因子 LOF 算法^[24]。采用 UCI 数据集作为实验数据, 详见表 1。

表 1 UCI 数据集

Name	Instances	Attributes	Outliers
Pima	768	8	268 (35%)
Cardio	1 831	21	176 (9.6%)
Satellite	6 435	36	2 036 (32%)
Satimage-2	5 803	36	71 (1.2%)
Optdigits	5 216	64	150 (3%)
Mnist	7 603	100	700 (9.2%)

在表 1 中, Pima 是印第安人糖尿病诊断数据集, 由 8 个医学预测变量和一个目标变量 Outcome 类标变量(0 或 1)组成, 预测变量包括患者的怀孕次数、BMI、胰岛素水平、年龄等; Cardio 是由心电图上的胎儿心率(FHR)和子宫收缩(UC)特征测量结果组成的分类数据集, 产科专家将其结果分为正常、可疑和病理三个类别, 将可疑类的数据对象丢弃后作为离群检测数据集; Satellite 是卫星图像的分类数据, 样本分为 7 类, 其中第 6 类没有实例, 实验将数量较少的第 2、4、5 类样本标记为离群数据; Satimage-2 是将多分类数据集

Statlog (Landsat Satellite) 的第 2 类数据对象作为离群数据, 其他类合并作为正常数据得到的; Optdigits 是光学识别手写数字数据集, 数字 1~9 的数据对象作为正常数据, 150 个数字 0 的数据对象作为离群数据; Mnist 将手写数字的原始 Mnist 数据集中数字 0 的图像作为正常数据, 将抽取的 700 幅数字 6 图像中作为离群数据, 并且数据集的特征是从图像总共 784 个特征中随机抽取了 100 个。

性能评估指标: AC 和 AUC。准确率 AC 表示给定测试集预测正确的样本数与总样本数之比, 准确率能够衡量预测结果总体的正确率。ROC 曲线是一个二维平面上的曲线, 平面的横坐标是实际负样本中被错误预测为正样本的概率(FPR), 纵坐标是实际正样本中被预测正确的概率(TPR), 可准确反映 FPR 和 TPR 的关系, 能更好地衡量样本不均衡的情况, 是检测准确性的综合代表。AUC (Area Under Curve) 是 ROC 曲线下方的面积。AUC 值可以体现算法的性能, AUC 越接近 1, 表明算法效果越好, AUC 低于 0.5, 表明算法效果比随机检测还差。

4.1 参数 ψ

为了实验验证参数抽样数 ψ 对算法性能的影响, 采用了表 1 中的 5 个数据集, 其实验结果详见图 1 和图 2。

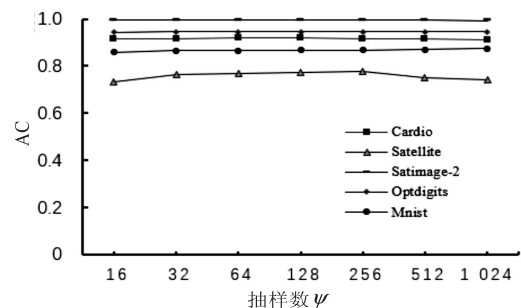


图 1 抽样数 ψ 对算法 AC 的影响

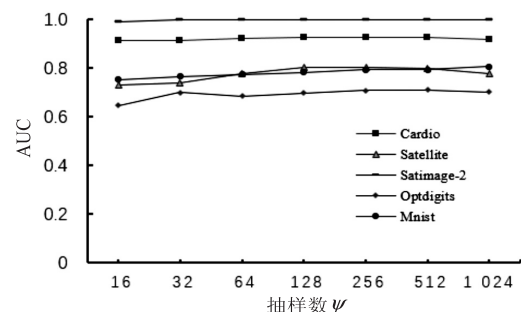


图 2 抽样数 ψ 对算法 AUC 的影响

从图 1 和图 2 可以看出, 随着 ψ 增大, AC、AUC 指标值逐渐增加, 并接近最优值, 后有所降低, 表明将 ψ 一般设置为 256 通常可取得好的检测效果, 无需再增大 ψ , 与文献[13]中的结论一致。其主要原因是随着 ψ 增大, 隔离树的树高增加, 可以更好地区分数据

对象的路径长度,得到数据对象的离群得分;当样本量过大时,包含正常数据对象过多,不利于离群对象被隔离出来,导致算法效果不好。

4.2 参数 t

为了实验验证参数隔离树的数量 t 对算法性能的影响,采用表 1 中的 4 个数据集,其实验结果详见图 3 和图 4。

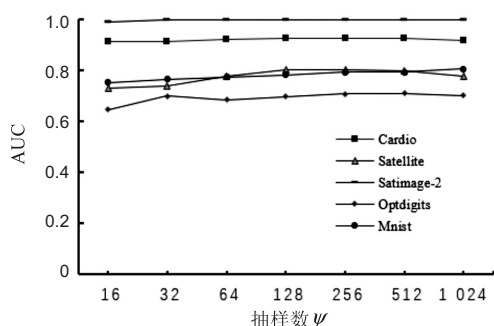


图 3 隔离树数量 t 对算法 AC 的影响

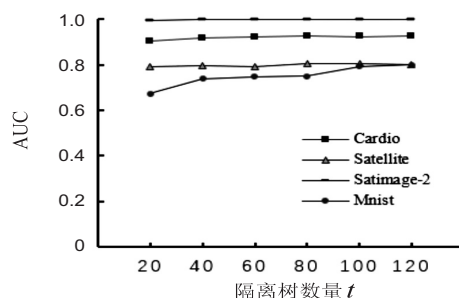


图 4 隔离树数量 t 对算法 AUC 的影响

从图 3 和图 4 可知,随着 t 的增加,AC、AUC 指标值逐渐增加,并趋于稳定,表明随着 t 值的增加,离群检测性能趋于稳定。其主要原因是隔离森林离群检测依赖于集成学习的聚合能力,多棵树集合更体现隔离森林离群检测优势。

4.3 近邻 k

为了实验验证参数近邻 k 对算法性能的影响,采用了表 1 中的 4 个数据集,其实验结果详见表 2。

表 2 近邻 k 对算法性能的影响

指标	数据集	5	10	15	20	30	40
AC	Cardio	0.915 019	0.916 985	0.911 742	0.906 718	0.892 736	0.883 779
AUC	Cardio	0.926 41	0.924 284	0.912 354	0.899 477	0.880 62	0.863 849
指标	数据集	10	20	40	60	80	100
AC	Satellite	0.782 471	0.763 761	0.753 846	0.775 851	0.755 369	0.764 724
	Optdigits	0.945 322	0.946 012	0.944 67	0.944 632	0.945 284	0.944 977
	Mnist	0.865 527	0.865 842	0.869 183	0.867 026	0.868 122	0.869 078
AUC	Satellite	0.796 289	0.801 738	0.791 23	0.803 898	0.796 945	0.796 524
	Optdigits	0.708 509	0.674 953	0.646 426	0.660 12	0.652 095	0.650 581
	Mnist	0.770 491	0.779 455	0.785 937	0.785 073	0.792 949	0.789 758

从表 2 可知,随着 k 的增加,AC、AUC 指标值逐渐增加,并接近最优值,随后逐渐趋于下降,其主要原因是随着近邻 k 的增加,数据对象的局部数据集可以更好地反映数据对象的分布情况,当 k 在 $(\frac{\sqrt{n}}{4} \sim \frac{\sqrt{n}}{2})$ 范围内,数据对象的局部数据集可较好地反映数据对象的分布情况,将数据集划分为稠密区域和稀疏区域,取

得较好的离群检测效果。

4.4 离群检测性能

为了实验验证算法的离群检测准确性,采用了表 1 中的数据集,以及 AC 和 AUC 指标,其实验结果详见表 3。RSGMM-EIF、IF 和 EIF 的默认参数:创建 100 棵隔离树,每棵树 256 个样本,树高限制为 8。

表 3 离群检测算法 AC 和 AUC 的比较

数据集	AC					AUC				
	RSGMM-EIF	IF	EIF	iNNE	LOF	RSGMM-EIF	IF	EIF	iNNE	LOF
Pima	0.717 4	0.673 6	0.640 3	0.673 4	0.572 9	0.756 1	0.675 7	0.640 3	0.713 5	0.558 6
Cardio	0.917	0.892 5	0.905 2	0.842 2	0.844 9	0.924 3	0.903 9	0.918 3	0.602 1	0.566 6
Satellite	0.775 9	0.718 7	0.766 9	0.586 7	0.626 4	0.803 9	0.702 3	0.724 9	0.512 2	0.571 4
Satimage-2	0.995 8	0.995 2	0.997 8	0.975 8	0.977 6	0.998	0.991 8	0.997 7	0.595 3	0.642 4
Optdigits	0.945 3	0.943 3	0.944 4	0.943 4	0.947 1	0.708 5	0.701 4	0.742 9	0.674 5	0.503 7
Mnist	0.869 2	0.861 6	0.876 5	0.860 9	0.882 2	0.785 9	0.773 4	0.825 1	0.698 4	0.786 3

从表 3 可知,在大部分数据集上,RSGMM-EIF 的 AC 和 AUC 指标值大,表明 RSGMM-EIF 优于其他算法的离群检测效果。在大部分数据集上,RSGMM-EIF 优于 IF 和 EIF 的离群检测效果,其主要原因是 RSGMM-EIF 利用了相关维系数 m_i ,将数据集划分为稠密区域和稀疏区域,优先在稀疏数据区域中选择分支切割的随机截距点,使离群数据快速地从稀疏数据区域中隔离出来,且在数据对象的相关子空间确定超平面的随机斜率,避免了无关维度的干扰,算法性能更佳。在少部分数据集上,EIF 是在数据集的全维度空间中确定超平面,利用了每个维度的信息,取得了良好的离群检测效果。iNNE 和 LOF 在大部分数据集上离群检测效果不佳,主要原因是 iNNE 的抽样数 ψ 对检测性能有显著影响,若 ψ 过大,会将分布密集的一些正常对象当成异常对象;LOF 没有考虑无关属性的影响,且对密度差异较大的簇边界上的数据对象评分不准确。

5 结束语

扩展隔离森林离群检测的超平面选取在数据集的密集区域或含有无关维度的区域,影响了离群检测效果。该文采用相关子空间,给出了一种扩展隔离森林离群检测算法 RSGMM-EIF。该算法利用数据对象的稀疏度和高斯混合模型,构造了相关子空间,并在扩展隔离树构建过程中,利用其相关子空间,确定超平面,从而使离群数据快速地从稀疏数据区域中隔离出来,避免了无关维度的干扰。

参考文献:

- [1] HAN J, KAMBER M, PEI J. 数据挖掘:概念与技术[M]. 第 3 版. 范明, 孟晓峰, 译. 北京:机械工业出版社, 2012: 351-374.
- [2] JIN F, CHEN M, ZHANG W, et al. Intrusion detection on internet of vehicles via combining log-ratio oversampling, outlier detection and metric learning[J]. Information Sciences, 2021, 579: 814-831.
- [3] 任家东, 刘新倩, 王倩, 等. 基于 KNN 离群点检测和随机森林的多层入侵检测方法[J]. 计算机研究与发展, 2019, 56(3): 566-575.
- [4] BEULAH J R, PUNITHAVATHANI D S. An efficient mixed attribute outlier detection method for identifying network intrusions[J]. International Journal of Information Security and Privacy, 2020, 14(3): 115-133.
- [5] MASSI M C, IEVA F, LETTIERI E. Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases[J]. BMC Medical Informatics and Decision Making, 2020, 20(1): 160.
- [6] KAIAFAS G, HAMMERSCHMIDT C, STATE R, et al. An experimental analysis of fraud detection methods in enterprise telecommunication data using unsupervised outlier ensembles[C]//2019 IFIP/IEEE symposium on integrated network and service management (IM). Arlington: IEEE, 2019: 37-42.
- [7] ZHANG J, XIE Y, PANG G, et al. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection[J]. IEEE Transactions on Medical Imaging, 2020, 40(3): 879-890.
- [8] ALAVERDYAN Z, JUNG J, BOUET R, et al. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening[J]. Medical Image Analysis, 2020, 60: 101618.
- [9] 陈庄, 黄勇, 邹航. 基于离群点挖掘的工业控制系统异常检测[J]. 计算机科学, 2014, 41(5): 178-181.
- [10] SAFAEI M, ISMAIL A S, CHIZARI H, et al. Standalone noise and anomaly detection in wireless sensor networks: a novel time-series and adaptive Bayesian-network-based approach[J]. Software: Practice and Experience, 2020, 50(4): 428-446.
- [11] DJENOURI Y, ZIMEK A, CHIARANDINI M. Outlier detection in urban traffic flow distributions[C]//2018 IEEE international conference on data mining (ICDM). Singapore: IEEE, 2018: 935-940.
- [12] 张继福, 蒋义勇, 胡立华, 等. 基于概念格的天体光谱离群数据识别方法[J]. 自动化学报, 2008, 34(9): 1060-1066.
- [13] LIU F T, TING K M, ZHOU Z H. Isolation forest[C]//2008 eighth IEEE international conference on data mining. Pisa: IEEE, 2008: 413-422.
- [14] HARIRI S, KIND M C, BRUNNER R J. Extended isolation forest[J]. IEEE Transactions on Knowledge & Data Engineering, 2021, 33(4): 1479-1489.
- [15] BANDARAGODA T R, TING K M, ALBRECHT D, et al. Isolation-based anomaly detection using nearest-neighbor ensembles[J]. Computational Intelligence, 2018, 34(4): 968-998.
- [16] KARCZMAREK P, KIERSZTYN A, PEDRYCZ W, et al. K-Means-based isolation forest[J]. Knowledge-Based Systems, 2020, 195: 105659.
- [17] ZHANG X, DOU W, HE Q, et al. LSHiForest: a generic framework for fast tree isolation based ensemble anomaly analysis[C]//2017 IEEE 33rd international conference on data engineering (ICDE). San Diego: IEEE, 2017: 983-994.
- [18] 李倩, 韩斌, 汪旭祥. 基于模糊孤立森林算法的多维数据异常检测方法[J]. 计算机与数字工程, 2020, 48(4): 862-866.
- [19] HARMAN R, LACKO V. On decomposition algorithms for uniform sampling from n-spheres and n-balls[J]. Journal of