基于信息熵的地学空间数据挖掘模型

周成虎 张健挺
(中国科学院资源与环境信息系统国家重点实验室，北京 100101)

摘 要 从信息熵的基本概念出发，认为地学空间数据子集划分产生的互信息或熵减源于子集划分，使得各个子集的不确定性或模糊性降低，并且子集之间的差异性增大。因此具有最大熵减的子集划分方案代表一定的地学模式和地学规律。以此为基础分别探讨了地学数据属性要素的子集划分产生多维属性关联规则，以及通过空间和时间的子集分割来进行聚类的方法。

关键词 信息熵 数据库挖掘 地学分析模型

0 引 言

日益丰富的地学数据在一定程度上已超过了地球科学家能够处理的能力，从这些海量数据中发现地学知识的需要使得地学数据分析与数据挖掘方法的结合成为必然。GIS 在空间数据的存储、表达和管理上已得到广泛的认可，增强 GIS 的分析功能，提高 GIS 解决地学实际问题的能力已得到共识。将地学空间数据挖掘工具与 GIS 紧密集成可以充分利用 GIS 存储、管理空间数据的功能，同时使得 GIS 中的有限数据变成无限的知识，使 GIS 成为智能的信息系统[1]。在地学数据分析领域引入数据挖掘与知识发现的概念、模式和方法，探讨适合地学数据挖掘的新方法，并与 GIS 紧密集成，对于有效地处理海量地学数据、提高地学分析的自动化和智能化水平、为全球变化和区域可持续发展提供有力的分析工具具有重要的意义。

1 信息熵在地学分析中的应用

自从 1948 年美国工程师 Shannon 在贝尔实验室杂志上发表的长篇论文《通信的数学理论》中定义了信息熵的基本形式以来，信息熵在各个领域得到广泛的应用，其中一个重要原因是它以简单的方式定义了系统的复杂性并具有明确的物理含义。

信息熵的基本定义如下：设 X 取值于集 A，$P_i=P[X=a_i](i=1,2,\cdots,n)$，则称 $1(a_i)=\log(1/p_i)=-\log(p_i)(i=1,2,\cdots,n)$ 为符号 a_i 所产生的信息量，$1(a_i)$ 又称为 a_i 的自信息。信息量的数学期望(平均值) $El(a_i)$ 称为信源的平均信息量或称为信源的信息熵，简称为信息熵并记为 $H(X)$，或 $H_m(P_1,P_2,\cdots,P_m)$。信息熵是在平均意义上来表征信息源总体特性的量[2]。

互信息定义为信源 Q 的信息熵与在信源 P 提供的信息条件下信源 X 的条件熵之差，即 $I(P;Q)=H(Q)-H(Q|P)$。互信息度量了一个信源从另一个信源获取的信息量的大小，通俗地把互信息称为熵减。

在地学数据分析中，已有许多利用信息熵的例子。如生态系统中的景观多样性描述[3]，地貌系统中流域水系在极大熵条件下高程的概率分布以及河流等级结构研究[4]，水文系统中复杂度的研究[5]，气象系统中在极大熵情况下暴雨的面深关系[6]，地图信息分析中关于信息传输增量问题的讨论[7]，遥感影象分类评估中不确定性的研究[8]等。这些研究都表明，信息熵是表征地学现象的一个普遍适用的特征量。由于信息熵与分形具有本质上的等价关系[9]，以信息熵为基础的信息分维数作为描述地学系统非线性的一个重要指标也得到广泛应用，如城镇等级结构[10]，城市人口分布[11]，流域水系结构[12]等。

以上的分析表明：目前地学领域中对于信息熵的应用主要有两个方向，一是利用信息熵的基本定义或其扩展定义计算信息熵，或基于信息熵的分形分维值作为度量地学非线形特征的一个指标；二是在自由或受限极大熵假设条件下推导出地学现象的

收稿日期：1999-10-27

结构特征或某些要素的分布特征并进行实际检验。

2　地学数据子集分割的互信息研究

2.1　子集分割熵减的计算公式

设按照某种标准将一个具有 m 种属性类型的地学数据集分割成 n 个子集,并设分割前的分布向量为 X_i,分割后的 n 个向量分别为 Y_{ij},其中 $i=1,m,j=1,n$(如图 1 所示)。

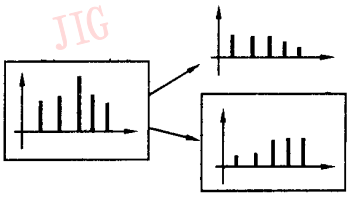

图 1　子集分割示意图

分割前的信息熵为:

$$Dis0=-\sum_{i=1}^{m}p_i\times\log(p_i) \tag{1}$$

分割后的信息熵有两种计算方式,分别为:

横向(子集内):$$Dis1=-\sum_{j=1}^{m}p_j\times\sum_{i=1}^{m}p_{ji}\times\log(p_{ji}) \tag{2}$$

纵向(子集间):$$Dis2=-\sum_{i=1}^{m}p_i\times\sum_{j=1}^{n}p_{ij}\times\log(p_{ij}) \tag{3}$$

其中,

$$X=\sum_{i=1}^{m}X_i,p_i=\frac{X_i}{X},p_{ij}=\frac{Y_{ij}}{X_i},X_i=\sum_{j=1}^{n}Y_{ij}$$

$$Y_j=\sum_{i=1}^{m}Y_{ij},p_j=\frac{Y_j}{X},p_{ji}=\frac{Y_{ij}}{Y_j}$$

根据最大熵定理,$Dis0$ 的最大值为 $\log(m)$,$Dis1$ 的最大值为 $\log(m)$,$Dis2$ 的最大熵为 $\log(n)$。对于熵减有两种定义方法:

子集内的熵减可以定义为:

$$Rec1=Dis0-Dis1 \tag{4}$$

子集间的熵减可以定义为:

$$Rec2=Dis0-Dis2 \tag{5}$$

因此归一化熵减有两种定义方法:

$$NRec1=\frac{(Dis0-Dis1)}{\log(m)} \tag{6}$$

$$NRec2=\frac{Dis0}{\log(m)}-\frac{Dis2}{\log(m)} \tag{7}$$

或

$$NRec2=1-\frac{Dis2}{\log(n)} \tag{8}$$

2.2　不同类型要素组合的子集分割及其熵减计算

熵减的产生是由于数据集有意义的子集分割。根据进行分割的属性类型和被分割子集数据类型的不同(连续或离散),分割方式的组合是多样的,图 2 列出连续变量分割和离散类型组合对数据集的划分。

对于被分割属性为连续变量的情况,一般要将其离散化后转化为离散类别再进行处理。因此下面我们只给出被分割属性为离散类别情况的实际例子。根据分割属性是连续属性还是离散类别又可以分为两种情况:

(1) 连续属性作为分割属性:首先按照属性值的大小对整个数据集进行排序,按照一定的步长从最小值或最大值开始顺序向相反的方向递增或递减确定一个分割点,即 $X_1=X_0+i\times\Delta X$,计算分割点 X_1 分隔成的两个子区中的总信息熵。如图 2(a)所示,分割前总记录数为 8,类别 1 的记录数为 4,类别 2 的记录数为 4,则信息熵为 $(-4/8)\times\log(4/8)+(-4/8)\times\log(4/8)=0.3010$;分割后子集 1 中记录数为 4,类别 1 的记录数为 3,类别 2 的记录数为 1,子集 2 中记录数为 4,类别 1 的记录数为 1,类别 2 的记录数为 3,则该划分的信息熵为 $(-4/8)\times[(3/4)\times\log(3/4)+(1/4)\times\log(1/4)]+(-4/8)\times[(1/4)\times\log(1/4)+(3/4)\times\log(3/4)]=0.2442$,熵减 $Rec1$ 为 $0.3010-0.2442=0.0568$。

(2) 离散类型作为分割属性:尝试各种类别的组合,对于具有 n 个类别的条件属性则不重复的有意义组合共有 $2^{n-1}-1$ 种。对于每一种组合计算该组合的信息熵。图 2(d)所示的总记录为 10 个。在同样假设情况下,分割前条件类 1 的记录为 2 个,对应的决策类 1 为 2 个记录而决策类 2 的记录为 0 个;分割前条件类 2 的记录为 3 个,对应的决策类 1 为 0 个记录而决策类 2 的记录为 3 个;分割前条件类 3 的记录为 3 个,对应的决策类 1 为 1 个记录而决策类 2 的记录为 2 个;分割前条件类 4 的记录为 2 个,对应的决策类 1 为 1 个记录而决策类 2 的记录为 1 个。因此分割前的系统总熵值为 $(-2/10)\times\log(2/10)+(-3/10)\times\log(3/10)+(-3/10)\times\log(3/10)+(-2/10)\times\log(2/10)=0.5933$。将条件类 1 和类 4 并入一类后对应的决策类 1 记录数为 3,决策类 2 对应的记录数为 1;条件类 2 和类 3 并入一类后对应的决策类 1 记录数为 1,决策类 2 对应的

记录数为 5。系统的信息熵为(4/10)×[(−1/4)×log(1/4)+(−3/4)×log(3/4)]+(6/10)×[(−1/6)×log(1/6)+(−5/6)×log(5/6)]=0.1586,熵减 *Rec*1 为 0.5933−0.1586=0.4347。

同理也可以计算 *Rec*2 和 *Nrec*1、*Nrec*2。显然如果以上这两个示例属性参与子集划分,采用熵减 *Rec*1 作为指标则离散指标,应作为第一分割属性。

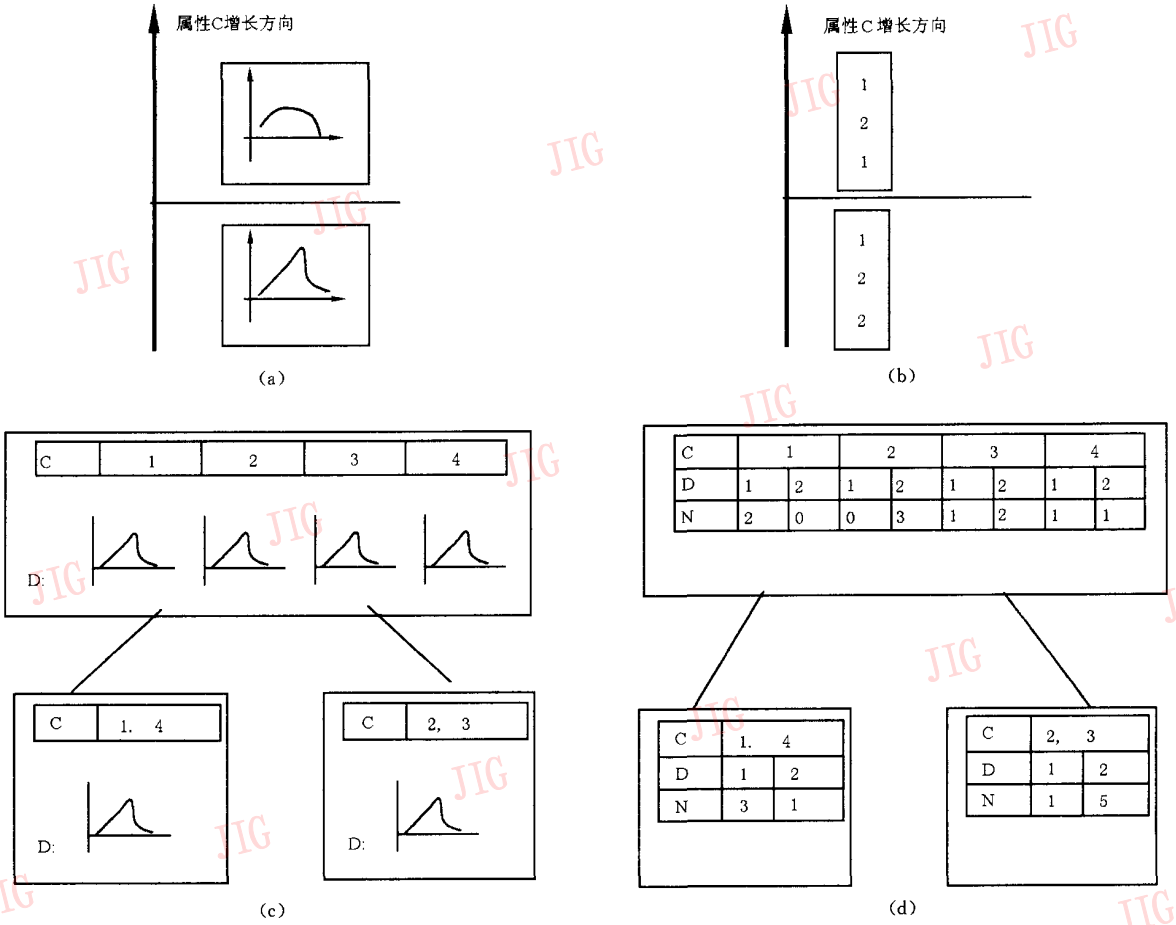

图 2 连续变量分割和离散类型组合对数据集的划分示意图

3 基于信息熵的决策树方法在地学数据挖掘中的应用

3.1 决策树方法及其在地学中的应用研究

决策树方法属于一种通过实际例子进行归纳学习、产生分类规则然后进行分类的机器学习方法,其关键在于属性分割标准的选择。决策树方法在地学数据分析中的应用包括:Michaelson 等给出了决策树在遥感影象分类中的综述和应用实例[13];Hansen 等对全球 1×1 度的 NDVI 数据集选择了包括最大 NDVI、平均 NDVI、生长季节总长度、最大土壤表层温度在内的 16 个指标,利用决策树方法进行遥感影象数据与 GIS 数据一体化的土地覆盖分类,建立了专家可理解的分类规则并进行分类[14];Eklund 等利用决策树方法建立了南澳大利亚 Adelaide 西北沿海地区的包括多波段遥感影象、地质条件、地下水水位、土壤导电率等在内的各要素与土壤二次盐碱类型之间的关联关系,产生的分类规则可以与专家分类规则相比较[15];Huang 等同样利用了决策树方法建立了美国南加利福尼亚 Savannah 河地区遥感影象、DEM、坡度、纹理等指标与土壤类型的关联规则[16]。与上面两个研究不同的是,Huang 的研究将 X 和 Y 坐标以及 ISODATA 的初步分类结果也作为条件属性参与建树过程,结果显示土壤类型与 ISODATA 的初步分类结果关系最大并将其作为第一分割属性。结果同时也表明,建立的规则与 X 和 Y 坐标也有一定的对应关系。

3.2　决策树方法在空间和时间域的扩展

地学现象在空间上的差异是地物内在相互作用的外在表现。但是在机理不清楚的情况下,通过空间分异的研究可以启示对地学现象机理的研究。在如图 3 所示的空间分布模式中,通过将 X 和 Y 两个方向作为条件变量,将类别作为决策变量,在理想作态下(如)可以利用决策树模型建立如下的规则:

IF $X<2$ THEN $CATA=1$

IF $X>2$ AND $Y>3$ AND $X<4$ THEN 2

IF $X>2$ AND $Y>3$ AND $X>4$ THEN 3

IF $X>2$ AND $Y<3$ THEN 4

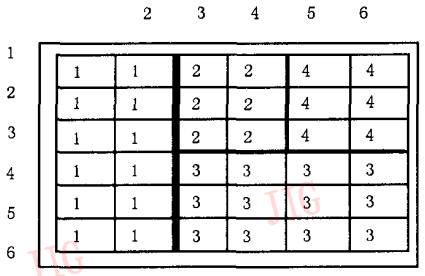

图 3　决策树方法在空间领域的扩展示意图

但是在实际中,一般类别都交错分布,直接采用决策树方法往往导致规则建立过细而缺乏实际意义。在这种情况下,可以采用空间分割聚类的方法,即分割线两边在当前情况下具有最大的对比度,子集内可以有多个类型分布而不再要求分割到单个的类别。

以上是对于空间位置的考虑。对于其它空间特征,如空间形态、空间邻近关系等可以通过提取空间特征指标,然后采用决策树模型进行数据挖掘和知识发现。

同样可以通过提取时间域的特征指标来进行分类。常用的类型是建立一种类型与一个时间序列的关联,其前提假设是不同的类型在时间序列上具有不同的特征,这些特征可以通过特征指标来体现。常用的指标包括时间序列中的 m 个样本点、M 个样本点的算术运算、时间序列的统计变量等。

4　基于互信息的地学数据时空分割聚类模型

4.1　基于空间要素的分割模型

决策树方法的分类过程中子集分割的熵减最大标准,实际上是使分割后的子集具有最大的对比度,它启示我们可以采用熵减标准作为聚类的依据。对于时间或空间要素,可以通过不断地尝试子集划分,寻找使分割后的子集具有最大的对比度的分割方式,这样递归循环直到子集内的样本全部或绝大部分为一种类型,或分布在一个取值区间,或者子集内的样本树木小于规定的门限值。分割的最终结果是具有层次特征的时间-空间-属性多维要素体,我们把这样分割成的多维体称为一类。显然这个类的概念与普通的聚类分析中的类的概念是不完全相同的。聚类分析中的属于同一类的样本可以在时间和空间上不相邻,属于某一类的样本与周围其它样本之间没有直接关系,数据集中的每一个样本都属于而且只能属于某一个类。这里所指的类是一个时间-空间-属性多维要素体,在这个多维体内存在使信息熵的计算具有统计意义的样本个数,这些样本可能原来属于不同的类或有不同的取值,但其中总有一些类或取值区间占据主导地位,以使该多维体的混乱度(熵值)较小。多维体的性质通过与同层次的其它多维体的比较而得以体现,即同划分层次的多维体之间具有最大的对比度(互信息最大)。

在图 4 所示的子集划分中,点线左边的区域(类)包含三种类型的地物分布,而右边包含两种类型的地物分布,尽管这两个区域都并非均一,但很明显 1、2、3 三种地物没有分布在点线右侧而 4、5 两种地物没有在点线左侧分布,因此点线左侧组成一类,而点线右侧组成另一类,从图中可以清楚地看出点线代表一种空间分布模式的划分。这种聚类属于一种概念聚类,无法采用一般的聚类方法,本文提出的这种方法为这类问题的解决提供了一种思路。

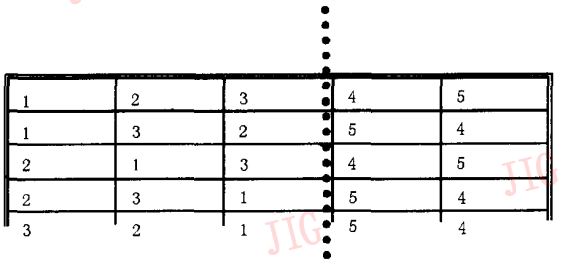

图 4　一种不能被传统聚类方法识别的地物空间分布模式(点线处相区分)

这种聚类实际上是寻找时间和空间变化上的突变线,在地学数据分析中具有重要的意义。如人口地理学中的"胡焕庸"线,自然地理中的 400mm 等强雨量线等。在地学现象中还存在许多像这样的地带

性或非地带性的分界线。目前对这些分界线还只能通过专家目视判断并结合大量的实践经验才能获得,本文提出的方法为从海量地学数据中自动获取这种类型的知识提供了一种分析方法。

与决策树分类方法相类似,这种方法寻找出来的分割线具有与坐标轴平行的特点。这个特点既是一个缺点也是一个优点,缺点在于地学中很少有严格意义上的直线分界线,因此不符合实际情况;优点在于平行坐标轴的直线分割实际上是对折线或曲线的一种近似,具有概括和综合的特点,在一定的情况下符合人类认识地学现象的特点,更便于认识和理解地学现象在空间和时间上突变的特点。

4.2 具有周期特征要素的分割模型

属性和空间要素一般都不具有周期循环的特点,但是在各种类型的时间要素中,如季、日等则存在这种循环往复的现象,其起始点和终结点都是人为划分的,并没有实际含义,不属于严格意义上单向递增的连续变量。因此在探测时间维的子集划分中,一般要通过移位(Shift)操作来消除人为分割点的影响。

首先做一些假设:设时间序列的长度为 N,每个时间点(段)要素分布向量的长度为 M,第 i 个时间点(段)的要素分布向量为 $\boldsymbol{X}_{ij}$,其中 $i=1\cdots n, j=1\cdots m$。

对于第 j 个要素进行具有周期特征分割的步骤如下:

① 取 N 的中点 K 作为初始分割点,$K=\text{int}(N/2)$作为初始分割长度。

② s 从 K 到 0 分别进行 3、4 两步的工作。

③ 顺次朝一个方向移动 s 个时间间隔,移动后的 N 个时间序列样本新序号记为$(s+N)\ \text{mod}\ N$,分别计算分割点 s 处的熵减,并将结果保存到 res[s]中。

④ 返回②;

⑤ 取 res 中的最大者做为最优分割方案,确定相应的 s 和 i。

可以看出,由于考虑了周期循环的特点,计算量大增。N 个样本序列的二分需要进行 $N(N-1)/2$ 次熵减值计算操作,是没有周期特征熵减值计算操作的 $N/2$ 倍。所幸的是,一年之中只有四季和 12 个月,而一天也只有 24 个小时,周期中的样本个数有限。同时,在进行了第一次分割之后,时间维要素将不再具有循环的特点,可以采用一般的计算方法。对于周期中的样本较大的情况,可以加大初始的移位步长然后逐渐缩小。

4.3 时空分割聚类方法的若干问题讨论

4.3.1 关于多维时空要素倾斜分割的问题

在图 5 中,显然按照斜线分割成的两个类具有最大的对比度和熵减。但是由于本文提出的方法具有平行于坐标轴的特点,因此分割出来的结果可能如点线所示。对于这个问题的改进方法是在平行坐标轴分割后进行分割线的旋转,以求得更大的熵减和更有意义的子集划分。Murthy 等指出,对于 n 个样本将 d 维属性空间划分成两个子集有 $2d\times(n^d)$ 个,属于一种 NP 困难问题[17]。尽管后来的一些研究人员对这个问题进行了大量的讨论,分别提出 CART-LC 算法、LMDT 算法、SADT 算法、OC1 算法[17]、Soft Entropy 算法[18]等方法加以改进,但目前在实际应用中尚存在困难。

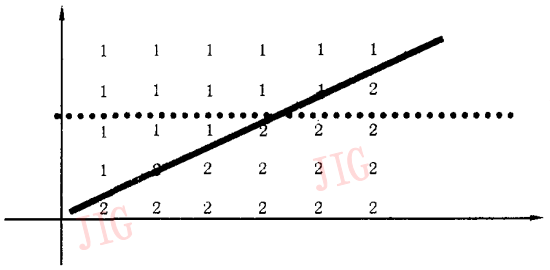

图 5 平行与倾斜子集分割

为此我们建议两种方法:一是人机交互的解决方法,即首先由挖掘模型寻找最有意义的平行分割线提交给领域专家修正,在此基础上进行再次分割。二是充分利用地学中大量的区划方案,指导挖掘模型进行分割。我们在第 3 节中提出的基于信息熵的空间对比和空间组合的方法就是在假设已知分区方案的基础上探讨区域之间在信息熵指标下的关系,在一定程度上也可以弥补平行分割带来的不足。随着对基于信息熵倾斜分割研究的深入,可以提高子集划分的自动化程度。

4.3.2 分割过程中的模糊性与不确定性

同一分割层次的两个子集中的样本分布不能完全区别时,$Rec2<1$。在图 6 所示的例子中,点线处的子集划分的互信息熵 $Rec1<1$ 而 $Rec2=1$。$Rec2$ 的门限值越小,则允许的子集间一致性就越大,允许的子集内不一致性也就越大,因此可以通过给定 $Rec2$ 的门限值来控制划分出来的多维体内部一致性程度,可见基于信息熵的时空分割聚类方法具有一定程度上模糊分割的能力。与一般针对属性的模糊聚

类方法不同,这种方法在兼容模糊性的同时保持了时间和空间的连续性和邻近性。在图 6 中,如果要求较大的 *Rec*2 门限值,则必须分割到左图所示的情况,而如果控制较小的 *Rec*2 门限值,则分割到右图所示的情况即可。具体门限值的确定需要在理论分析和实际检验的基础上由人机交互完成。

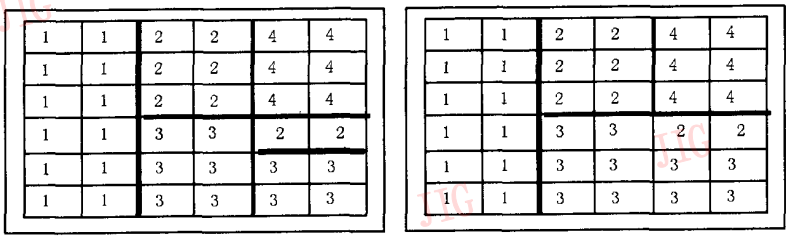

图 6　本文提出的方法对地物分布模糊性的兼容示意图

参 考 文 献

1 李德仁,程　涛. 从 GIS 数据库中发现知识. 测绘学报,1995,24(1):37～43.

2 沈永欢等编. 数学手册. 北京:科学出版社,1997.

3 肖笃宁,李秀珍. 当代景观生态学的进展和展望. 地理科学,1997,17(4):356～363.

4 Fiorentino M, Calps P, Singh V P. An entropy-based morphological analysis of river basin networks. Water Resources Research, 1993,29(4):1215～1224.

5 冯国章,宋松柏,李佩成. 水文系统复杂性的统计测度. 水利学报,1998,11:76～81.

6 张学文,杨秀松. 从熵原理得出的暴雨面积和雨量的关系. 高原气象,1991,10(3):225～232.

7 Neumann J. The topological information content of a map——An attempt at a rehabilitation of information theory in cartography. Cartographica, 1994,31(3):26～33.

8 Zhu A X. Measuring uncertainty in class assignment for natural resource maps under fuzzy logic. Photogrammetric Engineering & Remote Sensing, 1997,63(10):1195～1202.

9 Ryabko Y B. Problemy Peredaci. Informatsii, 1986,20(3):16～26.

10 刘继生,陈彦光. 城镇体系等级结构的分形维数及其测算方法. 地理研究,1998,17(1).

11 王益谦,王　放. 城市人口分布的多重分形特征刻划. 大自然探索,1997,16(62):72～77.

12 梁　虹,卢　娟. 喀斯特流域水系分形、熵及地貌意义. 地理科学,1997,17(4):310～315.

13 Michaelson J, Schmiel D S, Friedl M A *et al*. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. Journal of Vegetation Science, 1994,5:673～686.

14 Hansen M, Dubayah R, Defries R. Classification trees: An alternative to traditional land cover classifiers. Int J Remote Sensing, 1996,17(5):1075～1081.

15 Eklund P W, Kirkby S D, Sallim A. Data mining and soil salinity analysis, Int J GIS, 1998,12(3):247～268.

16 Huang Xuegiao, Jensen J R. A machine learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. Photogrammetric Engineering & Remote Sensing, 1997,63(10):1185～1194.

17 Murthy S K, Kasif S, Salzberg S. A system for induction of oblique decision trees. Journal of Artificial Intelligent Research, 1994,2:1～32.

18 John F Elder IV *et al*. A statistic perspective on knowledge discovery in database. In:Fayyad *et al* (eds), From data mining to knowledge discovery, AAAI Press/The MIT Press, 1997, 83～113.

Entropy-Based Model for Geo-Data Mining

Zhou Chenghu, Zhang Jianting

(*State Key Laboaratory of Resources and Environment Information System, Chinese Academy of Sciences, Beijing* 100101)

Abstract　More and more interest has been paid on Geo-data mining and knowledge discovery from large database with the rapid growth of Geo-data volume and eagerness for the Geo-knowledge. This paper presents a data-set partition model based on information entropy and mutual information. The author argued that the largest information entropy deduction is in accordance with the significant Geo-data pattern. With this kernel theoretical base, information-entropy-based decision-tree model and spatial-temopral clustering by partition model were developed.

Keywords　Information entropy, Data mining, Geo-analysis model