基于隔离森林集成策略的分类型属性分组离群检测

宋亦静，张继福

引用本文

宋亦静，张继福. 基于隔离森林集成策略的分类型属性分组离群检测[J]. 计算机科学，2026，53(1)：115-127.

SONG Yijing, ZHANG Jifu. Attribute Grouping-based Categorical Outlier Detection Using Isolation Forest Ensemble Strategy [J]. Computer Science, 2026, 53(1): 115-127.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000163

基于隔离森林集成策略的分类型属性分组离群检测

宋亦静　张继福
太原科技大学计算机科学与技术学院　太原 030024
(b202115310016@stu.tyust.edu.cn)

摘　要　属性分组是高维离群检测的有效途径之一,但现有的属性组离群检测集成策略仅利用了各属性组内的局部离群信息,忽略了属性组的全局离群信息,导致属性组离群信息集成出现偏差。为此,利用属性组局部与全局离群信息,提出了一种基于隔离森林集成策略的分类型属性分组离群检测方法。该方法根据属性之间的相关性,将属性自动划分为若干属性组,获得数据对象在各属性组中的离群信息;理论分析了现有离群信息集成策略存在集成偏差,并定义了属性组集成偏差系数;利用隔离森林设计了一种离群信息集成策略,有效地刻画了属性组局部与全局离群信息,降低了属性组离群检测集成偏差,并在此基础上提出了一种分类型属性分组离群检测算法。实验结果表明,与对比方法相比,该算法的 AUC 指标、效率分别平均提高了7.83%和48.43%。

关键词:离群检测;属性分组;集成偏差系数;隔离森林集成策略;全局离群信息路径
中图分类号　TP311

Attribute Grouping-based Categorical Outlier Detection Using Isolation Forest Ensemble Strategy

SONG Yijing and ZHANG Jifu
School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Abstract　Attribute grouping is one of the effective steps in high-dimensional outlier detection, but the current ensemble strategies in attribute grouping-based outlier detection only take into account the local outlier information within each attribute group, and ignore the global outlier information of all attribute groups, which can lead to a biased ensemble of attribute group outlier information. This paper proposes an attribute grouping outlier detection approach based on Isolated Forest ensemble strategy by using the local and global outlier information of attribute groups. Firstly, attributes are automatically divided into several attribute groups based on the local and global correlation among attributes, and the outlier information of data objects is obtained in each attribute group. Secondly, from the perspective of attribute grouping, the ensemble bias of the current outlier information ensemble strategy is theoretically analyzed, and the ensemble deviation coefficient are defined as the evaluation index of the outlier information ensemble strategy. Then an attribute grouping-based isolation forest ensemble strategy for categorical outlier detection is proposed, this strategy effectively depicts the local and global outlier information of attribute groups and lowers the ensemble bias of attribute group outlier detection. In the end, experimental results on the UCI validate that the ensemble strategy effectively alleviates the ensemble bias and improves the outlier detection performance. Importantly, compared with the competing methods, the algorithm bolsters the AUC index and the detection efficiency by averages of 7.83% and 48.43%.

Keywords　Outlier detection, Attribute grouping, Ensemble deviation coefficient, Isolated Forest ensemble strategy, Outlier information of global paths

1　引言

离群检测作为数据挖掘领域中的主要研究内容之一,旨在寻找明显偏离其他数据,不满足数据一般模式或行为的数据对象[1-2]。其已被广泛应用在城市交通[3]、工业故障[4]、医疗监测[5]、网络异常[6]等领域。目前,大多数离群检测都面向数值型数据[7-9],分类型数据离群检测方法相对较少,但随着实际应用的不断深入,出现了越来越多的高维分类型数据集[10]。属性分组是将高维数据的属性集划分为若干低维数据的属性组,将相关性强的属性划分在同一个组中,相关性低的属性划分在不同组中[11]。由于不同的属性组捕获了不同类型的离群信息,且降低了离群检测的搜索维度,因而属性分组是进一步提升离群检测性能和缓解"维度灾难"的有效途径之一[11]。

到稿日期:2024-10-29　返修日期:2025-02-14
基金项目:国家自然科学基金(62172293)
This work was supported by the National Natural Science Foundation of China(62172293).
通信作者:张继福(jifuzh@sina.com)

属性分组作为高维离群检测的有效途径之一，可有效降低“维灾”干扰。给定数据对象在各属性组中具有不同的离群程度，如何有效且合理地将数据对象在各属性组中的不同离群信息集成并最终检测出离群数据，是属性分组离群检测的关键步骤之一，影响着离群检测性能。现有的离群信息集成策略大多为平均值和最大值集成[11-15]，但其仅考虑了各属性组内数据对象的局部离群信息，而忽略了属性组全局离群信息，可能导致一些数据的离群程度偏高或偏低，从而产生属性组集成偏差，影响离群检测效果。本文利用隔离森林刻画属性组局部与全局离群信息，提出了一种基于隔离森林集成策略的分类型属性分组离群检测方法，有效降低了属性组离群信息集成偏差，提升了离群检测性能。本文的主要贡献如下：

1）利用属性相关性向量，给出了一种分类型属性自动分组算法；

2）理论分析了现有离群检测集成策略存在集成偏差，并设计了一种隔离森林集成策略；

3）提出了一种基于隔离森林集成策略的分类型属性分组离群检测算法。

本文第 2 章讨论了现有的分类型数据离群检测、属性分组与离群程度集成策略；第 3 章介绍了本文所采用的属性分组与离群检测方法；第 4 章从属性分组角度理论分析了现有常用离群信息集成策略存在属性组集成偏差，并定义了属性组集成偏差系数 EDC；第 5 章设计了隔离森林集成策略，提出了一种基于隔离森林集成策略的分类型属性分组离群检测算法；第 6 章描述了实验设置和结果；最后总结全文并展望未来。

2 相关工作

离群检测是数据挖掘中的重要研究内容之一。离群数据对象刻画了一些重要或特殊的行为模式，且可以获得一些有价值的信息[16]。不同于数值型数据，分类型数据具有取值无序与不可比等特点，是一类广泛出现在许多应用领域中的重要数据类型[17-20]。通过属性分组，可将高维数据划分为若干低维数据，并可有效地检测出隐藏在低维数据集中的离群数据对象[11]。

2.1 分类型数据离群检测

分类型数据离群检测作为一种重要的离群检测方法，可应用于入侵检测[17]、欺诈检测[18]和疾病早期检测[19]等领域。Dino 等[21]提出了基于上下文的分类属性的距离，首先根据属性间的相关性得出上下文属性；然后使用属性值的频率分布与每个上下文属性，推断出目标属性之间基于上下文的距离，并使用一种基于距离的方法进行离群检测。Pang 等[22]提出了一种耦合无监督离群点检测方法 CBRW，它根据属性取值之间的条件概率来刻画属性值之间的耦合，并通过对属性图上的偏置随机游动建模耦合来估计每个属性取值的异常值。Pang 等[23]提出了另一种耦合离群点检测方法 SDRW，其动机来自于 CBRW，但它用提升概念取代了条件概率对异常值的影响，有效地将值图转换为无向图，能够更好地区分离群值和正常值。Xu 等[24]提出了一种基于无监督嵌入的复杂耦合方法 SCAN，首先建模主值耦合，然后定义偏值耦合，突出异常值的本质，进而提出了双向选择值耦合学习方法来说明如何通过值耦合估计离群值。Zhang 等[25]使用相似哈希函数生成多叉隔离树，提出了 LSHiForest 算法，该算法展现出更好的数值型数据检测性能，并进一步将快速隔离机制扩展到分类型数据。Xiang 等[26]设计了最优隔离森林 OptIForest，建立了一个关于隔离效率的理论，并确定了隔离树的最优分支因子，解决了 LSHiForest 中关于隔离森林最优树结构的问题。

然而，上述分类型数据离群检测采用的是全维空间，针对高维分类型数据，容易出现“维灾”现象，难以检测出隐藏在高维数据中的离群数据对象。

2.2 属性分组与集成策略

属性分组将高维数据的属性集划分为若干低维数据的属性组，能够有效地发现隐藏在各个属性组中的离群数据对象。Au 等[27]提出了属性分组方法 ACA，该方法首先给出属性分组的个数 c，在所有属性中随机选 c 个属性作为每组的核心属性，然后根据属性组内的相似性不断地迭代更新核心属性和属性分组。Li 等[11]使用互信息和熵的比值来度量两个属性之间的相关性，首先通过优化组内相似性最大来确定属性分组的个数，然后不断迭代更新每个属性组的核心属性，每个组由高度相关的属性组成。Zheng 等[28]采用交互增益[29]作为属性之间的相似性度量，以属性作为节点，属性之间的相关性作为边的权重，构建一个无向图，并采用最小生成树的构造方法为其生成子图。该方法通过迭代地删除子图中权重最小的边来对属性进行分组，使得相似性小的属性出现在不同的组中。

在属性组离群检测中，给定的数据对象在不同属性组的离群程度不同，如何有效集成数据对象在各属性组中的离群信息，是关键问题之一。常用的离群信息集成策略有平均值集成策略与最大值集成策略。Li 等[11]分别在不同属性组中选取 k 个离群程度最大的数据对象，再从中将离群程度最高的数据对象作为离群数据，即最大值集成策略。Akanksha 等[30]提出的 AnD-SELECT 方法首先选择性地构建离群程度集合，以保持离群检测准确性和多样性的平衡，然后采用平均值集成的方法得到集成离群程度。Akanksha 等[12]提出的 FairComb 框架首先为所有检测方法的离群程度提供公平性处理方案，减小了离群程度在范围和尺度上的差异，提高了离群程度的可比性；然后采用平均值的方法集成离群程度，得到最终的离群检测结果。为了平衡不同离群检测方法之间的公平性与离群检测性能，Liu 等[31]提出了一个有效的框架，该框架首先通过叠加结构将任一离群集成策略转化为公平感知的离群集成，给出了平衡公平性和 AUC 度量优化问题的封闭解；然后生成离群值向量，并采用平均值方法集成离群程度。虽然上述属性组离群程度的集成策略都是有效的，但它们仅考虑了不同数据对象离群信息在数值上的大小差异，忽略了不同离群信息的分布信息，导致离群检测的准确度不高。

综上所述，在高维分类型数据离群检测中，属性分组是缓解“维灾”的有效途径之一，如何有效集成不同属性组的离群信息，是该方法的关键问题之一。现有的大多数离群信息集成策略采用最大值集成策略与平均值集成策略，此类方法仅包含各属性组内的局部离群信息，未包含所有属性组全局离群信息，因此平均值与最大值集成策略丢失了部分离群信息，影响了离群检测效果。

3 基础知识

在高维离群检测中，属性分组作为缓解“维灾”干扰的有效途径之一，其基本步骤是：首先，将所有属性分为若干属性组，且组内的属性相关性强，组间的属性相关性弱；然后，在各个组中分别度量数据对象的离群信息；最后，集成所有属性组的离群信息，得到数据对象的离群程度，并选取若干离群程度较大的数据对象作为离群数据[11]。

假设 $DS=\{x_1,x_2,\cdots,x_n\}$ 为分类型数据集，其中 n 为数据对象个数，$Y=\{y_1,y_2,\cdots,y_m\}$ 是 m 个分类属性的集合。对于任一 $x_i\in DS$，可以将其表示为一个向量$[x_{i1},x_{i2},\cdots,x_{im}]$，其中 $1\leqslant i\leqslant n$。表 1 列出了相关符号及其具体含义。

表 1 相关符号

Table 1 Symbols and description

符号	描述	符号	描述
DS	数据集	x_{ij}	第 i 个数据对象的第 j 个属性值
X	DS 中的数据对象集合	C_r	第 r 个属性组
Y	DS 中的属性集合	c	属性组的个数
x_i	X 中的第 i 个数据对象	q	属性组中包含属性的个数
y_j	Y 中的第 j 个属性	k	离群数据对象的个数
n	X 中数据对象的个数	m	Y 中属性的个数
OS	离群数据对象集合	G_Y	属性图
G_t	属性子图	U	属性相关性向量集合
S	属性组离群信息	S'	S 的子采样
$S(x_i)$	第 i 个数据对象的属性组离群信息	S_r	第 r 个属性组的离群信息
ψ	子采样数据对象的个数	t	属性组隔离树的个数

3.1 属性分组

属性分组将所有属性分为若干属性组，能够有效降低离群检测的搜索维度，减少“维灾”干扰，进一步提升高维离群检测效果。对于给定属性集 Y，依据属性之间的相关性，将所有属性 y_i 分配到一个属性组 C_r 的过程，被称为属性分组，其中 $i\in\{1,\cdots,m\}$，c 为分组个数，$r\in\{1,\cdots,c\}$，且任意两个属性组互不相交。参照文献[11]，下面给出相关概念的形式化描述。

对于任意两个给定属性 y_i 和 y_j，设 $d(y_i,y_j)$ 表示属性之间的相关性。若 $C_r=\{y_i\mid i=1,\cdots,q\}$ 为任意一个包含 q 个属性的属性组，则将属性 y_i 与 C_r 中的其他属性相关性之和称为多重关系 $Md(y_i)$，其表示如下：

$$Md(y_i)=\sum_{j=1}^{q}d(y_i,y_j) \tag{1}$$

式(1)表明，多重关系 $Md(y_i)$ 是属性 y_i 和属性组 C_r 中所有其他属性的关系和。$Md(y_i)$ 值越大，表明 y_i 与组内其他属性之间的相关性之和越大，因此选择 C_r 中的 Md 值最大的属性 σ_r 作为该组的核心属性。

属性分组就是将所有属性划分为若干属性组，其基本步骤是：1)选择 c 个核心属性；2)根据相关性 $d(y_i,y_j)$，将任意属性 y_i 分配给与其相关性最强的核心属性 c_r 所在的组 C_r 中；3)更新属性组中的核心属性，直到核心属性不再变化，即可将所有属性分为 c 个属性组。

3.2 离群检测

离群检测是识别与大多数数据对象具有明显差异的数据对象。针对分类型数据集，出现频率较少的属性取值包含了重要的离群信息。参照文献[11]，在第 r 个属性组 C_r 中，数据对象 x_i 的离群得分 $score_r(x_i)$ 表示如下：

$$score_r(x_i)=\frac{1}{q}\sum_{j=1}^{q}\begin{cases}0, & \text{if } n(x_{ij})=1\\ w(y_j)g(n(x_{ij})), & \text{else}\end{cases} \tag{2}$$

其中，q 为属性组 C_r 中包含的属性个数，x_{ij} 表示对象 x_i 的第 j 个属性的值；$n(x_{ij})$ 是 x_{ij} 出现的次数；$g(n(x_{ij}))=(n(x_{ij})-1)\log(n(x_{ij})-1)-n(x_{ij})\log(n(x_{ij}))$，函数 $g(n(x_{ij}))$ 刻画了出现次数较少的数据对象具有较高的离群得分；$w(y_i)$ 表示在第 r 个属性组中属性 y_j 的权重，$0\leqslant w(y_i)\leqslant 1$，其值越大，$y_i$ 越重要。式(2)体现了数据对象 x_i 在属性组中的离群信息。在第 r 属性组中，$score_r(x_i)$ 小于 0 且越接近于 0，表示数据对象 x_i 越有可能是离群数据。

式(2)仅体现了 x_i 在第 r 属性组中的离群信息。由于 DS 被划分为 c 个属性组，因此如何有效集成 x_i 在 c 个属性组的离群信息，充分体现 x_i 在 DS 中的离群程度，是属性分组离群检测方法的关键步骤之一。

4 属性分组与集成偏差

属性分组将相关性强的属性划分在同一个组中，相关性弱的属性划分在不同组中，且在各属性组中可以捕获到不同类型的离群信息。如何有效利用各属性组离群信息是设计集成策略的关键，影响着离群检测效果。

4.1 属性分组

不同的属性组可以捕获到不同类型的有用信息，利用属性相关性向量刻画属性之间的局部与全局相关性，将相关属性分配在同一属性组[32]。对于任意两个给定属性 y_i 和 y_j，采用属性相关性向量 $\vec{u}_i$ 有效地表征属性的局部与全局相关性。参照文献[8]，对于任意两个给定属性 y_i 和 y_j，互信息可以用如式(3)表示：

$$MI(y_i,y_j)=\sum_{k=1}^{d_i}\sum_{l=1}^{d_j}P_{ij}(y_i=v_{ik}\wedge y_j=v_{jl})\log\frac{P_{ij}(y_i=v_{ik}\wedge y_j=v_{jl})}{P_i(y_i=v_{ik})P_j(y_j=v_{jl})} \tag{3}$$

可通过式(3)与图结构来捕获属性之间的局部与全局关系。将 Y 中的任意两个属性 y_i 和 y_j 作为图节点，式(3)中的 $MI(y_i,y_j)$ 作为图边权重，构建属性图 G_Y。随着属性数量的增加，G_Y 的复杂度会增加，因此采用最小生成树构造方法 Kruskal[18] 来构造属性子图 $G_t=(Y,W_t)$。G_t 保留了 G_Y 中的属性间最大相关性，有效降低了属性图 G_Y 结构的复杂性。分

别采用属性之间的一阶相关度与二阶相关度[20]来刻画属性子图 G_t 中包含的局部和全局相关性。

一阶相关度刻画了直接用边连接的两个属性之间的相关度，描述了属性间的局部相关性。将 y_i 映射为 R^d 中的一个随机初始化属性向量 $\vec{u}_{i1}\in R^d$。采用属性向量 $\vec{u}_{i1}$ 之间的联合概率 $p_1(y_j,y_i)$ 刻画局部相关性，同时在 G_t 中采用节点 y_i 之间的经验概率 $\hat{p}_1(y_j,y_i)$ 刻画局部相关性，通过最小化属性 $p_1(y_j,y_i)$ 与 $\hat{p}_1(y_j,y_i)$ 之差，得到每个属性在 d 维空间中的一阶相关性表征向量 $\vec{u}_{i1}$，表示如下：

$$O_1=dif(\hat{p}_1(y_i,y_j),p_1(y_i,y_j)) \tag{4}$$

其中，$dif(\cdot,\cdot)$ 是距离函数，采用 KL 散度 $dif(\cdot,\cdot)$ 表示。采用负采样方法[20]逐步优化 $\vec{u}_{i1}$，使得 O_1 最小。

二阶相关度刻画了两个属性在邻域上的相关度，体现了属性间的全局相关性。将 G_t 中的每个属性映射为 R^d 中的两个随机初始化属性向量，即属性自身的表征向量 $\vec{u}_{i2}\in R^d$ 以及作为其他属性邻居时的表征向量 $\vec{u}'_{i2}\in R^d$。采用属性向量之间的条件概率 $p_2(y_j\mid y_i)$ 刻画属性向量 $\vec{u}_{i2}$ 周围存在邻居的全局特性，采用经验概率 $\hat{p}_2(y_j,y_i)$ 刻画 G_t 中属性 y_i 周围存在邻居属性 y_j 的全局特性。同样，通过最小化 $p_2(y_j\mid y_i)$ 与 $\hat{p}_2(y_j,y_i)$ 之差，得到每个属性在 d 维空间中的二阶相关性表征向量 $\vec{u}_{i2}$，表示如下：

$$O_2=-\sum_{i\in V}\lambda_i dif(\hat{p}_2(v_j\mid v_i),p_2(v_j\mid v_i)) \tag{5}$$

其中，λ_i 为控制属性的重要性因子，用 G_t 中属性节点 y_i 的度数表示；$dif(\cdot,\cdot)$ 为 KL 散度。采用负采样方法[20]逐步优化更新 $\vec{u}_{i2}$ 和 $\vec{u}'_{i2}$，使得 O_2 最小。

将属性的一阶相关性向量 $\vec{u}_{i1}=[u_1,u_2,\cdots,u_d]$ 与二阶相关性向量 $\vec{u}_{i2}=[u_{d+1},u_{d+2},\cdots,u_{2d}]$ 直接拼接，得到属性相关性向量 $\vec{u}_i$，表示如下：

$$\vec{u}_i=[u_1,u_2,\cdots,u_d,u_{d+1},u_{d+2},\cdots,u_{2d}] \tag{6}$$

其中，$\vec{u}_i$ 同时保留了属性之间的一阶相关性和二阶相关性，可有效体现属性之间的局部与全局相关性，有助于属性分组。

利用属性相关性向量 $\vec{u}_i$ 构造属性分组的基本步骤为：首先，以属性 y_i 为节点，属性之间的互信息 $MI(y_i,y_j)$ 为边的权重，构建属性图 G_Y；其次，生成属性子图 G_t，依据式(4)—式(6)，得出每个属性 y_i 的属性相关性向量 $\vec{u}_i$，根据 $\vec{u}_i$ 度量 3.1 节式(1)中的 $d(y_i,y_j)$；最后，根据轮廓系数值自动选取 c 个初始核心属性 σ_r，将任意属性 y_i 根据 $d(y_i,y_j)$ 分配到与其相关性最强的 σ_r 所在的组 C_r 中，将所有 m 个属性分为 c 个属性组。构造属性分组的伪代码描述如算法 1 所示。

算法 1 CAVAG()

输入：数据集 DS

输出：属性分组 C

1. 根据式(3)构建属性图 G_Y；
2. 根据 Kruskal algorithm 算法[18]，构建属性图 G_Y 的属性子图 G_t；
3. for(i=1,i≤m,i++) /* m 是 DS 中的属性个数
4. 根据式(4)计算属性的一阶相关性向量 $\vec{u}_{i1}$；
5. 根据式(5)计算属性的二阶相关性向量 $\vec{u}_{i2}$；
6. 根据式(6)得到属性相关性向量 $\vec{u}_i$；
7. end for
8. 构造属性相关性向量 U=$\{\vec{u}_1,\vec{u}_2,\cdots,\vec{u}_m\}$，对于 i,j∈{1,…,m}，采用马氏距离，计算式(1)中的 $d(y_i,y_j)$；
9. for(k=2; k≤m/2; k++)
10. 选择 k 个核心属性 σ_r; /* r∈(1,k) */
11. 对于 i∈{1,…,m}，根据 $d(y_i,\sigma_r)$，将属性 y_i 分配到离 y_i 最近的核心属性 σ_r 所在的属性组 C_r 中；
12. 采用文献[11]中的方法自动更新核心属性 σ_r；
13. 参照文献[33]计算 k 个属性组的轮廓系数；
14. end for
15. 从 m/2 个不同的属性分组中，利用最大轮廓系数值自动选择最优的属性组个数 c 及对应属性分组 C。
16. END CAVAG

4.2 离群信息集成策略

对于任意数据集 DS，假设所有属性 Y 划分为 c 个属性组，在各属性组中，可根据式(2)定义的离群得分值刻画其数据对象的离群程度。对于 DS 中的任意数据对象 x_i，其在 c 个属性组中的离群信息可描述为 $S(x_i)=\{score_1(x_i),score_2(x_i),\cdots,score_c(x_i)\}$，$x_i$ 在不同属性组中的离群程度是不同的，因此不同属性组的离群信息存在较大差异。利用 $S(x_i)$，得到 x_i 的集成离群程度 $Score(x_i)$。参照文献[14]，$Score(x_i)$ 的两种常用离群信息集成策略如下。

1)最大值集成策略：$S(x_i)$ 中的最大值为数据对象 x_i 的离群程度得分值。

$$Score(x_i)=\max[score_1(x_i),score_2(x_i),\cdots,score_c(x_i)] \tag{7}$$

2)平均值集成策略：$S(x_i)$ 中的平均值为数据对象 x_i 的离群程度得分值。

$$Score(x_i)=\mathrm{avg}[score_1(x_i),score_2(x_i),\cdots,score_c(x_i)] \tag{8}$$

由于 $score_r(x_i)$ 仅刻画了数据对象 x_i 在第 r 个属性组内的离群程度，因此属性组内的离群信息是一种局部离群信息，$score_1(x_i),score_2(x_i),\cdots,score_c(x_i)$ 分别是数据对象 x_i 在各属性组内的局部离群信息。由式(7)和式(8)可知，$Score(x_i)$ 仅体现了 x_i 在属性组内的局部离群信息，忽略了属性组全局离群信息。

4.3 离群信息集成理论分析

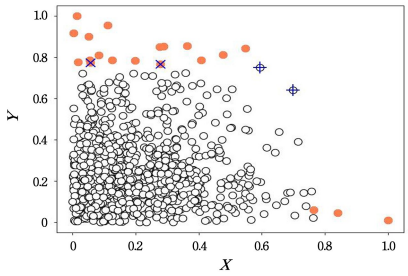
在属性分组离群检测中，首先，将数据集 DS 中的属性 Y 依据属性之间的相似性划分为若干不同属性组，并在各属性组中分别度量数据对象的离群信息；其次，集成各属性组的离群信息，得到该数据对象的离群程度；最后，选取离群程度最高的若干数据对象作为离群数据。离群信息集成策略是属性分组离群检测的重要环节之一，由式(7)和式(8)可知，不同的集成策略对离群检测效果具有不同影响。

定理 1 设数据集 DS 被分为 c 个属性组，x_i 是 DS 中的任意数据对象，且在第 r 个属性组中，x_i 是离群程度最高的若干数据对象之一的概率为 p_r，如果采用式(7)作为离群信息集成策略，则 x_i 成为 DS 中若干离群数据对象的概率为 $P=1-\prod_{r=1}^{c}(1-p_r)$。

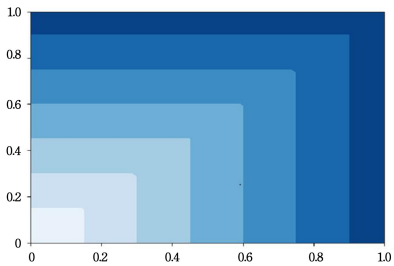
由定理 1 可知，当属性组个数 c 较大时，x_i 成为离群数据

的概率 P 较大。因此，采用最大值集成策略可能导致一些数据对象的离群程度偏高。

例 1　假设数据集含有 800 个数据对象，将其属性划分为两个属性组，其中第一属性组中的数据对象满足以 0 为标准差、1 为方差的半正态分布，第二属性组中的数据对象满足以 0 为标准差、30 为方差的半正态分布。针对 800 个数据对象，两个属性组中各数据对象的离群程度分别由 x 轴和 y 轴刻画，采用最大值集成策略的热图与其离群检测结果示意图如图 1 所示。图 1(a)中，圆圈代表正常数据，红色实型圆圈代表已检测出的离群数据，标记“＋”号的实型圆圈代表未检测出的离群数据，标记叉号的实型圆圈代表离群程度偏高的数据对象。在图 1(b)中，颜色越深，代表离群程度越高，且偏向坐标轴的数据离群程度偏高，因而图 1(a)中红色实型圆圈(已检测出的离群数据)大多偏向坐标轴，丢失了部分离群信息，未能检测出部分“＋”标记的离群数据，同时将“×”标记的某些数据检测为离群数据。



(a)最大值集成策略的离群检测结果



(b)离群得分的热图

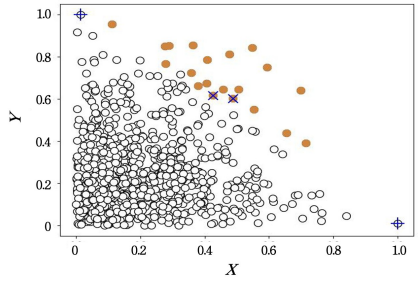
图 1　最大值集成策略(电子版为彩图)

Fig. 1　Maximum ensemble strategy

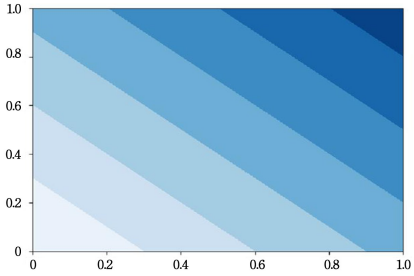
定理 2　设数据集 DS 被分为 c 个属性组，x_i 是 DS 中的任意数据对象，且在第 r 个属性组中，x_i 是离群程度最高的若干数据对象之一的概率为 p_r，若采用式(8)作为离群信息集成策略，则 x_i 成为 DS 中若干离群数据对象的概率 $P \leqslant \exp\left(-\frac{1}{2}c\,(1-2\bar{p})^2\right)$。

由定理 2 可知，当属性组个数 c 较大时，x_i 成为离群数据的概率 P 较小。因此，采用平均值集成策略可能导致一些数据对象的离群程度偏低。

例 2　利用图 1 生成的数据，采用平均值集成策略的热图与其离群检测结果示意图如图 2 所示。图 2(a)中，圆圈代表正常数据，棕色实型圆圈代表已检测出的离群数据，标记“＋”号的实型圆圈代表因离群程度偏低而未检测出的离群数据。图 2(b)中，偏向坐标轴的数据离群程度偏低，因而图 2(a)中棕色实型圆圈(已检测出的离群数据)大多偏离坐标轴，丢失了部分离群信息，未能检测出部分“＋”标记的离群数据，同时将“×”标记的某些数据检测为离群数据。



(a)平均值集成策略的离群检测结果



(b)离群得分的热图

图 2　平均值集成策略(电子版为彩图)

Fig. 2　Average ensemble strategy

综上所述，两种常用离群信息集成策略仅利用了各属性组内的局部离群信息，忽略了属性组的全局离群信息，可能导致一些数据对象的离群程度偏高或偏低，因而未能有效地刻画数据对象的离群程度，影响了离群检测性能。

4.4　属性组集成偏差

数据集 DS 采用属性分组，将相关性强的属性划分为一组，相关性弱的属性划分为不同组，各属性组内的局部离群信息存在较大差异。由 3.3 节可知，由于忽略了属性组间的离群信息，最大值集成策略可能导致部分数据对象的离群程度偏高，平均值集成策略可能导致部分数据对象的离群程度偏低，影响了离群检测性能。在属性分组离群检测中，离群信息集成策略可能导致部分数据对象的离群程度偏高或偏低，称为属性组集成偏差。

离群信息集成策略应同时刻画属性组内的局部离群信息及属性组间的离群信息，从而避免属性组集成偏差。在属性分组离群检测中，离群程度集成应合理客观地包含各分组的离群程度，不应偏好于特定属性组局部离群信息，降低离群程度的属性组集成偏差。参照文献[12]中的集成奇偶性 EP，定义了如式(9)所示的属性组集成偏差系数(Ensemble Deviation Coefficient，EDC)。

$$EDC=\frac{\mathrm{mean}(|OD_{G1}\cap OD|,|OD_{G2}\cap OD|,\cdots,|OD_{Gc}\cap OD|)}{\mathrm{var}(|OD_{G1}\cap OD|,|OD_{G2}\cap OD|,\cdots,|OD_{Gc}\cap OD|)} \tag{9}$$

其中，OD_{Gi} 表示在 i 个属性组中检测到的离群数据对象集合；c 为属性组个数，OD 表示 c 个属性组共同检测到的离群数据对象集合。$|OD_{Gi}\cap OD|$ 刻画了第 i 属性组的离群信息对离群检测的贡献量，其值越大，第 i 个属性组对离群检测的贡献越大。

属性组集成偏差系数 EDC 体现了各属性组对离群检测的贡献量均值与方差比值，其值越大，离群检测中包含越多的各属性组离群信息，且对属性组局部依赖越小，表示属性组集成偏差越小，反之亦然。因此，EDC 有效刻画了属性组集成偏差程度，可作为度量离群信息集成策略优劣的性能指标。

5 隔离森林集成策略与离群检测

现有的离群程度集成策略仅利用了各属性组内的局部离群信息，缺乏属性组的全局离群信息，导致属性组集成出现偏差。

5.1 属性组离群信息与隔离森林

离群检测是从海量数据集中识别或检测与多数数据对象具有明显差异的、潜在有用的若干数据对象[25]。对于任意数据集数据 DS，假设将属性 Y 划分为 c 个属性组，并在各属性组中分别刻画数据对象的离群程度。参照式(2)，$S_r=\{score_r(x_1),score_r(x_2),\cdots,score_r(x_n)\}$ 描述了第 r 属性组的离群信息，体现了第 r 个组内数据对象之间存在的差异，因此，$score_r(x_1),score_r(x_2),\cdots,score_r(x_n)$ 之间的差异刻画了属性组内的局部离群信息。属性组离群信息 $S=\{S(x_1),S(x_2),\cdots,S(x_n)\}$ 描述了 n 个数据对象在 c 个属性组中的离群信息，其中 $S(x_i)=\{score_1(x_i),score_2(x_i),\cdots,score_c(x_i)\}$ $(i=1,2,\cdots,n)$，$S(x_1),S(x_2),\cdots,S(x_n)$ 之间的差异刻画了属性组的全局离群信息。

在属性组离群检测中，离群信息 S 蕴含了数据对象在 DS 中的局部和全局离群信息。隔离森林是采用二叉树数据结构分离异常对象的有效途径之一[31-34]。一般情况下，由于异常对象个数少且明显不同于大多数对象，隔离森林隔离的路径长度较短，充分地体现与刻画了异常对象个数少且正常对象个数多的全局离群信息。在隔离森林中，利用隔离树的不同路径层，体现 S 所蕴含的不同局部离群信息；利用隔离路径长度或深度，刻画 S 所蕴含的全局离群信息。总之，采用隔离森林划分属性组离群信息 S，充分体现属性组局部与全局离群信息。

5.2 属性组隔离森林构建

在属性组离群检测中，离群数据对象个数较少，且明显不同于大多数正常数据对象。为了有效地避免绝大多数正常数据对离群数据检测的干扰，需将离群数据对象快速隔离出来，可从属性组离群信息 $S=\{S(x_1),S(x_2),\cdots,S(x_n)\}$ 中随机采样含有 ψ 个数据对象的子样本，并将其标记为 S'，其中 $S(x_i)=\{score_1(x_i),score_2(x_i),\cdots,score_c(x_i)\}$ 描述了任意数据对象 x_i 在 c 个属性组中的离群信息。随机采样可以使属性组离群程度的分散程度更加明显，减少正常数据的干扰[34]。递归地划分采样 S' 所包含的属性组离群信息，直到所有数据对象都被隔离。

设属性组离群信息为 S，T 是一棵二叉树，对于任意 $S=\{S_1,S_2,\cdots,S_c\}$ 中的子采样 $S'=\{S_1',S_2',\cdots,S_c'\}$（$c$ 为属性组的个数），采用下列步骤构造出的 T 称为属性组隔离树。

1) T 的根节点标记为 S'，$D=S'$；

2) 从 S' 随机选取第 a 属性组的离群信息 $S_a'=\{score_a(x_1),score_a(x_2),\cdots,score_a(x_\psi)\}(a\in(1,c))$，并从 S_a' 中随机选择一个数据对象 x_b 的离群程度值 $score_a(x_b)$ 作为分割依据；

3) 对于 D 中的所有数据对象，若其离群程度值大于或等于 $score_a(x_b)$，则将其数据对象划分到左子树 T_l，否则将其划分到右子树 T_r；

4) 左子树 T_l（右子树 T_r）的根节点标记为 $D=\{$划分到 $T_l(T_r)$ 的数据对象$\}$；

5) 针对 T_l 与 T_r，分别重复步骤 2)和 3)，直到树的预先深度限制或节点上仅标记一个数据对象为止。

步骤 3)和 4)中利用了第 a 属性组内的局部离群信息，被划分到左子树的数据对象离群程度高且划分到右子树的数据对象离群程度低，左子树中的数据对象成为离群数据的可能性较大。步骤 2)与 5)中，从 S' 中多次随机选取不同属性组的离群信息，迭代构造二叉树的下一层节点，充分地利用了数据对象的全部属性组离群信息，即全局离群信息。

离群数据明显偏离于大多数正常数据，离群数据对象个数较少且正常数据个数较多。由步骤 3)和 4)可知，左子树中的数据对象成为离群数据的可能性较大，因此随着二叉树节点的迭代构造，子树的左枝节点所标记的数据对象会迅速降低为一个离群数据对象，并终止构造二叉树的下一层节点(见步骤 5))，少数离群数据可以快速被隔离在位于树根较近的左子树中，大多数正常数据对象位于远离树根的右子树。从属性组离群信息 S 中多次采样 S'，构造的多个属性组隔离树被称为属性组隔离森林，充分地体现了 S 所蕴含的局部和全局离群信息。

根据上述属性组隔离树的构造步骤，从属性组离群信息 S 中采样 t 次 S'，构造出 t 个属性组隔离树，形成属性组隔离森林。构造属性组隔离森林的伪代码描述如算法 2 所示。

算法 2 AGIF

输入：数据集 DS，属性组隔离树的个数 t，子采样数据对象的个数 ψ

输出：一个包含 t 个属性组隔离树的属性组隔离森林

```
1. 由 CAVAG 算法，自动构建属性组集合 C；
2. for r in c / * c=|C| * /
3. 根据式(2)，计算第 c 个属性组的离群得分 score_c(x_i)，构建属性组
   离群信息 S_r；
4. end for
5. 构建属性值离群信息 S={S_1,S_2,…,S_c}；
6. 初始化隔离森林的预先深度限制 l=log_2 ψ；
7. for i in t
8.     S' ← 子采样(S,ψ)；
9.     if h≥l or |S'|≤1 / * 初始化 h=0 * /
10.        标记 S'为外部节点，且节点中包含数据对象的个数为|S'|；
11.    else
12.        随机选择第 a 个属性组离群信息，a ∈(1,c)；/ * c 是属性组
           个数 * /
13.        在 S_a={score_a(x_1),score_a(x_2),…,score_a(x_ψ)}中的最大值
           与最小值之间随机选择一个分割点 score_a(x_b)；
14.        for d in |S'|；
15.            if score_a(x_d) < score_a(x_b))
16.                将 S'中的数据对象 x_d划分到左子树 T_l；
17.            else
```

```
18.              将 S′中的数据对象 x_d划分到右子树 T_r；
19.          h=h+1；
20.          end if
21.       end for
22.    end if
23. end for
24. 构建由 t 个属性组隔离树构成的属性组隔离森林
25. END AGIF
```

5.3 隔离森林集成策略

由 5.2 节可知，依据属性组离群信息 S，属性组隔离森林将数据对象划分到不同节点，并快速在左枝节点隔离出离群数据，充分体现 S 所蕴含的属性组局部与全局离群信息。在属性组隔离森林中，与其他数据对象离群信息差异较大的离群数据对象可被快速地隔离到距离树根较近的节点，即距离树根越近的数据对象的离群信息越重要。

定义 1 对于任意一棵属性组隔离树 T，从 T 的根节点到包含数据对象 x_i 的叶子节点所遍历边的条数称为 x_i 的全局离群信息路径长度，记为 $h(x_i)$。

在属性组隔离树中，全局离群信息路径 $h(x_i)$ 刻画了数据对象 x_i 的离群程度与离群信息 S 的关联关系，蕴含着 x_i 的离群信息在全部属性组中的分布。$h(x_i)$ 越小，表明 x_i 所包含的离群信息越重要，反之亦然。参照文献[34]，在属性组隔离森林中，数据对象 x_i 的离群信息权重定义为：

$$W_D(x_i)=2^{-\frac{E(h(x_i))}{c(t)}} \tag{10}$$

其中，t 为隔离树棵数；$c(t)$ 为 t 个隔离树的平均路径长度，用于标准化数据对象 x_i 的路径长度 $h(x_i)$；$E(h(x_i))$ 为 t 个 $h(x_i)$ 的均值。

在式(10)中，当 $E(h(x_i))\to 0$ 时，$W_D(x_i)\to 1$，即当数据的 $W_D(x_i)$ 接近 1 时，数据对象 x_i 的离群信息权重越大；当 $E(h(x_i))\to t-1$ 时，$W_D(x_i)\to 0$，即当 $W_D(x_i)\ll 0.5$ 时，x_i 的离群信息权重越小。$W_D(x_i)$ 体现了在属性组隔离森林中任意数据对象 x_i 被快速隔离的平均路径长度，平均路径越短，表明 x_i 所包含的离群信息越重要。

由式(10)可知，若两个数据对象在隔离森林中被快速隔离的平均路径长度相同，则其离群信息权重也相同。然而，根据属性组隔离树步骤可知，左枝节点所标记的数据对象成为离群数据的可能性大，即依据属性组内的局部离群信息，离群程度越大的越可能为离群数据。

定义 2 对于任意一棵属性组隔离树 T，从 T 的根节点到包含数据对象 x_i 的叶子节点所遍历的包含 x_i 的节点，称为 x_i 的局部离群信息节点集，记为 $N(x_i)$。

在属性组隔离树 T 中，标记在叶子节点中的任意数据对象 x_i 对应一个属性组内的离群程度值 $score_r(x_i)$，$N(x_i)$ 为包含 x_i 的全部节点，将集合 $N(x_i)$ 中 x_i 对应的所有 $score_r(x_i)$ 的平均值记为 $s_h(x_i)$。在属性组隔离树中，x_i 的局部离群信息 $s_h(x_i)$ 越大，则 x_i 在不同属性组内的离群程度越高，反之亦然。参照文献[33]，数据对象 x_i 在属性组隔离森林中的离群程度定义如下：

$$S_N(x_i)=2^{E(s_h(x_i))} \tag{11}$$

其中，$E(s_h(x_i))$ 为 t 个属性组隔离树中的 $s_h(x_i)$ 平均值。

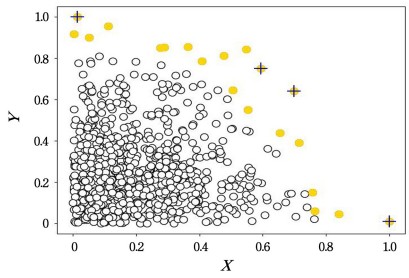
在式(11)中，当 $s_h(x_i)\to 0$，即 $E(s_h(x_i))\to 0$ 时，由式(2)可知，x_i 的局部离群程度越来越高，即当 $E(s_h(x_i))\to 0$ 时，$S_N(x_i)\to 1$，表明 x_i 的离群程度越来越高；当 $s_h(x_i)\to -\infty$，即 $E(s_h(x_i))\to -\infty$ 时，$S_N(x_i)\to 0$，x_i 的离群程度越来越低。总之，$S_N(x_i)$ 体现了 x_i 在各属性组内的离群程度。

综上，离群程度 $S_N(x_i)$ 是依据离群信息 $S(x_i)$ 刻画 x_i 的离群程度，而离群信息权重 $W_D(x_i)$ 是刻画离群信息 $S(x_i)$ 的重要程度。总之，$W_D(x_i)$ 权重越大，离群程度 $S_N(x_i)$ 越重要，反之亦然。结合式(10)和式(11)，在属性组离群信息 S 中，任意数据对象 x_i 的离群程度可重新定义为：

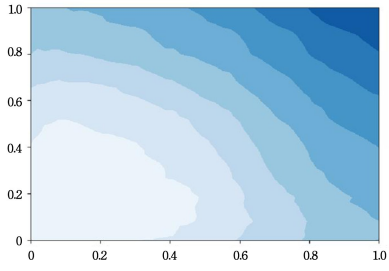
$$\begin{aligned} Score(x_i)&=W_D(x_i)\times S_N(x_i)=2^{-\frac{E(h(x_i))}{c(t)}}\times 2^{E(s_h(x_i))}\\ &=2^{-\frac{E(h(x_i))}{c(t)}+E(s(x_i))} \end{aligned} \tag{12}$$

其中，$W_D(x_i)$ 描述了在属性组离群信息 S 中，x_i 与其他数据对象之间的离群信息差异性，体现了 x_i 的离群信息重要程度；而 $S_N(x_i)$ 描述了在各属性组内的离群信息 S_r 中，x_i 与其他数据对象之间的离群信息异常程度，体现了 x_i 的局部离群信息。因此，$Score(x_i)$ 有效地刻画了 x_i 的离群程度。当 $Score(x_i)\to 1$ 时，x_i 的离群程度越来越高；当 $Score(x_i)\to 0$ 时，x_i 的离群程度越来越低。总之，由式(12)得到的 $Score(x_i)$ 不仅包含了 x_i 在各属性组内的局部离群信息，而且蕴含了 x_i 在所有属性组的全局离群信息。将此集成方法被称为属性组隔离森林(AGIF)集成策略。

例 3 利用图 1 生成的数据，采用 AGIF 集成策略的热图与其离群检测结果示意图如图 3 所示。图 3(a)中，圆圈代表正常数据，黄色实型圆圈代表已检测出的离群数据，标记“+”号的实型圆圈代表图 1(a)与图 2(a)中由于集成偏差未检出但由图 3(a)检测出的离群数据。



(a)AGIF 集成策略的离群检测结果



(b)离群得分的热图

图 3 隔离森林集成策略(电子版为彩图)

Fig. 3 AGIF ensemble strategy

在图 3(b)中，当数据对象在各属性组中的离群程度越大且越不同于其他数据对象时，集成离群程度越大，因而图3(a)中黄色实型圆圈(已检测出的离群数据)不偏向坐标轴，也不

偏离坐标轴，充分利用了数据对象的属性组离群信息，并有效避免了最大值与平均值集成策略造成的属性组集成偏差，从而可以检测出更有效的离群数据。

综上所述，现有的离群信息集成大多采用平均值[12]和最大值[11]离群信息集成策略，仅体现了各属性组内数据对象的局部离群信息，忽略了属性组全局离群信息，可能导致一些数据对象的离群程度偏高或偏低；AGIF 集成策略不仅体现了属性组内的局部离群信息，而且蕴含了属性组全局离群信息，降低了属性组离群信息集成偏差。

5.4 属性组离群检测算法

离群检测是识别或检测与多数常规对象具有明显差异的数据对象。依据前面章节描述，构建降低集成偏差的属性组离群检测方法的基本步骤为：首先，将数据 DS 分为 c 个属性组，并在各属性组中通过离群得分描述数据对象的离群信息；其次，根据所有属性组离群信息 S，构建属性组隔离森林；最后，在属性组隔离森林中，利用局部离群信息 $S_N(x_i)$ 与全局离群信息 $W_D(x_i)$ 度量每个数据对象的集成离群程度 $S(x_i)$，并选取若干离群程度最大的数据对象作为离群数据。其伪代码描述如算法 3 所示。

算法 3 AGEOD

输入：数据集 DS

输出：离群数据对象集合 OS

1. 由 CAVAG 算法自动构建属性组集合 C；
2. 根据 AGIF 算法获得包含 t 棵隔离树的属性组隔离森林；
3. for x_i in DS
4. for j in t
5. 根据对象 x_i 的属性组离群信息 $S(x_i)=\{score_1(x_i), score_2(x_i), \cdots, score_c(x_i)\}$，在 Forest 的第 j 棵树中隔离出 x_i；
6. 在第 j 棵隔离树中，根据数据对象 x_i 所在的叶子节点获得 x_i 的全局离群信息路径 $h(x_i)$；
7. 在第 j 棵隔离树中，从包含 x_i 的所有节点获得 x_i 所对应 $score_r(x_i)$ 的平均值，记为 $s_h(x_i)$；
8. end for
9. 根据式(10)计算数据对象 x_i 的离群信息权重 $W_D(x_i)$；
10. 根据式(11)计算数据对象 x_i 的集成离群程度 $S_N(x_i)$；
11. 根据式(12)更新数据对象 x_i 的集成离群程度 $Score(x_i)$；
12. end for
13. 选取 k 个离群程度 $Score(x_i)$ 最大的数据对象，构建离群数据对象集合 OS.
14. END AGEOD

设 n 为数据对象个数，m 为属性个数，c 为属性组个数，p 为各属性组中属性个数，t 为隔离树的棵数，ψ 为每棵隔离树中子采样的个数。AGEOD 算法包括 4 个部分：属性分组，度量各属性组中数据对象的离群程度，隔离森林构建，属性组离群信息集成。属性分组的时间复杂度为 $O(m^2\log m^2)+O(m^2c)$；度量各属性组中数据对象离群程度的时间复杂度为 $O(ncp)$；由于构建隔离森林具有随机性，其最高时间复杂度与平均时间复杂度是相同的，均为 $O(\psi t\log\psi)$，最低时间复杂度为 $O(\psi t)$；属性组离群信息集成的时间复杂度为 $O(nt\log\psi)$。综上可得，AGEOD 算法的最高复杂度、平均复杂度均为 $O(m^2\log m^2)+O(m^2c)+O(ncp)+O(\psi t\log\psi)+O(nt\log\psi)$，最低复杂度为 $O(m^2\log m^2)+O(m^2c)+O(ncp)+O(\psi t\log\psi)+O(\psi t)+O(nt\log\psi)$。

6 实验分析

6.1 实验环境与数据

实验环境为 Intel(R) Core(TM) i7-10750H，16 GB 内存，Windows 10 操作系统，采用 python 语言实现 AGEOD 算法以及对比算法。对比算法主要包括：WATCH[11]属性分组离群检测算法，ACA[25]和 GBFG[26]属性分组算法，最大值[11]与平均值[12]离群信息集成策略，SAnDCat[21]，CBRW[22]，SDRW[23]，SCAN[24]，OptIForest[26]分类型数据离群检测算法。

在实验验证隔离森林(AGIF)集成策略的对比实验中，由于 ACA 和 GBFG 仅为属性分组算法，因此将 AGEOD 中的属性分组算法 CAVAG 替换为 ACA 和 GBFG，分别记为 ACA_AGIF 和 GBFG_AGIF；再将其集成策略替换为最大值和平均值集成策略，分别记为 ACA_MAX，ACA_MEAN，GBFG_MAX，GBFG_MEAN。WATCH 为基于最大值集成策略的属性分组离群检测算法；将 WATCH 中的最大值集成策略替换为平均值和 AGIF 集成策略，分别标记为 WATCH_MEAN 和 WATCH_AGIF。AGEOD 为基于 AGIF 集成策略的属性分组离群检测算法；将其 AGIF 集成策略替换为最大值和平均值集成策略，分别标记为 AGEOD_MAX 和 AGEOD_MEAN。

在实验验证 AGEOD 算法的对比实验中，参照文献[20]，AGEOD 中一阶、二阶相似属性向量取值空间 d 均设置为 64；参照文献[35]，属性组隔离森林数通常在 $t=100$ 之前收敛，采用 $t=100$ 作为默认值，子采样个数 $\psi=256$ 时，其检测性能达到最优；参照文献[25]和文献[26]，在 ACA+和 GBFG+算法中，属性分组个数 c 设置为 $m/4$，其中 m 为数据集中属性的个数；参照文献[24]，SCAN 中离群值评分参数设置为 0.15，属性值向量长度设置为 128；参照文献[24]，POtIForest 中切割阈值 $\epsilon=512$。本文使用 UCI 机器学习库的 8 个分类型数据集[1)]，具体如表 2 所列。

表 2 分类型数据集

Table 2 Categorical data sets

数据集	类型	数据对象个数	属性个数	属性取值总数	离群数据对象个数
NIPS	UCI	10884	410	9598	140
USC	UCI	11742	67	389	284
Aps	UCI	11017	149	472850	100
Covtpye	UCI	120048	53	8846	311
Musk	UCI	3062	166	40382	97
Optdigits	UCI	5216	63	913	50
Ad	UCI	3278	1555	188	20
Connect	UCI	10302	41	109	2403

6.2 隔离森林集成策略

在属性分组离群检测中，各属性组中不同数据对象所包

1) http://archive.ics.uci.edu/ml/datasets.php

含的离群信息存在着较大差异，因此，在属性组离群信息集成过程中，需要客观地体现各属性组的离群信息，从而有效地给出数据对象的离群程度。为了验证评估隔离森林（AGIF）集成策略的集成效果与离群检测准确性，采用 GBFG[26]，WATCH[11]，ACA[25]，AGEOD 这 4 种不同的属性分组，将属性组集成偏差系数 EDC（见式（9））和离群检测准确性 AUC 作为评估指标，在表 2 所列的数据集上评估平均值集成策略、最大值集成策略、AGIF 这 3 种不同集成策略的集成效果和离群检测准确性，实验结果分别如表 3 和表 4 所列。

表 3 属性组集成偏差系数结果

Table 3 Results of EDC

数据集	NIPS	USC	Aps	Covtpye	Musk	Optdigits	Ad	Connect
GBFG_MEAN	6.2115	0.8051	1.0764	0.4082	1.0443	0.8245	0.3171	1.6233
GBFG_MAX	5.6073	0.8853	0.7285	0.4852	0.6253	0.9951	0.3616	0.6441
GBFG_AGIF	**8.7861**	**0.9736**	**1.0778**	**0.5811**	**1.0594**	**1.0642**	**0.5534**	**1.9595**
WATCH_MEAN	5.7686	1.2514	0.8610	0.4150	1.1667	0.9633	—	2.2991
WATCH	5.0344	0.5946	0.7071	0.3162	0.7081	0.8653	—	0.9128
WATCH_AGIF	**5.8094**	**1.3313**	**1.0691**	**0.4865**	**1.3230**	**0.9987**	—	**3.6382**
ACA_MEAN	6.0326	0.8470	1.0730	0.5374	1.1538	0.8039	0.3644	1.7283
ACA_MAX	5.7678	0.6469	0.9802	0.5080	0.7290	0.8463	0.3535	0.8528
ACA_AGIF	**6.0991**	**0.9235**	**1.0917**	**0.5765**	**1.2094**	**0.8638**	**0.3790**	**2.0315**
AGEOD_MEAN	6.1127	1.7092	1.2756	0.5201	1.1273	1.2793	0.3694	1.4233
AGEOD_MAX	4.7405	0.4287	0.4423	0.4143	0.9803	0.4745	0.3618	1.0096
AGEOD	**6.1626**	**1.8229**	**1.3000**	**0.5872**	**1.1927**	**1.3054**	**0.3722**	**1.5029**

表 4 集成策略准确性结果

Table 4 Results of ensemble strategy accuracy

数据集	NIPS	USC	Aps	Covtpye	Musk	Optdigits	Ad	Connect
GBFG_MEAN	0.9144	0.9678	0.8762	0.9253	0.9817	0.8294	0.5932	0.5278
GBFG_MAX	0.8330	0.9319	0.7120	0.6709	0.6919	0.5924	0.4271	0.5102
GBFG_AGIF	**0.9155**	**0.9732**	**0.8765**	**0.9321**	**0.9827**	**0.8639**	**0.7058**	**0.5420**
WATCH_MEAN	0.9107	0.7876	0.8488	0.9574	0.9610	0.7720	—	0.5319
WATCH	0.8044	0.8843	0.7068	0.9164	0.9484	0.7671	—	0.5242
WATCH_AGIF	**0.9148**	**0.8893**	**0.8651**	**0.9602**	**0.9622**	**0.8186**	—	**0.5488**
ACA_MEAN	0.9134	0.9375	0.8557	0.8787	0.9641	0.8137	0.5210	0.5167
ACA_MAX	0.7395	0.8393	0.6709	0.8250	0.7244	0.6411	0.4634	0.5111
ACA_AGIF	**0.9142**	**0.9469**	**0.8649**	**0.8836**	**0.9793**	**0.8146**	**0.6267**	**0.5218**
AGEOD_MEAN	0.9174	0.9637	0.8784	0.9688	0.9832	0.8970	0.7306	0.5515
AGEOD_MAX	0.8455	0.9324	0.8186	0.6680	0.8375	0.7401	0.6972	0.5253
AGEOD	**0.9180**	**0.9775**	**0.8871**	**0.9733**	**0.9854**	**0.9083**	**0.7389**	**0.5581**

由表 3 可知，针对 4 种不同的属性分组，AGIF 集成策略的属性组集成偏差系数值均为最大，表明 AGIF 集成策略可有效减少集成偏差，有效体现了属性组中的局部与全局离群信息。主要原因是：在各属性组离群信息 S 中，离群数据对象明显偏离于正常数据对象且离群程度较大，AGIF 集成策略同时利用了属性组局部与全局离群信息，减小了属性组集成偏差，而平均值与最大值集成策略仅利用属性组内的局部离群信息，属性组集成偏差较大。

由表 4 可知，无论采用哪种属性分组，AGIF 集成策略都可以将离群检测 AUC 值平均提高 9.20%，表明 AGIF 集成策略优于平均值与最大值集成策略。主要原因如下：

1）AGIF 集成策略同时刻画了在属性组离群信息 S 中离群数据对象明显偏离于正常数据对象且离群程度较大的局部与全局离群信息，而平均值与最大值仅考虑了在各属性组内离群程度较大的局部离群信息，丢失了全局离群信息。

2）在 AGEOD 算法中，属性分组采用了属性相关性向量刻画属性的全局与局部相似性，同时考虑了属性组内相关性大且组间相关性小；而 GBFG，WATCH 和 ACA 仅通过互信息刻画属性之间的局部相关性，且 GBFG 仅考虑了组间相关性大，WATCH 与 ACA 仅考虑了组内相关性小。

6.3 离群检测性能

为了评估离群检测算法 AGEOD 的准确性和效率，将 OptIForest，SDRW，WATCH，SCAN，SAnDCat，CBRW 作为对比算法，以准确性 AUC 与效率（运行时间）作为评估指标进行实验，结果如表 5 和表 6 所列。

由表 5 可知，AGEOD 算法的离群检测准确性优于所有对比算法，AUC 指标平均提高了 7.83%。主要原因在于：

1）OptIForest，SDRW，SCAN，SAnDCat 和 CBRW 均是全维属性离群检测算法，而 AGEOD 与 WATCH 是属性分组离群检测算法，有效缓解了“维灾”干扰。

2）WATCH 仅体现属性组内的相关性，并采用了最大值集成策略，其集成偏差较大；而 AGEOD 同时体现了属性组内与组间的相关性，并采用了属性组隔离森林集成策略，降低了集成偏差。

由表 6 可知，AGEOD 算法除了在 Aps 和 Ad 数据集上略

低于 OptIForest 算法，在 Connect，Ad，Covtype，USC 4 个数据集上略低于 CBRW 算法，在 Connect，Ad，Covtype 3 个数据集上略低于 SDRW 算法外，其离群检测效率均高于其他对比算法，且离群检测效率平均提高了 48.43%，尤其是对于属性取值较多的分类型数据集，其效率提升较为明显。主要原因如下：

1）OptIForest，SDRW，SCAN，SAnDCat 和 CBRW 均是所有属性维度的离群检测算法，AGEOD 和 WATCH 是属性分组离群检测算法。属性分组可有效降低离群检测过程中的搜索维度。

2）尽管 AGEOD 与 WATCH 都属于属性分组离群检测算法，但 WATCH 利用原始数据属性之间的相似性实现离群检测，而 AGEOD 利用属性相关性向量实现属性自动分组，且属性相关性向量之间的运算量远低于原始数据属性之间的相关性运算量。

3）OpiIForest 依据层次聚类形成一棵隔离树，该算法最初将每个数据对象视为一个簇，并将其依次合并，直到剩下单个簇，以自下而上的方式形成一棵层次树，计算复杂度受数据对象个数 n 的影响较大，因而 n 最大的数据集 Covtpye 运行时间最长。AGEOD 在各属性组中计算数据对象的离群得分（见式(2)）时，由于需计算每个属性取值的频率，因此相对于 OpiIForest，其计算复杂度受属性取值总个数影响较大，因而在数据集 Covtpye 上运行时间较长。

4）SDRW，SCAN 和 CBRW 算法利用了属性间耦合关系来度量数据对象的离群程度，在属性取值较少时，CBRW 和 SDRW 可直接构造属性值耦合，并度量数据对象的离群程度，而 AGEOD 通过属性分组、属性组内离群检测与离群信息集成策略，确定数据对象的离群程度。因此，当数据集（USC，Ad，Connect）的属性维度较低且属性取值较少时，SDRW 和 CBRW 算法效率略高。

5）当数据集的属性维度增高且属性取值较多时，CBRW，SDRW 和 SCAN 构造属性值间耦合的运算量和占用内存也增加，尤其是对于属性取值较多的数据集（Musk，Aps，NIPS），SDRW 和 SCAN 因属性值间耦合运算量较大或内存不足而无法获得运行结果。

表 5 离群检测准确率结果

Table 5 Results of outlier detection AUC

数据集	OptIForest	SDRW	WATCH	SCAN	SAnDCat	CBRW	AGEOD
NIPS	0.9093	0.6798	0.8044	—	0.9168	0.9162	**0.9180**
USC	0.9363	0.9726	0.8843	0.9565	0.9311	0.9462	**0.9775**
Aps	0.8790	—	0.7068	—	0.8746	0.8497	**0.8871**
Covtpye	0.9449	0.5525	0.9164	0.5494	0.9677	0.5263	**0.9733**
Musk	0.9373	—	0.9484	—	0.9805	0.9667	**0.9810**
Optdigits	0.7520	0.7089	0.7671	0.6625	0.8631	0.8907	**0.9083**
Ad	0.7246	—	—	0.7262	0.7196	0.6931	**0.7389**
Connect	0.5522	0.5490	0.5242	0.5342	0.5399	0.5432	**0.5581**

表 6 运行时间结果

Table 6 Results of efficiency time

(s)

数据集	OptIForest	SDRW	WATCH	SCAN	SAnDCat	CBRW	AGEOD
NIPS	733.71	4421.86	3753.58	—	1263.23	328.65	**257.09**
USC	20.82	14.78	141.81	56.08	44.27	**14.06**	24.01
Aps	**415.51**	—	2308.24	—	1102.46	1963.09	808.07
Covtpye	1586.78	2575.71	839.07	3474.03	875.48	**91.50**	219.89
Musk	145.87	—	173.79	—	123.36	168.38	**74.78**
Optdigits	10.65	15.39	54.60	100.92	18.96	**6.11**	10.07
Ad	**961.84**	—	—	5310.29	5314.13	1799.70	2648.03
Connect	19.90	**3.47**	24.04	18.92	15.42	3.48	14.33

6.4 统计检验

为了评价属性组隔离森林（AGIF）集成策略在 AUC 上的优越性，以及属性组离群检测算法 AGEOD 在 AUC 与效率上的优越性，采用了 SPSS 统计中的 Friedman 统计检验[11]。Friedman 统计检验是比较多个样本大小关系的非参数统计检验。

根据表 4 中的实验数据，集成策略 AUC 的统计检验结果如表 7 所列。其中，测试中的显著性水平 $\alpha=0.05$，$chi\text{-}square=70.484689$，$p=9.88055\times10^{-11}$。依据表 7，AUC 统计数据的 Friedman 检验图如图 4 所示。Friedman 检验图中，若两个算法的横线段没有交叠，则说明算法之间有显著性差异，反之亦然。

表 7 集成策略准确率的统计检验

Table 7 Statistical test of ensemble strategy accuracy

算法	平均值	标准差	最大值	最小值	排名
GBFG_MEAN	0.8269	0.1611	0.9817	0.5278	8.25
GBFG_MAX	0.6711	0.1532	0.9319	0.4271	2.12
GBFG_AGIF	0.8489	0.1415	0.9827	0.5420	**9.75**
WATCH_MEAN	0.8242	0.1383	0.9610	0.5319	5.37
WATCH	0.7930	0.1353	0.9484	0.5242	3.75
WATCH_AGIF	0.8512	0.1322	0.9622	0.5488	**7.62**
ACA_MEAN	0.8001	0.1681	0.9641	0.5167	5.25
ACA_MAX	0.6768	0.1269	0.8512	0.4634	1.87
ACA_AGIF	0.8190	0.1511	0.9793	0.5218	**6.75**
AGEOD_MEAN	0.8613	0.1390	0.9832	0.5515	10.75
AGEOD_MAX	0.7580	0.1195	0.9324	0.5253	4.50
AGEOD	0.8684	0.1388	0.9854	0.5581	**12.00**

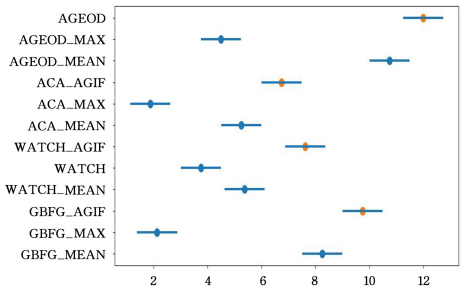


图 4 集成策略准确率的 Friedman 检验图

Fig. 4 Friedman test of ensemble strategy accuracy

由图 4 可知，当分别采用 GBFG，WATCH，ACA 和 CAVAG 这 4 种不同属性分组方法验证平均值集成策略、最大值集成策略和隔离森林（AGIF）集成策略的效果时，AGIF 效果均最优，且除采用 ACA 属性分组方法时 AGIF 与平均值集成策略没有明显区别外，在其余 3 种属性分组方法中，AGIF 集成策略均与平均值、最大值集成策略有明显区别，表明了 AGIF 集成策略在 AUC 上具有明显的优势。

根据表 5 中的实验数据，离群检测算法的 AUC 的 Friedman 统计检验结果如表 8 所列，测试中的显著性水平 $\alpha=0.05$，$chi\text{-}square=19.50$，$p=0.003$；依据表 8，图 5 给出了 AUC 统计数据的 Friedman 检验结果。

表 8 离群检测准确率的统计检验

Table 8 Statistical test of outlier detection AUC

统计数据	OptIForest	SDRW	WATCH	SCAN	SAnDCat	CBRW	AGEOD
平均值	0.8294	0.6925	0.7930	0.6857	0.8491	0.7915	0.8677
标准差	0.1318	0.1543	0.1353	0.1529	0.1395	0.1677	0.1383
最大值	0.9442	0.9726	0.9484	0.9565	0.9805	0.9667	0.9810
最小值	0.5522	0.5490	0.5242	0.5342	0.5399	0.5263	0.5581
排名	4.25	2.75	3.37	2.37	4.50	3.87	**6.87**

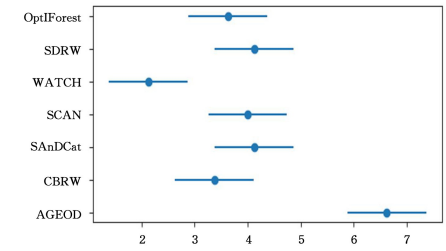


图 5 离群检测的 AUC 的 Friedman 检验

Fig. 5 Friedman test of AUC of outlier detection

由图 5 可知，AGEOD 与 5 种比较算法的横线段没有交叠区域。在 AUC 方面，AGEOD 算法与 OPtIForest，SDRW，WATCH，SCAN，SAnDCat，CBRW 都存在着显著性差异，且 AGEOD 的 AUC 最大，表明了 AGCOD 算法在 AUC 上具有明显的优势。

根据表 6 中实验数据，表 9 列出了离群检测算法的 Time 的 Friedman 统计检验结果，测试中的显著性水平 $\alpha=0.05$，$chi\text{-}square=24.80$，$p=0.0003$。依据表 9，图 6 给出了运行时间统计数据的 Friedman 检验结果。

表 9 离群检测运行时间的统计检验

Table 9 Statistical test of outlier detection time

(s)

统计数据	OptIForest	SDRW	WATCH	SCAN	SAnDCat	CBRW	AGEOD
平均值	486.88	2537.09	1381.08	2422.79	1094.66	546.87	507.03
标准差	533.82	2041.78	1541.32	1921.23	1667.67	778.22	846.26
最大值	1586.78	4421.86	3753.58	5310.29	5314.13	1963.09	2648.03
最小值	10.65	3.47	24.04	18.92	15.42	3.48	10.07
排名	3.12	4.87	5.37	6.0	4.25	2.12	2.25

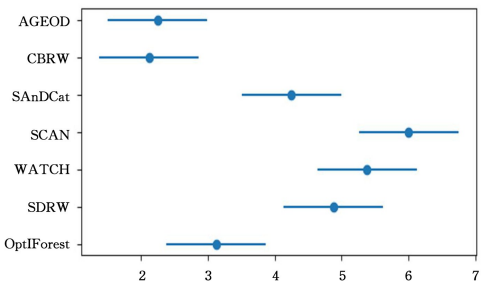


图 6 离群检测运行时间的 Friedman 检验

Fig. 6 Friedman test of outlier detection time

由图 6 可知，AGEOD 仅与 CBRW 和 OPtIForest 交叠区域较大，且排名略低于 CBRW，与其余 4 个对比算法的横线段没有明显交叠区域。在 Time 方面，AGEOD 算法与 SDRW，WATCH，SCAN，SAnDCat 都存在着显著差异，表明 AGCOD 算法在运行时间上具有优势。

结束语 本文利用属性组局部与全局离群信息，提出了一种基于隔离森林集成策略的分类型属性分组离群检测方法。隔离森林集成策略充分利用了离群数据对象明显偏离于正常数据对象的全局离群信息，以及离群数据对象在各属性组内离群程度较高的局部离群信息，并有效降低了属性组离群检测集成偏差。实验结果表明隔离森林集成策略不仅能降低集成偏差，而且离群检测的 AUC 指标平均提升了 9.20%；利用隔离森林集成策略，分类型属性分组离群检测算法 AGEOD 的 AUC 指标和效率分别平均提高了 7.83% 和 48.43%。下一步的研究工作为混合属性数据的属性分组离群检测。

参 考 文 献

[1] ZHANG J F, LI Y H, QIN X, et al. Related-Subspace- Based Local Outlier Detection Algorithm Using MapReduce[J]. Ruan Jian Xue Bao/Journal of Software, 2015, 26(5): 1079-1095.

[2] MAXIMILIAN T, BERNHARD C G, ROMAN K. CLUSTER Purging: Efficient Outlier Detection Based on Rate-Distortion Theory[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(2): 1270-1282.

[3] SAJANRAJ T, JASISON P M, RAGHAVENDRA S. Operational pattern forecast improvement with outlier detection in metro rail transport system[J]. Multimedia Tools and Applications, 2024, 83(4): 11229-11245.

[4] FANG J Z, WANG Z D, LIU W B, et al. A New Particle Swarm Optimization Algorithm for Outlier Detection: Industrial Data Clustering in Wire Arc Additive Manufacturing [J]. IEEE Transactions on Automation Science and Engineering, 2024, 21(2): 1244-1257.

[5] HUANG J Z, ZHAO Y, MENG B, et al. SEAOP: a statistical ensemble approach for outlier detection in quantitative proteomics data [J]. Briefings in Bioinformatics, 2024, 25 (3): bbae129.

[6] SINA D, ZEINAB T, NEGIN D. An outlier detection method based on the hidden Markov model and copula for wireless sensor networks[J]. Wireless Networks, 2024, 30(6): 4797-4810.

[7] HOSSEIN M, MOHAMMAD J, HAMID R D, et al. RODEO: Robust Outlier Detection via Exposing Adaptive Out-of-Distribution Samples [C] // Forty-first International Conference on Machine Learning. 2024: 21-27.

[8] ANTONELLA M, DAVID M, MANUELE B. Detecting outliers from pairwise proximities: Proximity isolation forests[J]. Pattern Recognition, 2023, 138, 109334.

[9] MAXIMILIAN T, BERNHARD C G, ROMAN K. Cluster Purging: Efficient Outlier Detection Based on Rate-Distortion Theory[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, l. 35(2): 1270-1282.

[10] PANG G S, XU H Z, GAO L B, et al. Selective Value Coupling Learning for Detecting Outliers in High Dimensional Categorical Data[C] // International Conference on Information and Knowledge Management. 2023: 807-816.

[11] LI J L, ZHANG J F, PANG N. Weighted Outlier Detection of High-Dimensional Categorical Data Using Feature Grouping [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(11): 4295-4308.

[12] AKANKSHA M, RAJEEV K. Combination fairness with scores in outlier detection ensembles[J]. Information Sciences, 2023, 645: 119337.

[13] AGGARWAL C C. Outlier ensembles: position paper[J]. SIGKDD Explorations, 2013, 14(2): 49-58.

[14] AGGARWAL C C, SATHE S. Theoretical foundations algorithms for outlier ensembles[J]. SIGKDD Explorations, 2015, 17(1): 24-47.

[15] ZIMEK A, CAMPELLO R, SANDER J. Ensembles for unsupervised outlier detection: challenges and research questions a position paper[J]. SIGKDD Explorations, 2013, 15(1): 11-22.

[16] HOU S Y, JIANG G X, WANG W J. A Label Noise Filtering Method Based on Relative Outlier Factor[J]. ACTA AUTOMATICA SINICA, 2024, 50(1): 1-15.

[17] CAI S H, HUANG R B, CHEN J F. An effificient outlier detection method for data streams based onclosed frequent patterns by considering antimonotonic constraints [J]. Information Sciences, 2021, 555: 125-146.

[18] JAVIER M, MARA C R, BERTRAND N. A review of recent approaches on wrapper feature selection for intrusion detection [J]. Expert Systems with Applications, 2022, 198: 116822.

[19] LIU C, PENG D Z, CHEN H M, et al. Attribute granules-based object entropy for outlier detection in nominal data[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108198.

[20] TANG J, QU M, WANG M Z. LINE: Large-scale Information Network Embedding[C] // Proceedings of the 24th International Conference on World Wide Web. 2015: 18-22.

[21] DINO I, RUGGERO G, ROSA M. A Semisupervised Approach to the Detection and Characterization of Outliers in Categorical Data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(5): 1017-1029.

[22] PANG G S, GAO L, CHEN L. Outlier Detection in Complex Categorical Data by Modeling the Feature Value Couplings [C] // International Joint Conference on Artificial Intelligence. 2016: 1902-1908.

[23] PANG G S, GAO L, CHEN L. Homophily outlier detection in non-IID categorical data[J]. Data Mining and Knowledge Discovery, 2021, 35(4): 1163-1224.

[24] XU H Z, WANG Y J, WU Z Y, et al. Embedding-Based Complex Feature Value Coupling Learning for Detecting Outliers in Non-IID Categorical Data[C] // AAAI Conference on Artificial Intelligence. 2019: 5541-5548.

[25] ZHANG X Y, DOU W H, HE Q, et al. Lshiforest: A generic framework for fast tree isolation based ensemble anomaly analysis[J]. IEEE International Conference on Data Engineering. 2017: 983-994.

[26] XIANG H L, ZHANG X Y, HU H S, et al. OptIForest: Optimal Isolation Forest for Anomaly Detection[C] // International Joint Conference on Artificial Intelligence. 2023: 2379-2387.

[27] AU W B, KEITH C C, ANDREW W, et al. AttributeClustering for Grouping, Selection, and Classification of Gene Expression Data[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4(1): 157.

[28] ZHENG L, CHAO F, PARTHALÁIN N M, et al. Feature grouping and selection: A graph-based approach[J]. Information Sciences, 2021, 546: 1256-1272.

[29] TANG X C, DAI Y W, SUN P, et al. Interaction-based feature selection using Factorial Design [J]. Neurocomputing, 2018, 281: 47-54.

[30] AKANKSHA M，RAJEEV K. Building outlier detection ensembles by selective parameterization of heterogeneous methods[J]. Pattern Recognition Letters，2021，146：126-133.
[31] LIU H Y，MA F D，HE S B，et al. Fairness-aware outlier ensemble[J]. arXiv：2103.09419，2021.
[32] CHEN X J，YE Y M，XU X F，et al. A feature group weighting method for subspace clustering of high-dimensional data[J]. Pattern Recognition，2012，45(1)：434-446.
[33] FENG Y，ZHAO S Y，ZHANG Y Z，et al. Noise-Tolerant Learning with Silhouette Coefficient for Unsupervised Person ReIdentification[C]// IEEE International Conference on Multimedia and Expo. 2022：1-6.
[34] SAHAND H，MATIAS C K，ROBERT J B. Extended Isolation Forest[J]. IEEE Transactions on Knowledge and Data Engineering，2021，33(4)：1479-1489.
[35] LIU F，TING K，ZHOU Z H. Isolation forest[C]// IEEE International Conference on Data Mining. 2008：413-422.

SONG Yijing，born in 1992，Ph.D candidate，is a member of CCF（No. I9481G）. Her main research interest is data mining.

ZHANG Jifu，born in 1963，Ph.D，professor，Ph.D supervisor，is a member of CCF（No. 05740D）. His main research interests include big data analysis and parallel computing.

（责任编辑：何杨）

又一所 985 院校成立学生分会

2025 年 12 月 11 日，CCF 大连理工大学学生分会成立大会在大连理工大学学生文化中心举行，这是 CCF 在全国高校中成立的第 90 个学生分会。

本次成立大会由大连理工大学计算机科学与技术学院团委书记张世凯主持。出席大会的嘉宾及人员包括：CCF 学生分会工作组组长、中山大学王昌栋教授（线上参会），大连理工大学计算机科学与技术学院党委书记葛宏伟教授，CCF 大连会员活动中心（以下简称 CCF 大连）主席、大连理工大学齐恒教授，CCF 大连监督委员会主席、大连海事大学王新年教授，大连理工大学计算机科学与技术学院张亮副教授以及 CCF 大连理工大学学生会员。

会上，王昌栋教授代表 CCF 学生分会工作组通过线上方式致辞。他衷心祝贺 CCF 大连理工大学学生分会的成立，并详细介绍了 CCF 学生分会的相关情况、组织的学术活动以及会员的权益等内容。

葛宏伟书记代表大连理工大学计算机科学与技术学院致辞，对参会嘉宾和会员的到来表示热烈欢迎，并对各位专家对学校工作的支持表示感谢。他鼓励学生会员积极参与分会活动，强化理论与实践的深度融合，借助 CCF 这个平台成长为计算机专业人才。

王新年教授代表 CCF 大连致辞，首先对大连理工大学学生分会的成立表示祝贺，对大连理工大学师生对 CCF 工作的支持表示感谢，随后对 CCF 文化进行介绍并希望学生会员们能够充分利用好 CCF 学生分会这一平台，不断提升自身的创新能力和组织能力。

在大会的竞选环节，通过候选人演讲、现场师生提问、无记名投票等一系列流程，差额选举产生了 CCF 大连理工大学学生分会首届执行委员会（名单附后）。之后，王新年教授宣读了首届执行委员会名单，并为齐恒教授颁发督导主任聘书，同时为首届执行委员会授旗。执行委员会主席钟剑辉现场宣读了就职誓词。

2025 年 12 月 13 日，一场主题鲜明、内容丰富的成立交流会在海创（大连）科技交流中心成功举办。本次交流会由 CCF 大连理工大学学生分会联合 CCF 大连会员活动中心共同策划组织，旨在搭建区域内计算机领域学子的交流桥梁，凝聚学术共识、共享发展经验。活动得到了大连地区兄弟院校的积极响应，CCF 大连海事大学学生分会、CCF 大连民族大学学生分会、CCF 大连大学学生分会的部分学生会员代表齐聚一堂，与 CCF 大连理工大学学生分会的会员们一道，围绕分会发展与学术成长展开深入探讨。

附：

CCF 大连理工大学学生分会首届执行委员会名单：

督导主任：齐恒

指导教师：张亮，王帅，张世凯

主席：钟剑辉（大连理工大学计算机科学与技术学院 2024 级本科生）

候任主席：刘致远（大连理工大学计算机科学与技术学院 2024 级本科生）

执委：宋佳音、牟学博、任英虹（大连理工大学计算机科学与技术学院 2024 级本科生）

据 CCF 微信公众号