文章编号:1672-3961(2011)06-0031-06

基于属性约简和相对熵的离群点检测算法

胡云[1,2],李慧[1],施珺[1],蔡虹[1]
(1.淮海工学院计算机工程学院,江苏 连云港 222000;2.南京大学计算机科学与技术系,江苏 南京 210000)

摘要:本研究结合信息熵与粗糙集理论中的属性约简技术,提出了一种新颖的离群点检测算法。这种方法通过在更小的属性子空间去获得相同或相近的离群数据集,使对离群数据的分析更加集中于较小的目标域。该算法对原属性空间进行划分,通过分析计算将具有最大相对熵与负相对势的对象集合判定为离群点集合。为了验证算法的有效性,还在通用数据集上进行了测试,理论分析和实验结果表明该离群点检测算法是有效可行的。

关键词:属性简约;相对熵;离群点检测
中图分类号:TP391 **文献标志码**:A

An outlier detection algorithm based on attribute reduction and relative entropy

HU Yun[1,2], LI Hui[1], SHI Jun[1], CAI Hong[1]
(1. School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222000, China;
2. Department of Computer Science and Technology, Nanjing University, Nanjing 210000, China)

Abstract: A new outlier detection algorithm combining a rough set and information entropy technology was proposed. This approach could obtain similar outlier sets by means of searching in an attributes subspace, which could lead the analysis of outlier detection to focus better on narrow and specific object fields. This algorithm divided the original attribute space into several segments, which filtered out those subjects with largest relative entropy negative relative cardinality as the outliers. To prove this algorithm's effectiveness, experiments on a real world dataset were conducted. Theoretical analysis and experimental results showed that this method of outlier detection was efficient and effective.

Key words: attribute deduction; relative entropy; outlier detection

0 引言

离群点检测[1-4]是数据挖掘领域研究的重要问题之一,着力于从数据集中发现与其他数据显著不同的一小部分对象,其目的是消除噪音或发现潜在的、有意义的知识。本研究结合属性约简与信息熵[5-9]技术,探讨在数据集属性域子集中利用信息熵进行离群点检测的方法(outlier detection based on relative information entropy, ODRIE)。该方法首先对知识系统进行属性约简、核集的求取;接着对原属性空间进行划分,通过分析计算将具有最大相对熵与负相对势的对象集合判定为离群点集合,实现基于信息熵度量的离群点检测。

1 问题描述与相关工作

粗糙集方法仅利用数据本身提供的信息,无须任何先验知识,能有效地分析不精确、不一致、不完整等各种不完备信息。粗糙集研究的对象是由一个多值属性集合描述的一个对象集合。

从信息系统的角度出发,一个数据集可以定义

收稿日期:2011-04-15
基金项目:江苏省自然科学基金资助项目(BK2008190)
作者简介:胡云(1977 -),女,江苏连云港人,讲师,博士研究生,主要研究方向为数据挖掘,智能信息处理. E-mail: huyunzhang@yahoo.com.cn

为如下的四元组：IS = (U, A, V, f)，其中 U 为全体对象的集合；A 为 U 中数据的全体属性的集合；V 为属性的值域；$f: U \times A \to V$ 是一个映射函数，使得 $\forall x \in U, a_i \in A$，$f(x, a_i) \in V_{a_i}$。对于给定对象 X，$f(x, a_i)$ 赋予对象 x 在第 i 个属性上的取值。信息系统也可以简记为 IS = (U, A)。

给定属性集合的子集 $B \subseteq A$，则由 B 可以决定数据集 U 上的一个不可区分关系 Ind(B)：

$$\mathrm{Ind}(B) = \{(x, y) \in U \times U : \forall a \in B, f(x, a) = f(y, a)\}。$$

显然，Ind(B)是 U 上的等价关系，且有：

$$\mathrm{Ind}(B) = \bigcap_{a \in B} \mathrm{Ind}(\{a\})。$$

等价关系 Ind(B)将数据集 U 分成若干个在属性集 B 上不可区分的子集，用 $U_{\mathrm{Ind}(B)}$ 表示。

文中，提出了一种新颖的利用粗糙集理论进行离群点检测的方法。该方法的主要思想描述如下，给定信息系统 IS = (U, A) 和一组在 U 上的不可区分的关系，则任意一个不可区分的关系Ind(B)将 U 划分为 $U_{\mathrm{Ind}(B)}$ 个分组。在此基础上，借助于特定的标准（采用 $U_{\mathrm{Ind}(B)}$ 中每个等价类的相对势作为划分标准），可将这些分组划分为两类，即 U 中属于多数的分组和属于少数的分组。因此，对于 U 中的任意对象 x，通过确定 x 在不可区分关系 Ind(B)下的分类并计算其在 Ind(B)下的相对熵则可以判定 x 在数据集中离群性。这是因为，相对熵作为对象不确定性的度量，可以反映数据对象的异常属性。这样，离群点检测的目标就是找出 U 中属于在多种不可区分关系下总是属于少数组并具有较高相对熵的数据对象。文献[10]基于上述思想，提出了一种可行的离群点检测算法。但是，其方法需要遍历所有可能的属性子集所生成的不可区分关系。对于高维数据集，过高的计算开销使上述方法难以实现。本研究认为，属性集 A 中存在的属性有核心属性与非核心属性之分。其中，核心属性是表述知识必不可少的属性。如果仅选取核心属性参与运算，不仅能够排除非核心属性对判断的干扰，还能够大大降低算法的复杂度。本研究运用属性约简的思想实现核心属性的提取，并构造基于核心属性和信息熵理论的离群点检测算法。

2 ODRIE 算法

2.1 核心属性集的择取及其算法

结合粗糙集理论中的划分[11-13]与属性约简技术[14-15]，探讨在核心属性子集上进行离群点检测的方法[16-18]。

在已知关于粗糙集研究成果中，Skowron 提出的可辨识矩阵为我们求取最佳属性约简提供了理论基础[7]。该方法将信息表（也称决策表，是指将真实世界的信息以条件属性与决策属性构成表的形式给出）中所有有关属性区分的信息都浓缩到一个矩阵中（称为可辨识矩阵，discernibility matrix），并通过该矩阵求得信息表的属性核（信息表中不可删除的属性）。本算法以可辨识矩阵为基础，重点研究矩阵中除属性核之外的其他属性组合，并利用析取范式进行属性约简。

Andrzej Skowron 于 1991 年提出一种用可知识矩阵表示知识的办法，这种表示有许多有利条件，特别是用它可以解释和便于计算数据核和约简。其定义如下：

定义 1 给定信息系统 IS = (U, A)，$U = \{x_1, \cdots, x_n\}$，将属性集 A 划分为条件属性集 $C = \{c_1, \cdots, c_m\}$ 和决策属性集 D：$A = C \cup D$，令 $c_i(x_j)$ 和 $D(x_j)$ 分别是数据点 x_j 在属性 c_i 上以及决策属性集 D 上的取值，则可辨识矩阵 $\boldsymbol{M}$ 中各元素的取值定义为

$$\boldsymbol{M}_{i,j} = \begin{cases} 0, & D(x_i) = D(x_j); \\ -1, \ \forall c \in C, \ c(x_i) = c(x_j), & D(x_i) \neq D(x_j); \\ \{c \in C : c(x_i) \neq c(x_j)\}, & D(x_i) \neq D(x_j), \end{cases}$$

$$i, j = 1, \cdots, n。$$

上述矩阵说明，当决策属性相同时，元素值为 0；当决策属性不同且可以通过某些条件属性加以区分时，矩阵元素为互不相同的属性组合，当决策属性不同而条件属性完全相同时，元素值为 -1，该情况表示数据有误或条件属性不足。

由可辨识矩阵的定义可知，矩阵中属性组合数为 1 时表示除该属性外其余条件属性无法将信息表中决策不同的两条记录区分出来，即该属性必须保留。因此，可辨识矩阵中所有属性组合数为 1 的属性均为决策表的核属性。用 C_0 表示核属性集，则有 $C_0 \subseteq A$。

考虑到可辨识矩阵包含了决策表中的所有属性区分信息，因此，核属性外的其余有用属性应该从属性组合数不为 1 的矩阵元素中分析获得。假设某信息表除 C_0 外剩余两个属性组合，分别用 $t_{11}, \cdots, t_{1e}$ 和 $t_{21}, \cdots, t_{2k}$ 表示。构造表达式：

$$P = (t_{11} \vee \cdots \vee t_{1e}) \wedge (t_{21} \vee \cdots \vee t_{2k})。$$

则该合取式代表的属性组合连同核属性即可将原决策表中的所有决策区分出来。如信息表除 C_0 外还剩余更多的属性组合，则其处理方法可依此类推。由于析取范式由多个合取式构成，究竟采用哪些属性组合可以根据需要而定，该属性组合与核属

性一起构成在指定要求下的最佳属性约简。在本研究中采用最精简的属性组合做为属性约简的最终结果。其具体算法如下:

算法 1　属性约简算法

输入　决策表 IS = (U,A,V,f),其中 $A = C \cup D$ 是属性集合。

输出　约简后的属性集合 Ω。

算法

(1) 计算决策表的可辨识矩阵 $\boldsymbol{M}$;

(2) 将矩阵中属性组合数为 1 的属性加入到核属性集 C_0 中;

(3) 将核属性列入属性约简后得到的属性集合 Ω 中,即 $\Omega = C_0$;

(4) 在可辨识矩阵中找出所有不包含核属性的属性组合 Q,即

$$Q = \{B_i \mid B_i \cap \Omega = \Phi, i = 1, \cdots, s\};$$

(5) 将属性组合 Q 表示为合取范式的形式,即

$$P = \wedge \{ \vee B_i, i = 1, \cdots, s\}_{k=1,\cdots,m};$$

(6) 将 P 转换为析取范式形式;

(7) 选择最精简的属性组合作为属性约简的结果加入到集合 Ω 中;

(8) 输出约简后的属性集合 Ω。

2.2　信息熵度量的离群点检测算法

给定信息系统 IS = (U,A,V,f),如果其某个数据对象 $x \in U$ 在某些属性上具有可度量的奇异取值,可以认为 x 是关于该信息系统的离群点。信息熵做为一种不确定性度量的有效方法被广泛地应用于机器学习领域中,本研究将信息熵应用于数据的离群性态的度量与检测,所涉及到的相关定义如下:

定义 2　信息熵　信息熵是对信息和随机变量的不确定性的一种度量。给定一个信息系统 IS = (U,A,V,f),其中 U 是非空对象集,A 是非空属性集。对于任意 $B \subseteq A$,以 $U_{\mathrm{Ind}(B)} = \{B_1, B_2, \cdots, B_m\}$ 表示由关系 Ind(B)将 U 划分后的 m 个分组。则数据集关于 B 的信息熵 $E(B)$ 定义为

$$E(B) = -\sum_{i=1}^{m} \frac{|B_i|}{|U|} \log_2 \frac{|B_i|}{|U|}, \tag{4}$$

其中 $|B_i|/|U|$ 表示元素 $x \in U$ 在等价类 B_i 中出现的概率,$|\cdot|$ 表示集合的势。

定义 3　相对熵　给定信息系统 IS = (U,A,V,f),对于任意 $B \subseteq A$,以 $U_{\mathrm{Ind}(B)} = \{B_1, B_2, \cdots, B_m\}$ 表示由关系 Ind(B)将 U 划分后的 m 个分组。对任意 $x \in U$,记 $[x]_B$ 为 x 在关系 Ind(B)下的等价类,$U_{\mathrm{Ind}(B)} - [x]_B = \{B'_1, \cdots, B'_m\}$。设

$$E_x(B) = -\sum_{i=1}^{m-1} \frac{|B'_i|}{|U| - |[x]_B|} \log_2 \frac{|B'_i|}{|U| - |[x]_B|}$$

为从 U 中移去 $[x]_B$ 等价类后的信息熵,则 x 在关系 Ind(B)下的相对熵 $\mathrm{RE}_x(B)$ 定义为

$$\mathrm{RE}_x(B) = \begin{cases} 1 - \dfrac{E_x(B)}{E(B)}, & \text{if } E(B) > E_x(B), \\ 0, & \text{else}。 \end{cases} \tag{5}$$

上式公式的含义可以解释如下:对于任意给定的 $B \subseteq A$ 和 $x \in U$,当从 U 中删除 x 的等价类 $[x]_B$ 中所有的对象后,如果信息熵的值显著下降,我们可以认为对象 x 在关系 Ind(B)下的不确定性很高。反之,如果信息熵的值变化稍有增加甚至不变,则可以认为对象 X 在关系 Ind(B)下的不确定性非常低甚至为 0。因此,在关系 Ind(B)下的相对熵 $\mathrm{RE}_x(B)$ 可以作为衡量对象 x 不确定性的重要指标,其值越大表示 x 的不确定性越高。

由于离群点检测的目标就是发现在目标 U 中具有特殊或不寻常属性的少数集合。不确定性正好可以用来描述这种特殊属性。因此,可以认为在 U 中那些具有较高相对熵值的对象就是异常对象。

定义 4　相对势　将等价类 $[x]_B$ 的相对势力定义如下:

$$\mathrm{RC}([x]_B) = |[x]_B| - \frac{|B'_1| + \cdots + |B'_{m-1}|}{m-1}。 \tag{6}$$

显然,当 $\mathrm{RC}([x]_B) > 0$ 时,就可以判定对象 x 属于多数组中,相反则可认为 x 属于少数组中。

离群点检测通常关注那些在集合中占少数对象,因为属于少数组的对象比属于多数组的对象更有可能是离群点。因此,如果 $\mathrm{RC}([x]_B) \leqslant 0$,则对象 x 属于 U 中的少数组,也就是说 x 比属于多数组中的其它对象更有可能是离群点。更进一步地,$\mathrm{RC}([x]_B)$ 的值越大,则 x 的离群性态越明显。因此可以将这 2 个条件做与操作,即可构造依据相对熵和相对势的离群点检测算法。

根据上述基本思想,下面给出基于属性简约和相对信息熵度量的 ODRIE 的流程:

算法 2　Algorithm ODRIE

输入　信息系统 IS = (U,A,V,f),其中 $A = \{a_1, \cdots, a_k\}$。

输出　离群点集合 O。

算法

(1) 计算得出 A 的属性约简集合 Ω;

(2) For $i = 1$ to k;

(3) 确定 $U_{\mathrm{Ind}}(\{a_i\})$ 划分的分组;

(4) For $i = 1$ to k;

(5) 计算相对熵 $\mathrm{RE}_{\{a_i\}}(u_i)$;

(6) 计算相对势 RC($[u_i]_{\{a_i\}}$);
(7) Endfor;
(8) Endfor;
(9) 将 $RE_{\{a_i\}}(u_i)$ 按从大到小排序,选取其值最大的前 p 个对象;
(10) For $i=1$ to p;
(11) If RC($[u_i]_{\{a_i\}}$) <0;
(12) 输出到离群点数据集 O;
(13) Endif;
(14) endfor。

ODRIE 算法流程分为4个步骤:(1) 计算属性的约简后的核集;(2) 确定分组(步骤2至步骤4);(3) 计算相对熵和相对势(步骤5至步骤8);(4) 按相对熵从大到小排序,取前 P 个对象,判断其相对势的正负性,从而输出离群点集合。本算法中,参数 P 的取值需要依据给定数据和问题的实际需求加以选择。如果 P 的值设置太小,会导致离群点不能完全被检测出来;如果 P 的取值过大,会增加算法的复杂度和额外开销。下述实验中,我们通过结果分析 P 的取值范围在15至20之间为最佳。

3 实验结果与分析

利用实际的数据集进行了 ODRIE 算法的详细实验。实验平台配置如下: INTEL1. 8 GHZ、512 MB, WINDOW S 2000 (SERVER 版),编程语言采用 VISUAL C ++(6.0) 实现。采用 AGGARWAL C 和 YU P 在文献[19]中提出的离群点检测算法有效性的评估标准,即检测结果中包含属于稀有类的点越多,则检测效果越好。为了验证 ODRIE 的有效性,取2组数据集对算法进行了测试,第1组测试 ODRIE 算法检测离群点的准确度;第2组测试算法效率。

3.1 算法精度测试

为了检测 ODRIE 算法对离群点的分类能力,我们对 UCI 中的 ZOO 数据集进行了测试。ZOO 数据集中有101个动物的记录,每条记录有18个属性其中包括1个动物名称, 16个条件属性(15个布尔属性, 1个腿个数的离散属性) 和一个决策属性, 所有动物被分成7个类别. 采用文献[20]中使用的方法,只取动物是哺乳动物和爬行动物2类。这样做的原因是: (1)本文关注离群点的检测问题, 当把数据集中所有记录都考虑进来时, 数据集中记录数目少的类别的离群特征不明显;(2)简化数据集以方便讨论.

对 ZOO 数据集,我们首先对某类别中某属性值上哪些动物是与众不同的数据进行了统计。其中与众不同的定义标准是:对象集中某个属性为某个属性值时,有小于15%的对象取该属性值时,则此对象在这个属性上是与众不同的。从客观角度分析和解释,如果某对象入选次数越多,则说明此对象成为离群点的可能性越大。表1给出了 ZOO 数据集中对哺乳类和爬行类各自内部与众不同属性的统计结果,以入选的次数为标准从多到少进行排序,表1中只给出前3名的结果(含并列)。每个类别均包括属性名称和出现次数。在相同的数据集上,运用提出的 ODRIE 算法检测离群点,在对 P 取不同值下,该算法检测到的前5个离群点如表2所示。

表1 ZOO 数据集中与众不同的对象入选次数统计
Table 1 The statistics for unusual animals

排名	哺乳类		爬行类	
	名称	次数	名称	次数
1	海豚,海豹	4	海蛇	3
2	鸭嘴兽	3	乌龟	2
3	鼠海豚	2	蝮蛇	2

表2 不同 P 值时 ODRIE 检测到的前5个离群点
Table 2 The top 5 outliers detected by ODRIE with various P values

P	ODRIE 检测到的前5个离群点				
	1st	2nd	3rd	4th	5th
5	熊	鸭嘴兽	水貂	土豚	负鼠
10	鸭嘴兽	水貂	乌龟	熊	土豚
15	鸭嘴兽	鼠海豚	海豚	海蛇	乌龟
20	鸭嘴兽	海蛇	乌龟	蝮蛇	熊
25	鸭嘴兽	乌龟	海蛇	海豹	负鼠
30	蝮蛇	乌龟	海蛇	鸭嘴兽	海豹
35	蝮蛇	乌龟	熊	鸭嘴兽	海豹

实验结果列于表2,由表2知,ODRIE 算法在 P 取值为15至25时,检测到的前3个离群点都出现在表1当中,说明应用 ODRIE 算法在 P 取值合适时,可以准确地找到数据集中的离群数据。图1给出了 P 在不同的取值时 ODRIE 算法的精确度。

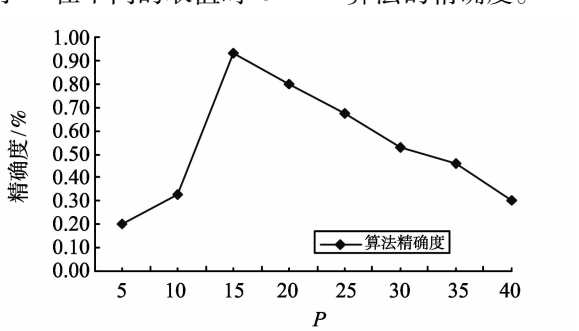

图1 P 在不同取值时的算法精度
Fig. 1 The algorithm precision with various P values

为了说明本算法能够在相对较少的数据集中快速准确的找到离群点,我们将 ODRIE 算法和 4 算法、KNN 算法在 ZOO 数据集上进行了精确度的对比实验,实验结果如图 2 所示。X 轴表示选取 ZOO 数据中的对象数量,Y 轴表示应用各算法查找到离群点的正确率。从图 2 中可以看出,ODRIE 可以在较小的对象集中快速准确地找到离群点,算法效率均高于 DIS 和 KNN 算法。

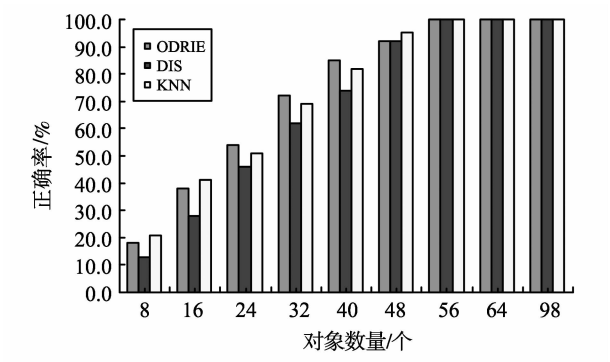

图 2　不同算法的准确度对比

Fig. 2　Accuracy of ODRIE, DIS and KNN

3.2　算法效率测试

该组实验数据来自某市移动通信业务数据库,取其一个子集 X 共 18 个属性 10 万条记录。对数据集分 2 种情况对算法效率进行测试:(1) 在属性全子空间搜索的算法执行时间;(2) 应用属性约简后的算法执行时间。

实验 1　首先选取客户号、呼叫时长、本地通话费、省内漫游费、国内漫游费等 5 个属性,对问题规模 $N=1,2,3,4,5$ 万条记录时进行测试,结果显示运行时间接近线性,然后增加呼叫次数等 5 个属性实验一次,再增加 GPRS 流量等 5 个属性在同样条件下比较实验,结果如图 3 所示,在不进行属性约简时,即在全部属性域上进行离群点检测时,算法的执行时间较问题规模而言随数据维数增长的更快。

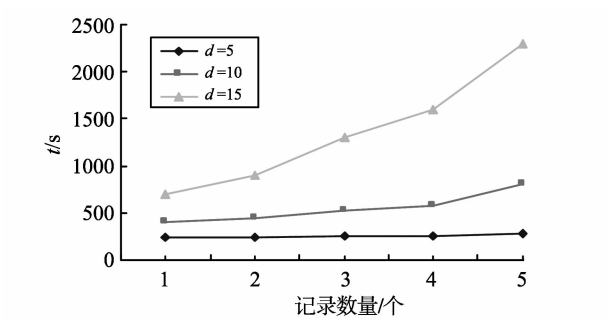

图 3　全子空间下不同维度时算法执行时间对比

Fig. 3　Comparison of execute time of the algorithm with all attribute

实验 2　取维数 $d=10$,问题规模 $n=90\,000$ 条记录分别在全子空间搜索和应用属性约简后的算法运行时间比较,图 4 给出了两者执行时间的比较。

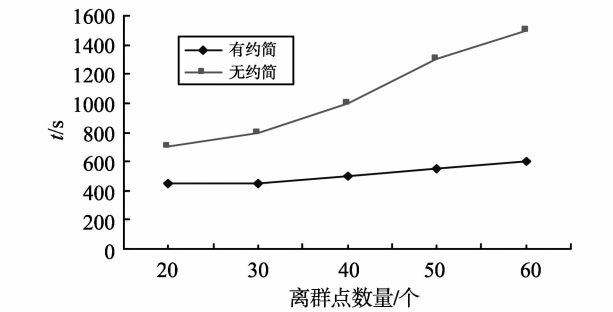

图 4　精确与近似算法的执行时间对比($d=10$,$n=90\,000$)

Fig. 4　Comparison of execute time of exact and approximate algorithm($d=10$,$n=90\,000$)

实验结果说明在应用属性约简后去除掉一些不会对数据离群有显著影响的非核心属性,这样执行时间可大大缩短而不会降低算法的有效性。

4　结语

结合信息熵与粗糙集理论研究了离群点的检测问题,提出了 ODRIE 算法。算法首先利用属性归约技术去除了冗余属性,在更小的属性子空间去获得相同或相近的离群数据集,然后通过相对熵与相对势的取值确定离群点集合。算法能够以最小的代价快速准确的找到离群点。实验结果表明,本研究提出的算法是可行而有效的,进一步提高了算法的实现效率。今后,将其扩展到数值属性和混合属性数据流以及与数据流中其他相关的数据挖掘算法(例如聚类)的结合等等,是下一步的研究内容。

参考文献:

[1] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density based local outliers[C]// Processings of the ACM SIGMOD International Conference on Management of Data. Dallas: ACM Press, 2000. 93-104.

[2] KNORR E M, NG R T. Algorithms of mining distance based outliers in large datasets[C]// Processings of VLDB'98. New York: ACM Press, 1998: 392-402.

[3] HE Z, XU X, DENG S. FP-outlier: frequent pattern based outlier detection[J]. ComSIS, 2005, 2(1): 103-118.

[4] ARNING A, AGRAWAL R, RAGHAVAN P. A linear method for deviation in large database[C]// Processings of the 2nd International Conference on Knowledge Discovery and Data Minning. Massachusetts: AAAI Press, 1996: 164-169.

[5] PEI J, JIANG B, LIN X M, et al. Probabilistic sklines on uncertain data[C]// Processings of VLDB'07. New

York: ACM Press, 2007: 15-26.

[6] 刘君强,王勋,孙晓莹. 多维多层关联规则挖掘的新算法[J]. 南京大学学报:自然科学版,2003, 39(2):205-210.
LIU Junqiang, WANG Xun, SUN Xiaoying. Effectively mining multi-dimension multi level association rules[J]. Journal of Nanjing University: Natural Sciences, 2003, 39(2):205-210.

[7] SKLWRON A. The discernibility matrices and functions in information systems[M]. [S. l.]: Kluwer Academic Publishers, 1992: 331-362.

[8] BARBARA D, LI Y, COUTO J. Coolcat: an entropy-based algorithm for categorical clustering[C]// Processings of ACM Conference on Information and Knowledge Management (CIKM). New York: ACM Press, 2002: 582-589.

[9] LI T, MA S, MITSUNORI O. Entropy based criterion in categorical clustering[C]// Processings of Internal Conference on Machine Learning (ICML), New York: ACM Press, 2004: 115-124.

[10] HE Z Y, XU X F, DENG S C. An optimization model for outlier detection in categorical data[C]// Processings of Lecture Notes in Computer Science of Advances in Intelligent Computing. Massachusetts: AAAI Press, 2005: 23-26.

[11] ZIARKO W. Probabilistic approach to rough sets[J]. International Journal of Approximate Reasoning, 2008, 49(1):272-284.

[12] PAWLAK Z, SKOWRON A. Rudiments of rough sets [J]. Information Sciences, 2007, 177(3):3-27.

[13] YAO Y Y, ZHAO Y. Attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2008, 178(17):3356-3373.

[14] YAO Y Y, ZHAO Y. Discernibility matrix simplification for constructing attribute reducts[J]. Information Sciences, 2009,179(5):867-882.

[15] YAO Y Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010,180(3):341-353.

[16] 于浩,王斌,肖刚. 基于距离的不确定离检测[J]. 计算机研究与发展,2010,47(3):474-484.
YU Hao, WANG Bin, XIAO Gang. Distance-based outlier detection on uncertain data[J]. Journal of Computer Research and Development, 2010, 47(3):474-484.

[17] FENG J, YUE F S, CUN G C. An information entropy-based approach to outlier detection in rough sets[J]. Expert Systems with Applications, 2010, 37(1):6338-6344.

[18] 薛安荣,鞠时光,何伟华. 局部离群点挖掘算法研究[J]. 计算机学报,2007, 30(8):1455-1460.
XUE Anrong, JU Shiguang, HE Weihua. Study on algorithm for local outlier detection[J]. Chinese Journal of Computers, 2007, 30(8):1455-1460.

[19] AGGARWAL C, YU P. An efective and eficient algorithm for high dimensional outlier detection[J]. The VLDB Journal, 2005, 14(2):211-221.

[20] 于绍越, 商琳. ENBROD: 基于信息熵的相对离群点的检测方法[J]. 南京大学学报:自然科学版, 2008, 44(2):212-218.
YU Shaoyue, SHANG Lin. An entropy-based algorithm to detect relative outliers: ENBROD[J]. Journal of Nanjing University: Natural Sciences, 2008, 44(2): 212-218.

(编辑:陈燕)