第13卷 第3期
1987 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 13, No. 3
May, 1987

复杂系统分划的熵方法

西 广 成
(中国科学院自动化研究所)

摘 要

本文在给出了反映系统变量(或子系统)之间关系强度的关联度定义之后,证明了关联度的一个基本性质——上可加性,提出了复杂系统分划的熵方法,并给出了检验复杂系统分划合理性的准则,最后给出了用随机抽样法所得数据的熵公式和系统分划的框图.

一、引 言

人类文明的高度发展特别是工农业生产和科学技术的高度发展,将现代控制理论引导到对复杂系统的研究. 复杂系统是由任意多个(≥ 1)子系统组成的具有任意结构的系统,这些子系统各自按照反馈原理形成闭路,按照不同最佳准则进行自调节,相互间发生各种关系,是由动力学的各种逻辑的和启发式的环节组成的. 复杂系统的复杂性至少包含以下四个方面的综合[1]:

1) 多维数性;

2) 多关系性(在同一等级上各个变量之间的关系以及在不同等级上各个变量之间的关系,或者两种关系的交叉);

3) 多判据性;

4) 所用知识的多学科性.

在对复杂系统的研究中,人们常常首先将系统分成相互关联不紧密的若干子系统. 这样,系统的复杂性会明显地减少,从而可将注意力集中于矛盾的主要方面,以便更有成效地进行研究. 本文讨论用熵分划复杂系统的方法.

二、用熵定义的关联度

对于一个复杂系统,如果除了知道描述其特征的一些变量外,什么也不知道,那么这个系统只能表示为矢量

$$\boldsymbol{s} = (X_1, X_2, \cdots\cdots, X_a, \cdots\cdots, X_p)^T. \tag{1}$$

其中 $X_a = (X_{a_i})$, $a = 1, 2, \cdots\cdots, p$; $i = 1, 2, \cdots\cdots, q$, 是描述系统某特征的变量. 令 $\boldsymbol{C}_a (a = 1, 2, \cdots\cdots, p)$ 为 X_a 分类的集合, $\boldsymbol{C}_a$ 的第 i 个元素 $\boldsymbol{C}_{a_i} = i$, 则有

本文于1984年6月19日收到.

$\boldsymbol{C}_a=\{1, 2, \cdots\cdots, i, \cdots\cdots k\}$, $k\leqslant q$, 并令 $n_i\left(\sum_{i=1}^{k} n_i=q\right)$ 为事件 X_a 属于 $\boldsymbol{C}_a$ 第 i 类的数量. 则变量 X_a 的熵定义为

$$H(X_a)=-\sum_{i=1}^{k} n_i/q\log n_i/q \tag{2}$$

X_a, X_b 的联合熵定义为

$$H(X_a\cup X_a)=-\sum_{i}\sum_{j} n_{ij}/q\log n_{ij}/q. \tag{3}$$

其中 n_{ij} 表示事件 X_a 属于 $\boldsymbol{C}_a$ 的第 i 类同时 X_b 属于 $\boldsymbol{C}_b$ 的第 j 类的数量.

以上关于联合熵的定义可推广到 p 个变量的情形. 为了便于应用,式(2)、(3)可分别表示成

$$H(X_a)=\log q-\frac{1}{q}\sum_{i=1}^{k} n_i\log n_i, \tag{2'}$$

$$H(X_a\cup X_b)=\log q-\frac{1}{q}\sum_{i}\sum_{j} n_{ij}\log n_{ij}. \tag{3'}$$

有了上述熵的定义,下面给出关联度的定义.

定义 1. 假定 $X_a\cap X_b=\phi$, 则称熵

$$I(X_a, X_b)=H(X_a)-H(X_a\cup X_b)$$

为 X_a 与 X_b 之间的关联度 $\mu(X_a, X_b)$,

$$\mu(X_a, X_b)\triangleq H(X_a)-H(X_a|X_b). \tag{4}$$

为方便起见,(4)式通常表示为

$$\mu(X_a, X_b)=H(X_a)+H(X_b)-H(X_a|X_b). \tag{4'}$$

定义 2. 假定对任意的 i, $j(i\neq j)$, $X_i\cap X_j=\phi$, p 为任意正整数,则称

$$\mu(X_1, X_2, \cdots\cdots, X_p)\triangleq\sum_{i=1}^{p} H(X_i)-H\left(\sum_{i=1}^{p} X_i\right) \tag{5}$$

为 X_1, X_2, $\cdots\cdots$, X_p 之间的关联度.

以上关于变量之间的关联度的定义完全适用于子系统之间关联度的定义. 事实上,变量本身作为系统的子集也是子系统.

假定系统 $\boldsymbol{s}$ 被分划成 m 个子系统 $s_1, s_2, \cdots\cdots, s_m$, 对任意子系统 $s_i=(X_i^1, X_i^2, \cdots\cdots, X_i^{n_1})$ 和 $s_j=(X_j^1, X_j^2, \cdots\cdots, X_j^{n_r})$ $(i\neq j)$, $s_i\cap s_j=\phi$, $\boldsymbol{s}=\sum_{i=1}^{m} s_i$, 定义

$$\mu(s_1, s_2, \cdots\cdots, s_m)\triangleq\sum_{i=1}^{m} H(s_i)-H\left(\sum_{i=1}^{m} s_i\right) \tag{6}$$

为 s_1, s_2, $\cdots\cdots$, s_m 之间的关联度.

考虑一个非空有限集 $\boldsymbol{X}$ 及由 $\boldsymbol{X}$ 的所有子集为元素组成的集族 $\boldsymbol{E}(\boldsymbol{X})$, 令 P 是定义在 $\boldsymbol{E}(\boldsymbol{X})$ 上的集函数,并满足

$$\text{(i) } P(\boldsymbol{A})\geqslant 0,\ \forall\boldsymbol{A}\in\boldsymbol{E}(\boldsymbol{X}),\ \text{(ii) } P(\boldsymbol{\phi})=0$$

与集函数的次可加性[2]相对应,引入下述定义.

定义 3. 如果对任意非空集 $\boldsymbol{S}_i\in\boldsymbol{E}(\boldsymbol{X})$, $\boldsymbol{S}_i\in\boldsymbol{E}(\boldsymbol{X})$, $i\neq j$, $\boldsymbol{S}_i\cap\boldsymbol{S}_j=\boldsymbol{\phi}$, 有

$$P(\boldsymbol{S}_i \cap \boldsymbol{S}_j) \geqslant P(\boldsymbol{S}_i) + P(\boldsymbol{S}_j), \tag{7}$$

则满足 (i)、(ii) 的集函数 P 称为上可加的.

为讨论关联度的性质,先看一个事实: 一对老夫妇有两个儿子——老大和老二,两人均已结婚并各有子女.当二人分家单过以后,他们的子女之间的关系(指经济等主要关系)便消失了. 一般来讲, 当把一个有限点集划分成两个不相交的子集后, 属于两个不同子集的任何两个元素之间的关联性便减少了, 消失了. 因此, 关联度所具有的唯一性质应该是上可加性, 即一个可分划成若干子集的有限集的关联度一定大于或等于各子集的关联度的和. 这一性质可表成如下的定理.

定理. 关联度 $\mu(s_1, s_2, \cdots\cdots, s_m)$ 是上可加的,并且是唯一的.

证. 由上述关联度的定义,唯一性是显然的.现证明上可加性. 设系统 $\boldsymbol{s}$ 被分划成 m 个子系统, 对任意的 $i, j(i \neq j), s_i \neq \phi, s_j \neq \phi, s_i \cap s_j = \phi, \boldsymbol{s} = \sum_{j=1}^{m} s_j = \sum_{s_j \in \boldsymbol{s}} s_j$, 只须证明

$$\mu\left(\sum_{j=1}^{m} s_j\right) \geqslant \sum_{j=1}^{m} \mu(s_j). \tag{8}$$

由关联度定义,有

$$\begin{aligned}\mu(\boldsymbol{s}) &= \mu\left(\sum_{j=1}^{m} s_j\right) = \mu(s_1, s_2, \cdots\cdots, s_m)\\ &= \mu(X_1, X_2, \cdots\cdots, X_p)\\ &= \sum_{i=1}^{p} H(X_i) - H\left(\sum_{i=1}^{p} X_i\right). \end{aligned}\tag{9}$$

$$\begin{aligned}\sum_{s_i \in \boldsymbol{s}} \mu(s_j) &= \sum_{s_j \in \boldsymbol{s}} \left(\sum_{X_j \in s_j} H(X_j) - H\left(\sum_{X_j \in s_j} X_j\right)\right)\\ &= \sum_{i=1}^{p} H(X_i) - \sum_{s_j \in \boldsymbol{s}} \boldsymbol{H}(s_j)\\ &= \sum_{i=1}^{p} \boldsymbol{H}(X_i) - \sum_{j=1}^{m} \boldsymbol{H}(s_j). \end{aligned}\tag{10}$$

由(9)式减去(10)式,得[3]

$$\begin{aligned}\mu\left(\sum_{j=1}^{m} s_j\right) - \sum_{j=1}^{m} \mu(s_j) &= \sum_{j=1}^{m} H(s_j) - H\left(\sum_{i=1}^{p} X_i\right)\\ &= \sum_{j=1}^{m} \boldsymbol{H}(s_j) - \boldsymbol{H}\left(\sum_{j=1}^{m} s_j\right) \geqslant 0. \end{aligned}$$

定理证完.

三、分划的要求和分划的方法

由以上对于关联度性质的讨论, 理想的分划要求的最低准则应该是, 对任意的 i, j $(i \neq j)$, $s_i \cap s_j = \phi$, 应有

(1) $\mu(s_i) \geqslant \mu(s_i, s_j)$, $\mu(s_i) \geqslant \mu(s_i, s_j)$.

(2) $\mu(s_i) \geqslant \mu(s_{i_1}) + \mu(s_{i_2})$, $\mu(s_j) \geqslant \mu(s_{j_1}) + \mu(s_{j_2})$.

其中 s_{il}, $s_{jl}(l = 1, 2)$ 分别为对 s_i, s_j 的分划. (1) 表示分划所得任意子系统本身的关联度大于或等于它与任意子系统之间的关联度; (2) 表示分划所得任意子系统的关联度具有上可加性.

当系统特征变量的数据量很大时,由于时间和空间的限制,不可能对数据进行全面的观测. 即使有观测的可能,也由于需要很长时间才能获得全部数据等原因,使所得数据失去价值. 在这种情况下,可采用统计方法获得数据[4].

设 $\boldsymbol{x}_a = (x_{a_1}, x_{a_2}, \cdots\cdots, x_{a_N})$ 是复杂系统特征变量 X_a 的数量指标,经随机抽样法测得 $\boldsymbol{x}_a = (x_{a_1}, x_{a_2}, \cdots\cdots, x_{a_p})$, $\boldsymbol{x}_a$ 是来自总体 X_a 的随机样本. 显然,任意 x_{ai} 必等于某一 $x_{a\theta_i}$,θ_i 是 $\boldsymbol{x}_a$ 的第 i 个个体的号码. 对任意 θ_i,有概率

$$p\{x_{a_1} = x_{a\theta_1}, x_{a_2} = x_{a\theta_2}, \cdots\cdots, x_{a_q} = x_{a\theta_q}\} = (N - q)!/N! \tag{11}$$

相应于式(2)′(3)′,有

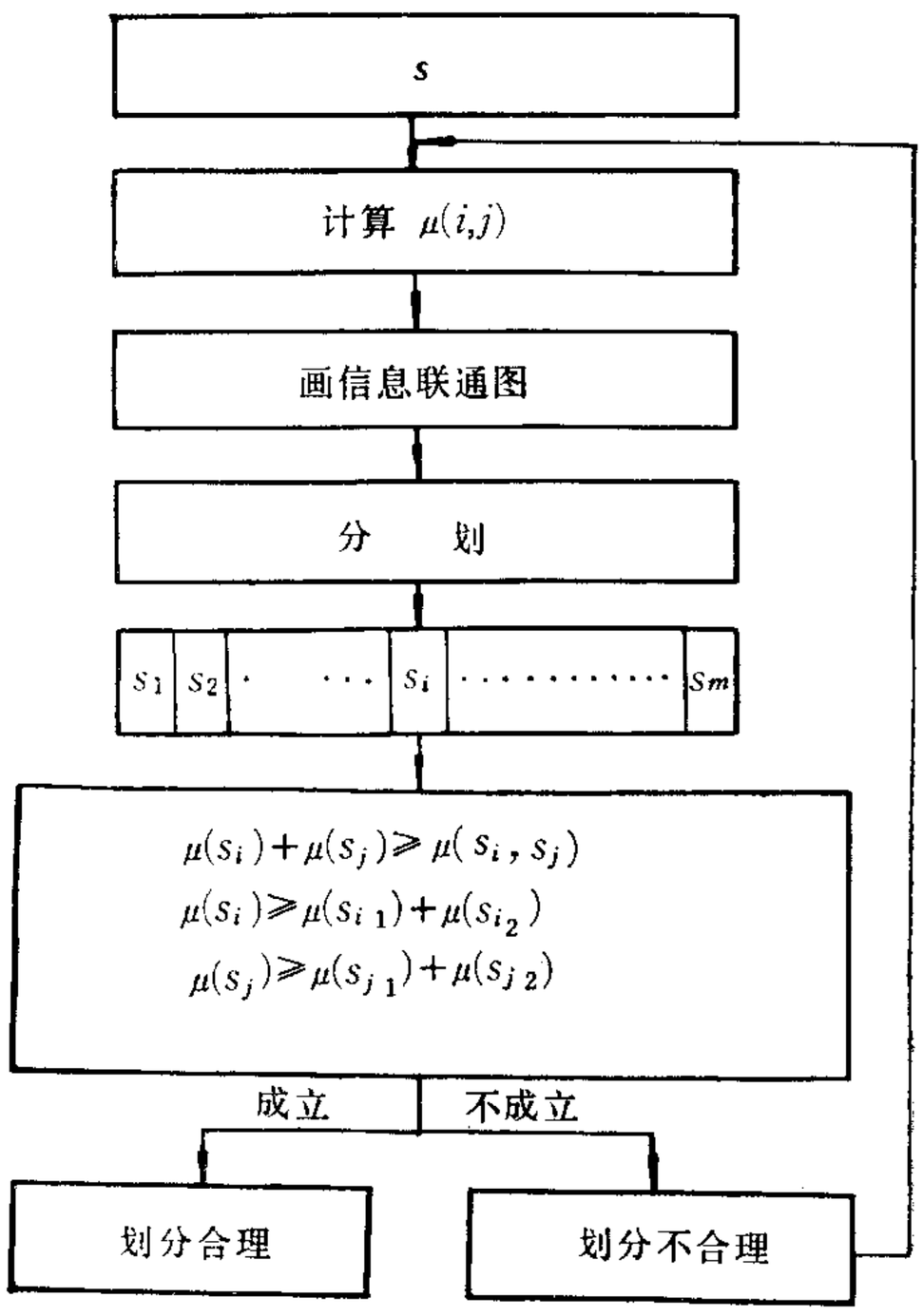
$$\boldsymbol{H}(x_a) = \frac{(N-q)!}{N!}\left[\log(N!q) - \log(N-q)! - \frac{1}{q}\sum_i n_i \log n_i\right], \tag{12}$$

$$\boldsymbol{H}(X_a \cup X_b) = \frac{(N-q)!}{N!}\left[\log(N!q) - \log(N-q)! - \frac{1}{q}\sum_i \sum_j n_{ij} \log n_{ij}\right]. \tag{13}$$

实际上,由观测数据具体地指出所划得的子系统时,常引进关联度系数 $\mu(i,j)$.

$$\mu(i, j) = \mu(X_i, X_j)/\boldsymbol{H}(X_j) \tag{14}$$

由于 $\mu(X_i, X_j)$ 要受到 X_i 与 X_j 所分类数或量级的影响，而使用关联度系数可以大大消



复杂系统熵分划实现框图

除这种影响. 显然, $\mu(i, j)$ 在 1 和 0 之间.

为了把一个复杂系统分划成若干子系统,须对所有的 i,j 计算 $\mu(i, j)$. 在所得到的矩阵(称为关联度矩阵)上, 任意 X_i, X_j 或 S_i, S_j 之间的关联度的大小是很明显的. 根据关联度矩阵 $M = [\mu(i, j)]$ 可画出信息联通图,方法是用宽度正比于 $\mu(i, j)$ 的带箭头的线(箭头指向 X_j), 将那些 $\mu(i, j)$ 大的 X_i, X_j 联在一起, 从而得到相应的若干子系统[5]. 这种分划复杂系统方法的计算机实现亦颇令人感兴趣.

为了反映 X_i, X_j 之间的动态关联性,可使用动态关联度系数 $\mu'(i, j)$,

$$\mu'(i, j) = \mu(X_i, X'_j)/H(X'_j). \tag{15}$$

其中, 若 $X_i = X_i(n)$, 则 $X'_j = X_j(n+1)$. 例如, $n = 1, 2, \cdots\cdots, N$ 是离散时间点.

上页图为复杂系统熵分划实现框图.

用这种方法研究我国某一大的区域(这是一个社会——经济——生态复杂系统)进行生态经济区划,计算机给出的结果是令人满意的.

参 考 文 献

[1] В. М. Глушков, и Т. Д., Сложные Системы Управления. Выиуск IV, Наукова Думак, Киев ——1968.

[2] 夏道行等,实变函数论与泛函分析,上册,人民教育出版社, 1978.11. 北京.

[3] 有本卓,现代情报理论,昭和 53 年 1 月 10 日.

[4] 复旦大学,概率论,第二册,1981.3.

[5] Roger C. Conant, Detecting Subsystem of a Complex System, *IEEE Transactions on System, Man, and Cybernetics*, V. SMC-**2**(1972), 550—553.

ENTROPY-METHOD OF PARTITION OF COMPLEX SYSTEM

XI GUANGCHENG

(*Institute of Automation, Chinese Academy of Sciences*)

ABSTRACT

Having given the definition of the measure of relation that represents the relational strength between variables (or sub-systems) of a system, a fundamental nature——super-additivity of the measure of relation is provided in this paper. An entropy-method of partition of complex system and a criterion for testing the reasonableness of the partition are proposed. The entropy-formula of data obtained by means of randomsampling method and the flow chart for realizing the partition of complex system are given at the end of the paper.