基于属性组权重的分类数据离群检测

张凯棋,宋亦静,陈 鑫
(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要:属性分组是高维离群检测中的有效手段之一,可以有效缓解“维度灾难”的干扰,但现有的属性分组离群检测方法未能体现属性组之间的差异性,以及属性组的偏离程度,严重影响了高维离群检测的效果与性能。该文采用信息熵累加和刻画与描述属性组之间的差异性,提出了一种基于属性组权重的分类离群检测方法。首先,根据数据模式频率和编码长度,定义了属性组偏离因子,并将其作为属性组之间的合并依据,有效地刻画了属性组的偏离程度,进一步提高了属性分组过程中的搜索效率;其次,利用信息熵累加和定义了属性组权重,有效地体现了不同属性组之间的差异性;然后,依据属性组权重,重新定义了离群得分函数,并提出了一种基于属性组权重的分类数据离群检测算法;最后,采用 UCI,NTU,KEEL 和人工合成数据集,实验验证了该离群检测算法不仅具有较高的检测精度和效率,而且也具有良好的可扩展性与伸缩性,可适用于高维海量分类属性数据集的离群检测任务。

关键词:离群检测;属性分组;分类数据;属性组权重;偏离因子
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2023)11-0020-08
doi:10.3969/j.issn.1673-629X.2023.11.004

Attribute Group Weight-based Outlier Detection for Categorical Data

ZHANG Kai-qi,SONG Yi-jing,CHEN Xin
(School of Computer Science and Technology,Taiyuan University of Science and Technology,Taiyuan 030024,China)

Abstract:Attribute grouping is one of the effective methods in high-dimensional outlier detection,which can effectively alleviate the interference of “the curse of dimensionality”. However,existing attribute grouping outlier detection methods fail to reflect the differences among attribute groups and the deviation degree of attribute groups,which have a significant negative influence on the efficiency and performance of high-dimensional outlier detection. We propose an attribute group weight-based outlier detection method for categorical data by using information entropy cumulative sum,which depicts and describes the difference among attribute groups. Firstly,the attribute group deviation factor is defined according to the data pattern frequency and code lengths,and used as a basis of merging attribute groups,which effectively portrays the deviation among attribute groups and further improves the search efficiency in the process of attribute grouping. Secondly,the information entropy cumulative sum is used to define the attribute group weights,which effectively reflects the difference among different attribute groups. Thirdly,the outlier score function is redefined based on the attribute group weights,and an outlier detection algorithm for categorical data is proposed on this basis. In the end,experimental results on UCI,NTU,KEEL and synthetic datasets validate that the outlier detection algorithm not only has high detection accuracy and efficiency,but also has good extensibility and scalability,which can be applied to the outlier detection task of high-dimensional massive categorical attribute datasets.

Key words:outlier detection;attribute grouping;categorical data;attribute group weight;deviation factor

0 引 言

离群数据(outlier)是指明显偏离其他数据对象,可能是由于一种不同于其他数据对象的机理而产生的数据对象[1]。由于离群数据蕴含着大量丰富的、有价值的潜在的信息,已广泛地应用在无线传感器网络[2]、肿瘤检测[3]、入侵检测[4]、欺诈检测[5]、网络安全[6]、数据清理[7]、工业系统[8]、地球科学[9]等领域。针对高维数据,由于出现了“维度灾难”,严重影响了离群检测效果。属性分组是指根据属性之间的相关性实现属性分组,并降低了离群检测维度,可有效地缓解“维度灾难”干扰。

属性分组离群数据检测作为一类高维离群检测方法,将依据属性之间的关联关系,将属性划分为若干子集,并利用各属性子集,识别或度量离群数据对象,从

收稿日期:2023-01-09 **修回日期**:2023-05-11
基金项目:山西省基础研究计划资助项目(202103021223267);山西省高等学校科技创新计划项目(2021L297);太原科技大学科研启动基金项目(20212053,20222107)
作者简介:张凯棋(1997-),男,硕士研究生,通讯作者,研究方向为数据挖掘;宋亦静(1992-),女,博士研究生,研究方向为数据挖掘。

而可挖掘出全局和局部离群数据[10-12]。但现有的属性分组离群检测未能体现属性组之间的差异性与属性组的偏离程度,影响了高维离群检测效果。该文利用信息熵累加和刻画属性组权重,提出了一种基于属性组权重的分类离群数据检测方法,有效地体现了属性组之间的差异性,进一步改善了高维离群检测效果,并缓解了"维度灾难"干扰。其主要贡献:

· 依据数据模式频率和编码长度,定义了属性组偏离因子及计算公式;

· 提出了一种基于信息熵的属性组权重计算方法;

· 提出了一种基于属性组权重的分类离群数据检测算法。

1　相关工作

分类数据作为一类广泛出现在多种应用领域中的重要数据类型,具有取值无序与不可比等特点。目前,分类数据离群检测主要包括:基于距离的方法[13-15]、基于密度的方法[16-17]、基于统计的方法[18-19]、基于聚类的方法[20-22]、基于频繁项集的方法[10-11,23]、基于子空间的方法[24-26]等。针对高维分类数据,由于"维度灾难",严重影响了分类数据离群检测性能;而子空间方法是将数据对象从高维空间投影到低维空间,可有效缓解"维度灾难"干扰,是高维分类离群检测的有效途径之一。

基于子空间的离群检测是将数据对象从高维空间投影到低维子空间,并在低维子空间中,度量与检测离群数据对象,主要包括相关子空间和稀疏子空间两类子空间方法。其典型研究成果:文献[27]采用遗传算法搜索稀疏子空间,但该算法受初始种群影响,离群挖掘结果的完备性和准确性等无法得到保证;文献[28]采用概念格作为子空间的描述工具,通过引入稠密度系数,在概念格内涵中确定稀疏子空间;文献[29]在概念格的基础上提出了约束概念格的方法,提高了概念格的构造效率,进一步提高了挖掘的效率;文献[30]提出了一种基于稀疏子空间并行搜索技术的类星体光谱异常数据并行检测算法,实现了对天体光谱中异常光谱的挖掘。稀疏子空间的方法的稀疏系数阈值需要人为设置,无法保证稀疏子空间划分的准确性;文献[31]利用局部稀疏差异因子得到相关子空间,解决了检测相关子空间时间复杂度较高的问题,但是维度的数据分布会使得到的相关子空间产生一定的误差;文献[24]提出一种轴平行子空间离群挖掘方法,该方法通过共享最近邻居,为数据对象寻找似子集以确定子空间,但该方法采用的欧氏距离的维平均方法度量离群具有明显不足;文献[32]提出一种在相关子空间进行离群检测的算法,算法使用 KNN 方法确定数据对象的局部数据集,然后生成稀疏因子矩阵和局部稀疏因子矩阵,再确定数据对象对应的子空间定义向量,计算数据对象的离群程度,但该方法的局部稀疏差异因子阈值的选取会影响子空间定义向量的确定和挖掘的精度。相关子空间的方法不能有效地在高维数据集中挖掘离群数据,因此离群挖掘的效率和准确性无法保证。

作为一种子空间离群检测,属性分组是将高维属性划分为若干不同属性组,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。其典型工作有:文献[10]提出了一种利用特征分组的加权离群检测的方法,该算法利用特征关系的概念对特征(即属性)进行分组,可以实现对分类数据中的全局和局部离群数据的挖掘,但是该算法组间离群数据的合并缺乏合理的解释性;文献[11]提出了一种基于压缩数据的离群检测算法,该算法根据最小描述长度定律最小化压缩长度实现属性的分组,但是由于建立代码表的开销过大,所以该算法的时间复杂度较高;文献[12]采用基于二叉搜索树的隔离算法实现属性分组,在属性组内实现数据对象的聚类过程,以实现对离群数据的挖掘。属性分组的方法对于分组组数的选取和分组时属性分配存在一定的难度,同时,对于全局离群数据的挖掘的组间合并缺乏合理的解释性。

综上所述,子空间离群检测可以有效地发现隐藏在子空间中的局部离群数据,但会丢失部分有用信息,影响了离群检测性能;属性分组离群检测将高维属性分为不同的属性组,有效地利用了全部离群信息,但现在的属性分组存在着组数选取敏感,离群检测未能有效体现属性组间的差异性等,从而影响离群检测效果。

2　属性分组与离群检测

在高维离群检测中,属性分组是缓解"维度灾难"的有效途径之一。所谓属性分组是依据属性之间的关联程度,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。参照文献[10-11],相关概念描述如下:

设 $D=\{x_i \mid 1 \leqslant i \leqslant n\}$ 为任意数据集,$C=\{f_k \mid 1 \leqslant k \leqslant d\}$ 为 D 的属性集,其中:x_i 表示第 i 个数据对象,n 为数据对象的个数,d 为属性个数,f_k 表示第 k 个属性。属性分组就是将 C 划分为 m 个互不相交的属性组 $P=\{F_j \mid 1 \leqslant j \leqslant m\}$,且每一属性组内的属性之间具有强相关性,分属任意 2 个属性组的属性之间具有弱相关性,其中:$\bigcup_{j=1}^{m} F_j = C$,$F_u \cap F_v = \varnothing$,$1 \leqslant u, v \leqslant m, u \neq v$。

信息增益描述了随机变量的概率分布的差异性，属性分组通过标准化信息增益（Normal Information Gain，NIG）来描述属性组间概率分布的差异性，可以有效地刻画属性组之间的相关性。对于任意 2 个属性组（F_u 和 F_v），其标准化信息增益的计算公式如下：

$$\mathrm{NIG}(F_u,F_v)=\frac{H(F_u)+H(F_v)-H(F_u,F_v)}{|F_u|+|F_v|} \tag{1}$$

其中，$H(F_u)$，$H(F_v)$ 分别为第 u 属性组的信息熵、第 v 属性组的信息熵，$H(F_u,F_v)$ 为 F_u 和 F_v 之间的互信息，$|F_u|$ 和 $|F_v|$ 分别表示第 u 属性组和第 v 属性组的属性个数。

在公式(1)中，$\mathrm{NIG}(F_u,F_v)$ 刻画了 F_u 和 F_v 之间的相关程度，其值越大，表明两组的相关性越强，合并为同一组的概率就越大，反之亦然。

数据模式是指数据对象（x_i）在属性组（F_j）中，属性的一个取值项集。在 F_j 中，x_i 由若干个互不相同的数据模式组成。代码表是为描述属性组中的数据模式和数据模式编码构建的一种模型结构。数据模式编码体现了数据模式在代码表中数据模式频率分布的差异性。在 F_j 中，设 p 为 x_i 的任一数据模式，CT 为描述数据模式和数据模式编码的代码表，则 p 的数据模式编码定义如下：

$$L(\mathrm{code}(p)\mid \mathrm{CT})=-\log_2\left(\frac{\mathrm{usage}(p)}{\sum_{p'\in \mathrm{CT}}\mathrm{usage}(p')}\right) \tag{2}$$

其中，usage（p）表示 p 出现的次数，$\sum_{p'\in \mathrm{CT}}\mathrm{usage}(p')$ 表示在 CT 中全部数据模式出现的总次数。

在公式(2)中，$L(\mathrm{code}(p)\mid \mathrm{CT})$ 刻画了数据模式在 CT 下的频率分布，其值越大，表明该数据模式在代码表中出现的频率越大，反之亦然。

数据编码是数据集（D）中所有数据对象（x_i）在 CT 中所有数据模式编码的总和，数据编码有效地刻画了 D 在 F_j 中的压缩效果，是评估 CT 模型质量的依据。设 CT 为代码表，A 为 x_i 在 F_j 中所有数据模式的集合，则利用公式(2)，D 的数据编码描述如下：

$$L(D\mid \mathrm{CT})=\sum_{x_i\in D}\sum_{p\subset A}L(\mathrm{code}(p)\mid \mathrm{CT}) \tag{3}$$

在公式(3)中，$L(D\mid \mathrm{CT})$ 刻画了数据集（D）在 CT 模型中的压缩效果，其值越大，表明压缩效果越差，无法适用于 D，反之亦然。

为了有效地度量构建代码表模型(CT)的代价，利用公式(2)，CT 的代价定义如下：

$$L(\mathrm{CT})=\sum_{p\in \mathrm{CT}}L(\mathrm{code}(p)\mid \mathrm{CT})+\sum_{i\in S}-r_i\log(p_i) \tag{4}$$

其中，$\sum_{i\in S}-r_i\log(p_i)$ 表示对 CT 中编码数据模式所需要的编码长度，S 表示 CT 中所有数据模式中全部单项的集合，$p_i=r_i/c$，r_i 表示单项 i 在数据模式中出现的次数，c 表示所有数据模式中单项出现的总次数。

在公式(4)中，$L(\mathrm{CT})$ 反映了构建代码表的代价，其值越大，表明构建该代码表所需的编码长度越长，代价成本就越高，反之亦然。

为了刻画属性分组（P）的压缩效果与体现属性分组的分组效果，利用 P 中的各个属性组对应的数据编码和代码表代价总和作为其度量依据。利用公式(3)和(4)，在 D 中，CT 和 P 的压缩消耗函数定义如下：

$$L(P,\mathrm{CT},D)=L(P)+\sum_{F\in P}L(\pi_F(D)\mid \mathrm{CT}_F)+\sum_{F\in P}L(\mathrm{CT}_F) \tag{5}$$

其中，$L(P)$ 表示描述分组情况所需要的编码长度。

在公式(5)中，$L(P,\mathrm{CT},D)$ 有效地刻画了属性分组的质量，其值越大，表明总的压缩效果越差，表明属性分组的效果越差，反之亦然。

依据上述定义，属性分组的基本步骤如下：

(1)将每个属性视为单独的一组，根据公式(2)，计算数据模式编码，构建对应的代码表；

(2)根据公式(1)计算两两属性组的 NIG，按照 NIG 降序的方式对属性组排序；

(3)根据公式(3)～(5)计算压缩消耗函数，根据压缩消耗函数的大小实现组间合并，更新属性组及代码表，若属性组全部遍历后，压缩消耗函数值仍未减小，分组结束，否则返回步骤 2，继续组间合并。

在上述属性分组的基础上，为了度量数据对象在其数据集中的离群程度，利用公式(2)、公式(3)，x_i 的离群得分计算公式定义如下：

$$\mathrm{score}(x_i)=\sum_{F\in P}\sum_{p\subset A}L(\mathrm{code}(p)\mid \mathrm{CT}_F) \tag{6}$$

在公式(6)中，score（x_i）刻画了 x_i 的离群程度，其值越大，表明数据对象的偏离程度越大，成为离群数据的概率越大，反之亦然。

3 属性组权重与分类离群检测

在上一章节的属性分组步骤中，NIG 作为一种搜索策略，是属性组之间合并顺序依据。由公式(1)可知，计算 NIG 的值与初始属性分组（P）中的属性组个数（d）有关，即：需要计算 d 个属性组中的任意两组间 NIG；依据公式(5)定义的压缩消耗函数，属性组迭代合并更新后，还需要重新计算属性组之间的 NIG。因此，属性分组中的搜索策略效率低下。

在属性分组中，数据模式出现的频率体现了属性取值项集疏密程度，数据模式编码长度体现了数据模式频率分布。数据模式编码长度体现了属性组的不确

定性,可以有效地衡量属性组的偏离程度。对于任意属性组(F_j),设 CT 为 F_j 对应的代码表, F_j 的偏离因子定义如下:

$$\mathrm{GDF}(\mathrm{CT}) = \log(1/p(r)) - \log(1/p(t)) \tag{7}$$

其中, r,t 分别为 CT 中的频次最小、频次最大数据模式, $p(r)$,$p(t)$ 分别表示 r,t 数据模式在 CT 中出现的频率。

在公式(7)中, $\log(1/p(r))$ 和 $\log(1/p(t))$ 分别刻画了 r,t 所含的信息量,数据模式出现的频次越小,包含的离群信息量越大,反之亦然。r 是属性组(F_j)中含离群信息量最大的数据模式,而 t 是属性组(F_j)中含离群信息量最小的数据模式,因此,GDF(CT)刻画了 F_j 的偏离程度。在属性分组中,采用偏离因子(GDF)作为属性组之间合并顺序依据,只需要利用代码表中已有的数据模式编码,计算 d 个属性组的 GDF。相较 NIG 而言,避免了 NIG 中任意两组间的互信息计算,因此有效地提高了搜索策略的效率。

在公式(6)中,离群得分计算仅体现了属性组(F_j)的数据模式编码长度,忽略了属性组之间所包含的离群信息量差异,影响离群检测效果。信息熵是根据系统内的分布差异,衡量系统所含信息量,信息熵越大表明该系统的不确定性或混乱程度越大[33]。因此,利用信息熵度量属性组,可以体现属性组内包含的离群信息量,信息熵越大表明所包含的离群信息越多,即:离群检测的重要程度也就越大,从而体现了属性组之间的差异性。设 $F_j = \{f_k \mid 1 \leqslant k \leqslant h\}$ 为任意属性组,则 F_j 的权重定义如下:

$$\omega(F_j) = \frac{\sum_{k=1}^{v} H(f_k)}{h} \tag{8}$$

其中, f_k 为 F_j 中的第 k 个属性, h 为 F_j 中的属性个数, $H(f_k)$ 表示第 j 属性组第 k 属性的信息熵。

在公式(8)中,属性组权重有效地度量了属性组的混乱程度,体现了属性组所包含的离群信息量,其值越大,所包含的离群信息量越大,反之亦然。属性组权重刻画了不同属性组对度量离群数据对象的重要程度,充分体现出属性组间的差异性。

由公式(7)和(8)可知,属性组偏离因子(GDF)利用数据模式编码频率分布,有效地刻画了属性组对度量离群数据的重要程度,改善了搜索策略的效率;属性组权重利用信息熵度量了属性组包含的离群信息重要程度,有效地刻画了属性组之间的差异。为了有效地度量数据对象(x_i)在其数据集(D)中的离群程度,参照公式(6)、公式(8), x_i 的离群得分重新定义如下:

$$\mathrm{score}(x_i) = \sum_{F_j \in P} \omega(F_j) \cdot \left(\sum_{p \subset A} L(\mathrm{code}(p) \mid \mathrm{CT}_F) \right) \tag{9}$$

在公式(9)中,离群得分值刻画了 x_i 的离群程度,其值越大,表明数据对象的偏离程度越大,成为离群数据的概率越大,反之亦然。由于该离群得分体现了属性组权重和数据模式编码长度,因而有效地刻画了属性组内和属性组间数据对象所包含的离群信息。

4　基于属性组权重的分类离群检测算法

依据上一章节描述,采用属性组权重,分类离群检测基本步骤描述如下:

(1)将每个属性视为单独的一组,根据公式(2),计算数据模式编码,构建对应代码表;

(2)根据公式(7)计算属性组偏离因子(GDF),按照 GDF 降序的方式对属性组排序;

(3)根据公式(3)~(5)计算压缩消耗函数,根据压缩消耗函数的大小实现组间合并,更新属性组及代码表,若属性组全部遍历后,压缩消耗函数值仍未减小,分组结束,否则返回步骤 2,继续组间合并;

(4)根据公式(8)计算属性组权重,根据公式(9)计算数据对象的离群得分,从中选取离群得分最高的 k 个对象作为离群数据。

算法伪代码如下:

算法:AGWODC(Attribute Group Weight Outlier Detection for Categorical data)

输入:数据集(D)(n 个数据对象× d 个属性)

输出: k 个离群数据对象

1. 初始化分组: $P = \{F_j \mid 1 \leqslant j \leqslant d\}$, $F_j = \{f_j\}$, $1 \leqslant j \leqslant d$

2. 根据公式(2)构建代码表: $\mathrm{CT} = \{\mathrm{CT}_j \mid 1 \leqslant j \leqslant d\}$

3. 根据公式(7)计算每个组的偏离因子(GDF),构成集合(OD)

4. For F_u in P :

5.　　for F_v in P :

6. 降序排序 OD

7. 根据公式(5),计算最小消耗函数 $L(P,\mathrm{CT},\mathrm{OD})$

　$\cos t = L(P,\mathrm{CT},\mathrm{OD})$

8. 根据公式(2)计算数据模式的编码长度,构建新的代码表(CT_{ulv})

9. $P = P \backslash (F_u \cup F_v) \cup F_{ulv}$

10. $\mathrm{CT}' = \mathrm{CT} \backslash (\mathrm{CT}_u \cup \mathrm{CT}_v) \cup \mathrm{CT}_{ulv}$

11.　　if $L(P',\mathrm{CT}',\mathrm{OD}) < \cos t$:

12. $P = P'$, $\mathrm{CT} = \mathrm{CT}'$

13. 根据公式(7)计算 CT' 的 CDF,更新 OD

14. 返回步骤(4)

15.　　endif;

16.　endfor;

17. Endfor;

18. For each j in P :

19. 根据公式(8)计算组权重 $\omega(F_j)$

20. For each $i \in (1,2,\cdots,n)$ in D :

21. 根据公式(9)计算数据对象的离群分数

22. Endfor;

23. 选择离群分数最高的 k 个离群数据对象

24. End AGWODC

时间复杂性分析:

在上述 AGWODC 算法中,主要包括属性分组、属性组权重与离群得分两个阶段。在属性分组中,主要包括了组间合并时寻找新的数据模式,合并过程中插入新的数据模式时,更新其他数据模式的频次等。组间合并时寻找新的数据模式所需时间复杂度约为 $O(n)$,更新其他数据模式的频次所需时间复杂度约为 $O(n)$,在最坏情况下,组间合并时间复杂度为 $O(d^2)$。因此,属性分组的时间复杂度为 $O(d^2n)$,其中 n 为数据对象的个数, d 为属性的个数。在属性组权重与离群得分中,主要包括了属性组权重的计算和计算数据对象离群得分,属性组权重的计算时间复杂度约为 $O(d)$,计算数据对象离群得分的时间复杂度约为 $O(n)$,因此该过程的时间复杂度约为 $O(n)$。总之,AGWODC 算法的时间复杂度约为 $O(d^2n+n)$。

5 实验结果

实验配置:Intel Core I5 6300HQ CPU,8G 内存,并采用 python 语言在 pycharm 平台上,实现了 AGWODC 算法及实验对比算法 Watch[10] 和 CompreX[11]。实验数据集包括 UCI, Kaggle, UIUC 公开数据集。另外,为了验证算法的可扩展性和伸缩性,采用 GAClust toolkit 软件[34] 生成人工数据集,将随机数发生器种子的值设置为 5,类别数设置为 2,根据实验需要设置相应的数据量、属性个数,分别从两类中选取符合实验需求的正常数据与离群数据,构成人工数据集:其中,data1-data4 为数据量相同,维数不同的人工数据集,用于验证算法的可扩展性;data5-data8 为维数相同,数据量不同的人工数据集,用于验证算法的伸缩性。其数据集详情如表 1 所示。

表 1 实验数据集

数据集	数据集类型	数据量	维数	离群数据	离群占比/%
W7a	NTU	34 704	300	933	2.69
connect-4	UCI	44 916	42	443	0.99
reuters	KEEL	12 897	100	237	1.84
covertype	UCI	581 012	44	2 747	0.47
census	UCI	194 629	34	12 380	6.36
loancleaned	Kaggle	100 000	61	10 000	1.00
phishing	Mendeley	60 000	109	4 812	8.02
Cowsgame	Kaggle	470 549	131	6 879	1.46
DSMLsurvey	Kaggle	10 727	339	1 503	14.01
diabetes	Kaggle	101 585	39	11 342	11.17
data1	人工	5 000	50	50	1
data2	人工	5 000	100	50	1
data3	人工	5 000	150	50	1
data4	人工	5 000	200	50	1
data5	人工	5 000	50	50	1
data6	人工	10 000	50	50	0.5
data7	人工	15 000	50	50	0.33
data8	人工	20 000	50	50	0.25

5.1 检测准确性

为了实验验证 AGWODC 算法的离群检测精度,采用了表 1 所示的 UCI, Kaggle, UIUC 等 10 个数据集,以及 2 个对比算法:Watch 和 CompreX,其 AUC[35] 指标的实验结果如表 2 所示。

由表 2 可知,AGWODC 的 AUC 指标值优于其对比算法,仅在 Cowsgame 数据集上略低于 CompreX 算法,在 diabetes 数据集上略低于 Watch 算法;尽管 AGWODC 的 ROC 曲线基本高于其他对比算法,但在 Cowsgame 数据集上略低于 CompreX 算法,diabetes 数据集上略低于 Watch 算法。因此,表明了 AGWODC 具有较高的离群检测精度。其主要原因:AGWODC 利用了基于信息熵度量的属性组权重,并在离群得分中充分体现出了属性组间的差异。此外,由于 Cowsgame 是一种猜数字游戏的数据集,其属性由玩家若干次猜测数字构成,每次玩家猜测的数字形成一

个属性组,从而形成有特定含义的属性分组,且与数据集分布无关,属性组权重会偏离了特定含义的属性分组;由于 diabetes 数据集的维数较低、属性的取值较少,信息熵无法体现出属性组之间的差异。

表 2　AUC 指标对比

Dataset	CompreX	Watch	AGWODC
w7a	0.53	0.48	0.54
connect-4	0.70	0.50	0.72
reuters	0.99	0.47	0.99
covertype	0.92	0.62	0.92
census	0.63	0.44	0.68
loancleaned	0.54	0.53	0.56
phishing	0.92	0.50	0.95
Cowsgame	0.56	0.40	0.55
DSMLsurvey	0.52	0.49	0.53
diabetes	0.51	0.53	0.52
Average AUC	0.69	0.50	0.70

5.2　检测效率

为了实验验证 AGWODC 算法的离群检测效率,采用了表 1 所示的 UCI, Kaggle, UIUC 等 10 个数据集,以及 2 个对比算法:Watch 和 CompreX,其实验结果如表 3 所示。

表 3　各算法运行时间比较　s

dataset	CompreX	Watch	AGWODC
w7a	28 895.18	18 921.10	8 097.13
connect-4	314.09	290.40	97.05
reuters	1 112.28	568.05	122.36
covertype	1 721.25	1 048.94	206.03
census	21 573.12	20 580.77	9 984.08
loancleaned	2 288.01	728.68	300.76
phishing	18 921.10	9 095.50	5 114.54
Cowsgame	28 760.52	16 342.84	8 721.25
DSMLsurvey	17 012.50	11 315.96	6 232.31
diabetes	2 316.18	1 757.84	563.45

由表3 可知,AGWODC 算法的耗时明显低于其他对比算法,因而表明 AGWODC 具有较高的检测效率。其主要原因:由于 Watch 算法采用特征关系(FR)等,实现其属性分组及离群得分,需要频繁计算属性之间的互信息;CompreX 算法在属性组之间合并过程中采用标准化信息增益(NIG)的搜索策略作为属性组间合并依据;AGWODC 采用基于偏离因子(GDF)的搜索策略,避免了 NIG 中任意两组间的互信息计算。

5.3　弗里德曼检验

弗里德曼检验是一种无参数检验方法,用于检测 3 组或者 3 组以上数据是否存在显著差异性。为了评价 AGWODC 与其他相关算法在 AUC 值和检测效率的优越性,采用统计检验中弗里德曼检验对 3 种算法在 10 个数据集上的 AUC 值和运行时间进行检验,比较 3 种算法是否存在差异性。根据表 2 和表 3 中的实验数据,弗里德曼检验统计结果如表 4 所示。

表 4　AUC 与运行时间的弗里德曼检验

指标	描述统计	CompreX	Watch	AGWODC
AUC	均值	0.69	0.50	0.70
	标准偏差	0.19	0.06	0.19
	最大值	0.99	0.62	0.99
	最小值	0.51	0.40	0.52
	秩平均值	2.10	1.20	2.70
运行时间	均值	12 291.42	8 065.01	3 943.90
	标准偏差	11 921.22	8 260.54	4 099.29
	最大值	28 895.18	20 580.77	9 984.08
	最小值	314.09	290.40	97.05
	秩平均值	3.00	2.00	1.00

由表4 可知,在显著性水平 $\alpha = 0.05$ 时,卡方值为 12.000,渐近显著性为 0.002,表明 AUC 值的秩平均值具有显著统计学差异,秩平均值越大,AUC 值越大;在显著性水平 $\alpha = 0.05$ 时,卡方值为 20.000,渐近显著性为 0.000 045,表明运行时间的秩平均值具有显著统计学差异,秩平均值越小,运行时间越短。由表 4 可知,AGWODC 的 AUC 值算法秩平均值最大,运行时间秩平均值最小,因此,从统计检验的角度而言,AGWODC 算法优于其他算法。

5.4　可扩展性与伸缩性

可扩展性与伸缩性是衡量离群检测算法的重要指标,可扩展性是当数据对象个数相同,维数变化对离群检测算法效率的影响;伸缩性是当维数相同,数据对象个数变化对算法效率的影响。为了验证 AGWODC 算法的可扩展性,采用了表 1 所示的 4 个人工数据集(data1 ~ data4),以及 2 个对比算法:Watch 和 CompreX,其实验结果如图 1 所示。

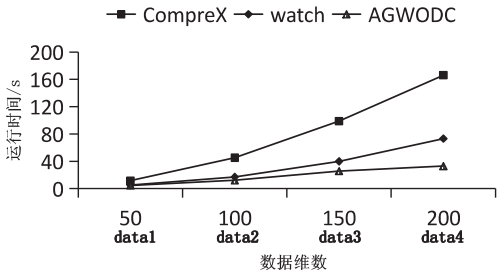

图 1　可扩展性

由图 1 可知,随着数据集的维数增加,3 个算法的耗时呈现了非线性增长,且 AGWODC 明显优于其他 2 个对比算法,表明 AGWODC 具有良好的可扩展性。

其主要原因:随着维数的增加,在属性分组过程中,属性组偏离因子(GDF)的计算随之增多,属性组代码表建立和属性组间合并随之增多,但避免了 CompreX 中 NIG 搜索策略与 Watch 中 FR 中属性之间的互信息计算,有效降低了算法的运行时间。

为了验证 AGWODC 算法的伸缩性,采用了表 1 所示的 4 个人工数据集(data5 ~ data8),以及 2 个对比算法:Watch 和 CompreX,其实验结果如图 2 所示。

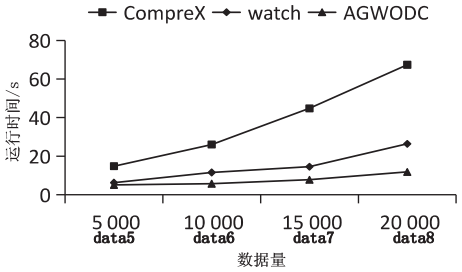

图 2 伸缩性

由图 2 可知,随着数据集的数据对象增加,3 个算法的耗时呈现了非线性增长,且 AGWODC 明显优于其他 2 个对比算法,表明 AGWODC 具有良好的伸缩性。其主要原因:随着数据对象的增加,属性组间合并时寻找新的数据模式、更新数据模式频次和数据对象离群得分的计算量也随之增加,但避免了 CompreX 中 NIG 搜索策略与 Watch 中特征关系(FR)中属性之间的互信息计算,大大缩短了算法的运行时间。

6 结束语

该文采用信息熵累加和刻画属性组权重,提出了一种基于属性组权重的分类离群数据检测方法,充分体现和刻画了属性组之间的差异性以及属性组的偏离程度,有效地改善了高维离群检测性能,并缓解了“维度灾难”干扰,可适用于海量高维分类数据离群检测任务。下一步研究工作是基于属性组权重的分类离群数据检测并行化,以及混合数据的高维离群检测等。

参考文献:

[1] KNOX E M,NG R T. Algorithms for mining distance based outliers in large datasets[C]//Proceedings of the international conference on very large data bases. San Francisco:Morgan Kaufmann Publishers Inc,1998:392-403.

[2] CAUTERUCCIO F, FORTINO G, GUERRIERI A, et al. Short-long term anomaly detection in wireless sensor networks based on machine learning and multi-parameterized edit distance[J]. Information Fusion,2019,52:13-30.

[3] JALALIFAR A, SOLIMAN H, RUSCHIN M, et al. A brain tumor segmentation framework based on outlier detection using one-class support vector machine[C]//2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC). Montreal: IEEE, 2020:1067-1070.

[4] SANDOSH S, GOVINDASAMY V, AKILA G. Enhanced intrusion detection system via agent clustering and classification based on outlier detection[J]. Peer-to-Peer Networking and Applications,2020,13(3):1038-1045.

[5] LAIMEK R, KAOTHANTHONG N, SUPNITHI T. ATM fraud detection using outlier detection [C]//International conference on intelligent data engineering and automated learning. Madrid:Springer,2018:539-547.

[6] CHAKHCHOUKH Y,LIU S,SUGIYAMA M,et al. Statistical outlier detection for diagnosis of cyber attacks in power state estimation[C]//2016 IEEE power and energy society general meeting (PESGM). Boston:IEEE,2016:1-5.

[7] LIU H,SHAH S,JIANG W. On-line outlier detection and data cleaning[J]. Computers & Chemical Engineering,2004,28(9):1635-1647.

[8] KAUPP L,BEEZ U,HÜLSMANN J,et al. Outlier detection in temporal spatial log data using autoencoder for industry 4.0[C]//International conference on engineering applications of neural networks. Xersonisos:Springer,2019:55-65.

[9] DUTTA H, GIANNELLA C, BORNE K, et al. Distributed top-k outlier detection from astronomy catalogs using the demac system[C]//Proceedings of the 2007 SIAM international conference on data mining. Minneapolis: Society for Industrial and Applied Mathematics,2007:473-478.

[10] LI J,ZHANG J,PANG N,et al. 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.

[11] AKOGLU L,TONG H,VREEKEN J,et al. Fast and reliable anomaly detection in categorical data[C]//Proceedings of the 21st ACM international conference on information and knowledge management. Maui: Association for Computing Machinery,2012:415-424.

[12] KARCZMAREK P, GAŁKA Ł, DOLECKI M, et al. Enhanced tree-based anomaly detection[C]//2022 IEEE international conference on fuzzy systems (FUZZ-IEEE). Padua:IEEE,2022:1-7.

[13] RAMASWAMY S, RASTOGI R, SHIM K. Efficient algorithms for mining outliers from large data sets[C]//Proceedings of the 2000 ACM SIGMOD international conference on management of data. New York:Association for Computing Machinery,2000:427-438.

[14] BHADURI K,MATTHEWS B L,GIANNELLA C R. Algorithms for speeding up distance-based outlier detection [C]//Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2011: 859-867.

[15] RADOVANOVIC M, NANOPOULOS A, IVANOVIC M.

Reverse nearest neighbors in unsupervised distance – based outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering,2014,27(5):1369–1382.

[16] BREUNIG M M,KRIEGEL H P,NG R T,et al. LOF:identifying density–based local outliers[C]//Proceedings of the 2000 ACM SIGMOD international conference on management of data. New York:Association for Computing Machinery,2000:93–104.

[17] REN D,WANG B,PERRIZO W. Rdf:a density–based outlier detection method using vertical data representation[C]// Fourth IEEE international conference on data mining (ICDM '04). Brighton:IEEE,2004:503–506.

[18] ROUSSEEUW P J,HUBERT M. Robust statistics for outlier detection[J]. Wiley Interdisciplinary Reviews:Data Mining and Knowledge Discovery,2011,1(1):73–79.

[19] ZHANG Y,HAMM N A S,MERATNIA N,et al. Statistics–based outlier detection for wireless sensornetworks[J]. International Journal of Geographical Information Science,2012,26(8):1373–1392.

[20] ELAHI M,LI K,NISAR W,et al. Efficient clustering–based outlier detection algorithm for dynamic data stream[C]// 2008 fifth international conference on fuzzy systems and knowledge discovery. Jinan:IEEE,2008:298–304.

[21] JAYAKUMAR G D S,THOMAS B J. A new procedure of clustering based on multivariate outlier detection[J]. Journal of Data Science,2013,11(1):69–84.

[22] PAMULA R,DEKA J K,NANDI S. An outlier detection method based on clustering[C]//2011 second international conference on emerging applications of information technology. Kolkata:IEEE,2011:253–256.

[23] HE Z,XU X,HUANG Z J,et al. FP–outlier:frequent pattern based outlier detection[J]. Computer Science and Information Systems,2005,2(1):103–118.

[24] KRIEGEL H P,KRÖGER P,SCHUBERT E,et al. Outlier detection in axis–parallel subspaces of high dimensional data[C]//Pacific–Asia conference on knowledge discovery and data mining. Berlin:Springer,2009:831–838.

[25] KRIEGEL H P,KRÖGER P,SCHUBERT E,et al. Outlier detection in arbitrarily oriented subspaces[C]//2012 IEEE 12th international conference on data mining. Brussels:IEEE,2012:379–388.

[26] RAHMANI M,ATIA G K. Randomized robust subspace recovery and outlier detection for high dimensional data matrices[J]. IEEE Transactions on Signal Processing,2016,65(6):1580–1594.

[27] AGGARWAL C C,PHILIP S Y. An effective and efficient algorithm for high – dimensional outlier detection[J]. The VLDB Journal,2005,14(2):211–221.

[28] 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.

[29] 张继福,张素兰,蒋义勇. 基于约束概念格的天体光谱局部离群数据挖掘系统[J]. 光谱学与光谱分析,2009,29(2):551–555.

[30] 马　洋,张继福,蔡江辉,等. 基于稀疏子空间的类星体光谱异常特征并行提取与分析[J]. 光谱学与光谱分析,2021,41(4):1086–1091.

[31] 张继福,李永红,秦　啸,等. 基于 MapReduce 与相关子空间的局部离群数据挖掘算法[J]. 软件学报,2015,26(5):1079–1095.

[32] 李永红,张继福,荀亚玲. 相关子空间中的局部离群数据挖掘算法研究[J]. 小型微型计算机系统,2015,36(3):460–465.

[33] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal,1948,27(3):379–423.

[34] CRISTOFOR D,SIMOVICI D A. Finding median partitions using information–theoretical–based genetic algorithms[J]. J. Univers. Comput. Sci.,2002,8(2):153–172.

[35] LOBO J M,JIMÉNEZ – VALVERDE A,REAL R. AUC:a misleading measure of the performance of predictive distribution models[J]. Global Ecology and Biogeography,2008,17(2):145–151.