文章编号: 1001-0920 (2013) 06-0867-06

一种基于信息熵的异常数据挖掘算法

陈玉明[1], 吴克寿[1], 李向军[1,2]

(1. 厦门理工学院 计算机科学与技术系，福建 厦门 361024；2. 南昌大学 计算机科学与技术系，南昌 330031)

摘 要: 信息熵是粒计算理论中度量不确定信息的重要工具之一, 已有的异常数据挖掘算法主要针对确定性的异常数据挖掘, 采用信息熵度量不确定性数据进行异常数据挖掘的研究报道较少. 鉴于此, 在引入信息熵概念的基础上, 定义基于信息熵的异常度来度量数据之间的异常程度, 并提出基于信息熵的异常数据挖掘算法, 该算法可有效进行异常数据的挖掘. 理论分析与实验结果表明, 所提出算法是有效可行的.

关键词: 粗糙集；粒计算；异常数据挖掘；信息熵

中图分类号: TP181 **文献标志码:** A

A kind of outlier mining algorithm based on information entropy

CHEN Yu-ming[1], *WU Ke-shou*[1], *LI Xiang-jun*[1,2]

(1. Department of Computer Science and Technology，Xiamen University of Technology，Xiamen 361024，China; 2. Department of Computer Science and Technology，Nanchang University，Nanchang 330031，China．Correspondent: CHEN Yu-ming，E-mail：cym0620@163.com)

Abstract: Information entropy is one of the important tool to measure the uncertainty information in the information theory. Many existing algorithms of outlier mining mainly aim at certainty data, and little work has been done for the uncertainty data aiming to outlier mining based on the information entropy. Therefore, after introducing information entropy concept, outlier degree based on information entropy is defined for measuring the outlier data. Furthermore, an algorithm for outlier mining based on information entropy is proposed, which can effectively obtain outliers from data set. Finally, theoretical analysis and experimental results show that the algorithm is efficient and feasible.

Key words: rough sets；granular computing；outlier data mining；information entropy

0 引 言

异常数据是数据集中偏离大部分对象的数据, 其表现与大多数常规对象有明显的差异, 甚至使人怀疑它们可能由另外一种完全不同的机制所产生[1]. 随着数据挖掘技术的飞速发展, 异常数据挖掘受到国内外学者的广泛关注, 成为数据挖掘领域的一个重要分支. 近年来, 异常数据挖掘在网络入侵检测、地质灾害预报、疾病诊断、故障检测、恐怖活动防范、信用卡欺诈检测等诸多领域得到广泛应用[2-3]. 现有的异常数据挖掘方法主要有基于距离的方法[4]、基于统计的方法[5]、基于密度的方法[6]和基于聚类的方法[7]. 国内外众多学者对这些方法进行了深入研究, 并取得丰硕成果, 但仍存在一些不足和缺陷. 基于距离的方法中, 距离函数与参数的选择存在一定困难; 基于统计的方法中, 要求预先知道数据的分布情况, 但数据的分布函数难以预先获得; 基于密度的方法中时间复杂度较大; 基于聚类的方法主要侧重于聚类问题. 这些问题极大地限制了异常数据挖掘方法的发展与应用, 且主要处理确定性数据, 对于不确定性的信息处理缺乏有效的理论模型和方法.

粒计算是人工智能领域新兴起的一个研究方向, 是信息处理的一种新的概念和计算范式[8], 主要用于处理不确定的、模糊的、不精确的、部分真的和海量的信息, 其基本思想是利用不同粒度上的信息进行问题求解. 粒计算模型主要包括模糊集模型[9]、粗糙集模型[10]和商空间模型[11]. 近年来, 该理论在知识获取、机器学习、数据挖掘、智能控制和模式识别等多个领域得到了广泛应用[8,12]. 在基于粒计算的不确定性理论研究中, 不确定性数据的度量是重要的研究内容之一, 也是知识获取的关键步骤. 国内学者提出的

收稿日期: 2012-02-13；**修回日期:** 2012-05-19.

基金项目: 国家自然科学青年基金项目(61103246).

作者简介: 陈玉明(1977−), 男, 博士, 从事粗糙集、粒计算的研究；吴克寿(1975−), 男, 副教授, 博士, 从事粗糙集、数据挖掘与加密算法等研究.

不确定性数据度量工具包括信息熵[13]、粗糙熵[14]和知识粒度[15-17]. 目前, 这些度量工具受到国内外学者的极大关注, 取得了一定成果, 但主要侧重于理论上的研究和探讨, 具体应用实例较为少见.

鉴于此, 本文引入不确定性数据的度量工具信息熵, 并将其应用于异常数据挖掘领域, 提出基于信息熵的异常数据挖掘理论和方法, 为异常数据挖掘处理不确定性数据提供一条新的途径, 拓展不确定性度量理论在数据挖掘领域的应用范围, 为不确定性理论开辟新的应用空间. 在此基础上, 进一步提出了基于信息熵的异常数据挖掘算法, 该算法不需要任何先验知识, 采用信息熵度量对象间的距离和异常度, 能够有效挖掘出异常程度较高的数据. UCI 数据集测试表明, 所提出方法具有较好的准确率.

1 信息熵概念

信息熵是数学上的一个抽象概念, 表示特定信息的出现概率, 在信息系统中可以表示信息的不确定性. 一个信息系统越是确定性的, 其信息熵越低; 反之, 信息熵越高. 信息熵也可以说是信息系统不确定性程度的一个度量.

定义 1 设 $\mathrm{IS}=(U,A,V,f)$ 为信息系统[12]. 其中: U 为非空有限集, 称为论域; A 为有限属性集; $V=\bigcup_{a\in A}Va$, Va 为属性 a 的值域; $f:U\times A\to V$ 为信息函数, 即对于 $\forall x\in U, a\in A$, 有 $f(x,a)\in Va$. 任意属性子集 $B\subseteq A$ 决定一个二元不可区分关系 $\mathrm{IND}(B)$, 有

$$\mathrm{IND}(B)=\{(x,y)\in U\times U|\forall a\in B, f(x,a)=f(y,a)\}. \tag{1}$$

$U/\mathrm{IND}(B)$ 构成了 U 的一个划分, 称为 U 上的一个知识, 其中每个等价类称为一个知识粒. 为了方便计算, 将 $U/\mathrm{IND}(B)$ 简记为 U/B. 特别地, 若

$$U/B=\omega=\{[x]_B|[x]_B=\{x\}, x\in U\},$$

则称为恒等关系; 若

$$U/B=\delta=\{[x]_B|[x]_B=U, x\in U\},$$

则称为全域关系.

定义 2 设 $\mathrm{IS}=(U,A,V,f)$ 为信息系统, $U/A=\{X_1,X_2,\cdots,X_m\}$, 则 A 的信息熵[8]定义为

$$H(A)=-\sum_{i=1}^{m}p(X_i)\log p(X_i). \tag{2}$$

其中: $p(X_i)=|X_i|/|U|$, $i=1,2,\cdots,m$, $|E|$ 为集合 E 的基数.

性质 1 $\forall B\subseteq A$, 有 $0\leqslant H(B)\leqslant\ln|U|$.

当 U/B 为恒等关系, 即 $U/B=\omega$ 时, B 的信息熵达到最大值 $\ln|U|$; 当 U/B 为全域关系, 即 $U/B=\delta$ 时, B 的信息熵达到最小值 0.

定义 3 设 $\mathrm{IS}=(U,A,V,f)$ 是一个信息系统, $B,C\subseteq A$, $U/B=\{X_1,X_2,\cdots,X_m\}$, $U/C=\{Y_1,Y_2,\cdots,Y_n\}$, 若 $\forall X_i\in U/B$, $\exists Y_j\in U/C(X_i\subseteq Y_j)$, 则称 B 比 C 更细, 记为 $B\leqslant C$.

定义 4 设 $\mathrm{IS}=(U,A,V,f)$ 是一个信息系统, $B,C\subseteq A$, $U/B=\{X_1,X_2,\cdots,X_m\}$, $U/C=\{Y_1,Y_2,\cdots,Y_n\}$, 若 $\forall X_i\in U/B$, $\exists Y_j\in U/C(X_i\subset Y_j)$, 则称 B 比 C 严格细, 记为 $B\prec C$.

定理 1 设 $\mathrm{IS}=(U,A,V,f)$ 是一个信息系统, $B,C\subseteq A$, $U/B=\{X_1,X_2,\cdots,X_m\}$, $U/C=\{Y_1,Y_2,\cdots,Y_n\}$, 若 $B\leqslant C$, 则 $H(B)\geqslant H(C)$.

证明 1) 由 $B\prec C$ 可知, $\forall X_i\in U/B$, $\exists Y_j\in U/C(X_i\subset Y_j)$; 反之, 对于任意 Y_j, 存在 s 个 X_i, 满足

$$\frac{|Y_j|}{|U|}=\frac{|X_1|}{|U|}+\frac{|X_2|}{|U|}+\cdots+\frac{|X_s|}{|U|}.$$

且有

$$\begin{aligned}&\frac{|X_1|}{|U|}\log\frac{|X_1|}{|U|}+\frac{|X_2|}{|U|}\log\frac{|X_2|}{|U|}+\cdots+\\&\frac{|X_s|}{|U|}\log\frac{|X_s|}{|U|}<\frac{|Y_j|}{|U|}\log\frac{|Y_j|}{|U|},\\&-\Big(\frac{|X_1|}{|U|}\log\frac{|X_1|}{|U|}+\frac{|X_2|}{|U|}\log\frac{|X_2|}{|U|}+\cdots\\&+\frac{|X_s|}{|U|}\log\frac{|X_s|}{|U|}\Big)>-\frac{|Y_j|}{|U|}\log\frac{|Y_j|}{|U|},\\&-\sum_{i=1}^{m}\frac{|X_i|}{|U|}\log\frac{|X_i|}{|U|}>-\sum_{i=1}^{n}\frac{|Y_i|}{|U|}\log\frac{|Y_i|}{|U|},\end{aligned}$$

因此有 $H(B)>H(C)$.

2) 由 $B=C$ 可知 $|X_i|/|U|=|Y_j|/|U|$, 因此有 $H(B)=H(C)$. □

定理 2 设 $\mathrm{IS}=(U,A,V,f)$ 是一个信息系统, $B,C\subseteq A$, 有:

1) 若 $B\supset C$, 则 $H(B)\geqslant H(C)$;

2) 若 $B\subset C$, 则 $H(B)\leqslant H(C)$.

由定理 1 可得证, 此略.

定理 1 反映了知识的粗细与信息熵的关系: 知识越粗, 信息熵越小, 不确定性越小; 知识越细, 信息熵越大, 不确定性越大. 定理 2 反映了属性与信息熵的变化: 属性增加, 划分更细, 信息熵变大, 不确定性增大; 属性减少, 划分更粗, 信息熵变小, 不确定性减少. 由此可见, 信息系统中属性的增减使得系统的粒度发生变化 (粗细程度), 同时信息熵和不确定性也相应变化. 因此, 信息熵能够度量信息系统的不确定性.

2 基于信息熵的异常数据挖掘

信息熵可以度量知识的不确定性, 近年来, 在数据挖掘等领域获得了广泛应用. 然而, 采用信息熵进行异常数据挖掘的研究并不多见, 下面结合信息熵的定义, 进一步定义对象相对于等价类的相对信息熵概

念, 并给出异常度概念来度量数据的异常.

2.1 基于信息熵的异常数据挖掘方法

在数据挖掘系统中, 数据集一般采用决策表或信息系统的方式来表示和处理. 本文基于信息系统讨论异常的定义和检测问题, 遵循 Hawkins 关于异常的定义[1], 对异常数据进行如下定义.

定义 5　设 IS $=(U,A,V,f)$ 是一个信息系统, $\forall x\in U$, 若对象 x 与所有非异常对象的距离较远, 且与所有异常对象的距离较近, 则称对象 x 为异常对象 (异常点或异常数据).

为了计算对象间的距离, 定义相对信息熵来表示距离函数, 对象与其他对象的距离之和表示该对象的异常程度, 异常程度高的对象即为异常对象.

定义 6　设 IS $=(U,A,V,f)$ 是一个信息系统, $U/A=\{X_1,X_2,\cdots,X_m\}$, 若

$$\forall x\in U,\ \{U-\{x\}\}/A=\{X_1^{'},X_2^{'},\cdots,X_m^{'}\},$$

$$H_x(A)=-\sum_{i=1}^{m}p(X_i^{'})\log p(X_i^{'}),$$

则对象 x 相对于 A 的对象相对信息熵定义为

$$RH_A(x)=H_x(A)/H(A). \tag{3}$$

其中: $H(A)$ 为 A 的信息熵, $H_x(A)$ 为删除对象 x 后 A 的信息熵. 信息熵可以表示不确定性信息的程度, 因此, 对象相对信息熵可以度量 x 的不确定性程度. 如果删去对象 x 后信息熵变化较小, 则 x 的不确定性程度较小; 反之, x 的不确定性程度较大.

定义 7　设 IS $=(U,A,V,f)$ 是一个信息系统, $A=\{a_1,a_2,\cdots,a_k\}$, 按信息熵从小到大排序, 形成序列 $S=\langle a_1^{'},a_2^{'},\cdots,a_k^{'}\rangle$, 其中 $H(\{a_i^{'}\})\leqslant H(\{a_{i+1}^{'}\})$, 称 S 为信息系统中单属性按信息熵递增序列.

定义 8　设 IS $=(U,A,V,f)$ 是一个信息系统, $S=\langle a_1^{'},a_2^{'},\cdots,a_k^{'}\rangle$ 为单属性按信息熵递增序列, 给定序列 AS$=\langle A_1^{'},A_2^{'},\cdots,A_k^{'}\rangle$, 其中 $A_1^{'}=A$, $A_k^{'}=\{a_1^{'}\}$, $A_{i+1}^{'}=A_i^{'}-\{a_i^{'}\}$, 称 AS 为信息系统中属性子集序列.

为了刻画数据集中每个对象的异常程度, 在对象相对信息熵的基础上引入异常度的概念来表示信息系统中每个对象的异常程度.

定义 9　设 IS $=(U,A,V,f)$ 是一个信息系统, $S=\langle a_1^{'},a_2^{'},\cdots,a_k^{'}\rangle$ 为单属性按信息熵递增序列, AS $=\langle A_1^{'},A_2^{'},\cdots,A_k^{'}\rangle$为属性子集序列, $\forall B\subseteq A$, $W_B(x)=|[x]_B|/|U|$ 表示 x 的权重, 对象 x 的异常度定义为

$$\mathrm{EOF}(x)=1-\Big(\sum_{i=1}^{k}RH_{\{a_i\}}(x)W_{\{a_i\}}(x)+\sum_{i=1}^{k}RH_{\{A_i\}}(x)W_{\{A_i\}}(x)\Big)\Big/|U|. \tag{4}$$

定义 10　设 IS $=(U,A,V,f)$ 是一个信息系统, 令 v 为给定的阈值, 对于任意 $x\in U$, 如果 $\mathrm{EOF}(x)>v$, 则称 x 为信息系统 IS 中一个基于信息熵的异常对象, 其中 $\mathrm{EOF}(x)$ 为对象 x 的异常度.

2.2 基于信息熵的异常数据挖掘算法

依据信息熵异常度的定义, 基于信息熵的异常数据挖掘算法描述如下.

算法 1　IEOM(information entropy based outlier mining).

输入: 信息系统 IS $=(U,A,V,f)$ ($A=\{a_1,a_2,\cdots,a_m\}$, $m=|A|$, $n=|U|$), 阈值 v;

输出: 异常对象的集合 O.

Step 1: 初始化 $m=|A|$, $n=|U|$, $O=\phi$.

Step 2: 对于信息系统 IS 中的每个属性 a_i ($1\leqslant i\leqslant m$), 循环执行如下操作:

Step 2.1: 根据 U 中对象在属性 a_i 上的取值, 进行基数排序;

Step 2.2: 求出划分 $U/\mathrm{IND}(\{a_i\})$;

Step 2.3: 计算信息熵 $H(\{a_i\})$.

Step 3: 根据定义 7 构造单属性按信息熵递增序列 $S=\langle a_1,a_2,\cdots,a_m\rangle$.

Step 4: 根据定义 8 构造属性子集序列 AS $=\langle A_1,A_2,\cdots,A_m\rangle$.

Step 5: 对于属性子集序列 AS 中的每个属性子集 A_i ($1\leqslant i\leqslant m$), 循环执行如下操作:

Step 5.1: 根据 U 中对象在属性子集 A_i 上的取值进行基数排序;

Step 5.2: 求出划分 $U/\mathrm{IND}(\{A_i\})$;

Step 5.3: 计算信息熵 $H(\{A_i\})$.

Step 6: 对于 U 中的每个对象 x_i ($1\leqslant i\leqslant n$), 循环执行如下操作:

Step 6.1: for $j=1$ to m, 循环执行如下操作:

Step 6.1.1: 计算对象 x_i 对于单个属性的相对信息熵 $\mathrm{RH}_{\{a_j\}}(x_i)$;

Step 6.1.2: 计算对象 x_i 对于属性子集的相对信息熵 $\mathrm{RH}_{\{A_j\}}(x_i)$;

Step 6.1.3: 计算权值 $W_{\{a_j\}}(x_i)$ 和 $W_{\{A_j\}}(x_i)$.

Step 6.2: 计算对象 x_i 的异常度 $\mathrm{EOF}(x_i)$.

Step 6.3: 若 $\mathrm{EOF}(x_i)>v$, 则 $O=O\bigcup x_i$.

Step 7: 输出异常对象的集合 O.

算法 IEOM 主要涉及信息熵的计算, 而信息熵来源于等价类的计算. 本文采用文献 [18] 中基数排序的思想计算等价类, 时间复杂度降为线性. 因此, 最坏情况下 IEOM 算法的时间复杂度为 $O(mn)$. 其中: m 为

属性的个数, n 为对象的个数.

2.3 示例说明

给定一个信息系统 $\mathrm{IS}=(U,A,V,f)$. 其中: $U=\{x_1,x_2,x_3,x_4,x_5,x_6\}$, $A=\{a,b,c\}$. 系统如表 1 所示, 设异常阈值 $v=0.75$.

表 1 信息系统

U	a	b	c
x_1	0	0	0
x_2	1	2	1
x_3	0	2	2
x_4	2	2	0
x_5	0	2	1
x_6	1	1	2

计算属性集 A 中每个属性的划分为

$U/\mathrm{IND}(a)=\{\{x_1,x_3,x_5\},\{x_2,x_6\},\{x_4\}\}$,

$U/\mathrm{IND}(b)=\{\{x_1\},\{x_2,x_3,x_4,x_5\},\{x_6\}\}$,

$U/\mathrm{IND}(c)=\{\{x_1,x_4\},\{x_2,x_5\},\{x_3,x_6\}\}$.

根据信息熵的定义, 单属性的信息熵分别为

$$H(\{a\})=-\sum_{i=1}^{3}p(X_i)\log p(X_i)=$$
$$-\left(\frac{1}{2}\log\left(\frac{1}{2}\right)+\frac{1}{3}\log\left(\frac{1}{3}\right)+\frac{1}{6}\log\left(\frac{1}{6}\right)\right)=0.439\,2,$$
$$H(\{b\})=-\sum_{i=1}^{3}p(X_i)\log p(X_i)=$$
$$-\left(\frac{1}{6}\log\left(\frac{1}{6}\right)+\frac{2}{3}\log\left(\frac{2}{3}\right)+\frac{1}{6}\log\left(\frac{1}{6}\right)\right)=0.376\,8,$$
$$H(\{c\})=-\sum_{i=1}^{3}p(X_i)\log p(X_i)=$$
$$-\left(\frac{1}{3}\log\left(\frac{1}{3}\right)+\frac{1}{3}\log\left(\frac{1}{3}\right)+\frac{1}{3}\log\left(\frac{1}{3}\right)\right)=0.477\,1.$$

去除某个对象后的信息熵为

$$H_{x1}(\{a\})=H_{x3}(\{a\})=H_{x5}(\{a\})=$$
$$-\left(\frac{2}{5}\log\left(\frac{2}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)\right)=0.458\,2,$$
$$H_{x2}(\{a\})=H_{x6}(\{a\})=$$
$$-\left(\frac{3}{5}\log\left(\frac{3}{5}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)\right)=0.412\,7,$$
$$H_{x4}(\{a\})=-\left(\frac{3}{5}\log\left(\frac{3}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)\right)=0.292\,3,$$
$$H_{x1}(\{b\})=H_{x6}(\{b\})=$$
$$-\left(\frac{4}{5}\log\left(\frac{4}{5}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)\right)=0.217\,3,$$
$$H_{x2}(\{b\})=H_{x3}(\{b\})=H_{x4}(\{b\})=H_{x5}(\{b\})=$$
$$-\left(\frac{1}{5}\log\left(\frac{1}{5}\right)+\frac{3}{5}\log\left(\frac{3}{5}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)\right)=0.412\,7,$$
$$H_{x1}(\{c\})=H_{x4}(\{c\})=$$
$$-\left(\frac{1}{5}\log\left(\frac{1}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)\right)=0.458\,2,$$
$$H_{x2}(\{c\})=H_{x5}(\{c\})=$$
$$-\left(\frac{1}{5}\log\left(\frac{1}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)\right)=0.458\,2,$$
$$H_{x3}(\{c\})=H_{x6}(\{c\})=$$
$$-\left(\frac{1}{5}\log\left(\frac{1}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)+\frac{2}{5}\log\left(\frac{2}{5}\right)\right)=0.458\,2.$$

根据相对信息熵的定义有

$$\mathrm{RH}_{\{a\}}(x_1)=\mathrm{RH}_{\{a\}}(x_3)=\mathrm{RH}_{\{a\}}(x_5)=$$
$$0.458\,2/0.439\,2=1.043\,3,$$
$$\mathrm{RH}_{\{a\}}(x_2)=\mathrm{RH}_{\{a\}}(x_6)=$$
$$0.412\,7/0.439\,2=0.939\,7,$$
$$\mathrm{RH}_{\{a\}}(x_4)=0.292\,3/0.439\,2=0.665\,5,$$
$$\mathrm{RH}_{\{b\}}(x_1)=\mathrm{RH}_{\{b\}}(x_6)=$$
$$0.217\,3/0.376\,8=0.576\,7,$$
$$\mathrm{RH}_{\{b\}}(x_2)=\mathrm{RH}_{\{b\}}(x_3)=\mathrm{RH}_{\{b\}}(x_4)=$$
$$\mathrm{RH}_{\{b\}}(x_5)=0.412\,7/0.376\,8=1.095\,3,$$
$$\mathrm{RH}_{\{c\}}(x_1)=\mathrm{RH}_{\{c\}}(x_4)=$$
$$0.458\,2/0.477\,1=0.960\,4,$$
$$\mathrm{RH}_{\{c\}}(x_2)=\mathrm{RH}_{\{c\}}(x_5)=$$
$$0.458\,2/0.477\,1=0.960\,4,$$
$$\mathrm{RH}_{\{c\}}(x_3)=\mathrm{RH}_{\{c\}}(x_6)=$$
$$0.458\,2/0.477\,1=0.960\,4.$$

根据信息熵大小, 得到单属性按信息熵递增序列 $S=\langle b,a,c\rangle$ 和属性子集序列 $\mathrm{AS}=\langle A_1,A_2,A_3\rangle=\langle\{a,b,c\},\{a,c\},\{c\}\rangle$. 针对属性子集序列求其对象相对信息熵为

$$\mathrm{RH}_{\{A1\}}(x_1)=\mathrm{RH}_{\{A1\}}(x_2)=$$
$$\mathrm{RH}_{\{A1\}}(x_3)=\mathrm{RH}_{\{A1\}}(x_4)=$$
$$\mathrm{RH}_{\{A1\}}(x_5)=\mathrm{RH}_{\{A1\}}(x_6)=0.898\,2,$$
$$\mathrm{RH}_{\{A2\}}(x_1)=\mathrm{RH}_{\{A2\}}(x_2)=$$
$$\mathrm{RH}_{\{A2\}}(x_3)=\mathrm{RH}_{\{A2\}}(x_4)=$$
$$\mathrm{RH}_{\{A2\}}(x_5)=\mathrm{RH}_{\{A2\}}(x_6)=0.898\,2,$$
$$\mathrm{RH}_{\{A3\}}(x_1)=\mathrm{RH}_{\{A3\}}(x_2)=$$
$$\mathrm{RH}_{\{A3\}}(x_3)=\mathrm{RH}_{\{A3\}}(x_4)=$$
$$\mathrm{RH}_{\{A3\}}(x_5)=\mathrm{RH}_{\{A3\}}(x_6)=0.960\,4.$$

对象 x 的权重为

$$W_{\{a\}}(x_1)=W_{\{a\}}(x_3)=W_{\{a\}}(x_5)=1/2,$$
$$W_{\{a\}}(x_2)=W_{\{a\}}(x_6)=1/3,\ W_{\{a\}}(x_4)=1/6,$$
$$W_{\{b\}}(x_1)=W_{\{b\}}(x_6)=1/6,$$
$$W_{\{b\}}(x_2)=W_{\{b\}}(x_3)=$$
$$W_{\{b\}}(x_4)=W_{\{b\}}(x_5)=2/3,$$

$$W_{\{c\}}(x_1)=W_{\{c\}}(x_2)=W_{\{c\}}(x_3)=$$
$$W_{\{c\}}(x_4)=W_{\{c\}}(x_5)=W_{\{c\}}(x_6)=1/3,$$
$$W_{\{A1\}}(x_1)=W_{\{A1\}}(x_2)=W_{\{A1\}}(x_3)=$$
$$W_{\{A1\}}(x_4)=W_{\{A1\}}(x_5)=W_{\{A1\}}(x_6)=1/6,$$
$$W_{\{A2\}}(x_1)=W_{\{A2\}}(x_2)=W_{\{A2\}}(x_3)=$$
$$W_{\{A2\}}(x_4)=W_{\{A2\}}(x_5)=W_{\{A2\}}(x_6)=1/6,$$
$$W_{\{A3\}}(x_1)=W_{\{A3\}}(x_2)=W_{\{A3\}}(x_3)=$$
$$W_{\{A3\}}(x_4)=W_{\{A3\}}(x_5)=W_{\{A3\}}(x_6)=1/3.$$

由计算得出对象 $x_1\sim x_6$ 的异常度分别为

$$\mathrm{EOF}(x_1)\approx 0.740<v,\ \mathrm{EOF}(x_2)\approx 0.669<v,$$
$$\mathrm{EOF}(x_3)\approx 0.660<v,\ \mathrm{EOF}(x_4)\approx 0.703<v,$$
$$\mathrm{EOF}(x_5)\approx 0.635<v,\ \mathrm{EOF}(x_6)\approx 0.775>v.$$

对象 x_6 的异常度大于异常阈值, 判断为异常对象.

3　实验分析

为了考察本文算法的有效性, 采用 UCI 数据集中的 Lymphography 数据集和 Wisconsin Breast Cancer 数据集进行测试. 在此基础上, 将本文算法 IEOM 与 DIS 算法[5]、FindCBLOF 算法[16]和 KNN 算法[20]进行性能比较.

3.1　Lymphography 数据集

首先对 Lymphography 数据集进行实验. 该数据集包括 148 个对象, 18 个条件属性, 1 个决策属性. 148 个对象分为 normal find(1.35%), metastases(54.73%), malign lymph(41.22%) 和 fibrosis(2.7%) 四类. normal find 和 fibrosis 两个类别所占比例小, 是稀少类, 可以看作异常对象. 实验结果如表 2 所示.

表 2　Lymphography 数据集实验结果

异常度前 k% 的对象/对象个数	属于稀少类的对象个数/属于稀少类的比例			
	IEOM	DIS	FindCBLOF	KNN
5%/7	6/100%	5/83%	4/67%	5/83%
6%/9	6/100%	6/100%	4/67%	5/83%
8%/12	6/100%	6/100%	4/67%	6/100%
20%/30	6/100%	6/100%	6/100%	6/100%

表 2 中, 对于 U 中每个对象 x, 分别采用 IEOM 算法、DIS 算法、FindCBLOF 算法和 KNN 算法计算对象 x 的异常度, 并根据异常度由高到低对 U 中的对象进行排序; 然后统计异常度前 k% 的对象覆盖了多少稀少类对象. "异常度前 k% 的对象/对象个数"表示在采用某种算法计算 U 中对象的异常程度并排序后, 异常度排在前 k% 的对象和这些对象的个数, "属于稀少类的对象个数"是指在由该算法所计算出的异常度排在前 k% 的对象中, 属于稀少类的对象个数. 稀少类可以作为异常数据, 因此, 属于稀少类对象的个数即为属于异常对象的个数.

由表 2 可见, 对于 Lymphography 数据集, IEOM 算法的性能明显好于其他方法. 在计算由各类方法所找出的异常对象中真正的异常对象所占的比例方面, IEOM 算法得到的结果均高于其他算法. 例如, 在各类方法所找出的异常度排在前 5% 的对象中 (共计 7 个对象), IEOM 算法检测出的真正的异常对象为 6 个 (占所有异常对象比例为 100%), DIS 和 KNN 算法检测出 5 个异常对象, FindCBLOF 算法检测出 4 个异常对象, 比例均低于 IEOM 算法.

3.2　Wisconsin Breast Cancer 数据集

Wisconsin Breast Cancer 数据集包含 699 个对象, 8 个条件属性, 1 个决策属性. 为了形成不均匀的分布, 参照文献 [21], 从数据集中移去一些属于"malignant"类的对象. 最终数据集包括 483 个对象, 其中 39 个属于"malignant"类, 444 个属于"benign"类. Wisconsin Breast Cancer 数据集中, "malignant"属于稀少类, 看作异常数据, 实验结果如表 3 所示.

表 3　Wisconsin Breast Cancer 数据集实验结果

异常度前 k% 的对象/对象个数	属于稀少类的对象个数/属于稀少类的比例			
	IEOM	DIS	FindCBLOF	KNN
1%/4	4/10%	4/10%	4/10%	4/10%
2%/8	7/18%	5/13%	7/18%	8/21%
4%/16	14/36%	11/28%	14/36%	16/41%
6%/24	21/54%	18/46%	21/54%	20/51%
8%/32	28/72%	24/62%	27/69%	27/69%
10%/40	32/82%	29/74%	32/82%	32/82%
12%/48	38/97%	36/92%	35/90%	37/95%
14%/56	39/100%	39/100%	38/97%	39/100%
16%/64	39/100%	39/100%	39/100%	39/100%
18%/72	39/100%	39/100%	39/100%	39/100%
20%/80	39/100%	39/100%	39/100%	39/100%
28%/112	39/100%	39/100%	39/100%	39/100%

由表 3 可见, 对于 Wisconsin Breast Cancer 数据集, DIS 算法、FindCBLOF 算法和 KNN 算法的性能非常接近, IEOM 算法的性能好于这 3 种算法.

4　结　论

本文将信息论中的信息熵概念应用于异常数据挖掘领域, 在粒计算的框架中提出基于信息熵的异常数据挖掘方法, 并给出了一种基于信息熵的异常数据挖掘算法. 该算法充分利用信息熵度量不确定性数据方面的优势, 可以在不确定性数据中高效地挖掘出异常对象. 目前, 采用信息熵的方法进行异常数据挖掘的研究还较为少见, 本文的研究拓展了信息论和粒计算理论研究的应用范围, 为异常数据挖掘研究提供了一条新的途径. 理论分析和实验表明, 本文所提出的算法是有效可行的.

参考文献(References)

[1] Hawkins D. Identifications of outliers[M]. London: Chapman and Hall, 1980: 1-2.

[2] Han J W. Data mining: Concepts and technologies[M]. San Francisco: Morgan Kaufmann, 2001: 381-394.

[3] 江峰, 杜军威, 眭跃飞, 等. 基于边界和距离的离群点检测[J]. 电子学报, 2010, 38(3): 700-705.
(Jiang F, Du J W, Sui Y F, et al. Outlier detection based on boundary and distance[J]. Acta Electronica Sinica, 2010, 38(3): 700-705.)

[4] Knorr E, Ng R, Tucakov V. Distance-based outliers: Algorithms and applications[J]. J of Very Large Databases, 2000, 8(3/4): 237-253.

[5] Rousseeuw P J, Leroy A M. Robust regression and outlier detection[M]. New York: John Wiley & Sons, 1987: 1-18.

[6] Johnson T, Kwok I, Ng R T. Fast computation of 2-dimensional depth contours[C]. Proc of the 4th Int Conf on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998: 224-228.

[7] Jain A K, Murty M N, Flynn P J. Data clustering: A review[J]. ACM Computing Surveys, 1999, 31(3): 264-323.

[8] 苗夺谦, 王国胤, 刘清, 等. 粒计算: 过去、现在与展望[M]. 北京: 科学出版社, 2007: 121-136.
(Miao D Q, Wang G Y, Liu Q, et al. Granular computing: Past, present and prospects[M]. Beijing: Science Press, 2007: 121-136.)

[9] Zadeh L A. Fuzzy sets and information granularity[C]. Advances in Fuzzy Set Theory and Applications. Amsterdam: North-Holland, 1979: 3-18.

[10] Pawlak Z. Rough sets[J]. Int J of Computer and Information Science, 1982, 11(5): 341-356.

[11] Zhang B, Zhang L. Theory and applications of problem solving[M]. New York: Elsevier Science Publishers, 1992: 86-99.

[12] Wang G Y. Granular computing based data mining in the views of rough set and fuzzy set[C]. Novel Developments in Granular Computing: Applications for Advanced Human Reasoning and Soft Computation. Hershey: IGI Global, 2010: 401-416.

[13] 苗夺谦, 王珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
(Miao D Q, Wang J. An information representation of the concepts and operations in rough set theory[J]. J of Software, 1999, 10(2): 113-116.)

[14] Pal Sankar K, Uma Shankar B, Pabitra M. Granular computing, rough entropy and object extraction[J]. Pattern Recognition Letters, 2005, 26(16): 2509-2517.

[15] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.
(Miao D Q, Fan S D. The calculation of knowledge granulation and its application[J]. Systems Engineering Theory & Practice, 2002, 22(1): 48-56.)

[16] Qian Y H, Liang J Y, Wu W Z, et al. Partial orderings of information granulations: A further investigation[J]. Expert Systems, 2012, 29(1): 3-24.

[17] 陈玉明, 苗夺谦. 基于幂图的属性约简搜索式算法[J]. 计算机学报, 2009, 32(8): 1486-1492.
(Chen Y M, Miao D Q. Searching algorithm for attribute reduction based on power graph[J]. Chinese J of Computers, 2009, 32(8): 1486-1492.)

[18] 徐章艳, 刘作鹏, 杨炳儒, 等. 一个复杂度为max{O(|C||U|), O(|C|^2|U|/|C|)}的快速属性约简算法[J]. 计算机学报, 2006, 29 (3): 391-399.
(Xu Z Y, Liu Z P, Yang B R, et al. A quick attribute reduction algorithm with complexity of max{O(|C||U|), O(|C|^2|U|/|C|)}[J]. Chinese J of Computers, 2006, 29(3): 391-399.)

[19] He Z Y, Deng S C, Xu X F. Discovering cluster based local outliers[J]. Pattern Recognition Letters, 2003, 24(9-10): 1651-1660.

[20] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large datasets[C]. Proc of the 2000 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2000: 427-438.

[21] Harkins S, He H X, Willams G J, et al. Outlier detection using replicator neural networks[C]. Proc of the 4th Int Conf on Data Warehousing and Knowledge Discovery. France: Springer Berlin Heidelberg, 2002: 170-180.