安徽大学
本科毕业论文

题　　目: 模糊最大熵模型及其应用

学生姓名: 覃浩蓝　　学号: A01714003

院（系）: 数学科学学院　　专业: 信息与计算科学

入学时间: 二〇一七年　九月

导师姓名: 吴涛　　职称/学位: 教授

导师单位: 数学科学学院

完成时间: 二〇二一年　四月

模糊最大熵模型及其应用

摘　要

本文先介绍了模糊数学中常用的模糊 C 均值（Fuzzy C-means.FCM）算法, 在模糊 C 均值算法中, 通过与模糊数学的融合, 给出了相比于 K-means 硬聚类更灵活的聚类结果. 在 FCM 算法的基础上, 我们介绍了一种基于最大熵准则的模型: 模糊最大熵模型. 最后我们在 UCI 的 iris 数据集上应用了 FCM 算法和我们的最大模糊熵模型, 并比较了它们的分类准确率.

关键字:　熵, 最大模糊熵, 模糊聚类

目 录

1 绪论

1.1 研究背景

进入信息时代后, 我们对信息的获取和加工能力在不断进步, 我们周围的信息越来越多, 五花八门各式各样的信息充斥在我们的生活中, 这些信息有的是确定的, 但更多的是不确定的、带有模糊性的信息. 所谓模糊性是指不确定的, 介于是和不是两者之间的性质. 例如, 对于优等生的判定, 有的人觉得 90 就可以了, 有的人却觉得需要达到 95 分以上才算优秀, 所以我们很难这种非此即彼的性质去衡量一个人是不是优等生.

我们所处的是一个复杂多变、时刻在运动的世界, 大到星系运动, 小到粒子碰撞, 里面蕴含的规律都是复杂多样的. 信息本身就包含着确定性和不确定性, 无所谓的好坏之分, 它取决于我们如何认识信息, 了解信息和使用信息. 比如, 我们在评价某一个菜品时, 会用"好吃"、"还行 "、" 难吃"来形容; 描述天气时说"多云"、"晴朗"; 说一个人的衣服搭配好看等等. 这些问题很难用统一的标准去衡量, 但是我们已经习惯在生活中运用模糊性所谓语言描述事物, 用模糊的方法认识生活中的事物. 虽然信息带着不确定性, 但是我们所处的客观世界是确定的, 所以我们需要一种方法研究模糊的信息, 得到清晰的结论. 于是数学诞生了一个新的分支: 模糊数学.

1965 年,L.A.Zadeh 在期刊 Information and Control 上发表了论文《Fuzzy Sets》[1], 标志着模糊理论的诞生. 随后,1968 年,L.A.Zadeh 又发表了《Probability Measures of Fuzzy Events》[2], 进一步补充模糊理论框架. 此后模糊理论开始进入广大学者的视野, 不断得到完善和改进, 并在控制理论领域得到广泛应用.

1.2 研究内容

最大熵模型是一种分类学习模型, 模糊熵是模糊数学里面的概念, 本文在模糊理论框架下, 将最大熵推广到模型信息情形, 在传统的模糊 C 均值聚类（FCM）上进行改进, 建立模糊最大熵模型并应用于实际的分类问题中, 通过进一步的研究探索模糊最大熵模型在实际问题中的应用.

1.3 研究意义

随着人工智能的大热, 机器学习开始迅速应用于我们的生活中, 比如商品推荐、语音识别和智能导航等. 其中, 分类问题是机器学习领域的一个重要问题. 生活中许多的分类问题是模糊的, 计算机无法直接处理这些模糊信息, 而我们的人脑却可以很好地从这些模糊信息中得到精确的结论. 随着熵理论和模糊数学的发展, 模糊数学和最大熵模型的应用范围也越来越广泛, 为了处理分类问题中的不确定性, 国内外的许多这方面的学者也进行了许多研究, 寻找分类问题模糊性的度量方式, 探寻新的实际应用. 本文将模糊熵与最大熵原理结合,

2 预备知识

聚类方法不仅是揭示给定数据集的基本结构的主要工具, 也是揭示复杂系统的局部输入-输出关系的有效工具. 这自然引起了许多学者的兴趣, 并由此产生了许多的聚类方法, 模糊聚类就是其中一种. 正如我们所知道的, 聚类问题是一个优化问题: 一组物体被分割成一个合理的具有某些特征的子群. 这往往是在主观选择的测量函数的基础上, 将一组物体分成合理数量的子组, 使子组内物体之间的距离小于属于不同子组的物体之间的距离. 对现有的聚类方法进行改进是一个非常困难的问题, 在这之前我们先了解一下模糊 C-均值算法 (FCM) 方法.

2.1 模糊集及其表示方法

在经典集合理论里面, 一个集合就是某一个概念的内涵. 对于论域上的一个对象, 它要么属于这个集合, 要么不属于这个集合, 两者只能选一个, 不能两者兼之, 也不能有模棱两可的情况. 而对模糊数学研究的对象来说, 我们不能简单地用是或否来描述一个对象是否属于一个集合. 由此, 我们把集合的特征函数的取值从 $\{0,1\}$ 这个集合扩充到 $[0,1]$ 这个区间上的连续取值. 越靠近 1, 说明该对象属于集合的程度越大, 反之, 越靠近 0 就越小. 这样我们就把经典集合扩充到带有模糊边界的模糊集了, 从而我们可以用这样的集合表示模糊概念.

2.1.1 模糊集的定义

定义 2.1.1 (模糊子集[1]). *设 U 为我们所研究的论域,*

$$\mu_{\tilde{A}} : \mathrm{U} \longrightarrow [0,1]$$

称 μ 确定了 U 上的一个模糊子集, 记为 $\tilde{A}$. μ 称为 $\tilde{A}$ 的隶属函数, 把 $\mu_{\tilde{A}}(u)(u \in \mathrm{U})$ 的值称为 u 对于模糊子集 $\tilde{A}$ 的隶属度. $\mu_{\tilde{A}}(u)$ 越大, 代表 u 隶属于 $\tilde{A}$ 的程度越高. 通常, 我们也把模糊子集简称为模糊集.

2.1.2 模糊集的表示方法

设有限集 $\mathrm{U} = \{u_1, u_2, \dots, u_n\}$, 则有限集可以用如下几种方法表示[3].

- Zadeh 表示法
$$\tilde{A} = \frac{\tilde{A}(u_1)}{u_1} + \frac{\tilde{A}(u_2)}{u_2} + \cdots + \frac{\tilde{A}(u_n)}{u_n}.$$
虽然我们以分式和的方式表示, 但是其中的 $\tilde{A}(u_i)/u_i$ 并不表示分数, "+" 也不表示和. $\tilde{A}(u_i)/u_i$ 表示的是元素 u_i 与对 $\tilde{A}$ 的隶属度的一一对应关系; "+" 表示的是 $\tilde{A}$ 在论域 U 上的整体.

- 序偶表示法
$$\tilde{A}=\{(\tilde{A}(u_1),u_1),(\tilde{A}(u_2),u_2),\dots,(\tilde{A}(u_n),u_n)\}.$$
序偶表示法是从例举法演变而来, 由元素的隶属度和对应的元素组成的有序对列出.

- 向量表示法
$$\tilde{A}=(\tilde{A}(u_1),\tilde{A}(u_2),\dots,\tilde{A}(u_n)).$$
向量表示法是用 n 维数组来实现的, 在论域中的元素按一定的顺序排列时, 按此顺序记录元素的隶属度. 此时也称 $\tilde{A}$ 为模糊向量.

2.2 模糊集的运算及其性质

我们先给出模糊幂集的定义:

定义 2.2.1. 论域 U 上的模糊子集的全体称为模糊幂集, 记为 $\mathcal{F}(U)$, 即

$$\mathcal{F}(U)=\{\tilde{A}\mid\tilde{A}(u):\mathrm{U}\to[0,1]\}$$

模糊集的包含与相等:

定义 2.2.2. 设 $\tilde{A},\tilde{B}\in\mathcal{F}(U)$, 如果对 $\forall u\in\mathrm{U}$ 都成立 $\tilde{B}(u)\geqslant\tilde{A}(u)$, 则称 $\tilde{B}$ 包含 $\tilde{A}(u)$, 记作 $\tilde{B}(u)\supseteq\tilde{A}(u)$.

定义 2.2.3. 设 $\tilde{A},\tilde{B}\in\mathcal{F}(U)$, 如果对 $\forall u\in\mathrm{U}$ 都成立 $\tilde{B}(u)=\tilde{A}(u)$, 则称 $\tilde{B}$ 等于 $\tilde{A}(u)$, 记作 $\tilde{B}(u)=\tilde{A}(u)$.

我们规定 $a\vee b=MAX(a,b),a\wedge b=MIN(a,b)$, 所以我们可以这样描述模糊集的并、交、余:

定义 2.2.4. 如果对于任意一个 $u\in\mathrm{U}$, 有 $\tilde{C}(u)=\tilde{A}(u)\vee\tilde{B}$, 则称 $\tilde{C}$ 为 $\tilde{A}$ 与 $\tilde{B}(u)$ 的并, 记为 $\tilde{C}=\tilde{A}\cup\tilde{B}$. 如果对于任意一个 $u\in\mathrm{U}$, 有 $\tilde{C}(u)=\tilde{A}(u)\wedge\tilde{B}$, 则称 $\tilde{C}$ 为 $\tilde{A}$ 与 $\tilde{B}(u)$ 的交, 记为 $\tilde{C}=\tilde{A}\cap\tilde{B}$.
它们的隶属度函数定义为:

$$(\tilde{A}\cup\tilde{B})(u)\stackrel{def}{=}\tilde{A}(u)\vee\tilde{B}(u),\forall u\in\mathrm{U}$$

$$(\tilde{A}\cap\tilde{B})(u)\stackrel{def}{=}\tilde{A}(u)\wedge\tilde{B}(u),\forall u\in\mathrm{U}$$

定义 2.2.5. 如果对于 $\forall u\in\mathrm{U}$, 有 $\tilde{B}(u)=1-\tilde{A}(u)$, 则称 $\tilde{B}$ 为 $\tilde{A}$ 的余, 记为 $\tilde{B}=\tilde{A}^c$.

2.3 模糊 C-均值算法

C-均值聚类是我们聚类经常用的方法之一, 通过迭代计算使得目标函数达到局部最小值的时候, 就是我们的最优分类. 在模糊 C-均值聚类中, 我们定义目标函数为:

$$J(A,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(a_{ij}\right)^{r}d_{ij}^{2} \tag{2-1}$$

$U=\{u_1,u_2,\dots,u_n\}, u_j=(x_{j1},x_{j1},\dots,x_{jm})\in R^m$ 为给定的 n 个样本的 m 维数据集, $A=(a_{ij})$ 是隶属度矩阵,r 是模糊数,$d_{ik}=\|u_k-v_i\|$ 是第个 k 个样本到第 i 个聚类中心的距离.

当 v_i 不变时问题等价于

$$\min L(A,\lambda)=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(a_{ij}\right)^{r}d_{ik} \tag{2-2}$$

$$s.t.\sum_{i=i}^{c}a_{ik}=1.\ \forall k \tag{2-3}$$

这是最优化问题, 我们引入拉格朗日乘子 λ, 于是变为

$$L(A,\lambda)=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(a_{ij}\right)^{r}\left\|u_j-v_i\right\|^{2}-\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}a_{ij}-1\right) \tag{2-4}$$

对式2-4求导, 局部最小值时必要条件为

$$\frac{\partial L(A,\lambda)}{\partial a_{ij}}=\left[r\left(a_{ij}\right)^{r-1}\left\|u_j-v_i\right\|^{2}-\lambda_j\right]=0 \tag{2-5}$$

$$\frac{\partial L(A,\lambda)}{\partial \lambda_j}=\sum_{i=1}^{c}a_{ij}-1=0 \tag{2-6}$$

由式2-5可得:

$$a_{ij}=\left(\frac{\lambda_j}{r\|u_j-v_i\|^2}\right)^{\frac{1}{r-1}} \tag{2-7}$$

将式2-7带入式2-6解得:

$$\left(\frac{\lambda_j}{r}\right)^{\frac{1}{r-1}}=\left[\sum_{i=1}^{c}\left(\frac{1}{r\|u_j-v_i\|^2}\right)^{\frac{1}{r-1}}\right]^{-1} \tag{2-8}$$

最后将式2-8代回式2-7得到隶属度的更新公式为:

$$a_{ij}=\left[\sum_{j=1}^{c}\left(\frac{\|u_j-v_i\|}{\|u_j-v_j\|}\right)^{\frac{2}{r-1}}\right]^{-1}\qquad 1\leqslant i\leqslant c,\quad 1\leqslant j\leqslant n \tag{2-9}$$

假设 a_{ij} 不变, 原问题就变成了无约束最优化问题, 必要条件为:

$$\frac{\partial J(A,V)}{\partial v_i}=-\sum_{j=1}^{n}2\left(a_{ij}\right)^r\left(u_j-v_i\right)=0 \tag{2-10}$$

解之得:

$$v_i=\frac{\sum\limits_{j=1}^{n}\left(a_{ij}\right)^r u_j}{\sum\limits_{j=1}^{n}\left(a_{ij}\right)^r},\quad 1\leqslant i\leqslant c \tag{2-11}$$

需要注意的是, 算法中要求 $v_i\neq u_j$, 因此, 在遇到只有一个样本的类别时, 需要将此类别排除, 在聚类结束时再加上.

2.4 最大熵原理

熵原本是物理学中的概念, 是由热力学第二定律引出的一个物质系统的状态参量, 是反映系统的混乱程度, 度量信息有效性的一个重要工具. 自从熵的概念被提出来之后, 很快引起了其他领域研究者的注意, 熵理论得以迅速传播, 渗透进各个领域.

2.4.1 信息熵

1948 年, 香农[4] 提出了信息熵的概念, 解决了信息的量化度量问题.

（1）离散模糊变量的信息熵

设某个离散型的随机变量 X, X 的分布率是 $\{p_i\}$, 且 $p_i=P\{X=x_i\},0\leq p_i\leq 1$, $\sum_{i=1}^{n}\mathrm{p}_i=1,\quad(i=1,2,\ldots,n)$ 则用信息熵来度量事件 X 的确定程度为:

$$H(X)=-\sum_{i=1}^{n}p_i\ln\mathrm{p}_i \tag{2-12}$$

（2）连续连续模糊变量的信息熵

假设某个连续型随机变量 X , 其概率密度函数是 $p(x)$, 则该连续随机变量 X 的信息熵为:

$$H(X)=-\int p(x)\ln p(x)dx \tag{2-13}$$

2.4.2 最大熵

1957 年,E.T.Jaynes 提出了最大熵原理, 通过把随机变量与信息熵联系起来, 然后最大化信息熵. 最大熵原理并不是某一固定的数学公式, 而是一种选择随机变量的准则. 最大熵的主要思想是, 在我们只掌握未知变量的部分知识的时候, 未知变量的分布可能有很多种, 此时我们应该选符合这些知识的情况下, 使得信息熵取得最大值的概率分布.

2.5 模糊熵

在自然界以及我们的日常生活中, 经常存在着带着模糊性质的不确定现象, 我们把它们定义成模糊变量. 自从模糊理论提出以后, 很多学者就在如何度量模糊变量的模糊程度

这一方面进行了许多研究, 比如 Li, Pingke 和 Liu, Baoding[5]、Aldo de Luca 和 Settimo Termini[6] 等.

定义 2.5.1 (模糊熵). 对于离散变量的模糊

$$H(\tilde{A}) = -\sum_{i=1}^{n} \tilde{A}(u_i) \ln \tilde{A}(u_i) \tag{2-14}$$

对于连续的模糊变量

$$H(\tilde{A}) = -\int_{-\infty}^{\infty} \tilde{A}(u) \ln \tilde{A}(u) du \tag{2-15}$$

3　模糊最大熵准则的模糊聚类

3.1　模糊最大熵模型

前面我们介绍了模糊聚类的一种常用方法: 模糊 C-均值聚类，介绍了模糊熵和最大熵原理。现在, 我们就可以将最大熵原理推广到模糊熵的情形, 在模糊 C-均值聚类中加入模糊最大熵进行约束. 设 a_{ij} 为第 j 个元素对于第 i 类的隶属度, 则我们模糊最大熵模型 (Fuzzy Maximum Entropy.FME) 的目标函数可以表示成

$$\max\left\{-\sum_{j=1}^{n}\sum_{i=1}^{c}a_{ij}\ln a_{ij}\right\} \tag{3-1}$$

我们首先得定义我们的损失函数, 在这里, 我们选取

$$\mathrm{L}=\sum_{i=1}^{c}\sum_{j=1}^{n}a_{ij}d_{ij}^{2}$$

作为我们的损失函数. 于是我们可以将问题归结为最优化问题, 用拉格朗日乘子法:

$$L(A,\beta,\lambda)=\sum_{i=1}^{c}\sum_{j=1}^{n}a_{ij}d_{ij}^{2}+\beta\sum_{i=1}^{c}\sum_{i=1}^{n}a_{ij}\ln a_{ij}+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}a_{ij}-1\right) \tag{3-2}$$

其中 β 是差异因子, 由数据集的分布情况来决定,λ_j 是 $\sum\limits_{i=1}^{c}a_{ij}$ 的拉格朗日乘子.

$$\frac{\partial L(A,\beta,\lambda)}{\partial a_{ij}}=d_{ij}^{2}+\beta(\ln a_{ij}+1)+\lambda_j=0 \tag{3-3}$$

$$\frac{\partial L(A,\beta,\lambda)}{\partial \lambda_j}=\sum_{i=1}^{c}a_{ij}-1=0 \tag{3-4}$$

最后解得

$$a_{ij}=\frac{\exp(\frac{-d_{ij}^{2}}{\beta})}{\sum\limits_{i=1}^{c}\exp(\frac{-d_{ij}^{2}}{\beta})} \tag{3-5}$$

$$v_i=\frac{\sum\limits_{j=1}^{n}a_{ij}x_j}{\sum\limits_{j=1}^{n}a_{ij}} \tag{3-6}$$

3.2 模糊最大熵模型的求解算法

在描述我们的算法之前, 我们首先来定义一些模型的输入输出:

$$\text{n 个样本的 m 维数据集}U:U=\{u_1,u_2,\dots,u_n\},u_i=\{x_1,x_2,\dots,x_m\}\forall i$$

$$\text{聚类数}c:2\leqslant c\leqslant n$$

$$\text{隶属度矩阵}A:n\times m\text{的矩阵}$$

$$\text{聚类中心}V:c\times m\text{的矩阵}$$

模糊最大熵模型的聚类算法步骤如下:
输入: 原始数据 U 和聚类数 c
输出: 隶属度矩阵 A 和聚类中心 V

1) 首先随机初始化隶属度矩阵并归一化;

2) 根据式3-6计算聚类中心;

3) 计算每一个元素到聚类中心的距离;

4) 根据式3-5更新隶属度矩阵;

5) 如果到达指定精度或迭代次数, 结束计算过程, 否则重复 $2\sim4$ 步;

6) 输出隶属度矩阵 A 和聚类中心 V.

4　模糊最大熵模型分类应用研究

4.1　IRIS 数据集

此数据是来自 UCI 的鸢尾花 (iris) 数据集. 这是一个经典的分类数据集, 由 Iris Setosa(山鸢尾)、Iris Versicolour(杂色鸢尾) 和 Iris Virginica(维吉尼亚鸢尾) 三种不同类别的鸢尾花组成, 每个样本由四个属性组成, 分别是 Petal.Length(花瓣长度)、Petal.Width(花瓣宽度)、Sepal.Length(花萼长度) 和 Sepal.Width(花萼宽度). 按两两属性绘制原始数据如下:

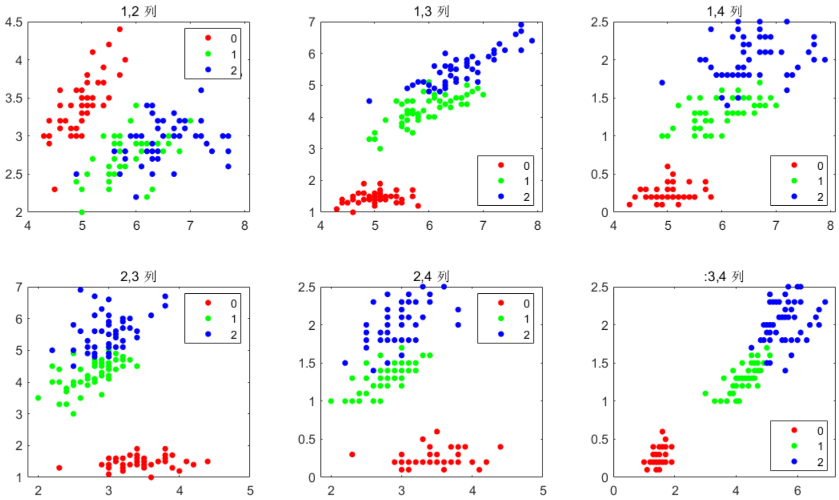

图 4.1　原始数据散点图

按 4.1 的算法对 iris 数据集进行聚类, 结果如下:

表 4.1　聚类中心

	sepal length	sepal width	petal length	petal width
v_1	5.0136	3.3903	1.5369	0.2781
v_2	6.4737	2.9437	5.1910	1.8012
v_3	6.0922	2.8186	4.6775	1.5731

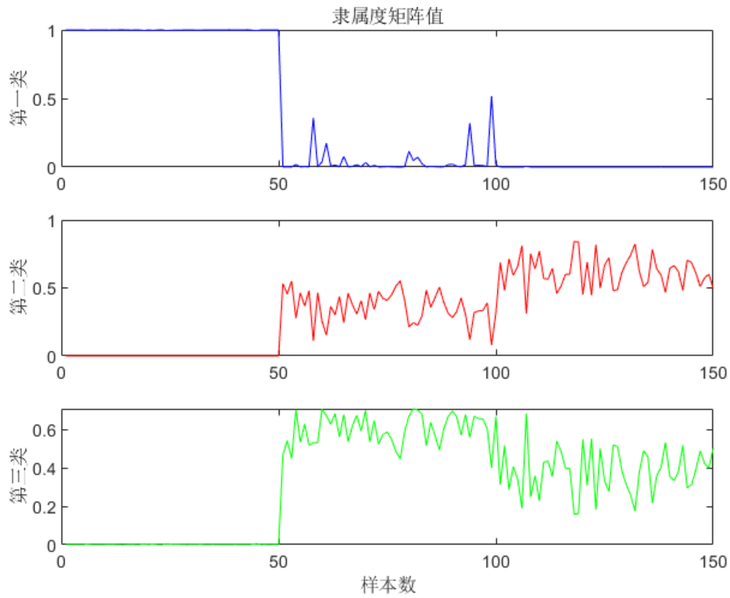

图 4.2　隶属度矩阵的值

与 FCM 算法分类结果比较 从数据的比较中, 可以看到，对于第一类，三个算法都做

表 4.2　准确率比较

算法/各类准确率	Setosa(山鸢尾)	Versicolour(杂色鸢尾)	Virginica(维吉尼亚鸢尾)	总样本
K-means	100%	96%	72%	89.4%
FCM	100%	76%	94%	90%
MFE	100%	86%	98%	94.7%

到了很好的识别，但是对于后两类非线性相关的类别，模糊最大熵模型比其他两个模型有着更好的表现.

4.2　红酒数据集

5 总结与展望

独立的系统演化是一个熵增的过程, 在没有外力的作用下, 熵是一直增加的, 即符合我们的最大熵原理. 而自然界也是一个巨大的系统, 时时刻刻产生信息, 有精确的, 但许多都是模糊的. 为了在已有的知识下, 使我们的结果更加准确, 我们趋向于使得信息的熵最大, 于是我们在 FCM 的基础上融入了最大熵模型, 形成了模糊最大熵模型. 最后我们将模糊最大熵模型应用于 iris 数据集的分类, 取得了相对于 FCM 算法更好的聚类效果.

现如今, 适逢新一代人工智能的浪潮, 各种智能算法、机器学习研究论文层出不穷, 而模糊聚类在图像分割、目标识别、故障诊断等方面也有广泛的应用, 相信在不久的未来, 关于模糊最大熵模型的应用于机器学习的结合会越来越多, 推动模糊分类算法变得越来越好.

参考文献

[1] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.

[2] ZADEH L A. Probability measures of fuzzy events[J]. Journal of mathematical analysis and applications, 1968, 23(2): 421-427.

[3] 李安贵, 张志宏. 模糊数学及其应用 [M]. （第 2 版）. 北京: 冶金工业出版社, 2005.

[4] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(4): 379-423.

[5] LI P, LIU B. Entropy of credibility distributions for fuzzy variables[J]. IEEE Transactions on Fuzzy Systems, 2008, 16(1): 123-129.

[6] DE LUCA A, TERMINI S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory[J]. Information and Control, 1972, 20(4): 301-312.

[7] 谭扬波, 陈光 (禹). 一种基于最大模糊熵的高斯聚类算法 [J]. 电子科技大学学报, 2000 (3): 269-272.

[8] QIAN P, SUN S, JIANG Y, et al. Cross-domain, soft-partition clustering with diversity measure and knowledge reference[J]. Pattern Recognition, 2016, 50.

[9] FU H J, WU X H, MAO H P, et al. Fuzzy entropy clustering using possibilistic approach[J]. Procedia Engineering, 2011, 15: 1993-1997.

[10] LI R P, MUKAIDONO M. A maximum-entropy approach to fuzzy clustering[C]// Proceedings of 1995 IEEE International Conference on Fuzzy Systems. [S.l.: s.n.], 1995.

致谢

谢谢！