# 基于模糊熵的聚类有效性函数*

范九伦 吴成茂

(西安邮电学院 信息与控制系 西安 710061)

**摘 要** 模糊熵描述了一个模糊集的模糊性程度.本文将模糊熵应用于聚类有效性的判决,指出用于聚类有效性判决的划分系数是一个基于模糊熵的判决标准.通过几个数据对不同模糊熵公式的判决功能进行了比较实验.

**关键词** 模糊熵,模糊 c-均值聚类,聚类有效性

**中图法分类号** O235 A

## 1 引 言

"熵"是信息论中非常重要的一个基本概念,它描述了一个随机分布的不确定性程度.1968 年 Zadeh[1]在模糊集理论中首次提出'熵'概念,之后 Deluca 和 Termini[2]进行了奠基性工作,提出了有限集上的模糊熵的公理化定义并模仿信息论中的香农熵,给出了第一个模糊熵公式.1992 年刘学成[3]首次提出模糊熵的一般化公理定义,讨论了模糊熵与贴近度的相互诱导关系.我们在此基础上作了更深入的探讨,讨论了模糊熵、贴近度、包含度的相互诱导关系[4].模糊熵描述了一个模糊集的模糊性程度,目前已提出了各种各样的模糊熵公式[4,5].在图像分割[6,7]、神经网络结构[8]等方面,模糊熵已有成功的应用.模糊熵在模式识别领域也有许多应用,如 1985 年,Pal[9]把模糊熵应用于确定数据集的大致聚类中心位置.

依据 Zadeh 的模糊集概念,Ruspini[10]在数据集上定义了模糊划分并对模糊聚类进行了研究.随后 Dunn 将硬 c-均值聚类算法推广到模糊情形,Bezdek[11]作了更一般性的工作.目前模糊 c-均值聚类算法及其推广形式在许多方面已有成功的应用.在应用模糊 c-均值聚类(FCM)算法时,一个首先需要给定的参数是数据集的分类数.确定数据集的分类数问题属于聚类有效性问题.Bezdek[11]定义了划分系数和划分熵来进行有效性的判决.我们曾基于数据集的模糊划分,依据包含度概念[12]提出了一个聚类有效性函数.本文,我们依据模糊熵概念,提出新的聚类有效性函数并指出划分系数实际上是一个基于模糊熵的聚类有效性公式.

## 2 模糊熵简介

模糊熵描述了一个模糊集的模糊性程度.分明集是不模糊的,因此要求分明集的模糊性为零;隶属度取值均为 0.5 的模糊集(记做$[\frac{1}{2}]$)是隶属度最难确认的模糊集,$[\frac{1}{2}]$的模糊性应最大;直观上看模糊集 $A$ 和补集 $A^c$(即 $A^c(x) = 1 - A(x)$)距$[\frac{1}{2}]$的远近程度是相同的,自然要求 $A$ 和补集 $A^c$ 的模糊性程度是一样的.另外,模糊集 $A$ 的模糊性应具有单调变化的性质,即 $A$ 越接近$[\frac{1}{2}]$,$A$ 的模糊性越大;$A$ 越远离$[\frac{1}{2}]$,$A$ 的模糊性越小.由此定义模糊熵如下.

**定义 1**[4] 实函数 $e: F(X) \to R^+$ 叫 $F(X)$ 上的模糊熵,若 $e$ 满足以下四个性质:

(1) $\forall D \in P(X)$, $e(D) = 0$;

(2) $e([\frac{1}{2}]) = \max_{A \in F(X)} e(A)$;

* 国家自然科学基金资助项目
收稿日期:2000 - 10 - 12;修回日期:2001 - 09 - 12

(3) $\forall A, A^* \in F(X)$,若 $A(x) \geqslant \frac{1}{2}$ 时 $A(x) \leqslant A^*(x)$; $A(x) < \frac{1}{2}$ 时 $A(x) \geqslant A^*(x)$, 则 $e(A) \geqslant e(A^*)$;

(4) $e(A) = e(A^c)$.

这里 $F(X)$ 表示论域 $X$ 上的全体模糊集之集,$P(X)$ 表示 $X$ 上的全体分明集之集,$R^+ = [0, +\infty)$.

以下取论域 $X = \{x_1, x_2, \cdots, x_n\}$,一些常用的模糊熵公式(参见文[4]) 如下:

$$e_1(A) = \frac{4}{n}\sum_{i=1}^{n} A(x_i)(1 - A(x_i)),$$

$$e_2(A) = -\frac{1}{n\ln 2}\sum_{i=1}^{n}[A(x_i)\ln A(x_i) + (1 - A(x_i))\ln(1 - A(x_i))],$$

$$e_3(A) = \frac{2}{n}\sum_{i=1}^{n} |A(x_i) - A_{near}(x_i)| = \frac{2}{n}\sum_{i=1}^{n} A(x_i) \wedge (1 - A(x_i)) = 1 - \frac{1}{n}\sum_{i=1}^{n} |1 - 2A(x_i)|,$$

$$e_4(A) = \frac{\sum_{i=1}^{n} A(x_i) \wedge (1 - A(x_i))}{\sum_{i=1}^{n} A(x_i) \vee (1 - A(x_i))} = \frac{\sum_{i=1}^{n} |A(x_i) - A_{near}(x_i)|}{\sum_{i=1}^{n} |A(x_i) - A_{far}(x_i)|},$$

$$e_5(A) = \frac{1}{n}\sum_{i=1}^{n} \frac{A(x_i) \wedge (1 - A(x_i))}{A(x_i) \vee (1 - A(x_i))};$$

这里

$$A_{near}(x) = \begin{cases} 1 & A(x) \geqslant \frac{1}{2} \\ 0 & A(x) < \frac{1}{2} \end{cases},$$

$$A_{far}(x) = \begin{cases} 0 & A(x) \geqslant \frac{1}{2} \\ 1 & A(x) < \frac{1}{2} \end{cases}.$$

## 3　基于模糊熵的有效性

模糊聚类问题可表示成下面的数学规划问题:

$$\min J_m(U, V) = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^m d_{ij}^2,$$

使得 $\sum_{j=1}^{c} u_{ij} = 1, i = 1,2,\cdots,n$; $u_{ij} \geqslant 0, i = 1,2,\cdots,n, j = 1,2,\cdots,c$; $\sum_{i=1}^{n} u_{ij} > 0, j = 1,2,\cdots,c$.

这里 $X = \{x_1, x_2, \cdots, x_n\} \subset R^s$ 是数据集,$n$ 是数据集中元素的个数,$c$ 是聚类中心数($1 < c < n$), $m$ 是权重系数($m > 1$), $d_{ij} = \| x_i - V_j \|$ 是样本点 $x_i$ 和聚类中心 $V_j$ 的欧氏距离,$V_j \subset R^s$. $u_{ij}$ 是第 $i$ 个样本属于第 $j$ 个中心的隶属度,$U = \{u_{ij}\}$ 是一个 $n \times c$ 矩阵,$V = [V_1, V_2, \cdots, V_c]$ 是一个 $s \times c$ 矩阵. 在文献[11] 中,Bezdek 给出了解决上述数学规划问题的迭代算法.

如何检验聚类的类数是否合理?一个好的分类应使具有相同特征的样本尽可能地分在一类,具有不同特征的样本尽可能地不在同一类. 基于数据集的模糊划分的聚类有效性函数建立在如下的思想:就模糊划分而言,划分的分明性越好,分类的不确定性就越小. 因此聚类有效性函数的定义是基于模糊划分的不确定性最小. 事实上,一个能很好分类的数据集,其模糊性是不能很大的. 模糊聚类的结果是对数据集进行模糊划分:$A = \{A_1, A_2, \cdots, A_c\}$. (这里 $c$ 是类数,$A_j$ 表示样本属于第 $j$ 类的模糊集,即隶属度矩阵 $U$ 的第 $j$ 列). 一个好的分类应使 $A_j$ 的模糊性尽可能地小. 基于以上分析,对于给定的聚类中心数 $c$ 和隶属度矩阵 $U$,我们给出公式 $E(U;c) = \sum_{j=1}^{c} e(A_j)$,作为一个好的分类应使 $E(U;c)$ 的取值最小. 即令 $\Omega_c$ 表示 $U \in M_{fnc}$ 的"最优"的有限集合,若存在 $(U^*;c^*)$ 满足

$$E(U^*;c^*) = \min_{c}\{\min_{\Omega_c} E(U;c)\},$$

则 $(U^*;c^*)$ 为最佳的有效性聚类,$c^*$ 为最好的分类数目. 这里 $M_{fnc}$ 表示满足约束条件:

$\sum_{j=1}^{c} u_{ij} = 1, i = 1,2,\cdots,n$; $u_{ij} \geqslant 0, i = 1,2,\cdots,n$, $j = 1,2,\cdots,c$; $\sum_{i=1}^{n} u_{ij} > 0, j = 1,2,\cdots,c$ 的隶属度矩阵之集.

## 4　划分系数的解释

**定义 2**[11]　对于给定的聚类中心数 $c$ 和隶属度矩阵 $U$,划分系数 $F(U;c)$ 定义为

$$F(U;c) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^2.$$

对于模糊熵公式 $e_1(A)$,其相应的有效性函数 $E_1(U;c) = \frac{4}{n}\sum_{j=1}^{c}\sum_{i=1}^{n} u_{ij}(1 - u_{ij})$,由于 $\sum_{j=1}^{c}\sum_{i=1}^{n} u_{ij} =$

$n$，于是有

$$\begin{aligned}E_1(U;c) &= \frac{4}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij} - \frac{4}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^2\\ &= 4 - \frac{4}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^2\\ &= 4\left(1 - \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^2\right)\\ &= 4(1 - F(U;c)).\end{aligned}$$

因此最小化 $E_1(U;c)$ 等价于最大化 $F(U;c)$，即从模糊熵的角度看，划分系数可看成是一个基于模糊熵公式的聚类有效性函数.

## 5 实验结果

既然划分系数是一个基于模糊熵的有效性函数，应用各种各样的模糊熵公式于聚类有效性的研究是可行的.余下的问题是确定哪些公式适合于有效性问题.由于划分系数在有效性问题的判决方面的效果是有限的[13]，我们对基于上述模糊熵公式 $e_1(A),\cdots,e_5(A)$ 的相应聚类有效性函数 $E_1(U;c),\cdots,E_5(U;c)$ 进行了实验比较，相比较而言，$E_3(U;c)$ 更适合于有效性问题.下面给出 $E_1(U;c)$ 到 $E_5(U;c)$ 对几个数据的检测效果.我们选取 $m$ 的三个典型值 $m = 1.5, m = 2.0$ 和 $m = 2.5$，统一取 $c_{max} \leqslant 10$.本节的表格中用黑体字表示最小值.

四维正态分布数据：该数据是 Pal 和 Bezdek 在文[13]中使用的人造数据.由四维空间坐标系中四个坐标轴上的点(3,0,0,0)、(0,3,0,0)、(0,0,3,0)、(0,0,0,3)为中心，方差为 $\sum = diag(1,1,1,1)$ 的正态分布数据，每类100个样本，共计400个数据.由表1和表2可见，$E_1(U;c)$ 和 $E_2(U;c)$ 仅在 $c = 1.5$ 时得到正确的分类；由表3可见，$E_3(U;c)$ 均得到正确的分类；由表4和表5可见，当 $m = 2.5$ 时，$E_4(U;c)$ 和 $E_5(U;c)$ 没有得到四类.

平面五类数据：该数据由平面上的五类数据构成，每类30个数据，共有150个样本.类中心分别在(0,0)、(0,2)、(2,0)、(2,2)、(1,1)，方差为 $\sum = diag(0.4,0.4)$.由表6和表7可见，当 $m = 2.5$ 时，$E_1(U;c)$ 和 $E_2(U;c)$ 没有得到五类.由表8、表9和表10可见，$E_3(U;c)$、$E_4(U;c)$ 和 $E_5(U;c)$ 均得到正确的分类.

表1 四维正态分布数据 $E_1(U;c)$ 值

| $C$ | $m = 1.5$ | $m = 2.0$ | $m = 2.5$ |
|---|---|---|---|
| 2 | 1.124001 | **1.783707** | **1.993474** |
| 3 | 1.031792 | 1.991591 | 2.421177 |
| 4 | **0.722469** | 1.894920 | 2.495172 |
| 5 | 1.052112 | 2.294414 | 2.808687 |
| 6 | 1.262650 | 2.517093 | 3.001291 |
| 7 | 1.422112 | 2.675773 | 3.127972 |
| 8 | 1.567804 | 2.791987 | 3.226865 |
| 9 | 1.619057 | 2.879509 | 3.314852 |
| 10 | 1.698784 | 2.974141 | 3.373391 |

表2 四维正态分布数据 $E_2(U;c)$ 值

| $C$ | $m = 1.5$ | $m = 2.0$ | $m = 2.5$ |
|---|---|---|---|
| 2 | 1.272261 | **1.837473** | **1.995283** |
| 3 | 1.242021 | 2.190582 | 2.557020 |
| 4 | **0.962672** | 2.231668 | 2.793750 |
| 5 | 1.342471 | 2.705232 | 3.219164 |
| 6 | 1.587592 | 3.005234 | 3.531566 |
| 7 | 1.773486 | 3.239086 | 3.769066 |
| 8 | 1.934267 | 3.420700 | 3.966117 |
| 9 | 2.023824 | 3.594035 | 4.164069 |
| 10 | 2.137922 | 3.766089 | 4.325482 |

表3 四维正态分布数据 $E_3(U;c)$ 值

| $C$ | $m = 1.5$ | $m = 2.0$ | $m = 2.5$ |
|---|---|---|---|
| 2 | 0.770771 | 1.419128 | 1.910034 |
| 3 | 0.689087 | 1.401474 | 1.825050 |
| 4 | **0.447258** | **1.248210** | **1.752152** |
| 5 | 0.676793 | 1.540398 | 1.899635 |
| 6 | 0.828856 | 1.695547 | 1.944059 |
| 7 | 0.948668 | 1.787516 | 1.962669 |
| 8 | 1.066461 | 1.851818 | 1.985951 |
| 9 | 1.086302 | 1.858848 | 1.988728 |
| 10 | 1.134341 | 1.886729 | 1.984760 |

表4 四维正态分布数据 $E_4(U;c)$ 值

| $C$ | $m = 1.5$ | $m = 2.0$ | $m = 2.5$ |
|---|---|---|---|
| 2 | 0.477371 | 1.099728 | 1.827813 |
| 3 | 0.391014 | 0.914445 | 1.311473 |
| 4 | **0.236928** | **0.739533** | 1.121784 |
| 5 | 0.365400 | 0.911996 | 1.172710 |
| 6 | 0.446604 | 0.987641 | 1.159976 |
| 7 | 0.509626 | 1.024699 | 1.141360 |
| 8 | 0.571762 | 1.047272 | 1.133719 |
| 9 | 0.578543 | 1.036572 | 1.117886 |
| 10 | 0.601810 | 1.042013 | **1.101730** |

表 5　四维正态分布数据 $E_3(U;c)$ 值

| $C$ | $m=1.5$ | $m=2.0$ | $m=2.5$ |
|---|---|---|---|
| 2 | 0.544976 | 1.143612 | 1.831968 |
| 3 | 0.479031 | 1.011111 | 1.407753 |
| 4 | **0.287767** | **0.851522** | 1.281652 |
| 5 | 0.452991 | 1.077472 | 1.335304 |
| 6 | 0.566733 | 1.194961 | 1.300089 |
| 7 | 0.660417 | 1.248421 | 1.262407 |
| 8 | 0.758417 | 1.282003 | 1.247816 |
| 9 | 0.761774 | 1.248829 | 1.213826 |
| 10 | 0.792187 | 1.242763 | **1.185047** |

表 6　平面五类数据的 $E_1(U;c)$ 值

| $C$ | $m=1.5$ | $m=2.0$ | $m=2.5$ |
|---|---|---|---|
| 2 | 0.555773 | 1.337484 | 1.629032 |
| 3 | 0.610626 | 1.312770 | 1.698372 |
| 4 | 0.506211 | 0.826896 | **1.322628** |
| 5 | **0.069204** | **0.571929** | 1.334885 |
| 6 | 0.188596 | 0.820653 | 1.633303 |
| 7 | 0.280104 | 0.995432 | 1.821899 |
| 8 | 0.408821 | 1.176118 | 1.956728 |
| 9 | 0.438976 | 1.249610 | 2.044302 |
| 10 | 0.568153 | 1.421987 | 2.154808 |

表 7　平面五类数据的 $E_2(U;c)$ 值

| $C$ | $m=1.5$ | $m=2.0$ | $m=2.5$ |
|---|---|---|---|
| 2 | 0.729208 | 1.482893 | 1.711364 |
| 3 | 0.757746 | 1.510526 | 1.911404 |
| 4 | 0.579297 | 1.040851 | **1.641123** |
| 5 | **0.114963** | **0.844832** | 1.745479 |
| 6 | 0.257119 | 1.1291013 | 2.090007 |
| 7 | 0.368736 | 1.326198 | 2.321444 |
| 8 | 0.515868 | 1.522149 | 2.489574 |
| 9 | 0.555042 | 1.629153 | 2.629753 |
| 10 | 0.706440 | 1.809001 | 2.756287 |

表 8　平面五类数据的 $E_3(U;c)$ 值

| $C$ | $m=1.5$ | $m=2.0$ | $m=2.5$ |
|---|---|---|---|
| 2 | 0.341610 | 0.891681 | 1.305942 |
| 3 | 0.390418 | 0.914978 | 1.187271 |
| 4 | 0.358005 | 0.529768 | 0.815392 |
| 5 | **0.037933** | **0.325001** | **0.802588** |
| 6 | 0.122072 | 0.506163 | 1.039040 |
| 7 | 0.182865 | 0.637288 | 1.187582 |
| 8 | 0.278755 | 0.776886 | 1.300490 |
| 9 | 0.299563 | 0.816566 | 1.344699 |
| 10 | 0.387442 | 0.947659 | 1.445483 |

表 9　平面五类数据的 $E_4(U;c)$ 值

| $C$ | $m=1.5$ | $m=2.0$ | $m=2.5$ |
|---|---|---|---|
| 2 | 0.186754 | 0.573738 | 0.969498 |
| 3 | 0.210146 | 0.546482 | 0.742236 |
| 4 | 0.187802 | 0.283744 | 0.453998 |
| 5 | **0.019065** | **0.168583** | **0.437811** |
| 6 | 0.062174 | 0.265720 | 0.570218 |
| 7 | 0.093084 | 0.335033 | 0.649783 |
| 8 | 0.142298 | 0.409129 | 0.708443 |
| 9 | 0.152822 | 0.428819 | 0.728099 |
| 10 | 0.197816 | 0.497996 | 0.780414 |

表 10　平面五类数据的 $E_5(U;c)$ 值

| $C$ | $m=1.5$ | $m=2.0$ | $m=2.5$ |
|---|---|---|---|
| 2 | 0.219596 | 0.604752 | 1.078190 |
| 3 | 0.258865 | 0.661363 | 0.855099 |
| 4 | 0.262262 | 0.348598 | 0.511859 |
| 5 | **0.020927** | **0.189427** | **0.495664** |
| 6 | 0.083098 | 0.326358 | 0.689070 |
| 7 | 0.125128 | 0.428360 | 0.810147 |
| 8 | 0.199804 | 0.539321 | 0.908522 |
| 9 | 0.214564 | 0.558760 | 0.929631 |
| 10 | 0.276557 | 0.658837 | 1.021632 |

## 6　结束语

本文从模糊划分的模糊熵角度定义了一类聚类有效性函数,指出划分系数实际上是一个基于模糊熵的有效性函数.本文只给出了二个数据的实验结果,我们还对其它的数据进行了实验,相比较而言,$E_3(U;c)$ 的总体性能最好.$E_4(U;c)$ 和 $E_5(U;c)$ 次之,$E_1(U;c)$ 和 $E_2(U;c)$ 没有多大区别.本文的实验比较分析是初步的,还需更多的理论分析和实验比较以确定各个模糊熵公式的分类性能和适用范围.

## 参考文献

[1] Zadeh L A. Probability Measures of Fuzzy Events. Journal of Mathematics Analysis and Applications, 1968, 23: 421 - 427

[2] Deluca A, Termini S. A Definition of Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. Information and Control, 1972, 20: 301 - 312

[3] Liu X C. Entropy, Distance Measure and Similarity Measure of Fuzzy Sets and Their Relations. Fuzzy Sets and Systems, 1992, 52: 305 - 318

[4] 范九伦.模糊熵理论.西安:西北大学出版社,1999

[5] Pal N R, Bezdek J C. Measuring Fuzzy Uncertainty. IEEE Trans on Fuzzy Systems, 1994, 2: 107 - 118

[6] Li X Q, Zhao Z W, Cheng H D. Fuzzy Entropy Threshold Approach to Breast Cancer Detection. Information Sciences, 1995, 4: 49 - 56

[7] Pal N R, Pal S K. Object-Background Segmentation Using New Definitions of Entropy. IEE Proc, 1989, 136: 284 - 295

[8] Ghosh A. Use of Fuzziness Measures in Layered Networks for Object Extraction: A Generalization. Fuzzy Sets and Systems, 1995, 72: 331 - 348

[9] Pal S K, Pramanik P K. Fuzzy Measures in Determining Seed Points in Clustering. Pattern Recognition Letters, 1986, 4: 159 - 164

[10] Ruspini E H. A New Approach to Clustering. Information and Control, 1969, 15: 22 - 32

[11] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981

[12] Fan J L, Xie W X. Subsethood Measure: New Definitions. Fuzzy Sets and Systems, 1999, 106: 201 209

[13] Pal N R, Bezdek J C. On Cluster Validity for the Fuzzy c-means Model. IEEE Trans on Fuzzy Systems, 1995, 3: 370 - 379

# CLUSTERING VALIDITY FUNCTION BASED ON FUZZY ENTROPY

Fan Jiulun, Wu Chengmao

(*Department of Information and Control, Xi'an Institute of Post and Telecommunications, Xi'an* 710061)



## ABSTRACT

Fuzzy entropy describes the degree of fuzziness of a fuzzy set . The first formula of fuzzy entropy was defined on finite set by Deluca and Termini, and the axiom definition of fuzzy entropy on general universal set is given by Liu. At present, many fuzzy entropy formulas are proposed. Fuzzy entropy has been applied to image segmentation and neural network. It also has been applied to determining seed points in clustering. Pal and Bezdek discussed some fuzzy entropy formulas from applied viewpoint. Detail statements about fuzzy entropy, relationship with containment degree , and subsethood measure are given in the book written by Fan . In this paper, fuzzy entropy is applied to clustering validity problem. Fuzzy clustering is obtained through a fuzzy partition on the data set and the minimum of the sum of each partition subset' s fuzzy entropy is regared as the clustering validity function . It is pointed out that partition coefficient , a commonly used clustering validity function , is equal to a function based on fuzzy entropy. The performances of different fuzzy entropy formulas are compared on several data sets.

**Key Words** Fuzzy Entropy, Fuzzy C-means, Clustering Validity