doi:10.3969/j.issn.1673-4785.2010.02.009

信息熵度量的离群数据挖掘算法

张 贺[1],蔡江辉[1],张继福[1],乔 衍[2]

(1.太原科技大学 计算机科学与技术学院,山西 太原 030024;2.北京航空航天大学 自动化科学与电气工程学院,北京 100191)

摘 要:离群数据挖掘是为了找出隐含在海量数据中相对稀疏而孤立的异常数据模式,但传统的离群数据挖掘方法受人为因素影响较大.通过引入基于信息熵的离群度量因子,给出一种离群数据挖掘新算法.该算法先利用信息熵计算每个数据对象的离群度量因子,然后通过离群度量因子来衡量每个对象的离群程度,进而检测离群数据,有效地消除了人为主观因素对离群检测的影响,并能很好地解释离群点的含义.最后,采用UCI和恒星光谱数据作为实验数据,通过对实验的分析,验证了该算法的可行性和有效性.

关键词:离群数据;信息熵;离群度量因子;数据挖掘

中图分类号:TP311 **文献标识码**:A **文章编号**:1673-4785(2010)02-0150-06

An outlier mining algorithm based on information entropy

ZHANG He[1], CAI Jiang-hui[1], ZHANG Ji-fu[1], QIAO Kan[2]

(1. School of Computer Science and Technology, Taiyuan University of Science & Technology, Taiyuan 030024, China; 2. Automation Science and Electrical Engineering College, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: The task of outlier mining is to discover patterns that are exceptional, interesting, and sparse or isolated even though they are concealed within tremendous volumes of data. Traditional outlier detection methods are easily influenced by man-made factors. A novel outlier mining algorithm based on information entropy has been formulated. It used an outlier measurement factor based on information entropy. In the algorithm, the outlier measurement factor of each record was calculated using information entropy. Outliers were then detected by analyzing the values of the outlier measurement factor. In this way the impact of man-made factors was eliminated in outlier mining. The definition of an outlier was based on an outlier measurement factor which could explain the meaning of the outliers. Experimental results proved the feasibility and effectiveness of the algorithm when it was used to analyze the UC Irvine (UCI) data set as well as high-dimensional star spectrum data.

Keywords: outlier; information entropy; outlier measure factor; data mining

离群数据(Outlier)是明显偏离其他数据,不满足数据的一般模式或行为,与存在的其他数据不一致的数据[1].但是,迄今为止,离群点还没有一个被普遍采纳的定义,统计学家Hawkins[2] 1980年给出的离群点定义在一定意义上揭示了离群点的本质:“离群点与其他点如此不同,以至于让人怀疑它们是由一个不同的机制产生的”.事实上,“一个人的噪声可能是另一个人的信号”[1],稀有事件比普通事件更有研究价值,这是由于数万个数据记录可能仅仅得出一个信息,而10个异常数据很可能得出10个不同的信息.离群数据的发现往往可以使人们发现一些真实的,但又出乎意料的知识;因此通过对离群数据的研究,发现异常的行为和模式,有着非常重要的意义.离群数据检测技术现已被广泛地应用于许多领域,如金融欺诈、电信计费、医疗保险、网络安全等.

目前,现有经典离群检测算法主要分为以下几类:基于统计(statistical-based)的方法[3]、基于深度(depth-based)的方法[4]、基于偏离(deviation-based)的方法[5]、基于距离(distance-based)的方法[6]与基于密度(density-based)的方法[7].这些方法存在以

收稿日期:2008-12-30.
基金项目:山西省青年科学基金资助项目(2008021028).
通信作者:张 贺. E-mail:zhanghe_helen@126.com.

下不足之处:1)需要人为事先给出一些参数和阈值,受人为因素影响较大,从而导致检测结果的客观性较差. 例如:基于距离的离群检测算法需要人为事先确定参数,当 pct、dmin 参数选择不当时,会产生错误结论;2)不能对非数值型数据进行处理,例如:基于统计和距离的离群检测算法较难对非数值属性数据进行挖掘;3)可解释性和可用性差,例如:基于统计的方法在解释时会发生多义性. 原因是:同一个离群点有可能是不同的分布模型检测出来的,即产生离群点的机制有可能不惟一,从而产生了多义性.

数据的维度是多少才能算高维数据? 10 维、100 维,还是1 000维. 实际上,高维数据拥有多少个属性并没有一个既定的界限,而是相对于某个算法而言. 例如,基于统计的方法:只能处理单变量数据集,即当数据维度为 2 时,算法不再有效;再如:基于深度方法,当数据维度大于 3 时,算法的可行性则非常差;基于距离和密度的方法,当数据的维度增加到一定程度,由于距离和密度对离群数据定义的局限性,使得方法执行效率也随之减弱. 因此离群检测算法在处理高维数据时,其可扩展性是尤为重要的.

信息熵可以用来度量一个系统无序和杂乱程度. 熵值越大,说明系统中的数据越无序,系统越"杂乱";反之,熵值越小,则说明系统中的数据越有序, 系统越"纯净"[8]. 出现在数据中的离群点是造成数据无序的主要原因之一,因此利用信息熵来度量、识别造成数据中无序的数据点[4],可以客观地识别出数据中的离群点[9]. 同时,利用信息熵来度量原始数据的无序特性,客观性比较强,受人为因素影响较小,不需要人为干预,从而得出更符合客观的结果. 信息熵也可以运用于非数值型属性数据集,例如标称属性数据集. 本文提出一种新的离群点检测方法——基于信息熵的离群数据挖掘算法(OMBIE). 通过引入离群数据度量因子量化地度量每个数据点的无序程度,即离群程度,并利用其挖掘造成数据无序的离群点,挖掘时无需人为事先设置参数或阈值,算法可以自动产生离群点,并能很好地解释离群点的含义.

1　基于信息熵的离群数据挖掘研究现状

2006 年,何曾友等人提出了基于信息熵的快速贪婪算法(GreedAlg)[10]. GreedAlg 算法事先人为设定期望产生的离群点个数 k,同时参数 k 用于发现一个势为 k 的离群数据集 $O(|O| = k)$;但此算法存在以下不足:1)需要人为事先给出期望产生的离群点个数 k, 这会有不能发现全部和多发现离群点的问题;2)文中提到 GreedAlg 算法需要全面扫描数据集 k 次,因此 I/O 代价通常比较高;3)因为使用贪婪算法的策略,计算过程中很容易陷入局部最小,而该算法未对此问题采取有效措施;4)作者在文中没有解释依据最大熵影响(maximal entropy impact)来识别离群点的原理.

2008 年,倪巍伟等人提出基于局部信息熵的加权子空间离群点检测算法(SPOD). 通过对数据点在各维进行邻域信息熵分析,生成数据点相应的离群子空间和属性权向量,对离群子空间中的属性赋以较高的权值,进一步提出子空间加权距离等概念. 采用基于密度离群点检测的思想,分析计算数据对象的子空间离群影响因子,判断是否为离群点. 算法能够有效地适应于高维数据离群点检测. 缺点是在处理高维数据时与 LOF 算法处于一个数量级 $O(n^2)$,而且还需要人为事先设置很多参数[11],从而影响了检测的结果.

同年,于绍越等人提出基于信息熵的相对离群点的检测方法(ENBROD). 文中首先引入一种新的信息熵增量的概念——去一划分信息熵增量,并在其基础上给出了每个对象所对应的相对离群点因子(ROF)的定义. 利用 ENBROD 算法来实现对 ROF 的计算,但 ENBROD 算法也需要人为事先设置参数,而这正影响了算法的运行效果[12].

2　信息熵

信息熵被用来度量一个系统的"无序"程度和"纯净"程度[8]. 信息熵是信息有用程度的一种表现形式.

定义 1　称四元有序组 $D = (U, A, V, f)$ 为数据集,其中:U 为所考虑对象的非空有限集合且 $|U| = m$,称为对象集;A 为属性非空有限集合,属性集的势为 $|A| = n$;$V = \bigcup_{a \in A} V_a$,而 V_a 为属性 a 的值域;$f: U \times A \to V$ 是一个映射函数, $\forall x \in U, a \in A$, $f(x,a) \in V_a$,对于给定对象 x,$f(x,a)$ 赋予对象 x 在属性 a 下的属性值. 数据集也可以简记为 $D = (U,A)$.

在本文中约定,数据集 $D = (U,A)$ 中的对象集的势为 $|U| = m$,属性集的势为 $|A| = n$;记录、数据点、对象是在不同范畴下表述的同一个事物.

定义 2　假设有一组离散的符号集 $\{v_1, v_2, \cdots, v_n\}$,每个符号具有相应的出现频率 P_i. 为了衡量用这组符号组成的特定序列的随机性(不确定性或不可预测性),定义离散分布的熵为

$$H = -\sum_{i=1}^{n} p_i \log_a p_i. \tag{1}$$

式中:对数的底 a 可为任何正数,一般取 2,此时熵的单位为“bit”. 规定当 $p_i=0$ 时,

$$\sum_{i=0}^{n} p_i \log_a \frac{1}{p_i} = 0. \tag{2}$$

这里要特别注意熵的值并不依赖于符号(对象、数据)本身,而只依赖于这些符号(对象、数据)的概率[13].

定义 3 如果 X 是一个离散的随机变量,$S(X)$ 是 X 可能取值的集合,$p(x)$ 是 X 的概率函数,那么信息熵 $H(X)$ 如式(3)所定义[8]:

$$H(X) = -\sum_{x \in S(X)} p(x) \lg(p(x)). \tag{3}$$

对于含有多个属性的记录 $\hat{X}=\{X_1,\cdots,X_n\}$ 的信息熵如式(4)计算:

$$H(\hat{X}) = -\sum_{x_i \in S(X_1)} \cdots \sum_{x_n \in S(X_n)} [p(x_1,\cdots,x_n) \cdot \lg p(x_1,\cdots,x_n)]. \tag{4}$$

如果记录的属性之间相互独立,式(4)可以转化成式(5). 为了简化对信息熵的计算,在本文中一律假设数据集中的记录的属性间是相互独立的.

$$H(\hat{X}) = -\sum_{x_1 \in S(X_1)} \cdots \sum_{x_n \in S(X_n)} [(p(x_1)\cdots p(x_n)) \cdot \lg(p(x_1)\cdots p(x_n))] = H(X_1) + H(X_2) + \cdots + H(X_n). \tag{5}$$

3 基于信息熵的离群数据挖掘算法

3.1 离群数据度量因子

信息熵可以用来度量一个系统无序和杂乱程度. 熵值越大,说明系统中的数据越无序,系统越杂乱;反之,熵值越小,则说明系统中的数据越有序,系统越纯净[8]. 如果将信息熵理论应用到离群数据挖掘中,根据 Hawkins[2] 对离群点定性地描述,出现在数据中的离群点是使系统不“纯净”、“杂乱”的原因,相当于系统中的“杂质”. 如果去除系统中的不“纯净”因素,那么系统则变得相对“有序”和“纯净”,熵值比去除前相对变小. 去除后,熵值相对减小地较大,说明去除的因素相对“杂乱”;熵值相对减小地较小,说明去除的因素相对“纯净”. 与此同时,从另外一个角度来讲,被去除的不“纯净”因素,也就是要寻找的离群数据,基于此理论基础,可通过测量熵值的变化来检测离群点. 为此定义了如下“离群数据度量因子”,来度量数据集中的离群数据.

定义 4 离群数据度量因子(outlier measure factor, OMF). 在数据集 $D=(U,A)$ 中,从对象集 U 中剔除对象 x_i 后,得到的新数据集,记作 $\overline{D}_i=\{\overline{U}_i,A\}$,其与原数据集 D 的信息熵的差 $H(D)-H(\overline{D})$ 定义为对象 x_i 的离群数据度量因子,记作 $\mathrm{OMF}(x_i)$.

$$\mathrm{OMF}(x_i) = H(D) - H(\overline{D}_i).$$

式中:对象 x_i 对应的离群数据度量因子 $\mathrm{OMF}(x_i)$ 的值越大,成为离群点的可能性越大.

通过离群数据度量因子定义的离群点与 LOF 算法中通过局部异常因子定义的离群点类似,即离群不再是一个二值属性,它摒弃了以前异常定义中非此即彼的绝对异常观念,更加符合现实生活中的应用. 离群数据度量因子 $\mathrm{OMF}(x_i)$ 可以量化地度量每个数据点 x_i 的离群程度,$\mathrm{OMF}(x_i)$ 的值越大,x_i 离群程度越强;反之,$\mathrm{OMF}(x_i)$ 越小,x_i 离群程度越弱. 因此,引进该因子既可以发现离群程度强的离群点,也可以发现离群程度弱的离群点. 离群数据度量因子 $\mathrm{OMF}(x_i)$ 是将数据集中的每个数据点看作一个有机整体并对其进行统一度量的,而不像 GreedAlg 算法把每个数据点孤立地度量. 此外,文中给出离群数据度量因子 $\mathrm{OMF}(x_i)$ 时,很好地利用了熵值并不依赖于符号(对象、数据)本身,而只依赖于这些符号(对象、数据)的概率[14]这一特性. 此方法不需要人为事先输入参数或设置阈值,从数据自身的本质和特征出发,更有利于挖掘隐藏在数据中的知识.

3.2 算法描述

根据上个小节的基本思想,图 1 给出了信息熵度量的离群数据挖掘算法(outlier mining based on information entropy, OMBIE)的流程.

```
Algorithm: 信息熵度量的离群数据挖掘算法OMBIE
Input: 数据集D
Output: 离群数据集Outliers
1)初始化将离散化的数据集存入数组Array[m][n]中;
2)计算数据集D的总信息熵totalInfoEtp;
3)计算每个数据点对应的离群数据度量因子OMF(x_i);
    For i=0 to m-1
        计算剔除第i记录后得到的新数据集的信息
        熵ElimInfoEtp[i];
        OMF[i]=totalInfoEtp - ElimInfoEtp[i];
    End For
4)将OMF[i]按大到小排序;
5)输出离群数据集.
```

图 1 算法 OMBIE 的描述

Fig. 1 The Description of OMBIE algorithm

OMBIE 算法的基本思想与 GreedAlg 算法相似,区别在于:1)不需要事先设置参数和阈值,从而避免 GreedAlg 算法不能找出尽可能多的离群点或多识别错误的离群点;2)GreedAlg 算法需要扫描数据集 k 趟(k 是人为事先输入的参数),大大地增加了算法的时间复杂度,而 OMBIE 算法只需对数据扫描一趟,从而大幅度地降低了算法的时间复杂度.

OMBIE 算法的复杂度主要受数据集中的记录

数(m)、每条记录的属性个数(n)、每个属性值的类别个数(c)影响. OMBIE 算法中,主要是 2 个步骤:1)计算每条记录对应的离群数据度量因子;2)将其排序找出离群点. 第 1 步最坏的情况是数据集中每个属性的属性值互不相等,时间复杂度 $O(m \times n \times m)$;但是实际情况下,每个属性的属性值的类别个数 c 远远的小于数据集的记录条数 m,因此,此步骤的时间复杂度应为 $O(m \times n \times c)$. 第 2 步就是一个简单的排序,可以选用一个时间复杂度在 $O(m\log m)$的排序算法. 所以,OMBIE 算法的时间复杂度是 $O(m \times n \times c)$.

4　实验分析

对 OMBIE 算法的性能进行实验分析. 实验平台配置如下:在 PentiumIV 3.0G CPU,512MB 内存,Windows XP 操作系统、DBMS 为 ORACLE9i,采用 Visual C + +6.0实现了 OMBIE、ENBROD [12]、LOF [7]和 GreedAlg 算法[10].

4.1　应用实例分析

选用 UCI 中的 ZOO 数据集,此数据集中有 101 条记录,每条记录拥有 18 个属性——由 1 个动物名称属性、15 个布尔属性、2 个数值属性组成. 其中,15 个布尔属性与动物腿个数的离散数值属性是条件属性;动物类别的离散数值属性是决策属性. 采用文献[12]中使用的方法,只取动物类别是哺乳动物和爬行动物 2 类. 这样做的原因是:1)使用数据集中的所有记录会使离群特征表现不显著;2)为了构造不平衡的分布,构造出来的新数据集中有 41 个哺乳动物(89%)和 5 个爬行动物(11%),其中将爬行动物视为离群数据. 选用此数据集和此方案来做实验是因为 ZOO 数据集的背景知识对于大家是熟知的,算法检测出来的离群数据,可以从客观角度去分析和检验算法的有效性和可行性.

表 1 列出的是从客观实际角度,统计属性集合中每个属性对应的对象集合中的对象是与众不同的次数. 其中,与众不同的评判标准是:对象集中某个属性为某个属性值时,有小于 15.22% 的对象取该属性值,则此对象在这个属性上是与众不同. 从客观实际角度分析和解释,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. 在表 2 中,参数取值是 ENBROD、LOF 和 GreedAlg 算法获得期望目标(将所有的爬行动物数据找出来)的较优取值;而 OMBIE 算法在进行挖掘离群点的过程中,不需要人为进行干预,即不需要事先输入任何参数和阈值. 从表 2 中,可以知道 OMBIE 算法在发现离群点的准确度上优于 ENBROD 和 LOF 算法. 通过表 1 和表 2 的对比,OMBIE 和 GreedAlg 算法更能挖掘出符合客观规律的离群点.

表 1　ZOO 数据集中与众不同的对象入选次数统计

Table 1　The number of distinct objects selected in ZOO data set

入选次数	与众不同的对象				
7	seasnake	/	/	/	/
6	pitviper	/	/	/	/
5	slowworm	/	/	/	/
4	tortoise	tuatara	seal	dolphin	porpoise

表 2　算法的检测准确度对比

Table 2　The contrast of algorithm accuracy

算法	参数	检测 5 个离群爬行动物数据					正确率/%
		1st	2nd	3rd	4th	5th	
ENBROD	MinPts = 45,46	seasnake	pitviper	tortoise	slowworm	seal	80
OMBIE	无需事先设置参数	seasnake	pitviper	slowworm	tortoise	tuatara	100
LOF	MinPts = 5	seasnake	pitviper	slowworm	tortoise	seal	80
GreedAlg	k = 5	seasnake	pitviper	slowworm	tortoise	tuatara	100

4.2　UCI 数据集[13]

为了测试算法对数据集维数的伸缩性,从 UCI 中选取 UCI_ZOO(18 维)、UCI_MUSHROOM(22 维)、UCI_CHESS(36 维)和 UCI_LUNGCANCER(56 维)4 个数据集,分别均匀地加入 3% 具有较大偏差的数据点作为离群点,得到测试数据集. 由图 2 可知,随着测试数据集维数的增加,OMBIE 算法的准确度变化不大并且比 LOF 算法和 ENBROD 算法有所提高,与 GreedAlg 算法的准确度相当.

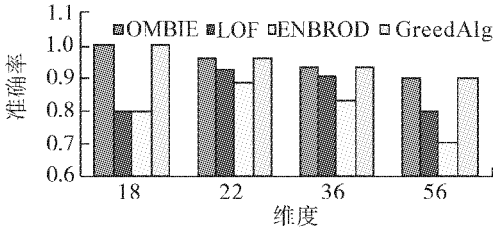


图2 不同算法对数据集维数的伸缩性对比

Fig.2 Scaling of precisions with dataset dimensionality

4.3 恒星光谱数据

采用国家天文台提供的恒星光谱数据,并使用文献[15]中的方法对其进行预处理,预处理后作为实验数据集. 预处理简述如下:1)选定间隔为20的200个波长为3 810,3 530,...,7 790作为属性集,共200个属性;2)依据每一波长处的流量、峰宽和形状,将其离散化为13种数值之一,并作为该波长处的取值. 然后,均匀地加入3%具有较大偏差的数据点作为离群点,得到测试数据集. 采取对测试数据集中的采样数据作预分析的实验方案. 离群检测的准确度的评估标准为

$$\text{准确率}=\frac{\text{正确离群点的个数}}{\text{期望得到离群点的个数}}.$$

图3是测试算法检测离群点准确度的实验结果,从图中可以知道OMBIE比LOF、ENBROD算法的准确度高,与GreedAlg算法的准确度相当. 这是因为OMBIE算法既可以发现离群程度最强的离群点,也可以发现离群程度最弱的离群点,所以它可以发现尽可能多的离群点. 而LOF算法不能发现全部的离群数据是因为高维空间中的数据具有高稀疏性和不规则性的特点,基于密度的异常意义应用到高维数据时失效了,使得LOF算法不能检测到一些离群点. ENBROD算法的准确度受输入参数的影响较大. 尽管实验结果表明GreedAlg算法的准确度与OMBIE算法相当,这是由于此实验方案事先知道数据集中有多少离群点;但是在实际应用中,事先并不知道数据集中有多少离群数据,GreedAlg算法的准确度则会有所降低的. OMBIE算法在挖掘离群点时,不需要人为设置参数,会自动地检测数据集中的离群点;所以不会因为事先不知道数据集中有多少离群数据而受到影响,能更有效地检测出离群点.

图4是测试数据集大小对算法影响的实验结果,OMBIE比LOF、ENBROD及GreedAlg算法(参数k设置为5时)挖掘效率要高. OMBIE算法在挖掘离群点时,无论用户期望产生多少离群点都只扫描一遍数据集,而GreedAlg算法需要扫描k遍数据集,每扫描一遍数据集只能发现1个离群点,因此GreedAlg算法的运行效率则降低了. LOF算法与ENBROD算法在处理高维数据时,索引结构失效,时间复杂度退化为$O(n^2)$.

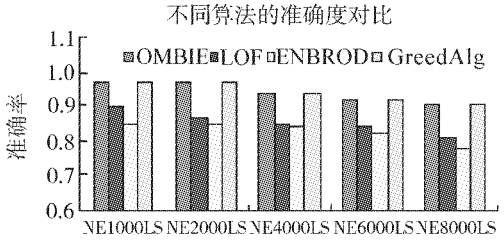


图3 不同算法的准确度对比

Fig.3 Accuracy of OMBIE, LOF, ENBROD and GreedAlg

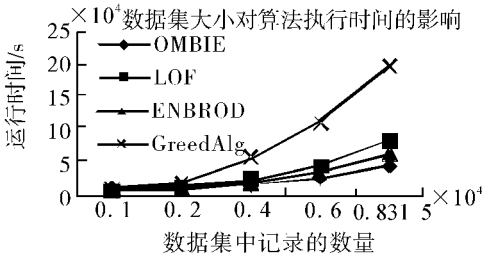


图4 数据集大小对算法协作时间的影响对比

Fig.4 Running time of OMBIE, LOF, ENBROD and GreedAlg

5 结束语

对于高维数据,传统的离群数据挖掘算法不再有效. 本文引入一个离群数据度量因子用来度量每一条记录的离群程度,与LOF算法中通过局部异常因子定义的离群点类似,即离群不再是一个二值属性(不是离群点,就是常规点),摒弃了以前异常定义中非此即彼的绝对异常观念,更加符合现实生活中的应用. OMBIE算法不需要事先人为设置参数和阈值,算法可以自动产生离群点. 由离群数据度量因子定义的离群点,可以对其做出解释(离群点就是使系统无序和杂乱的因素),此外OMBIE算法还可以应用于标称属性数据. 实验结果表明,OMBIE与ENBROD、GreedAlg和LOF算法相比,在发现高维空间的离群数据的能力和效率上都有提高.

参考文献:

[1] HAN Jiawei, KAMBER M. Data mining: concepts and techniques[M]. Bejing: China Machine Press, 2006: 254-255.

[2] HAWKINS D. Identification of outliers [M]. London: Chapman and Hall, 1980: 2-28.

[3] BARNETT V, LEWIS T. Outliers in statistical data[M]. New York: John Wiley & Sons, 1994: 7, 49.

[4] RUTS I, ROUSSEEUW P. Computing depth contours of bivariate point clouds[J]. Computational Statistics and Data

Analysis,1996,23(1):153-168.

[5] ARNING A, AGRAWAL R, RAGHAVAN P. A linear method for deviation in large database[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portlan,Oregon,USA,1996:164-169.

[6] KNORR E M, NG R T. Algorithms of mining distance-based outliers in large datasets[C]//Proc of Int Conf on Very Large Database (VLDB'98). New York,USA, 1998: 392-402.

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

[8] 熊家军,李庆华. 信息熵理论与入侵检测聚类问题研究[J]. 小型微型计算机系统,2005, 26(7):1163-1166.
XIONG Jiajun, LI Qinghua. Study on clustering problem for intrusion detection with information entropy[J]. Mini-micro Systems,2005,26(7):1163-1166.

[9] 薛　萍,金鸿章,王　双. 应用最大熵原理分析通信系统脆性风险[J]. 电机与控制学报,2007,11(2): 74-78.
XUE Ping, JIN Hongzhang, WANG Shuang. Application of the maximum entropy principle to brittleness risk analysis on communication system[J]. Electric Machines and Control, 2007, 11(2): 74-78.

[10] HE Zengyou, XU Xiaofei, DENG Shengchun. A fast greedy algorithm for outlier mining[C]//Proceedings of PAKDD' 2006 (LNAI3918). Berlin: Springer-Verlag, 2006:567-576.

[11] 倪巍伟,陈　耿,陆介平,等. 基于局部信息熵的加权子空间离群点检测算法[J]. 计算机研究与发展,2008,45(7):1189-1192.
NI Weiwei, CHEN Geng, LU Jieping. Local entropy based weighted subspace outlier mining algorithm [J]. Journal of Computer Research and Development, 2008, 45 (7): 1189-1192.

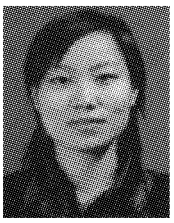
[12] 于绍越,商　琳. ENBROD:基于信息熵的相对离群点的检测方法[J]. 南京大学学报:自然科学版,2008,44(2):1189-1194.
YU Shaoyue, SHANG Lin. An entropy-based algorithm to detect relative outliers: ENBROD[J]. Journal of Nanjing University:Natural Sciences, 2008,44 (2):1189-1194.

[13] DUDA R O, HART P E, STOCK D G. Pattern classification[M]. 2nd ed. Beijing: China Machine Press, 2003: 317-356.

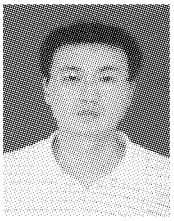
[14] NEWMAN D J, HETTICH S, BLAKE C L, et al. UCI repository of machine learning databases[DB/OL]. Irvine, CA: University of California, Department of Information and Computer Science, 1998. [2008-09-25] http://www.ics.uci.edu/~mlearn/MLRepository.html.

[15] 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,2007,34(9):1060-1066.
ZHANG Jifu, JIANG Yiyong, HU Lihua, et al. A concept lattice based recognition method of celestial spectra outliers [J]. Acta Automatica Sinica, 2007,34(9):1060-1066.

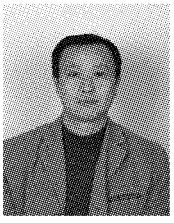
作者简介:



张　贺,女,1981 年生,硕士研究生. 主要研究方向为数据挖掘.



蔡江辉,男,1978 年生. 讲师,主要研究方向为离群数据挖掘.



张继福,男,1963 年生,教授,博士. 主要研究方向为数据挖掘、模式识别与智能信息系统. 已主持完成国家自然科学基金、国家"863"计划子课题等省部级以上科研项目 10 余项,发表学术论文 100 余篇,其中被 SCI、EI30 余篇.