Hans Journal of Data Mining 数据挖掘, 2021, 11(3), 150-166
Published Online July 2021 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2021.113014

Hans汉斯

基于反常项集的异常值处理算法

崔　晨，李　贵，李征宇，韩子扬，曹科研

沈阳建筑大学信息与控制工程学院，辽宁 沈阳
Email: 724079015@qq.com, ligui21c@sina.com

收稿日期：2021年4月27日；录用日期：2021年5月27日；发布日期：2021年6月7日

摘　要

异常值指的是数据中的噪声和不一致值。异常值检测与处理往往依赖于约束规则，通常的约束规则包括条件函数依赖、否定约束、编辑规则等。但对于特定领域，这些领域约束规则需要由领域专家制定，基于数据挖掘和机器学习算法，难以高效地发现这些领域约束规则。本文提出了一种用于数据清洗的反常项集的概念，与基于数据分布密度的异常值检测算法类似，反常项集是数据中不太可能出现的非常态取值组合。在此基础上，本文引入了加权调和提升度的概念及特性，利用改进的等价类变换算法挖掘低提升度的反常项集。并采用准反常项集对数据更正进行预计算，给出了一种类似于近邻插补算法的异常值更正算法，以保证异常值处理质量。在房地产信息数据集下的实验表明，基于反常项集的异常值检测与处理算法具有较高的精度，同时能够避免在数据修复中引入新的异常。

关键词

异常值处理，数据清洗，模式挖掘，反常项集

An Anomalies Processing Algorithm Based on Abnormal Itemsets

Chen Cui, Gui Li, Zhengyu Li, Ziyang Han, Keyan Cao

School of Information & Control Engineering, Shenyang Jianzhu University, Shenyang Liaoning
Email: 724079015@qq.com, ligui21c@sina.com

Received: Apr. 27th, 2021; accepted: May 27th, 2021; published: Jun. 7th, 2021

Abstract

Anomalies refer to the noise and inconsistent values in the data. The detection and processing of anomalies often depend on domain constraints, which usually include conditional functional de-

文章引用: 崔晨, 李贵, 李征宇, 韩子扬, 曹科研. 基于反常项集的异常值处理算法[J]. 数据挖掘, 2021, 11(3): 150-166.
DOI: 10.12677/hjdm.2021.113014

pendencies, negative constraints and editing rules, etc. However, for specific domains, these domain constraint rules need to be made by domain experts, and it is difficult to find these domain constraint rules efficiently based on data mining and machine learning algorithms. In this paper, a concept of abnormal itemset for data cleaning is proposed. Similar to the outlier detection algorithm based on data distribution density, abnormal itemset is an unlikely combination of abnormal values in data. Then, some characteristics of lifting degree are introduced to mine abnormal itemset with low lifting degree by using the improved equivalence class transformation algorithm. Furthermore, this paper proposes an anomalies repair algorithm similar to the nearest neighbor interpolation algorithm to ensure the repair quality. Experiments under the real estate information data set show that the anomalies detection and processing algorithm based on abnormal itemset have high accuracy and will not introduce new anomalies by data repairing.

Keywords

Anomalies Data Processing, Data Cleaning, Pattern Mining, Abnormal Itemset

1. 引言

随着大数据技术的发展，数据规模日渐增大，对数据质量的要求也越来越高，这给数据清洗工作带来了巨大的压力。异常值的检测与处理是数据清洗的主要任务之一。异常值的概念最初由 Hawkins 提出，异常值是一类数据观察值，它们与其他观察数据之间有很大的偏差，从而令人怀疑它们是由不同的机制产生的[1]。异常值，作为影响数据质量的一大重要因素，它包括缺失值、不一致值等，是数据中的主要噪声，对其的处理日渐重要。

最初，在异常值处理中需要大量的人工参与，要么由人工更改和校对，要么由领域专家制定规则。而在数据挖掘和机器学习技术的支持下，最常见的一种推理机制是统计推理，是从数据及其变化做出决定或者得出结论的过程，从而发现数据中的异常现象。有基于统计模型、数据距离和数据密度的异常值检测，也有使用聚类和分类查找异常值的替代方法。

下面提到的这几种方法则对异常给出了恰当的定义，让我们逐个考察下面的方法：1) 基于模型的方法：基于统计模型的方法假设异常值是在概率分布中具有低概率的对象。例如，如果对象服从高斯分布，则与平均值的距离超过阈值的对象就可以被视为异常值。2) 基于邻近度的方法：这种方法通过对象到 k-最近邻的距离来确定异常值。例如，k-最近邻的最小半径被视为对象的异常程度。所以，对象与周围邻居间的最小半径越大，则该对象是一个异常值的程度也越高。3) 基于密度的方法：此方法通过考察对象周围的密度，对密度取倒数来得到衡量参数，以确定对象是否是异常。该值越大，异常的程度就越高。4) 基于聚类的方法：如果对象不属于任何簇，则可以用此方法将该对象确定为异常值。在某些情况下，当对象属于小规模簇时也可以简单地被视为异常值。

在自动发现约束规则的算法中，存在着高维度问题，随着对象属性的数量和数据向量维度的增大，计算量可能会相应地呈指数增长，这样的问题称为维度灾难。

最初的基于规则(又称约束)的异常值处理概念由文献[2]提出，规则由复杂的逻辑所反映，以提供异常值检测与处理的依据。相应地，若符合规则中对异常值的定义，表示数据集中存在异常值，是“脏”的。当数据集中没有满足异常值条件的记录时，数据集是“干净的”。

传统上的基于统计的异常值检测算法，往往局限于发现单属性中的异常值，缺乏复杂性。如果多个属性共同确定的记录是异常记录，在传统的统计模型中是难以发现的。基于 K-means 的异常值检测算法，易受初始聚类中心的影响，且某些离散值属性难以量化，衡量距离的相似度函数有时也难以适应多维的复杂数据集。基于条件函数依赖的规则体系可以解决这一问题，但随着维度的增加，会遇到维度灾难问题。基于机器学习的异常值检测算法虽然可以解决多维度问题，但其发现的规则又过于抽象，难以被领域专家理解，不利于人机协同，难以提高质量。

在关联规则挖掘中，可以借助于项集的扩展来高效地发现复杂模式。项目组合的数量通常极为巨大。以子集作为其元素的集合被称为原始集合的幂集，幂集的大小是原始集合大小的指数函数量级，如果原始集合由 S 表示，则幂集通常由 2^S 表示。因为组合的总数变得非常大，而实际上并不是其中每个子集都有可能发生，所以不考虑零项集，仅考虑反常项集，反常项集是稀有项集的一个子集。在过去的研究中[3]，认为稀有项集与频繁项集是互补关系，它们共同构成幂集 2^S，但在异常值处理研究中，频繁项集仅是常态项集的一个子集，常态项集代表数据集中的“干净”数据记录。

在数据集中，基于频繁出现的项目组合可发现关联规则。关联规则对应着常态数据中频繁出现的部分，而反常项集(Abnormal Item Set, AbNI)指的是常态数据以外的部分。反常项集是一种特殊的稀有项集，用于描述多属性约束规则，它能够反映数据集是“干净”的，还是“脏”的。若反常项集挖掘算法在数据集中发现了新的反常项集，则数据集中仍含有异常值；若运行算法后，未发现反常项集，可认为数据集是“干净”的。而且，反常项集也可作为领域约束的一种，在较为可靠的低错误率数据集中发现的反常项集，可用于领域数据的更正。反常项集在形式上与条件函数依赖[4]和否定约束[5]相似，但用于挖掘条件函数依赖和否定约束的算法效率较低，扩展性较差。并且，反常项集不仅仅能够从数据本身之中发现约束规则，用于检测数据中是否含有异常值，与托梅克(Tomek)连接的思想类似，准反常项集可以用于异常值修复，且不会由修复产生新的反常项集。

本文主要贡献如下：1) 基于统计分析的异常值检测思想和稀有项集，提出了反常项集的概念，可以表示和用于挖掘多维度的复杂约束规则。将异常值检测与约束规则挖掘同步进行，从数据中获取约束规则，弥补了领域约束规则在来源上的不足。2) 提出了属性熵加权项集及相关概念，将提升度的特性推广到了加权项集之上，从而引入了适合挖掘反常项集的剪枝策略。3) 在等价类变换算法的基础上加以改进，将提升度用作项集挖掘的指标，采取逆序遍历的方式，并将加权调和提升度的性质用于剪枝和优化，使之能够高效且准确地挖掘反常项集。4) 在异常值更正的过程中，引入了准反常项集，确保在异常值更正的过程与更正的结果中不会引入新的不一致，同时也提高了效率。并给出了相应的异常值更正算法。

本文其他部分组织如下：第二节介绍了异常值处理与反常项集挖掘领域的相关研究工作；第三节基于信息熵为属性加权，给出了加权项集概念及理论，并引入了加权调和提升度的性质；第四节给出了反常项集挖掘算法；第五节介绍了准反常项集的概念、相关性质和挖掘算法，并给出了异常值更正算法；第六节进行了实验评估，在第七节中进行了总结。

2. 相关工作

在异常值检测与处理的研究中：文献[6]以函数依赖为基础，提出了一种通过数据集与函数依赖规则间的相对信任度来修复异常值或更正规则的方法。文献[7]综合了(近似)函数依赖与关联规则，提出了数据质量规则的概念，用于异常值检测。文献[8]利用编辑规则和主数据进行数据清洗，得到确定的数据修复，并利用条件函数依赖来填补缺失值。文献[9]将关联规则与最近邻算法相结合，用于数据的填补。但在规则获取与异常值处理的过程中，这些算法都有一定的局限性，(近似)函数依赖与条件函数依赖所涉及的维度较低，而且在挖掘过程中效率低下。现阶段编辑规则和主数据的获取仍主要依靠领

域专家的工作，目前在无监督下的约束规则挖掘算法研究较少。虽然以频繁项集为主的关联规则，挖掘效率较高，但难以表达数据中的异常，不能直接用于异常值处理。本文受到上述研究的启发，从而提出了反常项集的概念。

与本文提出的反常项集相类似的是稀有项集，文献[10] [11] [12] [13]分别提出了 4 种具有代表性的稀有项集挖掘算法，其中 Apriori-Rare 算法[10]先将所有的最小稀有项集挖掘出来，再自底向上地产生稀有项集；AfRIM 算法[11]在与关联规则挖掘的 Apriori 算法相反的方向上进行搜索，由稀有(k+1)-项集逐级生成稀有 k-项集，并在搜索空间的剪枝中采用 Apriori 算法思想；Walky-G 算法[12]为减少概念格空间中对频繁项集的搜索，采用了深度优先搜索；MRSP 算法[13]是一种基于大量事务的高效最小稀有模式挖掘算法。文献[1]指出了稀有项集与异常值检测间的联系。但稀有项集的范围仍旧太大，会将一部分频率较低的常态数据误判为异常。而且，这些算法中仅考虑了支持度，没有考虑提升度，后者揭示了项集中的相关性程度，而这正是异常值处理中应当重点考虑的性质之一。

3. 加权调和提升度及其性质

在本节中，首先利用信息熵为来自不同属性的项赋予不同的权重，接下来将传统的模式挖掘相关概念和理论推广到加权项集挖掘上，最后给出加权调和提升度的概念和性质。

3.1. 基于信息熵的属性权重

在传统的项集挖掘算法中，不同的项都处于平等的地位，而在实际的数据集当中，不同属性的重要性是不同的，因此需要将数据集中的不同属性赋予不同的权重。可由领域专家为属性规定权重值，但这么做易受主观因素影响。在本文中，利用属性信息熵，以无监督的方式为数据集当中的各属性分配权重。

信息熵(Entropy)用以度量随机变量序列中所包含的信息量大小，是一种衡量信息不确定程度的指标。信息熵越大，随机变量序列中的离散取值越多，随机变量的分布越不均匀，该随机变量序列中的不确定性越大。本文中，以属性作为随机变量序列，计算数据集当中各属性的信息熵，作为属性权重。

定义 1：属性信息熵 AEntropy 为属性中各取值提供的信息量之和，即

$$H\left(A_k\right)=\text{AEntropy}\left(A_k\right)=-\sum_{i=1}^{n}P\left(a_i\right)\log_2 P\left(a_i\right) \tag{1}$$

其中，A_k 表示数据集 $D=\left\{A_1\cup\cdots\cup A_k\cup\cdots\cup A_m\right\}$ 中的第 k 个属性$(1\le k\le m)$，a_i 是属性 A_k 值域中的第 i 个可取值$(1\le i\le n)$，$P\left(a_i\right)$为属性 A_k 中取值 a_i 出现的概率。

将属性信息熵标准化，得到属性权重 ω：

$$\omega_k=\frac{\text{AEntropy}\left(A_k\right)-\text{AEntropy}_{\min}}{\text{AEntropy}_{\max}-\text{AEntropy}_{\min}}+1 \tag{2}$$

其中，$\text{AEntropy}_{\max}$ 和 $\text{AEntropy}_{\min}$ 分别为 $D=\left\{A_1,\cdots,A_k,\cdots,A_m\right\}$ 中 m 个属性中属性信息熵的最大值和最小值，ω_k 为第 k 个属性的权重。信息熵的取值范围是 $0\le\text{AEntropy}\left(A_k\right)\le\log_2 n$，公式(2)将权重限制到了 $1\le\omega_k\le 2$ 范围内。

对于项集 M 的信息熵，可由属性联合信息熵来定义。将信息熵推广到两个及两个以上属性上，得到联合信息熵的概念：

$$H\left(M\right)=H\left(A_1,A_2,\cdots,A_n\right)=-\sum_{x\in A_1}\sum_{x\in A_2}\cdots\sum_{x\in A_n}P\left(a_1,a_2,\cdots,a_n\right)\log_2 P\left(a_1,a_2,\cdots,a_n\right) \tag{3}$$

其中，$P\left(x_1,x_2,\cdots,x_n\right)$为 n 个属性的联合概率分布函数。

当项集扩展到较高维度时，涉及的属性较多，属性联合信息熵所需的计算量将大幅提高。为此，采取一定程度的近似。

项集的属性联合信息熵具有如下性质，

$$\max\left[H(A_1),H(A_2),\cdots,H(A_n)\right]\le H(A_1,A_2,\cdots,A_n)\le H(A_1)+H(A_2)+\cdots+H(A_n) \tag{4}$$

由(4)可得，属性联合信息熵不小于该项集中任一属性的信息熵，不大于该项集中所有属性信息熵之和。可取的项集属性联合信息熵近似值为

$$H(M)^*=H(A_1,A_2,\cdots,A_n)^*=\max\left[H(A_1),H(A_2),\cdots,H(A_n)\right]+\sqrt[n]{\prod_{i=1}^{n}H(A_i)} \tag{5}$$

即项集 M 中的最大属性信息熵加上属性信息熵的几何平均数。

接下来，与公式(2)类似，可计算出项集的权重 ω_M。

3.2. 加权项集相关概念

首先，在异常值处理的应用场景下，对传统的模式挖掘相关概念做一下回顾。

异常值处理的研究对象，数据集 D 是由有限组对象构成的。其中的每个对象 o 用 $\langle \text{oid}, M\rangle$ 对来表示，其中 oid 为对象标识符(以自然数表示)，M 为项集。而其中的每个项集 M 又是由项 item 构成的集合，项为(属性，值)对，表示为 (A,v)，A 为属性，是属性集 $\mathcal{A}$ 中的元素，v 是 A 中有限类别值域 $\text{dom}(A)$ 中的已知值之一。项集中至多包含 $\mathcal{A}$ 中某属性值域中的一个元素。若项集 I 中的项 $(A,v)\in M$，将对象 $o=\langle \text{oid}, M\rangle$ 在属性 A 中的值定义为 v，记为 $v=o[A]$，否则令 $o[A]$ 未定义。并且，将 $\mathcal{I}$ 定义为属性-值对 (A,v) 的集合。

接下来，介绍一些模式挖掘中的基础概念和本文中所用的符号：

支持：若 $N\subseteq M$，即 N 包含于 M，则称对象 $o=\langle \text{oid}, M\rangle$ 支持项集 N。

包含集：N 在 D 中的包含集，为数据集 D 中支持 N 的对象所对应的 oid 集合，表示为 $\text{cov}(N,D)$。也就是说，符合某一规则的样本称为对该规则的包含。$\text{cov}(N,D)=\{i:1\le i\le |D|\wedge N\in D_i\}$，并且，包含集的基数为支持度计数 $\text{SupCnt}(N,D)=|\text{cov}(N,D)|$。

支持度：项集 N 在数据集 $\mathcal{D}$ 中的支持度是包含集中元素个数(即 oid)与数据集 D 中对象个数之比，$\sup(N,D)=|\text{cov}(N,D)|/|D|$。

提升度：基于概率论，提升度的定义源于这样的事实：若 $P(A\cup B)=P(A)P(B)$，则项 A 的出现完全独立于项 B 的出现，它们之间没有相关性，此时 $\text{lift}(A,B)=1$；若 $P(A\cup B)\ne P(A)P(B)$，则项 A 和项 B 是依赖的和相关的。A 与 B 间的提升度由下式得出：

$$\text{lift}(A,B)=\frac{P(A,B)}{P(A)P(B)}=\frac{P(B\mid A)}{P(B)}=\frac{\text{conf}(A\Rightarrow B)}{\sup(B)} \tag{6}$$

其中，$\sup(B)$ 为项 B 的支持度，$\text{conf}(A\Rightarrow B)$ 为置信度，$\text{conf}(A\Rightarrow B)=P(A\mid B)$。除了 $\text{lift}(A,B)=1$ 的情形外，若 $\text{lift}(A,B)>1$，则项 A 与项 B 间是正相关的，意味着其中任一个的出现都蕴含着另一个的出现；若 $\text{lift}(A,B)<1$，则项 A 与项 B 间是负相关的，意味着一个的出现可能导致另一个不出现。

垂直数据格式：数据集 $\mathcal{D}$ 可表示为垂直数据布局 D_{vert}。$D_{vert}=\{(i,\text{cov}(\{i\},D))\mid i\in\mathcal{I},\text{cov}(\{i\},D)\ne\Phi\}$。

基于上述理论，得出关于加权项集挖掘的一些概念，如下所述：

定义 2：给定项集 $M=\{m_1,m_2,\cdots,m_n\}$，项集 M 的加权支持度计数为 $\text{wSupCnt}(M,D)=\omega_M|\text{cov}(M,D)|$，其中 $|\text{cov}(M,D)|$ 为项集 M 的支持度计数，项集 M 的加权支持度为 $\text{wSup}(M,D)=\dfrac{\text{wSupCnt}(M,D)}{|D|}$。

在模式挖掘的算法中，尽管支持度是一种常用的关联度指标，可是并不适合于挖掘反常项集。除了频繁项集之外，数据集中的大多数项集的出现频度都远低于最小支持度阈值，使用过低的最小支持度阈值去挖掘长模式时，往往不能取得预期的效果。在本文中，采用相关分析方法，利用提升度来挖掘反常项集。

给定一组项单独发生的频率，提升度揭示了它们同时出现的可能性。前件和后件之间具有高提升度的规则被认为是最有趣的，提升度不仅仅用于关联规则的挖掘，还可用于约束发现中。继而，在所有单个项之间都具有“完全”的独立性的前提下，可将提升度从规则直接扩展到项集。也就是说，当 $\mathrm{wSup}(M,D)$ 远小于 $\mathrm{wSup}(\{m_1\},D)*\cdots*\mathrm{wSup}(\{m_n\},D)$ 时(其中 $m_1,m_2,\cdots,m_n$ 为 M 中的项)，M 即被视为不可能出现，将此项集 M 视为异常。但是，在许多情况下，具有轻微负相关的项目，与具有强烈负相关的项目相比，其提升度可能更低。这样，较大的项集，由于其在数据集中出现的频率相对较低，有被误判为异常的可能。

为了解决上述问题，采用分块独立来代替完全独立的概念。并引入调和提升度，调和提升度不受事务总个数的影响。而且，调和提升度的本质是支持度和提升度的几何平均数，这使得能够兼顾支持度与提升度的优点。在本文中，由于需要挖掘的是低提升度的反常项集，利用调和提升度可以避免较低支持度的项集被当作反常项集，从而提高所发现约束规则的质量。

定义 3：给定数据集 D，项集 M，N 为 M 的非空子集。项集 M 的加权调和提升度 $\mathrm{wHlift}(M,D)$，定义为：

$$\begin{aligned}\mathrm{wHlift}(M,D)&=\max_{\Phi\subset N\subset M}\left\{\frac{\mathrm{wsup}(M,D)}{\sqrt{\mathrm{wsup}(N,D)*\mathrm{wsup}(M\setminus N,D)}}\right\}\\&=\frac{\mathrm{wSupCnt}(M,D)}{\min\limits_{\Phi\subset N\subset M}\left\{\sqrt{\mathrm{wSupCnt}(N,D)*\mathrm{wSupCnt}(M\setminus N,D)}\right\}}\end{aligned}\tag{7}$$

其中，$M\setminus N$ 为集合减运算，目的是求出 N 在 M 中的差集，即所有属于 M 但不属于 N 的元素构成的集合。公式(7)的分母为项集 M 的所有分区方式中加权支持度计数中几何平均数 $\sqrt{\mathrm{wSupCnt}(N,D)*\mathrm{wSupCnt}(M\setminus N,D)}$ 中的最小值。本文在不引起混淆的情况下，将其简称为提升度。

定义 4：加权调和提升度 $\mathrm{wHlift}(I,D)\le\lambda$ 的项集称为反常项集。使用提升度做反常项集挖掘时，λ 通常很小，如数据质量较高，可取 $\lambda=0.01$。

3.3. 加权调和提升度的性质

在提出反常项集挖掘算法之前，先介绍一下加权调和提升度的一些性质。这些性质将用于算法 1 的优化和对深度优先搜索树的剪枝。

在模式挖掘算法的剪枝策略中，单调性和反单调性经常被加以利用。单调性是指：如果一个项集满足某个规则约束，则这个项集的所有超集也满足该规则约束。而反单调性是指：如果一个项集不满足某个规则约束，则它的所有超集也不能满足该规则约束。在模式挖掘的过程中，如果基于当前模式，某数据项不满足数据反单调性约束，则可以将其剪枝掉。

虽然加权调和提升度既不是单调的，也不是反单调的，它仍然有一些可用于在搜索树中进行大范围剪枝的性质：因为低提升度项集比其子集出现的频率低得多，可利用反常项集的支持度与其子集的支持度之间的关系来进行剪枝操作。

定义 5：给定项集 M，用 $\mathrm{SupCnt}_M^{\max}$ 表示 M 及其所有超集 L 中项 $\{i\}$ 的最高加权支持度计数，也就是说，$\mathrm{SupCnt}_M^{\max}$ 是深度优先搜索树中分枝 M 上的最高加权支持度计数，即

$$\mathrm{wSupCnt}_M^{\max} = \max\left\{\mathrm{wSupCnt}\left(\{i\},D\right)\right\},\left(i \in L, M \subseteq L\right)$$

在定义 5 的基础上，可将文献[14]中的提升度性质推广到加权项集上：

推论 1：给定两项集 M 和 L，若 M 是 L 的子集 $M \subset L$，且 L 是支持度上限为 λ 的反常项集，则 $\mathrm{wSupCnt}\left(M,D\right) \geq \dfrac{\mathrm{wSupCnt}\left(L,D\right)}{\left(\mathrm{wSupCnt}_M^{\max} * \lambda\right)^2}$。

当 L 为反常项集时，可以用 $\mathrm{wSupCnt}_M^{\max}$ 来确定 L 的子集加权支持度计数的下界。由于反常项集是在数据集中真实存在的，任意一个反常项集的加权支持度计数都符合 $\mathrm{wSupCnt}\left(L,D\right) \geq 1$。对于反常项集 L 的所有子集 M，($M \subset L$)，都有 $\mathrm{wSupCnt}\left(M,D\right) \geq \dfrac{1}{\left(\mathrm{wSupCnt}_M^{\max} * \lambda\right)^2}$。这意味着在深度优先搜索过程中，如项集 M 的支持度计数满足 $\mathrm{wSupCnt}\left(M,D\right) \geq \dfrac{1}{\left(\mathrm{wSupCnt}_M^{\max} * \lambda\right)^2}$，则项集 M 是可扩展的。此外，推论 1 还表明反常项集与其子集之间的支持度差距是极小的。

推论 2：对于 3 个项集 L、M 和 N，满足 $L \supseteq M \supset N$，如果 L 是反常项集，其加权调和提升度阈值为λ，则 $\sqrt{\mathrm{wSupCnt}\left(N,D\right)} - \sqrt{\mathrm{wSupCnt}\left(M,D\right)} \geq \sqrt{\dfrac{1}{\lambda}} - \sqrt{\mathrm{wSupCnt}_N^{\max}} > 0$。

特别地，如 M 为反常项集，则对它的所有子集，推论 2 中的不等式依然成立。在深度优先搜索过程中，由项集 N 扩展到项集 M 时，如果不满足该式，则 M 和它的所有超集 L 都应被剪枝掉。此外，推论 2 还意味着反常项集具有生成元的性质，即为，如果 M 为反常项集，则对于任何 $N \subset M$，$\sqrt{\mathrm{wSupCnt}\left(N,D\right)} > \sqrt{\mathrm{wSupCnt}\left(M,D\right)}$。由于生成元的所有子集都是生成元，这意味着如果在深度优先搜索过程中遇到非生成元，则可以剪枝掉整个子树。

推论 3：给定项集 N 和 M，若 N 为 M 的子集(即 $N \subset M$)，则：

$$\mathrm{wHlift}\left(M,D\right) \geq \frac{\mathrm{wSupCnt}\left(M,D\right)}{\min\limits_{\Phi \subset P \subset M}\left\{\sqrt{\mathrm{wSupCnt}\left(P,D\right) * \mathrm{wSupCnt}\left(N \setminus P,D\right)}\right\}}$$

由于 $\mathrm{wSupCnt}\left(M,D\right) \geq 1$，项集 M 的加权调和提升度与其子集加权支持度计数间的关系为 $\mathrm{wHlift}\left(M,D\right) \geq \dfrac{1}{\sqrt{\mathrm{wSupCnt}\left(P,D\right) * \mathrm{wSupCnt}\left(N \setminus P,D\right)}}$。随着项集向更高维度的扩展，项集的基数随之增大，分母 $\min_{\Phi \subset P \subset M}\left\{\sqrt{\mathrm{wSupCnt}\left(P,D\right) * \mathrm{wSupCnt}\left(N \setminus P,D\right)}\right\}$ 也相应地单调增大。

推论 4：如果 N 是反常项集，则其加权支持度 $\mathrm{wsup}\left(N,D\right) \leq \dfrac{\lambda}{2-\lambda}$。

4. 反常项集挖掘算法

在上述理论的支持下，给出在数据集 D 中挖掘反常项集的算法。该算法是在频繁项集挖掘领域著名的等价类变换(Eclat)算法基础上改进而来的。

Eclat 算法是由 J. Zaki 在 2000 年提出的，与传统的 Apriori 和 FP-growth 算法不同，Eclat (Equivalence Class Transformation，等价类变换)算法以概念格[15]理论为基础，利用倒排索引的思想进行数据统计，将数据转换为垂直数据表示方式，在等价类划分的思想下执行深度优先搜索来挖掘频繁项集。在垂直数据表示下，由于 TID 集中保存了计算支持度的完整信息，仅需对数据集进行一次数据扫描得出数据项的事务编号，再对其进行求交集操作，通过交集中的元素个数即可得出项集的支持度。倒排索引是一种更适

合于做关系运算的数据结构，Eclat 算法将记录划分到每个项之下，使得利用集合的简单运算就能获得所需的支持度。

为挖掘反常项集，对经典 Eclat 算法做出了如下改进[16]：

1) 在 Eclat 算法中集成了加权调和提升度计算。在深度优先搜索概念格空间时，采取了逆序遍历的方式，使得项集 M 的所有子集先于项集 M 被遍历到，并利用额外的前缀树来存储支持度。而且，若项集 M 的支持度没有存储于前缀树中，意味着项集 M 的子集被剪枝掉了，项集 M 的子集不是反常项集，那么项集 M 和它的所有超集也不是反常项集。从而将项集 M 所在的分枝剪枝掉。

2) 将 3.3 节中加权调和提升度的性质应用于 Eclat 算法中，以进行剪枝和优化。

算法 1：基于改进 Eclat 的反常项集挖掘算法

输入：数据集 D 的垂直数据表示形式 $D_{\text{vert}}[P]$，提升度阈值 λ

输出：反常项集 AbNI

procedure AbNIMiner$\left(D_{\text{vert}}[P], P \subseteq \mathcal{I}, \lambda\right)$

  AbNI $\Leftarrow \varnothing$

  for all $m \in \mathcal{I}$ occurring in $D_{\text{vert}}[P]$ in reverse order **do**//depth-first search

    $N \Leftarrow P \cup \{n\}$ //expanding the itemset N

    if not isGenerator(N) **then**

      continue

    storeGenerator(N)

    if $|N|>1$ and $\text{wSup}(N,D) \le \dfrac{\lambda}{2-\lambda}$ **then**

      if a subset of Nhas been pruned **then**

      continue

      if $\text{wHlift}(N,D) \le \lambda$ **then**

      AbNI $\Leftarrow$ AbNI $\cup \{N\}$

      If $\lambda^{-1} > \min\limits_{S \subset N}\left\{\sqrt{\text{wSupCnt}(S,D) * \text{wSupCnt}(N \setminus S,D)}\right\}$ **then**

      continue

    if $\text{wSupCnt}(N,D) \ge \dfrac{1}{\left(\text{wSupCnt}_N^{\max} * \lambda\right)^2}$ **then**

      continue

    $D_{\text{vert}}[N] \Leftarrow \varnothing$ //generating the next layer of the searching tree

    for all $m \in \mathcal{I}$ in Dsuch that $m > n$ **do**

      $C \Leftarrow \text{cov}(\{n\},D) \cap \text{cov}(\{m\},D)$

      $M \Leftarrow N \cup \{m\}$

      $\text{wSupCnt}(M,D) \Leftarrow \omega_C |C|$

      if $\sqrt{\text{wSupCnt}(N,D)} - \sqrt{\text{wSupCnt}(M,D)} \ge \lambda^{-\frac{1}{2}} - \sqrt{\text{wSupCnt}_N^{\max}}$ **then**

Continued

23	**if** wSupCnt$(M,D)>0$ **then**
24	$D_{\text{vert}}[N] \Leftarrow D_{\text{vert}}[N] \cup \{(m,C)\}$
25	AbNI $\Leftarrow$ AbNI $\cup$ AbNIMiner$(D_{\text{vert}}[N], N, \lambda)$
26	**return** AbNI

反常项集挖掘算法如算法 1 所示。算法 3-25 行的循环为在概念格空间中进行深度优先搜索。在搜索树中扩展项 N 时(第 4 行)时，仅需将项 N 扩展为数据集中所有出现在 N 包含集当中的项(第 18~24 行)，就可以生成新的项集。此外，这些新加项是依据项的总排序添加的，即，仅当新项处于包含 N 的项集之后才添加该项(第 18 行)。而为提高反常项集挖掘的效率，项是依据支持度升序来排列的。算法 1 的第 9 行表示如果项集的子集在之前已被剪枝掉，导致其加权支持度未被存储，无法计算该项集的加权调和提升度，将该项集连同其超集一同剪枝掉。第 13 行利用了推论 3 进行剪枝，第 15 行对应于推论 1，第 22 行将推论 2 用于剪枝，第 8 行利用推论 4 减少了加权调和提升度的计算量。

由推论 2 可知反常项集具有生成元的性质，即反常项集的支持度应严格低于其所有子集的支持度，算法 1 第 5~7 行通过过程 isGenerator()和 storeGenerator()实现了对非生成元处理，由于非生成元的所有超集都是非生成元，可以将整个子树剪枝掉。这两个过程与 Talky-G 算法[14]和 Charm 算法[15]类似，采用了基于散列的方法，将对象的 oid 之和作为哈希值。

5. 异常值更正算法

第 4 节中的算法用于在数据集当中挖掘反常项集，以发现约束规则和检测出异常值。当数据集中存在反常项集时，数据集 D 被认为是脏的。接下来将提出异常值更正的算法。假设对数据集 D 作出的异常值更正为 D^*，异常值更正应遵循如下原则：1) D^*是干净的，也就是说，在 D^*中不应当存在反常项集；2) D^*与 D 的差异应当很小，即 D^*仅对 D 做最小化的修改。

本文的异常值更正算法的思想与托梅克(Tomek)连接相似。托梅克(Tomek)连接，由数学家 Tomek 提出，指一个样例对，x 和 y，如果同时满足如下 3 个要求，则称它们形成了一个托梅克连接：1) x 是 y 的最近邻。2) y 是 x 的最近邻。3) x 和 y 的类别不同。托梅克(Tomek)连接指出了边界样例所应具备的条件，对于异常样例的处理，通常会遇到困难，因为处理掉原有托梅克连接后，有可能会在数据集当中产生新的托梅克连接，使得该过程需要迭代运行多次。这与本文中的异常值处理相似，正如前文所述，通过反常项集发现了由多个属性共同确定的异常取值组合，可以实现异常值检测。但在异常值更正中，情况更为复杂，需要找到反常项集与正常项集的边界，即准反常项集，通过与异常值差别最小的正常值来进行异常值更正。

本节首先提出准反常项集和最小可能提升度定义，再由推论 1 到推论 4 提出用于在挖掘准反常项集时进行剪枝优化的特性，最后提出异常值更正算法。

5.1. 准反常项集

反常项集可以被看做是由多个属性确定的异常取值组合，具体取到异常值的属性是未知的。异常值更正算法通过在与反常项集最为相似的干净数据(正常项集)中找到针对个别值的更正建议，来进行异常值更正。将数据集当中每个反常项集视为异常，并将它们构成的集合视为异常值集合 D_{dirty}。给定数据集 D 及 D_{dirty}，D_{dirty} 在 D 中的补集为 D_{clean}。经过异常值更正，数据集 D 变为 D^*。异常值更正过程可能引入新的异常值，为了避免这一不足，其中一种做法是每进行一次候选更正，都对新的 D^*运行一次反常项集挖掘算法，若发现了与更正操作有关的反常项集(更正后仍为反常项集或因更正引入了新的反常项集)，就将

该更正操作视为无效更正。但这种做法的直接后果是大幅提高计算量。为解决这一问题，可在 D_{dirty} 与 D_{clean} 的边界处寻找准反常项集，准反常项集指数据集 D 中经最多 k 次更正后可能变为异常的项集。准反常项集的本质还是数据集中的“干净”数据，而且，对准反常项集的挖掘过程仅涉及到数据集 D 及其中的异常数据，而与对异常值的具体更正操作无关。

定义 6：准反常项集。对数据集 D，经过至多 k 次更正后，得到 D^*，如果 D^*中出现了由更正操作引入的新反常项集，则称这样的反常项集为准反常项集，用 QAbNI(Quasi AbNI)表示。

通过准反常项集，可以预计算出针对数据集 D 的至多 k 次更正。准反常项集挖掘算法与算法 1 类似，但采用了较为宽松的提升度概念，即经 k 次更正后项集 M 的最小可能加权调和提升度，用 mpwHlift 表示。接下来，将提出 mpwHlift 的定义，并由推论 1 到推论 4 得出其性质。

定义 7：(最小可能提升度)给定数据集 D，其中的项集 M，设

$$\text{wHlift}(M,D)=\text{wSupCnt}(M,D)\Big/\min_{\Phi\subset N\subset M}\sqrt{\text{wSupCnt}(M,D)*\text{wSupCnt}(M\setminus N,D)}$$

且 $\text{wSupCnt}(N,D)\le\text{wSupCnt}(M\setminus N,D)$，其中 $N\subset M$。则经 1 次数据更正后，

$$\text{mpwHlift}(M,N,1)=\min\left\{\frac{\text{wSupCnt}(M,D)-1}{\sqrt{\text{wSupCnt}(N,D)*\left[\text{wSupCnt}(M/N,D)-1\right]}},\frac{|D|*\text{wSupCnt}(M,D)}{\sqrt{\left[\text{wSupCnt}(N,D)+1\right]*\text{wSupCnt}(M/N,D)}}\right\}$$

经迭代计算 k 次，可将其推广到 k 次更正后的最小可能提升度，表示为 $\text{mpwHlift}(M,N,k)$。

由于干净的对象不应被修改，所以在任一 D^*中对项集的支持度都有严格的限制，设 D^*由 D 经过最多 k 次更正后得到，对于任何项集 N，都应当符合：

$$\text{wSupCnt}(N,D_{\text{clean}})\le\text{wSupCnt}(N,D^*)\le\text{wSupCnt}(N,D_{\text{clean}})+k$$

并且，还必须考虑 $\text{wSupCnt}(N,D_{\text{clean}})=0$ 的项集 N *[16]*。

推论 5：如定义 7，已知项集 M，其子集为 N，即 $N\subset M$，对数据集 D 经最多 k 次修改后得到已修复数据集 D^*。若 $M\in\text{AbNI}(D^*,\lambda)$，有 $\text{mpwHlift}(M,N,k)\le\lambda$。

推论 6：对任意两个符合 $N\subset M$ 的项集 N 和 M，若 M 是数据集 D^*中的反常项集，有

$$\text{wSupCnt}(N,D_{\text{clean}})\ge\frac{\text{wSupCnt}(M,D^*)}{\left(\text{wSupCnt}^{\max}_{N,D^*}*\lambda\right)^2}-k$$

与反常项集挖掘算法的剪枝策略类似，这里取加权支持度计数下限 $\text{wSupCnt}(M,D^*)=1$。对任何更正后的数据集 D^*，$\text{wSupCnt}^{\max}_{N,D^*}$ 满足不等式 $\text{wSupCnt}^{\max}_{N,D^*}\le\text{wSupCnt}^{\max}_{N,D_{\text{clean}}}$。为了在 $\text{wSupCnt}^{\max}_{N,D^*}$ 未知的情况下进行剪枝，由推论 6，当 $\text{wSupCnt}(N,D_{\text{clean}})<\frac{|D|}{\left(\text{wSupCnt}^{\max}_{N,D_{\text{clean}}}+k\right)^2*\lambda^2}-k$ 时，可将项集 N 的超集剪枝掉。

推论 7：对任意三个符合 $N\subset M\subset L$ 的项集 N、M 和 L，若 L 是 D^*中的反常项集，有

$$\sqrt{\text{wSupCnt}(N,D_{\text{clean}})}-\sqrt{\text{wSupCnt}(M,D_{\text{clean}})}\ge\sqrt{\frac{1}{\lambda}}-\sqrt{\text{wSupCnt}^{\max}_{M}}-\frac{k}{|D|}$$

由推论 7，可得到基于生成元的剪枝策略，若项集 N 符合条件 $\sqrt{\frac{1}{\lambda}}-\sqrt{\mathrm{wSupCnt}_{N,D^*}^{\max}}>\frac{k}{|D|}$，可将项集 N剪枝掉。在具体的数值计算中，不必为每个项集 N 计算 $\mathrm{wSupCnt}_{N,D^*}^{\max}$，可用 $\mathrm{wSupCnt}_{\varnothing,D^*}^{\max}$ 去估计，且 $\mathrm{wSupCnt}_{\varnothing,D^*}^{\max}=\mathrm{wSupCnt}_{\varnothing,D_{\mathrm{clean}}}^{\max}+k$。而且，若 M 是 D^*经过 k 次更正后得到的反常项集，对于 M 的子集 N ($N\subseteq M$)，必有 $\mathrm{wSupCnt}\left(N,D_{\mathrm{clean}}\right)-\mathrm{wSupCnt}\left(M,D_{\mathrm{clean}}\right)\geq\sqrt{\frac{1}{\lambda}}-\sqrt{\mathrm{wSupCnt}_{\varnothing,D_{\mathrm{clean}}}^{\max}+k}-\frac{k}{|D|}$。

推论 8：对于任两个符合 $N\subset M$ 的项集 N 和 M，都有

$$\mathrm{wHlift}\left(M,D^*\right)\geq\frac{\mathrm{wSupCnt}\left(N,D^*\right)}{\min\limits_{S\subset I}\left\{\sqrt{\mathrm{wSupCnt}\left(S,D_{\mathrm{clean}}\right)+k}*\sqrt{\mathrm{wSupCnt}\left(N\setminus S,D_{\mathrm{clean}}\right)+k}\right\}}$$

推论 8 由推论 3 而来，以将可能的更正考虑在内。由于最小可能调和提升度 mpwHlift 与 $\mathrm{wSupCnt}\left(M,D^*\right)$ 及对项集的划分方式相关，mpwHlift 的分母不具有反单调性。推论 8 确定了 mpwHlift 的分母在更正效果最差情况下的增量。

5.2. 准反常项集挖掘算法

准反常项集挖掘算法是在反常项集挖掘算法的基础上，采取较为宽松的提升度和不同的剪枝策略后得到的。在反常项集挖掘算法的框架下，算法 2 以推论 5 到推论 8 为基础，来进行剪枝和优化。

算法 2：准反常项集挖掘算法

输入：数据集 D 的垂直数据表示形式 $D_{\mathrm{vert}}[P]$，提升度阈值 λ，更正操作数量 k

输出：准反常项集 QAbNI

procedure QAbNIMiner$\left(D_{\mathrm{vert}}[P],P\subseteq\mathcal{I},\lambda\right)$
  QAbNI $\Leftarrow\varnothing$
  for all $m\in\mathcal{I}$ occurring in $D_{\mathrm{vert}}[P]$ in reverse order **do**//depth-first search
    $N\Leftarrow P\cup\{n\}$ //expanding the itemset N
    if $k\leq|D|\left(\lambda^{-\frac{1}{2}}-1\right)$ **then**
    if not isGenerator(N) **then**
      continue
    storeGenerator(N)
    if $|N|>1$ and $\mathrm{wSup}\left(N,D_{\mathrm{clean}}\right)\leq\frac{\lambda}{2-\lambda}$ **then**
      if a subset of N has been pruned **then**
      continue
      if $\mathrm{mpwHlift}\left(N,D\right)<\lambda$ **then**
      QAbNI $\Leftarrow$ QAbNI $\cup\{N\}$
      If $\frac{1}{\lambda}>\min\limits_{S\subset N}\left\{\sqrt{\mathrm{wSupCnt}\left(S,D_{\mathrm{clean}}\right)+k}*\sqrt{\mathrm{wSupCnt}\left(N\setminus S,D_{\mathrm{clean}}\right)+k}\right\}$ **then**

Continued

15 **continue**

16 **if** $\mathrm{wSupCnt}(N, D_{\mathrm{clean}}) \geq \frac{|D|}{\left(\mathrm{wSupCnt}^{\max}_{N, D_{\mathrm{clean}}} + k\right)^2 * \lambda^2} - k$ **then**

17 **continue**

18 $D_{\mathrm{vert}}[N] \Leftarrow \varnothing$ //generating the next layer of the searching tree

19 **for all** $m \in \mathcal{I}$ in Dsuch that $m > n$ **do**

20 $C \Leftarrow \mathrm{cov}(\{n\}, D) \cap \mathrm{cov}(\{m\}, D)$

21 $M \Leftarrow N \cup \{m\}$

22 $\mathrm{wSupCnt}(M, D) \Leftarrow \omega_C |C|$

23 $\mathrm{wSupCnt}(M, D_{\mathrm{clean}}) \Leftarrow \mathrm{wSupCnt}(M, D) - \mathrm{wSupCnt}(M, D_{\mathrm{dirty}})$

24 **if** $\sqrt{\mathrm{wSupCnt}(N, D)} - \sqrt{\mathrm{wSupCnt}(M, D)} \geq \lambda^{\frac{1}{2}} - \sqrt{\mathrm{wSupCnt}^{\max}_{M}} - \frac{k}{|D|}$ **then**

25 $D_{\mathrm{vert}}[N] \Leftarrow D_{\mathrm{vert}}[N] \cup \{(m, C)\}$

26 $\mathrm{QAbNI} \Leftarrow \mathrm{QAbNI} \cup \mathrm{QAbNIMiner}(D_{\mathrm{vert}}[N], N, \lambda, k)$

27 **return** QAbNI

算法 2 与算法 1 类似，第 14 行对应于推论 6，第 16 行利用了推论 8 进行剪枝，第 24 行将推论 7 用于剪枝。由推论 7 可知准反常项集具有生成元的性质，但仅当 $k \leq |D|\left(\lambda^{-\frac{1}{2}} - 1\right)$ 时，才可对非生成元进行剪枝，算法 2 第 6~8 行通过过程 isGenerator()和 storeGenerator()实现了对非生成元处理。

k 表示进行预更正的次数，其取值与算法 2 的性能有关。接下来，对 k 的取值进行讨论。

由于不考虑对数据集的具体更正操作，该算法也是无监督的，其输入仅为数据集 D 和更正操作数量上限 k。显然 k 最多为 $|D_{\mathrm{dirty}}|$。预期的更正次数 k 对准反常项集挖掘算法的剪枝能力有直接影响。若 k 取最大值，即 $k = |D_{\mathrm{dirty}}|$，推论 6 和推论 7 确定的界限将无法用于剪枝。而且，由于最小可能加权调和提升度 mpwHlift 与k高度相关，mpwHlift 会将较多项集判定为准反常项集。虽然算法只需要运行一次就可以获得准反常项集 Q，并且所有脏对象都可以基于集合 Q 进行修复，但是这个集合会很大，并且由于缺乏剪枝，算法的效率会很低。另一方面，当 $k = 1$ 时，可进行剪枝操作，Q 的大小也是合理的。而此时，若要更正所有异常值，就需要每处理一个异常值运行一次算法。

为使准反常项集挖掘算法在运行时间和运行次数上得到平衡，将 D_{dirty} 划分为大小为 r 的块，并对分块大小 r 进行参数优化。上文中已经讨论了块大小 $r = 1$ 和 $r = |D_{\mathrm{dirty}}|$ 的情形，接下来确定 r 的范围，使剪枝成为可能。

给定符合 $N \subset M$ 的项集 N 和 M，假设 D^* 由数据集 D 经至多 r 次更正后获得，推论 6 仅适用于 $\frac{\mathrm{wSupCnt}(M, D^*)}{\left(\mathrm{wSupCnt}^{\max}_{N, D^*} * \lambda\right)^2} > r$ 的情形，而推论 7 仅适用于 $\sqrt{\frac{1}{\lambda}} - \sqrt{\mathrm{wSupCnt}^{\max}_{N, D^*}} > r$ 的情形。显然，由于 $\frac{\mathrm{wSupCnt}(M, D^*)}{\left(\mathrm{wSupCnt}^{\max}_{N, D^*} * \lambda\right)^2} \geq \sqrt{\frac{1}{\lambda}} - \sqrt{\mathrm{wSupCnt}^{\max}_{N, D^*}}$，$r = \frac{\mathrm{wSupCnt}(M, D^*)}{\left(\mathrm{wSupCnt}^{\max}_{N, D^*} * \lambda\right)^2}$ 为支持剪枝的最大分块大小。另外，

当$\sqrt{\frac{1}{\lambda}}-1>r$时，有$\frac{1}{\sqrt{\text{wSupCnt}_{N,D^*}^{\max}}}\geq 1$，可得块大小$r=\sqrt{\frac{1}{\lambda}}-1$时，算法中将可以进行大规模剪枝。

当然，全局最优的块大小r是不存在的。数据的细节，甚至对象在r个分块中的分布情况，都会影响剪枝能力。同样值得注意的是，上面导出的分块大小仅保证了相关剪枝策略的适用性，但随着分块大小增大，剪枝能力仍会降低。当数据集中的属性数量较少，且取分块大小$r=\frac{\text{wSupCnt}\left(M,D^*\right)}{\left(\text{wSupCnt}_{N,D^*}^{\max}*\lambda\right)^2}$时，算法的运行时间会大大增加，在实验部分，取$r=(2\lambda)^{-1}$。

5.3. 异常值更正算法

正如前文所述，本文的异常值更正算法是在对准反常项集的预计算基础上实现的。与传统的异常值处理算法相比，本算法避免了由更正操作产生新的异常值，无需进行迭代计算。设数据集D中的异常值为o_d，对其所做的修复为o_d^{mod}。对于两种特殊的异常值，采取如下措施：

旧异常值：这些项集不应该出现在修复后的数据集当中，因为它们在D中本来就是异常值。这分为两种情况，包含修复结果o_d^{mod}的项集在反常项集中，即$C\in Q\bigcap\text{AbNI}(D,\lambda)$，其中$C$为包含$o_d^{\text{mod}}$的项集；或者更正产生的项集在原始数据集$D$中不存在，即包含更正结果$o_d^{\text{mod}}$的项集$C\in Q$，但$\text{wSupCnt}(C,D)=0$，也就是说，若它们存在于$D$中，就会成为反常项集。不严谨的修复可能会将这样的项集引入D^*中，因此更正结果o_d^{mod}中不应该存在任何被判为或可能被判为反常的项集Q_{old}。

潜在异常值：设修复后的对象o_d^{mod}包含于项集C且$C\in Q$，其提升度满足$\text{wHlift}(C,D)>\lambda$且$\text{wHlift}\left(C,D_{\text{mod}}^*\right)<\text{wHlift}(C,D)$，称由这样的项集组成的集合为$A_{\text{potential}}$。应避免这样的更正操作，因为该项集属于$D$中的准反常项集，而不是反常项集，更正后的结果与更正前相比，其提升度反而减小了，出现了新的反常项集。其中D_{mod}^*是原始D经修复后的数据集，修复操作仅用o_d^{mod}去替换o_d，而原反常项集中的其他项保持不变。

处理方式：对于旧异常值，当预修复结果中出现了Q_{old}时，所做更正将被视为无效，放弃该更正结果。对于潜在异常值，由于无法计算$Q_{\text{potential}}$中所有项集的加权调和提升度，这里选择了一种更有效的方法，以确保项集$Q_{\text{potential}}$的提升度不会下降，通过断言：1) 不应当将M删除掉(M的存在会减少提升度中的分子)；2) 不应当加入M的真子集(这将增加提升度中的分母)。确保$\text{wSupCnt}(M,D)\geq\text{wSupCnt}\left(M,D^*\right)$且$\text{wSupCnt}(N,D)\leq\text{wSupCnt}\left(N,D^*\right)$，$N\subsetneq M$，使得$\text{wHlift}\left(M,D^*\right)\geq\text{wHlift}(M,D)$。

通过上述针对特殊异常值的处理措施，可以确保在异常值更正后不会出现新的反常项集，使得候选更正有效。

在异常值更正中不会更改D_{clean}，以确保每次修复都是基于真实存在的可信值。对于每个异常值组合，通过从D_{clean}中寻找类似的对象，替换异常值组合中的某些项，从而生成候选更正。对象间的相似度由Linsim相似度[17]来衡量，

$$\text{Linsim}(o,o')=\frac{\sum_{A\in\mathcal{A}}S\left(o[A],o'[A]\right)}{\sum_{A\in\mathcal{A}}\log\left(\text{wSup}\left(\left\{\left(A,o[A]\right)\right\},D\right)\right)+\sum_{A\in\mathcal{A}}\log\left(\text{wSup}\left(\left\{\left(A,o'[A]\right)\right\},D\right)\right)}$$

其中，$S\left(o[A],o'[A]\right)$由

$$S\left(o[A],o'[A]\right)=\begin{cases}2*\log\left(\text{wSup}\left(\left\{\left(A,o[A]\right)\right\},D\right)\right), & o[A]=o'[A]\\ 2*\log\left(\text{wSup}\left(\left\{\left(A,o[A]\right)\right\},D\right)\right)+\log\left(\text{wSup}\left(\left\{\left(A,o'[A]\right)\right\},D\right)\right), & \text{其他}\end{cases}$$

确定。Linsim 相似度同时考虑实际数据中匹配和不匹配的情况，对频繁值的匹配赋予较高的权重，对非频繁值的不匹配赋予较低的权重。

异常值更正算法如算法 3 所示[16]。

算法 3：异常值更正算法

输入： $D_{\text{dirty}}, D_{\text{clean}}, \text{linsim}, \lambda$

输出： D^{*}, D^{**}

1 **for all** $M_i \in \text{blocks}\left(D_{\text{dirty}}, m\right)$ **do**

2 $m := \left|M_i\right|$

3 $D^{*} := \varnothing$ ； $D^{**} := \varnothing$

4 $Q_i = \text{QAbNIMiner}\left(D, m, \lambda, k\right)$

5 **for all** $o_{di} \in M_i$ **do**

6 success:=False

7 **for all** $o_c \in D_{\text{clean}}$ in $\text{sim}\left(o_c, o_{di}\right)$ with descend order **do**

8 $o_{di}^{\text{mod}} = \text{MODIFY}\left(o_{di}, o_c\right)$

9 **if** TRUSTED ($o_{di}, o_{di}^{\text{mod}}, Q_i$) **then**

10 success:=True

11 $D^{*} := D^{*} \bigcup o_{di}^{\text{mod}}$

12 **break**

13 **if** not success **then**

14 $D^{**} = D^{**} \bigcup o_{di}$

15 **return** $\left(D^{*}, D^{**}\right)$

算法 3 的第 7 到第 8 行为对异常值的更正操作，对于每个异常值组合 o_{di}，该算法按照它们与 o_{di} 的相似性顺序逐个匹配干净对象。随后通过过程 Modify(o_{di}, o_c)去更正 o_{di}，在更正中仅替换包含于反常项集中的项。

在算法 3 中的第 12 行，过程 TRUSTED ($o_{di}, o_{di}^{\text{mod}}, A_i$)用于检查候选更正是否是可信的，可信与否是相对于准反常项集而言的。若更正结果可信，将其添加到 D^{*}中；若不可信，将检查下一个候选更正，循环的终止条件是找到可信更正(第 11 行)或所有候选更正均被拒绝(第 13 行)，若无候选更正可用，o_{di} 被添加到无法修复的对象集 D^{**}中(第 14 行)。

6. 实验

6.1. 实验设置

本文用 Python 对“面向领域的 Web 数据抽取”项目所抽取到的房地产数据进行实验，实验所用到的数据集如表 1 和表 2 所示。程序运行环境为 Inteli7-7700HQ，16GB 内存，运行于 Windows10 操作系统。实验结果取 5 次独立运行的平均值。

Table 1. Experimental data set A
表 1. 实验数据集 A

项目名称	区县	环线	商圈	板块	物业类型	产品形态	建筑结构	交屋标准	供暖方式	……
博曼领仕郡	和平区	一环以内	太原街	西塔	商业	高层	钢混	毛坯	地热	……
枫景名城	和平区	一环以内	太原街	太原街	普通住宅	高层	框剪	毛坯	热网	……
皇城酒店公寓	和平区	一环以内	太原街	南市	商住楼	高层	框架	精装修	热网	……
金阳 SOHO	和平区	一环以内	太原街	太原街	写字间	高层	框剪	精装修	中央空调	……
……	……	……	……	……	……	……	……	……	……	……

Table 2. Experimental data set B
表 2. 实验数据集 B

项目名称	使用用途	设计用途	楼房结构	楼栋性质	楼栋类型	建筑类型	电梯	区县	街	路	社区	……
梧桐花园	住宅	住宅	框架	单位自建房	楼房	多层建筑	1	龙岗区	横岗街	梧桐路	红棉	……
安华小区	综合	综合	框架	非自建房	楼房	小高层	0	福田区	沙头街	泰然八路	天安	……
梅兴苑	商业	住宅	框架	非自建房	楼房	小高层	1	福田区	梅林街	梅华路	新兴	……
立新花园	商业	商业	框架	非自建房	楼房	小高层	0	罗湖区	东门街	立新路	立新	……
……	……	……	……	……	……	……	……	……	……	……	……	……

在实验中，由领域专家评估反常项集在异常值检测中的准确性，和在异常值更正中的合理性。当给定的λ值较小时，异常值处理的精度很高。在实验中，具体的 λ 值依数据质量而定，取值介于 0.01 到 0.05 之间，本文实验中取 $\lambda = 0.01$。较小的 λ 值不仅能降低算法的计算量，还可避免误报。

6.2. 评价指标

为了评估本文中反常项集挖掘算法及异常值更正算法的可行性和有效性，并与其他方法进行比较，本文采用精确度(precision)、召回率(recall)、f1-score 作为评价指标。由领域专家评估算法所发现约束规则的准确性、合理性，并为实验提供一部分经验规则作为参考标准。在实验中，对算法所发现的约束规则和异常值更正的结果分别进行评估，并在最后去计算综合 f1-score，评价指标公式如表 3 所示。

Table 3. Evaluation index
表 3. 评价指标

评价指标	定义
约束规则精确度	$\text{Precision}_{\text{rule}} = \frac{\text{算法挖掘出的正确约束规则数量}}{\text{算法挖掘出的所有约束规则数量}}$
约束规则召回率	$\text{Recall}_{\text{rule}} = \frac{\text{算法挖掘出的正确的约束规则数量}}{\text{领域专家给出的约束规则数量}}$
约束规则 f1-score	$\text{f1score}_{\text{rule}} = \frac{2\text{Precision}_{\text{rule}} * \text{Recall}_{\text{rule}}}{\text{Precision}_{\text{rule}} + \text{Recall}_{\text{rule}}}$

Continued

数据更正精确度	$\text{Precision}_{data} = \frac{\text{算法做出正确修改的记录数量}}{\text{异常值更正算法修改的记录数量}}$
数据更正召回率	$\text{Recall}_{data} = \frac{\text{算法做出正确修改的记录数量}}{\text{领域专家给出的错误记录总数}}$
数据 f1-score	$\text{f1score}_{data} = \frac{2\text{Precision}_{data} * \text{Recall}_{data}}{\text{Precision}_{data} + \text{Recall}_{data}}$
综合 f1-score	$\text{f1score}_{total} = \frac{1}{2}(\text{f1score}_{data} + \text{f1score}_{rule})$

准确率越高，越能保证规则或更正结果的正确性；召回率越高，越不会错过正确的规则或更正结果。召回率用以衡量领域约束规则的完备性。为公正地考量算法的性能，需要在这两项指标间进行折衷。试想，如果算法仅得到很少一部分结果，就可以达到接近 100%的精确度，而只用算法得到的结果去衡量时，缺乏领域专家的评估时，也会使召回率接近 100%。因此，需要由 f1-score 使两者得到综合。

6.3. 实验结果

在实验中，将本文算法与其他两种可用于异常值处理算法进行对比，这两种算法分别为：基于条件函数依赖的异常值处理算法 CFDM [18]和基于罕见项集[3]的异常值处理算法。其中，关于罕见项集的文献仅介绍了罕见项集的挖掘算法，为使罕见项集能够用于异常值处理，基于本文的准反常项集思想，在文献[18]的基础上，本文加入了数据更正环节。运行结果如图 1 所示，无论是在约束规则发现方面，还是在数据更正质量方面，本文所提出的方法都能获得较好的精确度、召回率和 f1-score。基于函数依赖的 CFDM 算法次之，这是由于该算法仅考虑了属性之间的语义联系，难以发现复杂的约束规则。RIM 算法的约束规则能力和数据更正能力均较低，这是因为罕见项集的概念范围过于笼统，其中不仅包括了数据中的非频繁项集，还包括了没有在数据集中出现过的零项集，该算法将它们全部视为异常，造成了误报。

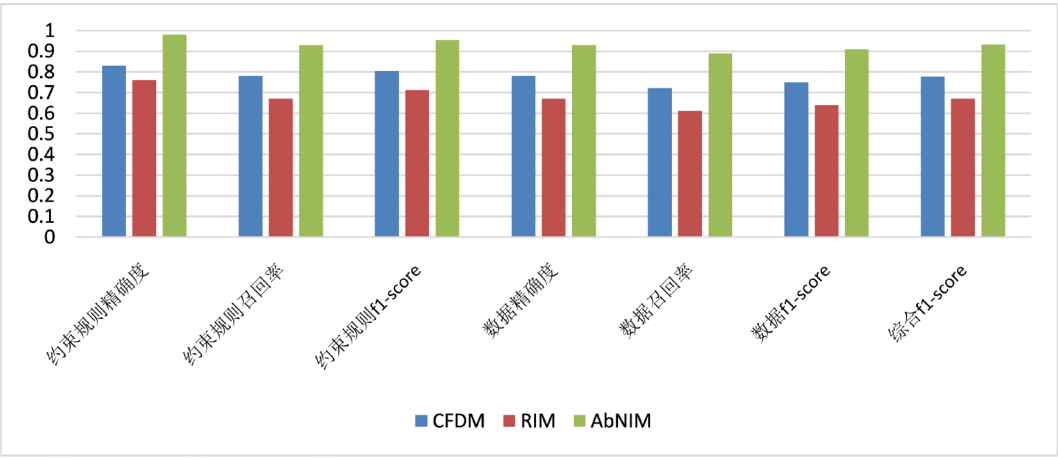

Figure 1. The comparison of experimental results
图 1. 实验结果对比

7. 总结

本文首先分析了传统的离散值异常处理算法的不足，由于基于统计分析的异常值检测算法无法发现

涉及多属性的复杂约束规则，仅能发现一些较低维度的异常值，本文提出了反常项集的概念，并提出了一种基于信息熵的属性加权方法，由等价类变换算法出发，利用加权调和提升度挖掘低提升度反常项集。为避免由数据更正引入新的异常值，本文采用准反常项集对可能的更正结果进行预计算，利用数据集中的可信部分来更正异常值。最后，在房地产领域数据集上的对比实验中，验证了本文算法的准确性和可靠性，表明本文所给出算法可在无监督下高效且准确地发现领域约束规则和更正异常值。

参考文献

[1] Hemalatha, C.S., Vaidehi, V. and Lakshmi, R. (2015) Minimal Infrequent Pattern Based Approach for Mining Outliers in Data Streams. *Expert Systems with Applications*, **42**, 1998-2012. https://doi.org/10.1016/j.eswa.2014.09.053

[2] Duncan, K. and Wells, D. (1999) A Rule Based Data Cleansing. *Journal of Data Warehousing*, **4**, 146-159.

[3] Szathmary, L., Napoli, A. and Valtchev, P. (2007) Towards Rare Itemset Mining. 19*th IEEE International Conference on Tools with Artificial Intelligence*, Patras, 29-31 October 2007, 305-312. https://doi.org/10.1109/ICTAI.2007.30

[4] Kolahi, S. and Lakshmanan, L. (2009) On Approximating Optimum Repairs for Functional Dependency Violations. *Database Theory-ICDT, International Conference*, St Petersburg, March 2009, 53. https://doi.org/10.1145/1514894.1514901

[5] Chu, X., Ilyas, I.F. and Papotti, P. (2014) Discovering Denial Constraints. *Proceedings of the Vldb Endowment*, **6**, 1498-1509. https://doi.org/10.14778/2536258.2536262

[6] Beskales, G., Ilyas, I.F., Golab, L. and Galiullin, A. (2012) On the Relative Trust between Inconsistent Data and Inaccurate Constraints. 2013 *IEEE* 29*th International Conference on Data Engineering* (*ICDE*), Brisbane, 8-12 April 2013, 541-552. https://doi.org/10.1109/ICDE.2013.6544854

[7] 刘波, 耿寅融. 数据质量检测规则挖掘方法[J]. 模式识别与人工智能, 2012(5): 835-844.

[8] 林印华, 张春海, 刘洁. 基于清洗规则和主数据的数据复算法实现[J]. 计算机科学, 2012, 39(11): 174-176.

[9] 王凤梅, 胡丽霞. 一种基于近邻规则的缺失数据填补方法[J]. 计算机工程, 2012, 38(21): 53-55.

[10] Szathmary, L., Valtchev, P. and Napoli, A. (2010) Generating Rare Association Rules Using the Minimal Rare Item Sets Family. *International Journal of Software and Informatics*, **4**, 219-238.

[11] Adda, M., Lei, W. and Yi, F. (2007) Rare Itemset Mining. *International Conference on Machine Learning & Applications*, Cincinnati, 13-15 December 2007, 73-80. https://doi.org/10.1109/ICMLA.2007.106

[12] Szathmary, L., Valtchev, P. and Napoli, A. (2010) Finding Minimal Rare Itemsets and Rare Association Rules. *International Conference on Knowledge Science, Engineering and Management*, Belfast, 1-3 September 2010, 16-27. https://doi.org/10.1007/978-3-642-15280-1_5

[13] Ouyang, W.M. and Huang, Q.H. (2013) Mining Indirect Temporal Sequential Patterns in Large Transaction Databases. *Applied Mechanics and Materials*, **385-386**, 1362-1365. https://doi.org/10.4028/www.scientific.net/AMM.385-386.1362

[14] Szathmary, L., Valtchev, P., Napoli, A. and Godin, R. (2009) Efficient Vertical Mining of Frequent Closures and Generators. *Advances in Intelligent Data Analysis VIII*, 8*th International Symposium on Intelligent Data Analysis, IDA* 2009, Lyon, 31 August-2 September 2009, 393-404. https://doi.org/10.1007/978-3-642-03915-7_34

[15] Zaki, M.J. and Hsiao, C.J. (2002) CHARM: An Efficient Algorithm for Closed Itemset Mining. *Proceedings of the Second SIAM International Conference on Data Mining*, Arlington, 11-13 April 2002, 457-473. https://doi.org/10.1137/1.9781611972726.27

[16] Rammelaere, J., Geerts, F. and Goethals, B. (2017) Cleaning Data with Forbidden Itemsets. *IEEE International Conference on Data Engineering*, San Diego, 19-22 April 2017, 897-908. https://doi.org/10.1109/ICDE.2017.138

[17] Boriah, S., Chandola, V. and Kumar, V. (2008) Similarity Measures for Categorical Data: A Comparative Evaluation. *Proceedings of the SIAM International Conference on Data Mining, SDM* 2008, Atlanta, 24-26 April 2008, 243-254. https://doi.org/10.1137/1.9781611972788.22

[18] Chiang, F. and Miller, R.J. (2008) Discovering Data Quality Rules. *Proceedings of the Vldb Endowment*, **1**, 1166-1177. https://doi.org/10.14778/1453856.1453980