基于 Gauss 分布函数的区间值数据的模糊聚类算法

吕泽华,金 海,袁平鹏,邹德清
(华中科技大学计算机学院服务计算技术与系统教育部重点实验室,湖北武汉 430074)

摘 要: 通过分析投票模型中中立者的思想倾向,对区间值数据进行二次特征提取,给出了一种区间值数据的 Gauss 函数表示法,利用这种方法对区间值数据进行相似度量,从而导出一种新的区间值数据的距离度量公式.将该距离度量公式运用于区间值数据的模糊 c 均值聚类算法(FCM)中,得出一种新的基于 Gauss 分布函数的区间值数据的模糊聚类算法,试验表明该方法比传统的区间值数据的模糊聚类算法能获得更好的分类效果.

关键词: 投票模型; Gauss 分布函数; 区间值数据; 相似度量; 模糊聚类
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112 (2010) 02-0295-06

A Fuzzy Clustering Algorithm for Interval-Valued Data Based on Gauss Distribution Functions

LÜ Ze-hua, JIN Hai, YUAN Ping-peng, ZOU De-qing
(*Key Laboratory of Services Computing Technology and System, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei* 430074, *China*)

Abstract: According to analyzing the tendency of neutrals in a vote model, a new kind of similarity measure between interval-valued data based on Gauss distribution functions is proposed and the distance between two interval-valued data is given, and then, a novel fuzzy clustering algorithm for interval-valued data is presented. Examples show that this algorithm can get better performance than other existing methods.

Key words: Gauss distribution functions; interval-valued data; similarity measure; fuzzy clustering

1 引言

聚类分析是数据预处理的一种重要手段,要对某个事物进行研究,就必须首先进行数据采集,获取其特征信息.在数据采集过程中可能获得两种类型的信息:一是来自测量仪器的数值信息,二是来自人类专家的语言信息.语言信息是一种典型的不精确信息,而测量信息由于测量误差和噪声污染等原因使得测量值并不可靠,像约为 10(模糊数)或者 8 到 12 之间(区间数)这样的测量结果十分常见,它们均为不精确数据.为了研究这种不精确数据集的分类问题,人们首先从最基本的数据形式——区间值数据着手:Ishibuchi 提出了基于线性感知器的区间值数据的神经网络分类方法[1];Mandal 提出了可以处理区间数的基于 IF-THEN 规则的分类器[2];文献[11]则给出了一种基于 Gauss 分布函数的区间值数据的相似度量方法.在无监督分类方面,近年许多学者对其进行了深入地讨论[3~8,18],具有代表性的方法是文献[6]利用区间值模糊推理提出了基于模糊逻辑神经元的聚类网络;文献[17]则给出了一种符号数据的模糊聚类算法,文献[19]则讨论了区间值模糊集(Intuitionistic Fuzzy Sets, IFSs)的聚类算法.文献[9]从 c 均值算法角度提出了两种区间值数据的 FCM 算法,称为范方法;高则对范方法进行了进一步推广[10,12],给出了范方法的一般形式,并且提出了两种扩展的 FCM 算法,一种是针对模糊数的 FCM 算法,另外一种是基于特征加权的 FCM 算法;Antonio Irpino 等则提出了一种基于 Wasserstein 距离的区间值数据的动态聚类算法[13];Marie-Helene 等利用信任函数对区间值数据进行聚类[14].

在范方法的两种扩展方法中,方法一是借助区间数的距离度量公式定义聚类目标函数,通过隶属函数和聚类原型之间的迭代来实现聚类,但是由于距离度量形式的差异使得该算法不能直接调用成熟的 FCM 算法工具箱,而且算法的收敛性尚未证明;方法二是先将区间数转化为区间中值,利用传统的 FCM 算法求解隶属函

收稿日期:2008-11-04;修回日期:2009-10-10
基金项目:国家 973 重点基础研究发展计划(No.2003CB317003)

数和区间中值的聚类中心,再通过后处理获得区间型聚类原型,尽管从 FCM 算法的收敛型可以导出范方法二的收敛型,但是由于没有考虑区间大小对分类的影响,使得对具有相同区间中值的区间数均赋予相同的隶属函数,这是不太合理的.高方法[10]不但考虑区间的中值,而且还考虑区间的大小对聚类的影响,首先对区间数据进行二次特征提取,获得区间数 $\bar{x}$ 的区间中值 $\dot{x}$ 和区间大小 $\hat{x}$,把区间数 $\bar{x}$ 投影到 $\dot{x}$ 和 $\hat{x}$ 张成的空间 span($\dot{x}$,$\hat{x}$)中,变成特征空间中的一个点,然后调用传统的 FCM 算法求解,再经过后处理来获得区间值数据的类原型模式,这种方法虽然将区间中值和区间大小都作了考虑,但是将二者割裂开来,没有挖掘出两者之间的内在关联.

本文在对以上几种区间值数据聚类方法分析的基础上提出了一种基于 Gauss 分布函数的区间值数据的模糊聚类算法,通过对投票模型中中立者的倾向性进行分析,利用 Gauss 分布函数对区间值数据进行二次特征提取,得到一个新的特征空间,然后在该特征空间中采用 FCM 算法进行模糊聚类,最后经过后处理获得区间值的原型模式,试验表明用该方法进行聚类可以得到比现存聚类算法更好的分类效果.

2 区间值数据的二次特征提取及距离度量

2.1 区间值数据的 Gauss 分布函数表示方法

令 $I(R^+)=\{\bar{x}\mid\bar{x}=[x^-,x^+]\subset R^+\}$,则 $I(R^+)$中的元素 $\bar{x}\in I(R)$即为区间值数据,简称为区间数,其中 x^-为区间左值,x^+为区间右值,定义 $\bar{x}$ 的区间中值 $\dot{x}$ 和区间大小 $\hat{x}$ 分别为:

$$\dot{x}=\frac{x^-+x^+}{2},\hat{x}=x^+-x^-$$

对于区间数的运算及度量,有如下定义:设两个区间数 $\bar{x}=[x^-,x^+]$,$\bar{y}=[y^-,y^+]$,其加法定义为:

$$\bar{x}+\bar{y}=[x^-+y^-,x^++y^+] \tag{1}$$

减法定义为:

$$\bar{x}-\bar{y}=[x^--y^-,x^+-y^+] \tag{2}$$

常数 $\lambda\in R$ 与区间数 $\bar{x}=[x^-,x^+]$的数乘定义为:

$$\lambda\cdot\bar{x}=\begin{cases}[\lambda x^-,\lambda x^+] & ,\lambda\geqslant 0\\ [\lambda x^+,\lambda x^-] & ,\lambda<0\end{cases} \tag{3}$$

两个区间数 $\bar{x}=[x^-,x^+]$和 $\bar{y}=[y^-,y^+]$的大小比较定义为:

$$\bar{x}\leqslant\bar{y}\quad\text{iff}\quad x^-\leqslant y^-,x^+\leqslant y^+ \tag{4}$$

两个区间数 $\bar{x}=[x^-,x^+]$和 $\bar{y}=[y^-,y^+]$的距离定义为:

Euclid 距离: $D_E^2=(x^--y^-)^2+(x^++y^+)^2$ (5)

Hausdorff 距离: $D_H(\bar{x},\bar{y})=\max(|x^--y^-|,|x^+-y^+|)$ (6)

区间值数据的研究始于对 IFSs 的讨论,而 IFSs 是经典 Fuzzy 集理论之后最具代表性也运用最为广泛的一种 2 型模糊集合.和经典模糊集合相比,IFSs 中元素的隶属度被定义为[0,1]的一个子区间,即元素的隶属度仍然是模糊的,只能确定元素的隶属度在[0,1]的某个子区间内,这一性质使得区间值模糊集合能更准确地刻画各种模糊性信息,一个典型的例子就是能很好的描述存在中立者的投票模型.

但 IFSs 在表达模糊信息时仍然存在缺陷,虽然 IFSs 能表达存在中立者的投票模型,但不能对投票模型中中立者的思想倾向性进行描述.在一次投票中,中立者虽然投了中立票,但是他(她)未必就是一个绝对的中立者,仍然存在投支持票或者反对票的思想倾向,只是这种倾向并不足以让他(她)做出支持或者反对决定的.例如,给出了两个区间值模糊集合 A_1 和 A_2,$A_1=\{[0.1,0.3]/x\}$,$A_2=\{[0.7,0.9]/x\}$,用投票模型来分别进行解释,A_1 表示在投票中有 1 个支持者 2 个中立者 7 个反对者,A_2 表示 7 个支持者 2 个中立者 1 个反对者,虽然这两次投票中均有 2 个中立者,很显然,A_1 中的 2 个中立者会倾向于投反对票,而 A_2 中的 2 个中立者则倾向投赞成票,但这种倾向性在 IFSs 并不能被体现出来.也就是在 IFSs 中只知道元素的隶属度在[0,1]的某个子区间中,而隶属度在子区间内部的分布情况是不确定的,这就是区间值 2 型模糊集在表达模糊信息时所存在的缺陷.对比可以得到区间值数据有类似的性质,例如,在一次测量中得到了区间值数据[8,10],初步判断测量的精确值应该在 8 和 10 之间,但是精确值在区间内部的分布状况是怎样的? 哪个值最有可能作为该次测量的精确值? 甚至,区间之外的值是否也有某种可能是精确值呢? 这都是值得深入探讨的问题.一种直观合理的结果是区间的中值 9 是精确测量结果的可能性最大,区间中其它值随着与中值距离的变大,作为精确值的可能性随之下降.在这种思路的启发下,本文给出一种基于 Gauss 分布函数的区间值数据的二次特征提取方法.

设 $\bar{x}=[x^-,x^+]\in I(R)$,根据 $\bar{x}$ 的中值 $\dot{x}$ 和区间大小 $\hat{x}$ 分别作为 Gauss 分布函数的期望和方差来生成一个 Gauss 分布函数,其生成方法如下:

$\bar{x}$ 所对应的 Gauss 分布函数记为 $\varphi_{\bar{x}}(z)$,则 $\varphi_{\bar{x}}(z)$满足

$$\mu=E(\varphi_{\bar{x}}(z))=\dot{x},\sigma=D(\varphi_{\bar{x}}(z))=\frac{1}{\sqrt{2\pi}(1-\frac{\hat{x}}{2})} \tag{7}$$

随着 $\hat{x}$ 的变大,σ 也变大,对于有相同区间中值不同区间大小的区间值数据,所对应的 Gauss 分布函数示意图如图 1 所示,为了方便比较,图 1 中所给出的三个区间

值数据 $\bar{x}_1,\bar{x}_2,\bar{x}_3$ 满足 $\hat{x}_1=\hat{x}_2=\hat{x}_3,\hat{x}_1>\hat{x}_2>\hat{x}_3$.

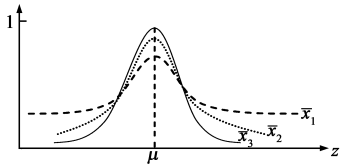


图1　同中值不同区间长度的区间值数据对应的Gauss分布函数图像对比

2.2　基于 Gauss 分布函数的区间值数据的相似度量

迄今已经提出了多种区间值类型数据的距离度量方法,最有代表性的是 Euclid 距离和 Hausdorff 距离,而距离和相似性是相辅相成的.这里给出一种基于 Gauss 分布函数的区间值数据的相似度量方法,设两个区间值数据分别为 $\bar{x}=[x^-,x^+]$和 $\bar{y}=[y^-,y^+]$,其对应的 Gauss 分布函数分别为 $\varphi_x(z),\varphi_y(z)$.

(1)当$\dfrac{(x^-+x^+)}{2}<\dfrac{(y^-+y^+)}{2}$,如图 2($a$)所示,则 $\bar{x}$ 与 $\bar{y}$ 之间的相似度 $S_N(\bar{x},\bar{y})$为:

$$S_N(\bar{x},\bar{y})=\frac{\int_{-\infty}^{t}\varphi_y(z)\mathrm{d}z+\int_{t}^{+\infty}\varphi_x(z)\mathrm{d}z}{\int_{-\infty}^{t}\varphi_x(z)\mathrm{d}z+\int_{t}^{+\infty}\varphi_y(z)\mathrm{d}z} \tag{8}$$

这里 t 是$\varphi_x(z),\varphi_y(z)$图像交点的横坐标.

(2)当$\dfrac{(y^-+y^+)}{2}<\dfrac{(x^-+x^+)}{2}$,如图 2($b$)所示,则 $\bar{x}$ 与 $\bar{y}$ 之间的相似度 $S_N(\bar{x},\bar{y})$为:

$$S_N(\bar{x},\bar{y})=\frac{\int_{-\infty}^{t}\varphi_x(z)\mathrm{d}z+\int_{t}^{+\infty}\varphi_y(z)\mathrm{d}z}{\int_{-\infty}^{t}\varphi_y(z)\mathrm{d}z+\int_{t}^{+\infty}\varphi_x(z)\mathrm{d}z} \tag{9}$$

这里 t 与前相同.

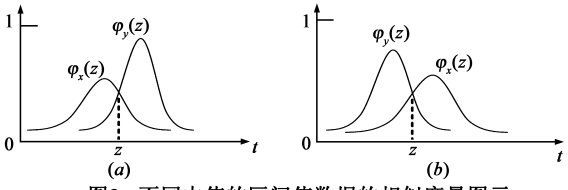


图2　不同中值的区间值数据的相似度量图示

(3)当$\dfrac{(x^-+x^+)}{2}=\dfrac{(y^-+y^+)}{2}$,若 $\hat{x}>\hat{y}$,如图 3(a)所示,交点的横坐标分别为 t_1 和 t_2,则 $\bar{x}$ 与 $\bar{y}$ 之间的相似度 $S_N(\bar{x},\bar{y})$为:

$$S_N(\bar{x},\bar{y})=\frac{\int_{-\infty}^{t_1}\varphi_y(z)\mathrm{d}z+\int_{t_1}^{t_2}\varphi_x(z)\mathrm{d}z+\int_{t_2}^{+\infty}\varphi_y(z)\mathrm{d}z}{\int_{-\infty}^{t_1}\varphi_x(z)\mathrm{d}z+\int_{t_1}^{t_2}\varphi_y(z)\mathrm{d}z+\int_{t_2}^{+\infty}\varphi_x(z)\mathrm{d}z} \tag{10}$$

若 $\hat{x}<\hat{y}$,如图 3(b)所示,则 $\bar{x}$ 与 $\bar{y}$ 之间的相似度 $S_N(\bar{x},\bar{y})$为:

$$S_N(\bar{x},\bar{y})=\frac{\int_{-\infty}^{t_1}\varphi_x(z)\mathrm{d}z+\int_{t_1}^{t_2}\varphi_y(z)\mathrm{d}z+\int_{t_2}^{+\infty}\varphi_x(z)\mathrm{d}z}{\int_{-\infty}^{t_1}\varphi_y(z)\mathrm{d}z+\int_{t_1}^{t_2}\varphi_x(z)\mathrm{d}z+\int_{t_2}^{+\infty}\varphi_y(z)\mathrm{d}z} \tag{11}$$

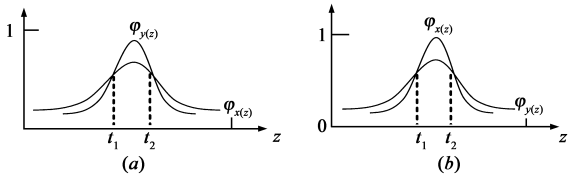


图3　同中值的区间值数据的相似度量

从几何图形上,可以得出两个区间值数据的相似度就是其所对应的两个 Gauss 分布函数曲线和 x 轴所围成的公共区域面积与总面积之比.

2.3　区间值矢量数据的距离度量

由于相似度和距离可以相互定义,因此可以用区间值数据的相似度来定义区间值数据之间的距离.设两个区间值数据分别为 $\bar{x}=[x^-,x^+]$和 $\bar{y}=[y^-,y^+]$,$\bar{x}$ 和 $\bar{y}$ 之间的距离记为 $d_N(\bar{x},\bar{y})$,则 $d_N(\bar{x},\bar{y})$定义为:

$$d_N(\bar{x},\bar{y})=1-S_N(\bar{x},\bar{y}) \tag{12}$$

假设有两个观测样本 $\bar{\boldsymbol{x}}_1=(\bar{x}_{11},\bar{x}_{12},\cdots,\bar{x}_{1s})^{\mathrm{T}},\bar{\boldsymbol{x}}_2=(\bar{x}_{21},\bar{x}_{22},\cdots,\bar{x}_{2s})^{\mathrm{T}}$,其中每个特征矢量 $\bar{x}_{kj}$由区间数来表示,则 $\bar{\boldsymbol{x}}_1$ 和 $\bar{\boldsymbol{x}}_2$ 的距离定义为:

$$D_N(\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_2)=\frac{1}{s}h\sum_{i=1}^{s}d_N(\bar{x}_{1i},\bar{x}_{2i}) \tag{13}$$

其中 $d_N(\bar{x}_{1i},\bar{x}_{2i})$根据(7)~(10)式定义.

3　基于 Gauss 分布函数的区间值数据的 FCM 算法

假设有一组观测样本 $\bar{X}=\{\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n\}$,其中每个样本为一个 s 维的特征矢量,$\bar{x}_k=(\bar{x}_{k1},\bar{x}_{k2},\cdots,\bar{x}_{ks})^{\mathrm{T}}$,每个特征 $\bar{x}_{kj}\in I(R^+)$由区间值数据描述,假定已知这类数据集中存在 c 个自然结构,聚类的目的就是把这 n 个样本按照"物以类聚"的准则把它们划分到 c 个子集中,使得相似的样本尽量归为一类.但是现实的分类往往伴随着模糊性,一个事物是否属于一个子类并不是泾渭分明,有一个程度的问题,定义区间值数据的模糊 c 划分为

$$M_{fc}=\{U\in R^{cn}\mid\mu_{ik}\in[0,1],\ \sum_{i=1}^{c}\mu_{ik}=1,\forall k,0<\sum_{k=1}^{n}\mu_{ik}<n,\forall i\}$$

其中 $\mu_{ik}=\mu_{\bar{X}_i}(\bar{x}_k)$表示第 k 个区间值特征矢量 $\bar{x}_k$ 隶属于第 i 个模糊子集 $\bar{X}_i$ 的程度,用类内加权误差平方和准则定义区间值数据的模糊聚类目标函数为:

$$\begin{cases}J(U,\bar{P})=\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^mS(\bar{x}_k,\bar{p}_i),m\in[1,+\infty)\\ \text{s.t.}\quad U\in M_{fc}\end{cases} \tag{14}$$

式(14)中 $\bar{P}=\{\bar{p}_1,\bar{p}_2,\cdots,\bar{p}_c\}$ 为每个聚类的原型模式，$S(\bar{x}_k,\bar{p}_i)$ 表示样本 $\bar{x}_k$ 与聚类原型 $\bar{p}_i$ 间的距离，下面给出具体算法来优化目标函数，以获得区间值数据的最优模糊划分.

算法：

预处理 将区间值数据集 $\bar{X}$ 中的元素 $\bar{\boldsymbol{x}}_k=(\bar{x}_{k1},\bar{x}_{k2},\cdots,\bar{x}_{kn})^{\mathrm{T}}$ 的每一维特征值采用式(7)的方法进行变换，使得每一维特征值都用一个 Gauss 分布函数来表示；

初始化 设定迭代阈值 ε，初始化原型模式 $\boldsymbol{P}^{(0)}$，设置迭代计数器 $b=0$；

步骤 1 用公式(15)计算和更新分类矩阵 $\boldsymbol{U}^{(b)}$：

$$\mu_{ik}{}^{(b)}=\left\{\sum_{j=1}^{c}\left[\left(\frac{D_N^{(b)}(\boldsymbol{x}_k,\boldsymbol{p}_i)}{D_N^{(b)}(\boldsymbol{x}_j,p_i)}\right)^{\frac{2}{m-1}}\right]\right\}^{-1} \tag{15a}$$

如果 $\exists\, i,r$，使得 $D_N^{(b)}(x_r,p_i)=0$，则有

$$\mu_{ir}^{(b)}=1,\text{且对于 } j\neq r,\mu_{ij}^{(k)}=0 \tag{15b}$$

这里 $D_N^{(b)}(x_k,p_i)$ 根据式(13)来定义.

步骤 2 用公式(16)更新聚类原型模式矩阵 $\boldsymbol{P}^{(b+1)}$

$$\boldsymbol{p}_i^{(b+1)}=\frac{\sum_{k=1}^{n}(\mu_{ik}^{(b)})\cdot\boldsymbol{x}_k}{\sum_{k=1}^{n}(\mu_{ik}^{(b)})^m},\ i=1,2,\cdots,c \tag{16}$$

步骤 3 如果 $D_N(\boldsymbol{P}^{(b)},\boldsymbol{P}^{(b+1)})<\varepsilon$，则算法停止并输出划分矩阵 $\boldsymbol{U}^*$ 和最佳聚类原型 $\boldsymbol{P}^*$，否则令 $b=b+1$，转向步骤 1；

后处理：根据式(17)获得区间值聚类原型：

$$\boldsymbol{p}_i^{\pm}=\boldsymbol{\mu}_i\pm\frac{\boldsymbol{\sigma}_i}{2} \tag{17}$$

4 试验结果和分析

为了便于比较，先定义两个性能参数：平均失真度和平均相似度.

定义 1 平均失真度定义为样本相对于聚类原型间的平均距离，即

$$\varepsilon=\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ki}\cdot D_N(\bar{x}_i,\bar{p}_k) \tag{18}$$

定义 2 平均相似度定义为样本与聚类原型间的平均模糊相似的程度，即

$$\eta=\sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ki}\cdot S_N(\bar{x}_i,\bar{p}_k) \tag{19}$$

试验 1

本试验是采用一组实际数据作为测试集，表 1 所示的为一组关于 Fat-Oil 的实际数据[9]，包含 8 个四维特征矢量，各维特征的取值均为区间值数据. 首先对表 1 中的数据进行归一化处理，将每一维数据映射到单位区间中，归一化方法如下：

表 1 有关 Fat-Oil 的实测数据集

样本	比重(g/cm³)	冰点(℃)	io 值	sa 值
亚麻油	0.930～0.935	－27～－8	170～240	118～196
紫苏油	0.930～0.937	－5～－4	192～208	188～197
棉籽油	0.916～0.18	－6～－1	99～113	189～198
芝麻油	0.920～0.926	－6～－4	104～116	187～193
山茶油	0.916～0.917	－21～－15	80～82	189～193
橄榄油	0.914～0.919	0～6	79～90	187～196
牛 油	0.860～0.870	30～38	40～48	190～199
猪 油	0.858～0.864	20～－32	53～77	190～204

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{20}$$

其中，$x_{\max}$ 是所有区间中最大的区间右值，$x_{\min}$ 是所有区间中最小的区间左值，x 为待归一化的区间的左值或者右值，x' 为归一化后的结果. 表 1 归一化后的结果见表 2.

表 2 归一化后的数据集

	1	2	3	4
亚麻油	0.911～0.975	0.000～0.292	0.774～0.976	0.000～0.929
紫苏油	0.911～1.000	0.338～0.354	0.893～1.000	0.833～0.940
棉籽油	0.732～0.759	0.323～0.400	0.351～0.435	0.845～0.952
芝麻油	0.785～0.861	0.323～0.354	0.381～0.452	0.821～0.983
山茶油	0.734～0.747	0.092～0.185	0.238～0.230	0.845～0.893
橄榄油	0.749～0.772	0.415～0.508	0.232～0.298	0.232～0.298
牛 油	0.025～0.152	0.877～1.000	0.000～0.048	0.857～0.964
猪 油	0.000～0.076	0.754～0.908	0.077～0.220	0.857～1.000

聚类性能比较

(1)采用基于神经网络的区间值聚类算法所得到的聚类中心如表 3 所示. 平均距离 ε_1 和平均相似度 η_1 分别为：$\varepsilon_1=0.261$，$\eta_1=0.795$.

表 3 区间值神经网络所得到的聚类中心

	1	2	3	4
V_0	0.912～0.980	0.111～0.313	0.805～0.974	0.278～0.932
V_1	0.725～0.775	0.313～0.394	0.281～0.337	0.830～0.917
V_2	0.015～0.109	0.794～0.936	0.050～0.159	0.857～0.986

(2)将区间值转化为区间中值，然后直接调用传统的 FCM 算法，所得到的聚类中心如表 4 所示.

表 4 范算法 1 所得的聚类中心

	1	2	3	4
V_0	0.910～0.981	0.104～0.311	0.807～0.979	0.259～0.932
V_1	0.755～0.792	0.296～0.367	0.321～0.379	0.832～0.942
V_2	0.016～0.117	0.813～0.951	0.041～0.136	0.857～0.982

平均距离 ε_2 和平均相似度 η_2 分别为：$\varepsilon_2=0.254$，$\eta_2=0.797$.

(3)综合考虑区间中值和区间长度对聚类的综合影响，提取区间数 $\bar{\boldsymbol{x}}$ 的中值 $\dot{\boldsymbol{x}}$ 和区间的大小 $\hat{\boldsymbol{x}}$，把区间数 $\bar{\boldsymbol{x}}$ 投影到 $\dot{\boldsymbol{x}}$ 和 $\hat{\boldsymbol{x}}$ 张成的空间 $span(\dot{\boldsymbol{x}},\hat{\boldsymbol{x}})$，变成特种空间中的一个点，即，即范方法 2，所得到的聚类中心如表

5 所示,平均距离 ε_3 和平均模糊相似程度 η_3 分别为:$\varepsilon_3 = 0.258$, $\eta_2 = 0.796$.

表 5　范算法 2 所得的聚类中心

	1	2	3	4
V_0	0.910 ~ 0.985	0.158 ~ 0.321	0.826 ~ 0.983	0.392 ~ 0.934
V_1	0.752 ~ 0.788	0.295 ~ 0.367	0.309 ~ 0.365	0.831 ~ 0.942
V_2	0.014 ~ 0.116	0.814 ~ 0.953	0.039 ~ 0.153	0.857 ~ 0.982

(4)采用本文基于 Gauss 分布函数的转化方法对区间值数据进行预处理,然后调用传统 FCM 算法所得到的聚类中心如表 6 所示:

表 6　归一化处理后得的聚类中心

	1	2	3	4
V_0	0.911 ~ 0.985	0.159 ~ 0.311	0.836 ~ 0.979	0.493 ~ 0.917
V_1	0.754 ~ 0.778	0.289 ~ 0.354	0.310 ~ 0.364	0.847 ~ 0.937
V_2	0.012 ~ 0.115	0.823 ~ 0.953	0.042 ~ 0.132	0.857 ~ 0.976

其平均距离 ε 和平均模糊相似度 η 分别为:$\varepsilon = 0.248$, $\eta = 0.801$.

通过以上的性能指标的比较发现基于 Gauss 分布函数的聚类算法能得到更好的聚类效果.

(5)上述的 4 种处理方法中,都必须对数据进行归一化处理,但是归一化处理会对数据集的结构产生影响,特别是像冰点这样的数据特征,将一个包含了 0 点的区间也映射到单位子区间,必然对数据集的本来性质和数据结构带来改变.而本文所提出的聚类方法对区间值数据的聚类是不须要进行归一化的,下面对 Fat-Oil 数据集不进行归一化,直接进行 Gauss 变换后作聚类,表 7 给出的是该处理方法下所得出的聚类中心.平均距离和平均模糊相似度分别为:$\varepsilon = 0.231$, $\eta = 0.856$

表 7　不归一化处理得出的聚类中心

	1	2	3	4
V_0	0.930 ~ 0.936	- 16.80 ~ 5.35	178.77 ~ 105.15	150.93 ~ 196.46
V_1	0.917 ~ 0.920	- 7.826 ~ - 3.15	91.92 ~ 100.32	187.81 ~ 196.96
V_2	0.859 ~ 0.867	25.91 ~ 34.95	46.55 ~ 62.68	189.99 ~ 200.32

试验 2

本试验的的目的在于检验本文所提出的聚类算法的分类性能.试验用著名的 IRIS 数据集作为测试数据,已知该数据集中包含三个 IRIS 种类 Setosa, Versicolor 和 Virginica,每类有 50 个样本,其中 Setosa 与其他两个类间完全分离, Versicolor 和 Virginica 间有交叉.分别用传统 FCM 算法,基于特征加权的 FCM 算法[9]和本文提出的算法对 IRIS 样本分类处理,比较这三种算法的样本误分率和类内误差平方和等指标,比较算法的性能.IRIS 数据经常被用作检验聚类算法的分类性能的标准测试数据, Hathaway 给出的 IRIS 数据的实际中心位置为[16]:

$$p_1 = (5.00, 3.42, 1.46, 0.24)$$
$$p_2 = (5.93, 2.77, 4.26, 1.32)$$
$$p_3 = (6.58, 2.97, 5.55, 2.02)$$

表 8　三种 FCM 算法的分类性能的比较

聚类算法	误分数	误分率	聚类原型短矢量	误差平方和
传统的 FCM	10.67	10.67%	$p_1 = (5.0062, 3.4242, 1.4684, 0.2492)$ $p_2 = (5.0946, 2.7460, 4.4154, 1.4273)$ $p_3 = (6.8484, 3.0750, 5.7283, 2.0741)$	0.1554
特征加权 FCM	7	4.67%	$p_1 = (5.0060, 3.4278, 1.4624, 0.2461)$ $p_2 = (5.9378, 2.7450, 4.3438, 1.3315)$ $p_3 = (6.6274, 3.0151, 5.5673, 2.0642)$	0.0145
基于 Gauss 分布的 FCM	4	2.67%	$p_1 = (5.0045, 3.4235, 1.4627, 0.2458)$ $p_2 = (5.9341, 2.7630, 4.3302, 1.3401)$ $p_3 = (6.6034, 2.981131, 5.5608, 2.0541)$	0.0098

分别用传统 FCM 算法、基于特征加权的 FCM 算法以及本文所提出的基于 Gauss 分布函数的 FCM 算法对 IRIS 数据进行分类,对模糊划分矩阵进行去模糊处理,也就是把样本硬划分到最大隶属度值所对应的类中,得到的分类结果如表 8 所示.为了比较试验的结果,在该试验中,算法的参数均为:加权指数 $m = 1.25$,分类数目 $c = 3$,迭代阈值 $\varepsilon = 10^{-6}$,对特征加权的 FCM 算法还有 $\gamma = 1$, $G_{\max} = 1000$.从表 8 中可以看出基于 Gauss 分布函数的区间值数据的 FCM 算法不但误分率小,而且得到的聚类原型模式更接近实际中心位置,误差平方和较小.

5　结论

本文通过分析投票模型中中立者的思想倾向性问题,给出了一种区间值数据的 Gauss 分布函数表示方法,根据两个 Gauss 分布函数在图像上的重叠比得出区间值数据的相似度量方法,然后通过该相似度给出了区间值数据的距离度量,改进现有的基于区间值数据的 FCM 算法,基于 Gauss 分布函数的区间值数据的模糊聚类算法可以较好地处理区间值数据的聚类分析问题,并用标准的 IRIS 数据进行聚类测试,实验结果表明与传统区间值数据的 FCM 算法相比,基于 Gauss 分布函数的区间值数据的模糊聚类算法在误分率和误差平方和等技术指标上均显示了良好的分类效果.

参考文献:

[1] Ishibuchi H, Nozaki N, Tanaka H. Efficient fuzzy partition of pattern space for classification problems[J]. FSS, 1993, 59(3): 295 - 304.

[2] Mandal P. Partitioning of feature space for pattern classification [J]. Pattern Recognition, 1997, 12(30): 1971 - 1990.

[3] 孟丹. 基于区间值的模糊聚类分析[J]. 辽宁师范大学学报(自然科学版), 2003, 26(2): 113 - 116.

D. Meng. Fuzzy aggregation analysis method based on interval value[J], Journal of Liaoning Normal University (Natural Science Edition), 2003, 26(2): 113 – 116. (in Chinese)

[4] 陆建江,徐宝文.区间数据的并行模糊聚类算法[J].东南大学学报,2003,33(4):406 – 409.
J Lu, B Xu. Parallel fuzzy clustering algorithm for interval data [J]. Journal of Southeast University (Natural Science Edition), 2003, 33(4): 406 – 409. (in Chinese)

[5] Xu E, Liangshan Shao, Wendong Tan. A clustering algorithm based on discretized interval value[A]. IMACS Multiconference on "Computational Engineering in Systems Applications" (CESA)[C]. October 4 – 6, 2006, Beijing, China, 864 – 868.

[6] 李文华.模糊聚类新算法与模糊聚类神经网络[D].西安:西安电子科技大学,1995.

[7] 李洁,高新波,焦李成.基于特征加权的模糊聚类新算法[J],电子学报,2006,34(1):89 – 92.
Li J, Gao X, Jiao L. A new feature weighted fuzzy clustering algorithm. Acta Electronica Sinica, 2006, 34(1): 89 – 92. (in Chinese)

[8] 董红斌,黄厚宽,周成,何军,尚文倩.基于模糊权和有效性函数的演化聚类算法[J],电子学报,2007,35(5):964 – 970.
Dong H, Huang H, Zhou C, He J, Shang W. A fuzzy weighted sum validity function for clustering with a mixed strategy evolutionary algorithm[J]. Acta Electronica Sinica, 2007, 35(5): 964 – 970. (in Chinese)

[9] 范九伦.模糊聚类新算法和聚类有效性问题研究[D].西安:西安电子科技大学,1998.

[10] 高新波.模糊聚类算法的优化及应用研究[M].西安:西安电子科技大学,1999.

[11] Z Lv, C Chen, W Li. A new method for measuring similarity between intuitionistic fuzzy sets based on normal distribution functions[C]. Fourth international conference on Fuzzy systems and knowledge discovery, Haikou, China, 2007, (2): 108 – 113.

[12] 高新波.模糊聚类分析及其应用[M],西安,西安电子科技大学出版社,2004.

[13] Antonio Irpino, Rosanna Verde. Dynamic clustering of interval data using a Wasserstein-based distance[J]. Pattern Recognition Letters, 2008, 29(11): 1648 – 1658.

[14] Marie-Helene Masson, Thierry Denoeux. Clustering interval-valued proximity data using belief functions [J]. Pattern Recognition Letters, 2004, 25(2): 163 – 171.

[15] Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms[C]. Plenum Press, New York, 1981, 176 – 185.

[16] Hathaway R J, Bezdek J C. Nerf c-means: non-Euclidean relation fuzzy clustering[J]. Pattern Recognition, 1994, 27(3): 429 – 437.

[17] Francisco de A. T., de Carvalho. Fuzzy c-means clustering methods for symbolic interval data[J]. Pattern Recognitions Letters, 2007, 28(4): 423 – 437.

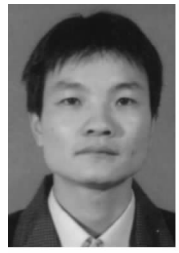
[18] 武小红,周建江.可能性模糊 c-均值聚类新算法[J].电子学报,2008,36(10):1996 – 2000.
Wu X, Zhou J. A novel possibilistic fuzzy c-means clustering [J]. Acta Electronica Sinica, 2008, 36(10): 1996 – 2000. (in Chinese)

[19] Z Xu, J Chen, J Wu. Clustering algorithm for intuitionistic fuzzy sets[J]. Information Sciences, 2008, 178(19): 3775 – 3790.

作者简介:

吕泽华 男,1976 年生,博士后,主要研究方向为近似计算、模式识别、语义网.
E-mail: lzhhust@gmail.com

金 海 男,1966 年生,博士,教授,博士生导师,华中科技大学计算机学院院长,华中科技大学"服务计算技术与系统"教育部重点实验室&"集群与网格计算"湖北省重点实验室主任.主要研究方向为虚拟化技术、网格计算、集群计算、网络安全、分布式存储等.



袁平鹏 男,1972 年生,博士,副教授,主要从事语义网与知识管理、分布式计算技术及系统等方面的研究.

邹德清 男,1975 年生,博士,副教授,华中科技大学计算机学院信息安全系副主任,研究兴趣包括系统安全、可信计算、虚拟化技术及集群/网格计算.