一种基于最大模糊熵的高斯聚类算法*

谭扬波** 陈光禹

(电子科技大学自动化系 成都 610054)

【摘要】 介绍了一种新的模糊聚类方法，定义了模糊熵，提出了基于最大模糊熵的模糊聚类的方法，得到了一种新的聚类算法—— GCM算法。该算法的物理意义清晰，有明确的数学含义，相对于传统的FCM聚类算法，其聚类效果更好。

关 键 词 熵； 最大模糊熵； 模糊聚类； 高斯聚类算法

中图分类号 O236； TP182

聚类分析是近十几年发展十分迅速的一种新的数学方法，采用这种方法可以定量地确定研究对象之间的亲疏关系，从而达到对其合理分类的目的。在模式识别、图像分割及边缘特征提取等领域中，常被作为一种无导师学习的算法样本数据进行分类[1]。本文在模糊熵的基础上，给出了基于最大模糊熵的高斯聚类算法，相对于传统的FCM聚类算法，其聚类效果更好，而且算法的物理意义清晰，有明确的数学含义。

1 模糊熵的定义

对模糊集 $A=\{x_1,x_2,\cdots,x_n\}$，假设其隶属度分别为 $\mu_A(x_1)$， $\mu_A(x_2),\cdots,\mu_A(x_n)$，令其模糊熵为 $e(\mu_A(x_i))$，则 $e(\mu_A(x_i))$ 应满足：1) $e(\mu_A(x_i))$ 随 $\mu_A(x_i)$ 的增加而减少；2) 当 $\mu_A(x_i)$ 为1时，$e(\mu_A(x_i))$ 为0；3) 两个独立的模糊集合的熵应满足可加性。这里可加性是一个很严格的条件，只有满足可加性模糊集合的熵才能唯一确定。如果忽略此条件，将会有很多不确定函数能满足条件1)、2)[2]。对两个独立模糊集合A、B，其积为

$$AB \Leftrightarrow \mu_{AB}(x)=\mu_A(x)\mu_B(x) \tag{1}$$

集合AB的熵定义为

$$e(\mu_{AB}(x_i))=e(\mu_A(x_i))+e(\mu_B(x_i)) \tag{2}$$

类似随机熵的证明可以得到满足以上三项条件的模糊熵的表达式为[3]

$$e(\mu_A(x))=-K\ln\mu_A(x) \tag{3}$$

式中 K是一个大于0的数。于是模糊集合A的平均不确定度，即模糊熵为

$$S(A)=-K\sum_{i=1}^{n}\mu_A(x_i)\ln\mu_A(x_i) \tag{4}$$

2 最大模糊熵在模糊聚类中的应用

最大熵原理已经在测量理论的误差处理、谱估计、图像恢复及协同宏观分析中得到了广泛的应用。本文把最大熵原理推广到模糊领域，讨论最大模糊熵方法，并将其应用到模糊聚类之中。在已知模糊集中各元素的隶属度函数后，采用最大模糊熵对输入数据进行处理，其数学表达式为

$$\max\left\{-K\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_i(\bar{x}_k)\ln\mu_i(\bar{x}_k)\right\} \quad \text{s.t.}\ \ c_1,c_2,\cdots,c_m \tag{5}$$

1999年10月13日收稿
* 电子部预研基金资助项目
** 男 27岁 博士生

式中 $\mu_i(\bar{x}_k)=\mu_{ik}$，$c_1,c_2,\cdots,c_m$为$m$个约束条件。

下面运用最大模糊熵原理对输入数据进行聚类。首先定义损失函数L为

$$L=\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik}d_{ik}^2 \tag{6}$$

式中 $d_{ik}^2=\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^2$，$c$、$n$分别为聚类个数及输入数据的个数，$\boldsymbol{x}_k$为输入$k$维向量，$\boldsymbol{v}_i$为第$i$类的聚类中心，$\mu_{ik}$为第$k$个输入向量属于第$i$类的隶属度函数，且$\mu_{ik}$须满足

$$\sum_{i=1}^{c}\mu_{ik}=1 \qquad \forall k \tag{7}$$

类似FCM聚类算法，需进行如下迭代：1) 假设聚类中心$\{v_i\}$已知，并由此计算$\{\mu_{ik}\}$；2) 固定$\{\mu_{ik}\}$，计算出新的聚类中心$\{v_i\}$。

在迭代1)中，每一个输入向量均利用式(6)进行计算，需满足式(7)和式(8)的约束条件，式(8)的约束条件η为

$$\sum_{i=1}^{c}\mu_{ik}d_{ik}^2=\eta \tag{8}$$

式中 η应尽可能地小。于是该问题就成为

$$\max\left\{-K\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik}\ln\mu_{ik}\right\} \tag{9a}$$

s.t.

$$\sum_{k=1}^{n}\mu_{ik}d_{ik}^2=\eta \tag{9b}$$

$$\sum_{i=1}^{c}\mu_{ik}=1 \qquad \forall k \tag{9c}$$

为得到聚类结果，可采用Lagrangian乘法。由于隶属度函数采用了高斯形式，故令

$$\mu_{ik}=\lambda\exp(-d_{ik}^2/2\sigma^2) \tag{10}$$

式中 λ、σ均为Lagrangian乘子。由式(9c)约束可得

$$\lambda=\frac{1}{\sum_{j=1}^{c}\exp(-d_{jk}^2/2\sigma^2)}$$

代回式(10)可得

$$\mu_{ik}=\frac{\exp(-d_{ik}^2/2\sigma^2)}{\sum_{j=1}^{c}\exp(-d_{jk}^2/2\sigma^2)} \qquad i=1,2,\cdots,c;\ j=1,2,\cdots,n \tag{11}$$

参数σ通过式(9b)的约束与κ建立联系，在实际运用中只须给出σ即可，因为可由σ直接得到κ，即损失函数L的值。

步骤2)是计算式(6)中损失函数L的极限值。为得到该值，应有

$$\frac{\partial}{\partial\boldsymbol{v}_i}\left(\sum_{k=1}^{n}\mu_{ik}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^2\right)=0 \quad \forall i \tag{12}$$

即$\sum_{k=1}^{n}\mu_{ik}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|=0$，于是有

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} \mu_{ik} \boldsymbol{x}_k}{\sum_{k=1}^{n} \mu_{ik}} \qquad \forall k \tag{13}$$

由此可以得到采用高斯最大模糊熵法(GCM)进行聚类的算法步骤为：

1) 给定 $\sigma > 0$，$\varepsilon > 0$，$2 < c \leqslant n$ 及最大迭代次数；

2) 对每一个 i，k 随机初始化 $\mu_{ik} \in [0,1]$；

3) 对 t=1,2,…,T，k=1,2,…,n，由式(13)计算可得 $\{v_i(t)\}$，从式(11)计算得 $\{\mu_{ik}(t)\}$，循环 k；

4) 若 $\max_{i,k} |\mu_{ik}(t) - \mu_{ik}(t-1)| < \varepsilon$ 或 $t > T$，停止，否则循环 t。

采用高斯最大模糊熵聚类方法后，利用模拟退火法，式(6)可以得到全局最小和 σ 的最优值[4]，相对于FCM聚类法来说其聚类效果更好。另外，参数 σ 不仅决定了损失函数 L 的值，还决定聚类的有效个数。若 σ 接近于0，聚类结果就接近于“硬”聚类结果。在实际运用中，参数 σ 的取值取决于使用者的要求。

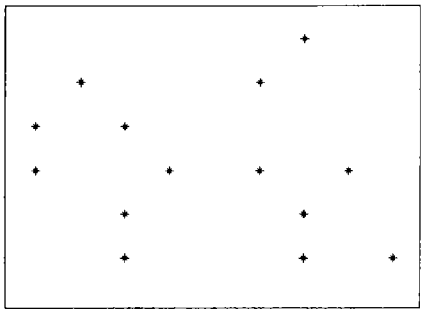

图1　输入数据集

3　聚类结果

本文利用GCM算法对输入数据进行聚类。设有数据集 $\boldsymbol{X}=\{x_1, x_2, \cdots, x_{16}\}$，如图1所示[5]。对该数据集同时采用FCM算法和GCM算法，在FCM算法中令 $c=2$，$m=2.0$，$\varepsilon=10^{-3}$；在GCM算法中令 $c=2$，$\sigma=1.5$，$\varepsilon=10^{-3}$，其聚类结果如表1所示，其中由FCM算法得到的聚类中心为 v_1=(1.43, 2.82)，v_2=(6.14, 3.15)；由GCM算法得到的聚类中心为 v_1=(1.41, 2.76)，v_2=(6.17, 3.24)。从表1可以看出，采用GCM算法得到的聚类结果更好。

表1　GCM与FCM聚类结果对比

输入数据	FCM算法结果		GCM算法结果	
x_k	μ_1	μ_2	μ_1	μ_2
(0,4)	0.92	0.08	1.00	0.00
(0,3)	0.95	0.05	1.00	0.00
(1,5)	0.86	0.14	1.00	0.00
(2,4)	0.91	0.09	0.97	0.03
(3,3)	0.80	0.20	0.84	0.16
(2,2)	0.95	0.05	0.98	0.05
(2,1)	0.86	0.14	0.99	0.01
(1,0)	0.82	0.18	1.00	0.00
(5,5)	0.21	0.79	0.05	0.95
(6,5)	0.12	0.88	0.01	0.99
(7,6)	0.18	0.82	0.01	1.00
(5,3)	0.01	0.99	0.07	0.93
(7,3)	0.02	0.98	0.00	1.00
(6,2)	0.06	0.94	0.01	0.99
(6,1)	0.16	0.84	0.01	0.99
(8,1)	0.15	0.85	0.00	1.00

4 结 论

本文在模糊熵的基础上提出了基于最大模糊熵的方法，并将其应用于模糊聚类中，得到了一种新的聚类算法——GCM算法。从聚类结果可以看出，采用GCM得到的聚类中心比采用FCM得到的聚类中心更接近，即其聚类结果更好，且GCM算法的物理意义更加清晰，并有明确的数学含义，因此基于模糊最大熵的GCM算法有着广阔的应用范围。

参 考 文 献

1 Arakawa K, Arakawa Y. A nonlinear digital filter using fuzzy clustering. ICASSP92,1992,4: 309~312

2 Zadeh L A, The concept of a linguistic variable and its application to approximte reasoning. I, II, III. Inform Sci, 1975, 8,9: 199~245，301~357, 43~80

3 Aczel J，Daroczy Z. On measures of information and their characterizations. New York: Academic Press，1975

4 Geman S，Geman D. Stochastic relaxation，Gibbs distributions，and the Bayesian restoration of image. IEEE Trans Pattern Anal Machine Intell，1984，PAMI-6: 721~741

5 Bezdek J C. Pattern recognition with fuzzy objective function algorithms. New York: Plenum，1981

6 谭扬波，陈光禹. 固定极Reed-Muller展开式在布尔函数等效性的应用. 电子科技大学学报，1999, 28(2): 216~218

A Gaussian Clustering Method Based on Maximum Fuzzy Entropy

Tan Yangbo Chen Guangju

(Dept. of Automation, UEST of China Chengdu 610054)

Abstract This paper proposes a new method for fuzzy clustering. Fuzzy entropy is defined at first, then a new fuzzy clustering method based on maximum fuzzy entropy is proposed, that is GCM method. Compared to traditional FCM method, this method has clearer physical meaning and well-defined mathematical features. It also has better clustering results.

Key words entropy; maximum fuzzy entropy; fuzzy clustering; gaussian clustering method