基于全息熵的空间离群点挖掘算法研究*

薛安荣, 何 峰, 闻丹丹
(江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)

摘 要: 基于距离和基于密度的离群点检测算法受到维度和数据量伸缩性的挑战,而空间数据的自相关性和异质性决定了以属性相互独立和分类属性的基于信息理论的离群点检测算法也难以适应空间离群点检测,因此提出了基于全息熵的混合属性空间离群点检测算法。算法利用区域标志属性进行区域划分,在区域内利用空间关系确定空间邻域,并用 R*-树进行检索。在此基础上提出了基于全息熵的空间离群度的度量方法和空间离群点挖掘算法,有效解决了混合属性的离群度的度量和离群点的挖掘问题。由于实现区域划分有利于并行计算,从而可适应大数据量的计算。理论和实验证明,所提算法在计算效率和实验结果的可解释性方面均具有优势。

关键词: 全息熵; R*-树; 空间离群点; 离群点检测; 混合属性
中图分类号: TP311; TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2014)02-0369-04
doi:10.3969/j.issn.1001-3695.2014.02.012

Spatial outlier detection based on holographic entropy

XUE An-rong, HE Feng, WEN Dan-dan
(*School of Computer Science & Communication Engineering*, *Jiangsu University*, *Zhenjiang Jiangsu* 212013, *China*)

Abstract: The outlier detection algorithms based on distance and density are faced with the challenges of both the dimensions and the amount of data scalability, and the autocorrelation and heterogeneity of spatial data determines that outlier detection algorithm which is characterized by attribute independent of each other and categorical attributes based on information theory is difficult to adapt to the spatial outlier detection. Hence, this paper proposed a spatial outlier detection algorithm based on mixed attributes of holographic entropy. The algorithm partitioned the region by regional identity property, determined the spatial neighborhood using spatial relationships in the region and then retrieved it by R*-tree. On this basis, it proposed spatial outlier degree based on holographic entropy and spatial outlier mining algorithm; it solved the outlier degree of the mixed attributes and the problems of outliers mining effectively. It could adapt to the large volume of data calculation because partitioning the region was conducive to parallel computing. Theoretical and experimental results show that the algorithm proposed has advantage in terms of the computational efficiency and the interpretative aspects.

Key words: holographic entropy; R*-tree; spatial outlier; outlier detection; mixed attributes

随着采集设备性能的提高和数量的增加,采集数据的维数和数量均呈上升趋势,有些数据的维数高达数百维,数据点的数量高达 TB 级,这对已有离群点挖掘算法是一个挑战,对空间离群点的挖掘尤为突出[1,2]。因为现有的挖掘方法大多是基于距离或密度的,而在高维情况下数据十分稀疏,数据点之间的距离及区域密度不再具有直观的意义;并且现有的挖掘算法大多具有 $O(n^2)$(n 为数据对象数目)的计算复杂度。因此上述算法对高维大数据离群点的挖掘不再有效,提高度量的有效性及计算的高效性是当前研究的热点[1,2]。

针对空间离群点挖掘,许多学者已经提出了基于距离[3]和基于密度等算法[4,5],但这些算法始终受维度和计算复杂度的限制,不适合高维大数据离群点的挖掘。Aggarwal 和 Keller 等人[6,7]提出了基于子空间的离群点挖掘算法,但在子空间的选择和挖掘结果的解释上仍然存在复杂度高和难以解释等问题;近年来 Wu 等人[8]提出了基于信息理论的离群点挖掘算法,比较好地解释了离群点的含义。但现有算法大多数假设属性间相互独立且其属性为分类属性,这在很多情况下受到限制。空间数据具有自相关性和异构性,上述假设显然不符合空间数据,不能直接应用于空间数据,正因为如此,到目前为止基于信息熵理论的空间离群点检测还未见报道。文献[8]提出了全息熵的概念和基于全息熵的离群度的度量方法,从而有效解决了属性的关联性和基于信息熵的离群点挖掘问题,解决了属性间相互独立的假设与实际相悖的问题,但仅限于分类属性,不适合连续数据。本文将在文献[8]成果的基础上,结合空间数据的特点,综合考虑信息熵和关联性来解决空间离群点的挖掘问题,提出基于全息熵的空间离群度的度量方法以及空间离群点的挖掘方法,有效解决了空间数据的关联性和混合属性的挖掘问题。

1 问题描述与定义

1.1 信息熵

信息熵可用于度量数据集的无序和杂乱程度。熵值越大,说明数据集无序和杂乱程度越高;反之,说明数据集越有序和

收稿日期: 2013-05-02; **修回日期**: 2013-06-24 **基金项目**: 国家自然科学基金资助项目(61300228);高校博士点基金资助项目(20093227110005)

作者简介: 薛安荣(1964-),男,教授,博士,主要研究方向为数据挖掘、机器学习(xuear@mail.ujs.edu.cn);何峰,男,硕士,主要研究方向为数据挖掘;闻丹丹,女,硕士研究生,主要研究方向为数据挖掘.

越纯净[9]。出现在数据集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,可以客观地识别出数据中的离群点[8]。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。

按照离群点的定义,离群点是数据集中的杂质,如果剔除数据集中的离群点,可显著降低信息熵值。基于此,可通过测量熵值的变化来检测离群点。此外,对于多维或高维属性,属性之间存在关联,在已有的基于信息熵的离群点检测算法中,大多忽略了属性之间的关联,导致检测结果不精确。属性之间的关联可通过属性之间的互信息来度量。因此,本文将基于文献[8]中的信息理论,综合利用信息熵和属性间关联的全息熵概念,给出基于全息熵的离群程度的度量方法和混合属性的空间离群点检测方法。

设数据集 $X=\{x_1,x_2,\cdots,x_n\}$ 由 n 个对象组成,对象 $x_i(1\leqslant i\leqslant n)$ 对应分类属性矢量 $Y=[y_1,y_2,\cdots,y_d]^{\mathrm{T}}$ 由 d 维属性组成,y_j 的属性值域为 $[y_{1,j},y_{2,j},\cdots,y_{n_j,j}](1\leqslant j\leqslant d)$,属性 y_j 有 n_j 个不同属性值。将每个 y_j 看做是随机变量,随机矢量 $[y_1,y_2,\cdots,y_d]^{\mathrm{T}}$ 用 Y 表示。x_i 可表示为 $(x_{i,1},x_{i,2},\cdots,x_{i,d})^{\mathrm{T}}$。下面给出信息熵 H、互信息 I 和总的关联 C 定义。

定义 1 信息熵。随机矢量 Y 的信息熵为

$$H(Y)=H(y_1,y_2,\cdots,y_d)=\sum_{i=1}^{d}H(y_i|y_{i-1},\cdots,y_1)=H(y_1)+H(y_2|y_1)+\cdots+H(y_d|y_{d-1},\cdots,y_1) \tag{1}$$

其中:$H(y_d|y_{d-1},\cdots,y_1)=-\sum_{y_d,y_{d-1},\cdots,y_1}p(y_d,y_{d-1},\cdots,y_1)\log p(y_d|y_{d-1},\cdots,y_1)$。

定义 2 y_1 与 y_2 间的互信息。变量 y_1 与 y_2 间的互信息为

$$I(y_1;y_2)=\sum_{y1,y2}p(y_1,y_2)\log\frac{p(y_1,y_2)}{p(y_1)p(y_2)}=H(y_1)-H(y_1|y_2) \tag{2}$$

定义 3 条件互信息。变量 y_1 与 y_2 间的互信息为

$$I(y_1;y_2|y_3)=H(y_1|y_3)-H(y_1|y_2,y_3) \tag{3}$$

定义 4 Y 间互信息。随机矢量 Y 间的互信息为

$$I(Y)=\sum_{i=2}^{d}\sum_{|r_1,\cdots,r_i|\subset|1,\cdots,d|}I(y_{r_1};\cdots;y_{r_i})=\sum_{|r_1,\cdots,r_i|\subset|1,\cdots,d|}I(y_{r1};y_{r2})+\cdots+I(y_{r1};\cdots;y_{rd}) \tag{4}$$

定义 5 Y 间关联信息。随机矢量 Y 间的关联信息为

$$C(Y)=\sum_{i=2}^{d}\sum_{|r_1,\cdots,r_i|\subset|1,\cdots,d|}I(y_{r_1};\cdots;y_{r_i})=\sum_{|r_1,\cdots,r_i|\subset|1,\cdots,d|}I(y_{r_1};y_{r_2})+\cdots+I(y_{r_1};\cdots;y_{r_d}) \tag{5}$$

其中:$r_1\cdots r_i$ 是 $1\sim d$ 的属性成员;$I(y_{r_1};\cdots;y_{r_i})=I(y_{r_1};\cdots;y_{r_{i-1}})-I(y_{r_1};\cdots;y_{r_{i-1}}|y_{r_i})$ 是多变量 $y_{r_1}\cdots y_{r_i}$ 的互信息,$I(y_{r_1};\cdots;y_{r_{i-1}}|y_{r_i})=E(I(y_{r_1};\cdots;y_{r_{i-1}}|y_{r_i}))$ 是条件互信息。总关联度表示数据集之间的相互依赖程度或共享信息量。

属性 y_1 与 y_2 的关联性为 $C(y_1;y_2)=I(y_1;y_2)$,其值越大,说明彼此之间的相同属性值越少。设 $O\subset X$,$X'=X\setminus O$,若 $C_{X'}(Y)$ 越小,则子集 O 为离群点集的可能性越大。

综合信息熵和关联信息可以得到全息熵,其定义如下:

定义 6 随机矢量的全息熵。随机矢量的全息熵是其熵和总的关联信息之和,即

$$HL(Y)=H(Y)+C(Y)=\sum_{i=1}^{d}H(y_i) \tag{6}$$

当 Y 的成分相互独立或仅有一个成分时,$HL(Y)=H(Y)$。

1.2 空间数据特点

空间离群点挖掘是空间数据挖掘的主要研究内容之一,在交通、生态、公共安全、公共健康、气候和基于位置的服务等应用领域中有着广泛的应用[1]。空间数据与一般的关系数据相比在很多方面有其独特性,其中最主要的两个特性就是空间自相关性和空间异质性[1],这两个特性决定了空间数据的相关性和多样性与局部性。正因为如此,以数据独立或统一数据分布的假设应用于空间离群点挖掘是不合适的。

空间对象的数据属性可分为标志属性、空间属性和非空间属性[1]。空间属性包括位置、形状、方向、空间邻接关系和其他几何或拓扑性质,非空间属性包括名称、年代、长度和高度等属性。非空间维属性是对象固有的,从本质上刻画了数据对象;而空间维属性并非对象所固有,但提供了对象的位置索引。由于空间数据具有空间自相关性和空间异质性等特点,因此,空间邻居在空间数据的分析中扮演着重要的角色。例如,在一个新兴大都市的老社区中,一栋新房屋就是基于非空间属性房屋年代的空间离群点。

1.3 标志属性特性

现有文献大多忽略了标志属性的使用,该属性既不参与空间邻域的确定,也不参与离群度的度量,仅起标志作用。而在实际空间数据库的应用中,很多标志属性包含采用层次编码方式表示特定对象所处的地域信息,如邮政编码、行政区划、域名等[1]。这类变量有下列特点:

a)变量可划分为多个部分,这些部分之间具有层次的关系。

b)相同前缀越多的变量之间彼此相似程度越高,反之变量之间相似程度越低。

由于像人口统计这类数据库的数据对象的标志中一般都含有地域标志属性,有些地域标志属性是独立的,如邮政编码,有些与其他信息混合在一起,如我国的身份证号中的前 6 位数字编码代表了地域信息。因此,具有相同地域编码信息的对象属于同一地区,具有相同前缀地域编码对象在相应层次上处于同一地区。地域编码可粗略地表述一个对象所处地区。利用地域编码信息也可实现按地域关系的数据对象的划分。

1.4 层次编码结构

在现实生活中,邮政编码、行政区划、车辆类型、域名等都是层次编码变量,其特点是可以分为 m 个互不相交的部分,记为 $P=(P_1,P_2,\cdots,P_m)$,其中每个 $P_i(i=1,2,\cdots,m)$ 是一个标称型变量,同时任意 $P_i(i=1,2,\cdots,m)$ 可进一步划分为 m' 个互不相交的部分 $P_i=(P_{i1},P_{i2},\cdots,P_{im'})$,$P_{ij\,i}(j=1,2,\cdots,m')$ 的父节点是 P_i,P_i 的值域由 P_{ij} 决定。因此,可以将层次编码转换为一棵平衡树 T,而任意层次编码变量 P 可以看做是从树 T 的根到叶子的路径上经过的节点集合。由于树 T 的每个叶子与路径具有一一对应关系,因此可以用叶子来代表任意层次编码变量 P,记为 $T(P)$。图 1 是一个层次编码平衡树示例,图 1 中根节点表示省和直辖市的编码,第二层表示省辖市或直辖市内区编码,第三层表示县(区)编码。例如江苏省编码是 32,镇江市编码是 11,京口区编码是 02,$P=\{32,11,02\}$ 即江苏省镇江市京口区。

1.5 空间邻居与空间邻域

对象的空间邻域是指基于空间属性与空间关系与对象邻接的所有空间邻居的集合。设对象集 $X=\{x_1,x_2,\cdots,x_n\}$ 由 n 个对象组成,对象 $x\in X$ 的标志属性函数是 $id(x)$,空间属性函数是 $s(x)$,非空间属性函数是 $f(x)$,$f(x)$ 的维度为 d-维,σ_c 表示在指定条件 c 下的空间邻接关系。d-维非空间属性 $f(x)$ 表

示为($f(x_1)$,$f(x_2)$,…,$f(x_d)$)。

对数据对象集 X 用标志属性建立如图 1 所示的层次编码平衡树,基于平衡树进行区域划分,在此基础上,在区域内,基于空间属性和空间邻接关系确定空间邻域。由空间数据的自相关性和异构性决定了空间数据的局部性,空间邻域的确定和基于邻居的离群点检测可以在各自区域内完成。这样,可以用分布式完成各自区域内的并行计算,然后进行汇集,从而使得算法适应大数据量的计算。

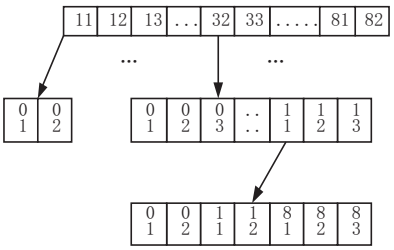

图1　层次编码平衡树示例

数据对象集 X 基于标志属性函数 $id(x)$ 划分为 m 个互不重叠的区域,即 $X=(X_1,X_2,\cdots,X_m)$,$X_i=\{x_1^i,x_2^i,\cdots,x_{n_i}^i\}$,$X=X_1\cup X_2\cup\cdots\cup X_m$,且 $X_i\cap X_j=\varnothing$,$1\leqslant i,j\leqslant m$,$n=\sum_{i=1}^{m}n_i$。

定义 7　空间邻居。对象 x 的空间邻居是指与对象 x 在指定条件 c 下存在空间邻接关系 σ_c 的对象,即 $X_i\subset X$,$\forall x\in X_i$,$\exists p\in X_i\backslash\{x\}$,使得 $s(p)\sigma_c s(x)$ 为真,则对象 p 是对象 x 的空间邻居。

定义 8　空间邻域。对象 x 的空间邻域 $N(x)$ 是指对象 x 的所有空间邻居的集合,即 $X_i\subset X$,$\forall x\in X_i$,$N(x)=\{p\mid s(p)\sigma_c s(x)=\text{true},p\in X_i\backslash\{x\}\}$。

2　基于全息熵的算法

2.1　算法思想

由于现有的基于信息理论的离群点检测算法主要是基于分类属性和假设属性间相互独立,这两点决定了现有算法难以适应空间离群点检测。下面将研究基于全息熵的混合属性的空间离群点检测算法。算法将在上述已经划分的区域和基于区域的邻居(域)的基础上进行。

文献[8]提出了全息熵的概念,全息熵的计算如式(6)所示。从式(6)可以看出全息熵为每维信息熵之和,这样在考虑属性之间关联性的同时又简化了计算。

为了有效解决混合属性计算,首先考虑对象 x_i 与 x_j 之间的相异度计算。假设数据对象包含 d-维非空间属性,其数据类型分为离散型和连续性;离散型又可分为区间标度类型、二元类型、分类类型、序数类型和比例标度型。

定义 9　第 k 维属性的相异度。对象 x_i 与 x_j 之间第 k 维属性的相异度 $s_k(x_i,x_j)$ 计算分为以下几种情况:

a)如果是二元类型或分类变量类型:若 $x_i^{(k)}=x_j^{(k)}$,则 $s_k(x_i,x_j)=0$;否则 $s_k(x_i,x_j)=1$。

b)如果是连续型变量,$s_k(x_i,x_j)=\dfrac{|x_i^{(k)}-x_j^{(k)}|}{\max_h x_h^{(k)}-\min_h x_h^{(k)}}$,这里 h 遍取 k 维属性上所有对象中的非空缺值,$\max_h x_h^{(k)}$ 和 $\min_h x_h^{(k)}$ 分别表示第 k 维属性的最大和最小值。

c)如果是序数型或者比例标度型变量,秩的范围为[0,M_k-1],计算秩 $r_i^{(k)}$ 和 $r_j^{(k)}$,$s_k(x_i,x_j)=\dfrac{|r_i^{(k)}-r_j^{(k)}|}{M_k-1}$。

定义 10　第 k 维信息熵。对象 x_i 的第 k 维信息熵定义为

$$X_i\subset X,\forall x_i\in X_i,H_{N(x_i)}(y_k)=-\sum_{x_j\in N(x_i)}s_k(x_i,x_j)\log s_k(x_i,x_j)\quad(7)$$

定义 11　对象的信息熵。对象 x_i 的信息熵定义为

$$X_i\subset X,\forall x_i\in X_i,HL(x_i)=H(x_i)+C(x_i)=\sum_{k=1}^{d}H_{N(x_i)}(y_k)\quad(8)$$

定义 12　对象的属性权值。对象 x_i 的属性权值定义为

$$w_k(x_i)=2\left(1-\frac{1}{1+\exp(-H_{N(x_i)}(y_k))}\right)\quad(9)$$

其中:该逆 Sigmod 递减函数范围在(0,2)区间。因为熵是正的,故加权系数在[0,1]范围内。

定义 13　对象的加权全息熵。对象 x_i 的加权全息熵定义为

$$HL(x_i)=\frac{\sum_{k=1}^{d}w_k(x_i)\delta_k H_{N(x_i)}(y_k)}{\sum_{k=1}^{d}\delta_k}\quad(10)$$

其中:w_k 为第 k 维属性的权重,δ_k 为指示项。如果第 k 维属性上两个对象的值都是 0,或者一个对象的第 k 维属性具有遗漏值,则 $\delta_k=0$,否则 $\delta_k=1$。

定义 14　空间离群系数。对象 x_i 的基于加权全息熵的空间离群系数定义为

$$\text{HSOF}(x_i)=\frac{HL(x_i)}{\dfrac{\sum_{x_j\in N(x_i)}HL(x_i)}{|N(x_i)|}}\quad(11)$$

定义 15　空间离群点　给定 n 个空间对象集 X,希望挖掘 k 个离群点,计算每个对象的 HSOF,HSOF 最大的 k 个对象就是空间离群点。

2.2　算法描述

算法主要步骤包括:a)根据 $id(.)$ 确定分区,根据 $s(.)$ 和指定的空间关系确定邻域;b)计算每个对象的空间离群系数;c)输出离群点集。

MAHSOF 算法:混合属性的空间离群点挖掘算法

输入:对象集 $X=\{x_1,x_2,\cdots,x_n\}$;对象 $x_i(id(x_i),s(x_i),f(x_i))$ 的空间属性为 $s(x_i)$,非空间属性为 $f(x_i)$,d-维非空间属性 $f(x_i)$ 表示为 $(f(x_{i1}),f(x_{i2}),\cdots,f(x_{id}))$;$\sigma_c$ 表示在指定条件 c 下的空间邻接关系,离群点个数 k。

输出:空间离群点集 O。

a)根据 $id(x)$ 确定分区 $P_1,P_2,\cdots,P_m$,P_i 分区包含数据对象集 X_i

b)for($O=\varnothing$,$q=1$;$q\leqslant m$;$q++$)

{//对每个分区分别进行计算,得到离群点集 O_q

c)O_q = computing_HSOF(Xq, σ_c, k);//分别对每个分区计算离群点集合

d)$O=O\cup O_q$;//分别对每个分区计算离群点集合

e)}

f)Sort(HSOF);//对 HSOF 按降序排列

g)Top_k_Set = Get_Top_k(HSOF, O);//前 k 个对象

h)Output(Top_k_Set);//输出离群对象

MAHSOF 算法是基于标志属性进行划分,然后分别对每个分区进行计算,从而可进行并行计算。实际上这些分区本身就是分布的,所以利用分布计算非常有利,并适应大数据量的计算。

每个分区计算算法:

computing_HSOF(Xq, σ_c, k)

输入:Xq, σ_c, k。

输出:O_q,返回分区的离群点集合。

a)for($i=1$;$i\leqslant|Xq|$;$i++$){//$|Xq|$为 Xq 中对象数量

b) x_i = Get_One_Object(i, Xq); //从 Xq 中选取一个对象
c) $N(x_i)$ = Find_Neighbor_Nodes_Set(x_i, Xq, σ_c);
//从 Xq 中寻找与 x_i 存在空间关系 σ_c 的所有邻居
d) for(r = 1; $r \leqslant d$; r ++) {//d 为非空间属性维数
e) for($H(r)$ = 0, j = 1; $j \leqslant |N(x_i)|$; j ++) {
//$|N(x_i)|$ 为 x_i 的邻域对象数量
f) 按定义 9 计算 $s_r(x_i, x_j)$;
g) $H(r)$ -= $s_r(x_i, x_j)\log(s_r(x_i, x_j))$;
//按定义 10 计算第 r 维信息熵 H
h) }
i) 按定义 12 计算 $w_r(x_i)$;
j) }
k) $HL(x_i)$ += $w_r(x_i)$ * $H(r)$;//按定义 13 计算全息熵
l) 按定义 14 计算 HSOF(x_i);
m) }
n) Sort(HSOF); //对 HSOF 按降序排列
o) Top_k_Set = Get_Top_k(HSOF, O_q); //取前 k 个对象
p) return O_q; //返回离群对象

2.3 算法分析

MAHSOF 算法中:行 a)是分区划分;行 b) ~ e)是收集每个分区的离群点挖掘结果;行 f) ~ h)是排序并输出最终的挖掘结果。假设空间数据对象数目为 n,非空间属性维度为 d 维,分区数为 m,离群点数量为 k,那么行 a)的计算复杂度为 $O(n)$,行 b) ~ e)的计算复杂度为 $O(m)$,行 f) ~ h)的计算复杂度为 $O(mk\ \log(mk))$,总的计算复杂度为 $O(n + m + mk\log(mk))$。由于 $m \ll n$, $k \ll n$,因此,MAHSOF 算法的计算复杂度为 $O(n)$。

算法 computing_HSOF(Xq, σ_c, k)中,行 a) ~ m)是针对每个空间对象的基于信息熵的空间离群系数的计算。其中,行 c)是确定对象的空间邻域,行 d) ~ j)是计算每维信息熵和权值,行 k)是计算全息熵,行 l)是计算离群系数,行 n) ~ p)是排序并返回挖掘的分区离群点集。算法中在空间邻居的确定上采用了空间索引技术,即利用 R*-树索引技术[10]来加快空间邻域的确定,降低了计算复杂度。行 a) ~ m)的计算复杂度为 $O(|Xq|\log|Xq|)$,行 n) ~ p)的计算复杂度为 $O(|Xq|\log|Xq|)$,总的计算复杂度为 $O(|Xq|\log|Xq|)$。由于 $|Xq| \leqslant n$,所以计算复杂度不大于 $O(n\ \log\ n)$,只有在单个分区时相等,所以本算法更适合分布并行计算。

3 实验结果与分析

本文用 VC 6.0 编写了以 R*-tree 为索引结构的 MAHSOF 算法的程序,测试采用 UCI 网站上的森林火情监测数据[1]和美国 2000 年的人口统计数据[1],下面分别加以说明。用 VC ++ 编写了以 R*-tree 为索引结构的 MAHSOF 算法程序,运行机器配置为奔腾双核 1.6 GHz CPU、1.5 GB 内存、操作系统为 Windows XP SP3、数据库为 SQL Sever 2005。

实验 1 采用 UCI 网站上的森林火情监测数据作为测试数据。测试数据共含 517 个数据对象,每个数据对象包含 13 维属性,其中 11 维数值型属性,2 维分类型属性。2 维分类属性表示采集数据发生的时间,这里作为分类数据进行测试,11 维数值型属性中包含 2 维空间属性,表示数据对象所在的区域,其余属性为非空间属性。实验分别测试了纯分类属性、纯数值属性和混合属性情况。

实验中以每个数据分区为一个空间对象,以与其相邻的所有邻居构成邻域,即假设空间对象的坐标为 (x, y),则坐标为 $(x-1, y-1) \sim (x+1, y+1)$ 之间的在规定范围 $(1,2) \sim (9,9)$ 内的空间对象就是 (x, y) 的直接邻居,每个空间对象最多有八个邻居,这些邻居构成其邻域。对于纯分类属性挖掘的前五个最离群的空间对象分别是(7,6)、(7,3)、(9,5)、(4,6)、(3,3);对于纯数值属性挖掘的前五个最离群的空间对象分别是(8,4)、(5,5)、(2,3)、(3,3)、(9,6);对于混合属性挖掘的前五个最离群的空间对象分别是(8,4)、(5,5)、(3,3)、(7,6)、(9,5)。通过对原数据的进一步分析可以发现,测试结果是正确的,与文献[1]测试结果相符,限于篇幅这里省去原数据。

实验 2 以美国 2000 年的人口统计数据作为测试数据,并与 SLOF[10]、SLOM[4]和 LCK 算法[5]进行比较,求得最离群的五个离群点。表 1 为四种检测结果比较。

从表 1 中可以看出,MAHSOF 与 SLOF 算法有一个相同,四个不同;SLOM 和 LCK 算法求得的五种离群点中有四个是相同的,与前两种算法完全不同。进一步分析可知,MAHSOF 与 SLOF 算法有类似的结果,主要差别是 MAHSOF 算法仅考虑了区域内邻居,即仅限于州内邻居,没有考虑州之间的县级邻居,而 SLOF 等算法同时考虑了州之间的邻居,不同地区可能在政策上有所不同,因此跨地区直接比较有时不太合适。SLOF 与 MAHSOF 算法四个不同的就是跨地区的比较,所以 MAHSOF 与 SLOF 算法有类似结果,优于另外两种算法[10]。

在执行时间上,由于最费时的是根据空间属性确定空间邻域,四种算法均采用相同算法和数据,但 MAHSOF 算法在分区内确定邻域,所以 MAHSOF 算法在确定邻域执行效率上优于其他三种;在计算比较上耗时处于同一数量级,执行时间分别是 SLOM 3.8 s、LCK 4.8 s、SLOF 3.5 s、HSOF 2.2 s。但随着非空间属性维数的增加,由于 LCK 算法需要计算庞大的协方差矩阵及其逆矩阵,算法将变得非常困难,效率也急剧下降,所以 HSOF、SLOM 与 SLOF 算法具有更好的伸缩性和更高的性能,并且 HSOF 算法由于采用分区方法,计算效率更高。

在可解释性上,由于 MAHSOF 的每维属性权值是根据数据属性的信息熵计算得到的,所以可有效解释离群原因。此外,MAHSOF 算法在高维属性的适应性和分布并行计算上具有优势。

表 1 四种算法检测结果比较

序号	MAHSOF			SLOF		
	编号	县名	HSOF 值	编号	县名	SLOF 值
1	13153	Houston	4.59	48029	Bexar	17.24
2	48479	Webb	4.22	48303	Lubbock	13.37
3	55025	Dane	4.19	40143	Tulsa	11.44
4	48029	Bexar	4.06	21111	Jefferson	11.23
5	36091	Saratoga	4.01	18003	Allen	10.36

编号	SLOM		LCK		
	县名	SLOM 值	编号	县名	LCK 值
06037	Los Angeles	1.36	06037	Los Angeles	1611
17031	Cook	1.17	17031	Cook	942
48201	Harris	0.48	04013	Maricopa	809
04013	Maricopa	0.39	12086	Miami-Dade	770
48113	Dallas	0.33	48201	Harris	667

4 结束语

由于现有的基于信息理论的离群点检测算法主要是基于分类属性和假设属性间相互独立,这两点决定了现有算法难以适应空间离群点检测。本文提出了基于全息熵的混合属性的空间离群点检测算法。该算法利用区域标志属 (下转第 397 页)

表 4 中检验结果“S +”表示 SAGBA 在该行所对应的问题上的计算结果要显著性优于该列所对应的算法,“S -”表示该列所对应的算法在该行所对应的问题上显著性优于 SAGBA,“~”表示 SAGBA 在该行所对应的问题上的性能与该列所对应的算法的性能是相等的。

从表 4 中可以直观地看出,与 BA 相比,在 Griewank 函数上,SAGBA 的显著性差于 BA;在 Booth、Hartmann、Matyas、Shubert 和 Zakharov 函数上,SAGBA 的显著性均与 BA 相同;除此之外,对于其他的函数,SAGBA 的显著性均要优于 BA。与 SAPSO 算法相比,除了在 Griewank、Hartmann 和 Matyas 上 SAGBA 的显著性与 SAPSO 算法相同以外,在其他的函数上,SAGBA 的显著性均优于 SAPSO 算法。从统计的角度可以得出以下结论:SAGBA 在大部分的测试问题上都显著性优于 BA 和 SAPSO 算法。

5　结束语

将模拟退火算法与蝙蝠优化算法相结合,对蝙蝠个体进行高斯扰动,进一步搜索以保留个体“精英”。随着进化过程的推进,温度逐渐降低,接收较差解的几率逐渐减小,从而提高算法的收敛性能。仿真实验结果表明,该算法的收敛性能在不同程度上优于其他两种算法。下一步将研究 SAGBA 与其他优化算法的比较,并进一步研究 SAGBA 在多目标规划问题中的应用。

参考文献:

[1] YANG Xin-she. A new met heuristic bat-inspired algorithm [C]// Nature Inspired Cooperative Strategies for Optimization. 2010: 65-74.

[2] YANG Xin-she. Bat algorithm for multiobjective optimization [J]. International Journal Bio-Inspired Computation, 2011, 3(5): 267-274.

[3] 李枝勇,马良,张惠珍. 遗传变异蝙蝠算法在 0-1 背包问题上的应用[J/OL]. [2012-10-11]. http://www.cnki.net/kcms/detail/11.2127.TP.20121011.1019.027.html.

[4] LEMMA T A,BIN M H F. Use of fuzzy systems and bat algorithm for energy modeling in a gas turbine generator [C]//Proc of IEEE Colloquium on Humanities, Science and Engineering. 2011: 305-310.

[5] YANG Xin-she, GANDOMI A H. Bat algorithm: a novel approach for global engineering optimization [J]. Engineering Computations, 2012,29(5): 464-483.

[6] MISHRA S, SHAW K, MISHRA D. A new metaheuristic classification approach for microarray data[J]. Procedia Technology,2012, 4(1): 802-806.

[7] KHAN K, NIKOV A, SAHAI A. A fuzzy bat clustering method for ergonomic screening of office workplaces, S3T 2011[C]//Advances in Intelligent and Soft Computing. 2011: 59-66.

[8] KHAN K, SAHAI A. A comparison of BA, GA, PSO, BP and LM for training feed forward neural networks in e-learning context [J]. International Journal of Intelligent Systems and Applications, 2012, 4(7): 23-29.

[9] ALTRINGHAM J D. Bats: biology and behaviour [M]. Oxford: Oxford University Press, 1996: 37-64.

[10] KENNEDY J, EBERHART R. Particle swarm optimization [C]// Proc of IEEE International Conference on Neural Networks. 1995: 47-53.

[11] 康立山. 非数值并行算法——模拟退火算法[M]. 北京:科学出版社,1997.

[12] 吴志远,邵惠鹤,吴新余. 遗传退火进化算法[J]. 上海交通大学学报,1997,31(12): 69-71.

[13] 王雪梅,王义和. 模拟退火算法与遗传算法的结合[J]. 计算机学报,1997,20(4): 381-384.

[14] 赵世安,黄敢基. 模拟退火并行粒子群优化算法程序设计与研究[J]. 百色学院学报,2006,19(6): 9-12.

[15] 龚纯,王正林. 精通 MATLAB 最优化计算[M]. 北京:电子工业大学,2009.

[16] HEDAR J. Test functions for unconstrained global optimization [EB/OL]. http://www-optima. amp. i. kyoto-u. ac. jp/member/student/hedar/Hedar_files/TestGO_files/Page364. htm.

[17] 龚鲁光. 概率论与数理统计[M]. 北京:清华大学出版社,2006.

(上接第 372 页)性进行区域划分,在区域内利用空间邻域关系和 R*-树确定空间邻域,在此基础上提出了全息熵空间离群度的度量方法和基于全息熵的空间离群度的空间离群点挖掘算法。所提算法有效解决了混合属性的基于全息熵的空间离群度的度量方法。由于实现区域划分,算法可进行并行计算,从而适合大数据量的计算。理论和实验证明,本算法在计算效率和实验结果的可解释性方面均具有优势。

参考文献:

[1] 薛安荣. 空间离群点挖掘技术的研究[D]. 镇江:江苏大学,2008.

[2] 薛安荣,姚林,鞠时光,等. 离群点挖掘方法综述[J]. 计算机科学, 2008,35(11):13-18.

[3] SHEKHAR S, LU C T, ZHANG Pu-sheng, *et al*. A unified approach to spatial outliers detection[J]. GeoInformatica,2003,7(2): 139-166.

[4] CHAWLA S, SUN Pei. SLOM: a new measure for local spatial out-liers [J]. Knowledge and Information Systems,2006,9(4): 412-429.

[5] LU C T, CHEN De-chang, KOU Yu-feng. Detecting spatial outliers with multiple attributes[C]//Proc of the 15th Conference on Tool Artificial Intelligence. Washington DC: IEEE Computer Society, 2003: 122-128.

[6] AGGARWAL C C, YU P. Outlier detection for high dimensional data [C]//Proc of ACM SIGMOD International Conference on Management of Data. 2001:37-46.

[7] KELLER F, MÜLLER E, BÖHM K. HiCS: high contrast subspaces for density-based outlier ranking[C]//Proc of the 28th IEEE International Conference on Data Engineering. 2012: 1037-1048.

[8] WU Shu, WANG Sheng-rui. Information-theoretic outlier detection for large-scale categorical data[J]. IEEE Trans on Knowledge and Data Engineering,2013,25(3):589-602.

[9] COVER T M, THOMAS J A. Elements of information theory[M]. 2nd ed. New Jersey: Wiley & Sons,2006.

[10] 薛安荣,鞠时光,何伟华,等. 局部离群点挖掘算法研究[J]. 计算机学报,2007,30(8): 1455-1463.