基于模糊聚类和遗传算法的具备解释性和精确性的模糊分类系统设计

邢宗义[1],张 永[1],侯远龙[1],贾利民[2]
(1. 南京理工大学机械学院,江苏南京 210094; 2. 北京交通大学交通运输学院,北京 100044)

摘 要: 提出一种基于模糊聚类和遗传算法的模糊分类系统的设计方法. 首先定义了模糊分类系统的精确性指标,给出解释性的必要条件. 然后利用聚类有效性分析确定模糊规则数目,利用模糊聚类算法辨识初始的模糊分类系统. 随后利用模糊集合相似性分析与融合对初始的模糊分类系统进行约简,提高其解释性;利用遗传算法对约简后的模糊分类系统进行优化,提高其精确性,该过程反复迭代直至满足中止条件. 最后利用该方法进行 Iris数据样本分类,仿真结果验证了该方法的有效性.

关键词: 模糊分类系统; 模糊聚类; 遗传算法; 解释性; 精确性
中图分类号: TP273 **文献标识码**: A **文章编号**: 0372-2112 (2006) 01-0083-06

Design of Interpretable and Precise Fuzzy Classification System Based on Fuzzy Clustering and Genetic Algorithm

XING Zong-yi[1], ZHANG Yong[1], HOU Yuan-long[1], JIA Li-min[2]
(1. *School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu* 210094, *china*;
2. *School of Traffic and Transportation, Beijing Jiaotong University, Beijing* 100044, *china*)

Abstract: An approach of constructing interpretable and precise fuzzy classification system based on fuzzy clustering and genetic algorithm is proposed. First, the precision index is defined, and the necessary conditions of interpretability are analyzed. Second, the number of fuzzy rules is determined by cluster validity indices, and the initial fuzzy classification system is identified using a fuzzy clustering algorithm. Subsequently, the method of merging similar fuzzy sets is used to enhance the interpretability of the initial model. A genetic algorithm is used to improve the precision of the model. The process continues iteratively until the stop criteria are satisfied. The proposed approach is applied to the Iris benchmark classification problem, and the results show its validity.

Key words: fuzzy classification system; fuzzy clustering; genetic algorithm; interpretability; precision

1 引言

由于具备不确定或模糊信息的处理能力,并能融合专家经验,模糊理论在分类问题上得到了广泛的应用[1]. 模糊分类系统可以由专家根据经验构造,但在很多情况下这种经验并不存在或不完备,而相关数据却相对容易获得,因此,如何从数据中自动构造模糊分类系统,在近年来成为研究的热点,其主要研究方法可归纳为以下三种:基于模糊聚类的方法[2]、基于遗传算法的方法[3]、和基于模糊神经网络的方法[4].

上述方法仅仅利用了模糊分类系统的万能逼近器功能,追求精确性指标,而忽略了模糊系统的另外一个重要特性,即解释性. 而实际上,模糊分类系统的解释性,是模糊分类系统区别于神经网络等其他分类系统的最重要特性. 在诸如医学、金融等领域,解释性有时是构建分类系统时的首要目标. 因此,近年来,诸多学者对如何提高模糊分类系统的解释性进行了研究.

文献[5,6]给出了模糊系统解释性的一些必要条件. 文献[7]利用遗传算法,以最大精确性、最少模糊规则和隶属函数数目为进化目标,在栅格初始化的模糊规则库中选择模糊规则,得到包含少量语义解释性的模糊规则. 文献[8]利用决策树初始化模糊分类系统,然后利用相似性分析和多目标优化来提高分类系统的精确性和解释性. 文献[9]利用 VISIT算法产生模糊规则和隶属函数,利用进化算法提取最优模糊规则,为避免动态权值的进化多目标优化的缺点,采用模糊专家系统作为进化算法的目标函数.

本文提出一种基于数据的模糊分类系统的设计方法,

收稿日期: 2005-02-28;修回日期: 2005-07-25
基金项目:国家自然科学基金 (No. 60332020);南京理工大学科研发展基金资助计划项目 (2005)

同时考虑精确性和解释性. 首先利用聚类有效性函数确定规则数目,采用基于模糊聚类的方法辨识初始的模糊分类系统. 然后利用模糊集合相似性分析与融合对初始的模糊分类系统进行约简,提高其解释性;采用遗传算法对约简后的系统进行优化,提高其精确性,该过程迭代进行直至满足中止条件. Iris数据分类的仿真,验证了该方法的有效性.

2 预备知识

2.1 模糊分类系统

考虑 n维 M类 N样本的分类问题,其中 $x\in X\subseteq R^n$, $x=(x_1, x_2, \cdots, x_n)$为特征变量, $\{C_1, C_2, \cdots, C_M\}$为输出类的标号值,则典型的模糊分类系统规则形式如下:

$$\begin{aligned} R_i:&\ \text{if } x_1 \text{ is } A_{i1} \text{ and } x_2 \text{ is } A_{i2} \text{ and} \cdots \text{and } x_n \text{ is } A_{in}\\ &\text{Then the pattern}(x_1, x_2, \cdots, x_n) \text{ belongs to}\\ &\text{class } C_i \text{ with } CF=CF_i \end{aligned} \tag{1}$$

其中 CF_i为第 i条规则的置信度, $A_{i1}, \cdots, A_{in}$为定义在特征论域中的隶属函数,可以取三角形、高斯型、梯形或者钟型等. 本文采用高斯型隶属函数:

$$A_{ij}(x_{jk})=\exp\left\{-\frac{1}{2}\frac{(x_{jk}-v_{ij})^2}{\sigma_{ij}^2}\right\} \tag{2}$$

其中 v_{ij}, σ_{ij}分别代表函数的中心和方差.

对未知样本 x_k,模糊分类系统的输出采用"赢家通吃"的策略,即系统的输出为具备最大激励强度的规则所对应类的标号值:

$$x_k\in C_{i^*},\ i^*=\arg(\max(\beta_i(x_k))) \quad 1\le i\le M \tag{3}$$

其中 β_i为第 i条规则的激励强度:

$$\beta_i(x_k)=CF_i\prod_{j=1}^{n}A_{ij}(x_{jk}) \tag{4}$$

2.2 精确性与解释性

给定特征变量 x_k,其分类误差定义为:

$$e_k=\begin{cases}1 & \text{如果 } x_k \text{ 被正确分类}\\ 0 & \text{如果 } x_k \text{ 被错误分类}\end{cases} \tag{5}$$

则定义衡量模糊分类系统的精确性的指标为:

$$J=\frac{1}{N}\sum_{k=1}^{N}e_k \tag{6}$$

与精确性等可以量化的性能指标不同,模糊分类系统的解释性,目前尚无明确的标准和定义,但是一般认为,模糊系统的解释性,与模型结构、特征变量和模糊规则数目、隶属函数特性等密切相关,现将主要因素陈述如下[5, 6]:

①模糊分类系统的特征变量和模糊规则数目越多,其解释性越低,因此模糊分类系统应该采用尽可能少的特征变量和模糊规则. ②隶属函数必须是凸的;隶属函数划分必须是完备和可区分的的. ③模糊规则库应该是完整的、一致的和精简的.

本文在构造模糊分类系统时,在解释性上主要考虑隶属函数特性,使得到的模糊分类系统容易赋予语义项,并同时通过模糊集合的融合来尽可能地剔除冗余的特征变量和模糊规则.

3 解释性与精确性的模糊分类系统

3.1 聚类有效性分析

模糊规则数目的确定,即在模糊聚类中确定聚类的数目,是构建模糊分类系统的一个首要问题. 聚类有效性分析就是寻找最优的聚类数目,使得数据划分贴近实际情况,聚类目标函数最小. 聚类有效性分析一般通过聚类有效性函数来实现.

聚类有效性函数主要分为两类:第一类仅仅利用了模糊划分矩阵的隶属函数信息;第二类同时利用了隶属函数信息和数据本身的信息.

文献[11]中提出的划分系数函数(*PC*: Partition Coefficient)和划分熵函数(*PE*: Partition Entropy)是第一类聚类有效性函数的典型代表:

$$PC(c)=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^2 \tag{7}$$

$$PE(c)=-\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}\log_a\mu_{ik} \tag{8}$$

其中 $PC(c)\in[1/c, 1]$, $PE(c)\in[0, \log_a c]$. 随着聚类数目的增大, *PC*和 *PE*分别呈现减小或增大的趋势,一般取其第一显著拐点的对应值作为最优聚类数目.

上述聚类有效性函数对模糊指数 m的鲁棒性差,当 $m\to 1$时,不同聚类数目对应的函数值均相同,当 $m\to\infty$时,函数都得到最优聚类数目为 2的结果,从而无法判断最优的聚类数目.

文献[12]提出的紧密/分离性函数(*XB*: Xie-Beni index)属于第二类聚类有效性函数,其最小值对应最优的聚类数目,对模糊指数 m的鲁棒性强,融合了数据的几何结构信息,从而得到了广泛的应用:

$$XB(c)=\frac{\sum_{i=1}^{c}\sum_{k=1}^{N}\mu_{ik}^m\|x_k-v_i\|^2}{n\cdot\min_{i,k}\|v_i-v_k\|^2} \tag{9}$$

单独的聚类有效性函数不一定能给出准确的聚类数目,往往需要综合多个聚类有效性函数的结果,确定最优的聚类数目. 因此本文除采用上述函数外,尚采用如下聚类有效性函数:文献[13]提出的非模糊指数(*NFI*: Non-Fuzzy Index),文献[14]提出的最小硬趋势(*MinHT*: Min Hard Tendency)、平均硬趋势(*MeHT*: Mean Hard Tendency),文献[15]提出的模糊超容积函数(*FHV*: Fuzzy Hyper Volume)、平均划分密度(*PA*: Average Partition Density)、划分密度指数(*PD*: Partition Density Index),文献[16]提出的可能性分布划分(*PP*: Possibility Partition). 其中 *NFI*、*MinHT*、*MeHT*、*DPA*和 *PD*的最大值对应最优的模糊聚类数目, *XB*、*FHV*和 *PP*的最小值对应最优的模糊聚类数目.

本文采用上述 10个聚类有效性函数确定最优的聚类

数目.如果结果不一致,则根据"从众"策略,选取对应最多有效性函数的结果为最终解.

3.2 基于聚类的模糊分类系统

目标函数的模糊聚类是基于数据模糊建模最常用的方法之一.本文采用 Gustafson-Kessel (GK)聚类算法离线辨识模糊分类系统的前件参数[10].

GK聚类算法的目标函数为

$$J(Z;U,V)=\sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^{m}D_{ik}^{2} \tag{10}$$

其中 $Z=[z_1,z_2,\cdots,z_N]$为数据集,$U=[\mu_{ik}]_{c\times N}$为数据集的模糊划分矩阵,$V=[v_1,v_2,\cdots,v_c]$是聚类中心,即隶属函数的中心,c是聚类数目,N是样本数目,m是模糊指数,D_{ik}是第 i个聚类和第 k个数据间的距离范数,μ_{ik}是第 k个数据相对于第 i个聚类中心的隶属度,且满足以下条件

$$\mu_{ik}\in[0,1];\ \sum_{i=1}^{C}\mu_{ik}=1;\ 0<\sum_{k=1}^{N}\mu_{ik}<N \tag{11}$$

第 i个聚类和第 k个数据间的距离范数:

$$D_{ik}^{2}=\|z_k-v_i\|_{A_i}^{2}=(z_k-v_i)^{\mathrm{T}}A_i(z_k-v_i) \tag{12}$$

其中

$$A_i=(\rho\det(F_i))^{1/n}F^{-1} \tag{13}$$

$$\det(A_i)=\rho \tag{14}$$

F_i是模糊协方差矩阵:

$$F_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^{m}(z_k-v_i)(z_k-v_i)^{\mathrm{T}}}{\sum_{k=1}^{N}(\mu_{ik})^{m}} \tag{15}$$

利用拉格朗日乘法可以求得使目标函数最小的必要条件为:

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{c}(D_{ik}/D_{jk})^{2/(m-1)}} \tag{16}$$

$$v_i=\frac{\sum_{k=1}^{N}(\mu_{ik})^{m}z_k}{\sum_{k=1}^{N}(\mu_{ik})^{m}} \tag{17}$$

高斯型隶属函数的方差为:

$$\sigma_{ij}^{2}=\frac{\sum_{k=1}^{N}(\mu_{ik})^{m}E_k}{\sum_{k=1}^{N}(\mu_{ik})^{m}} \tag{18}$$

为确定规则的后件类的标号值,首先定义如下函数:

$$M_{ij}=\frac{\sum_{k=1}^{N}u_{ik}f_j(k)}{\sum_{k=1}^{N}f_j(k)} \tag{19}$$

其中

$$f_j(k)=\begin{cases}1 & \text{如果 } x_k\in C_j\\ 0 & \text{如果 } x_k\notin C_j\end{cases} \tag{20}$$

则对于第 i条规则,其后件类的标号值 C_{i^*}为:

$$i^{*}=\arg(\max(M_{ij}))\ j=1,2,\cdots,M \tag{21}$$

规则置信度为:

$$CF_i=\max_j(M_{ij}) \tag{22}$$

3.3 模糊集合的相似性分析与融合

经过模糊聚类得到的初始的模糊分类系统,其隶属函数(模糊集合)可能存在冗余,表现为模糊集合间存在过度的交叉或重叠,从而难以赋予相应的语义值,降低了系统的解释性,因此需要对每个变量的隶属函数进行相似性分析和融合,从而实现模糊分类系统的约简.

模糊系统的隶属函数存在三种类型的冗余:第一种是两个模糊集合相似,重叠区域过大,这是最常见的模糊集合冗余形式;第二种是模糊集合以较大值覆盖整个论域;第三种是模糊集合接近于单点集合.

第二种和第三种的隶属函数冗余,由于其不存在解释性的实际意义,在满足精度的前提下,一般在对应的规则前件中直接去除即可.对于第一种冗余,本文采用相似性测度来评判两个隶属函数的相似性.

对模糊集合 A和 B,定义相似性测度如下[17]:

$$S(A,B)=\frac{|A\cap B|}{A\cup B} \tag{23}$$

其中$|\cdot|$表示集合的基数,$\cap$和$\cup$算子分别表示集合的交和并.

对于离散论域 $X=\{x_k\mid k=1,\cdots,N\}$,式(23)表述如下:

$$S(A,B)=\frac{\sum_{k=1}^{N}[\mu_A(x_k)\wedge\mu_B(x_k)]}{\sum_{k=1}^{N}[\mu_A(x_k)\vee\mu_B(x_k)]} \tag{24}$$

其中$\wedge$和$\vee$分别为最小最大算子.S为定义在$[0,1]$间的相似性测度,$S=1$表示两个集合完全相等,而 $S=0$意味着两个集合没有交叉或重叠.

如果两个模糊集合 A和 B的相似性测度大于预先设定的阈值,那么集合 A和 B可以融合为新的集合 C.对于本文所采用的高斯型隶属函数,由集合 A和 B融合生成的新集合 C的参数如下:

$$\begin{cases}v_c=(v_A+v_B)/2\\ \sigma_c=\sqrt{\sigma_A^{2}+\sigma_B^{2}/2}\end{cases} \tag{25}$$

阈值的大小直接影响模糊分类系统的性能,阈值越小,得到的分类系统的精度越低而解释性越高,一般阈值取$[0.4\sim0.7]$.

模糊集合融合过程需要反复迭代进行.在每一次迭代过程中,对每一个变量的相邻模糊集合进行两两相似性分析,相似性测度大于阈值的两个模糊集合融合为新的集合.迭代反复进行,直到没有任何两个模糊集合的相似性测度大于阈值,然后再将第二类和第三类模糊集合删除,从而完成整个模糊集合的相似性融合过程.

在模糊集合融合过程中,如果某特征变量的隶属函数经过融合后,仅表示为一个模糊集合,则在不影响精确性的前提下,可以在模糊分类系统中删除该变量.如果某规则的所有隶属函数,均可以同其他规则的模糊集合相融

合,则该规则同样可以删除. 特征变量和模糊规则的删除,均可以提高模糊分类系统的解释性.

如果对模糊分类系统的精确性要求较高,则在进行模糊集合的相似性分析和融合时,需要分析融合前和融合后模糊分类系统精确性的变化. 如果某两个模糊集合的融合严重恶化了精确性,则该两个模糊集合不能进行融合,即使其相似性测度大于预先设定的阈值.

3.4 遗传算法优化

利用遗传算法优化模糊分类系统,就是对模糊分类系统的参数,包括前件参数和后件置信度,编码成染色体,然后模拟自然界的进化过程,对染色体进行选择、交叉和变异操作,使染色体不断进化,最终产生代表问题最优解的染色体,再经过反编码得到优化的模糊分类系统.

3.4.1 **染色体编码** 相对于二进制编码,实数编码减轻了遗传算法的计算负担,提高了运算效率,能够更好地保持种群多样性,并且编码方式直接自然,因此本文采用实数编码方式.

对于式(1)所示的模糊分类系统,待编码的变量为隶属函数的中心 v_{ij}和方差 σ_{ij},以及模糊规则的置信度 CF_i,因此每条染色体共有 $c(2n+1)$个实数(以未进行约简的模糊分类系统为例,对约简后的模糊分类系统,在编码中直接删除对应的实数即可),其编码方式为:

v_{11}	…	v_{cn}	σ_{11}	…	σ_{cn}	CF_1	…	CF_c

给定种群大小为 L,染色体为 H_p, $p=1, 2, \cdots, L$,将得到的模糊分类系统编码为第一条染色体,即

$$H_1 = (v_{11}, \cdots, v_{cn}, \sigma_{11}, \cdots, \sigma_{cn}, CF_1, \cdots, CF_c) \quad (26)$$

给定搜索空间 $[H^{\min}, H^{\max}]$:

$$H^{\min} = (v_{11}^{\min}, \cdots, v_{cn}^{\min}, \sigma_{11}^{\min}, \cdots, \sigma_{cn}^{\min}, 0, \cdots, 0) \quad (27)$$

$$H^{\max} = (v_{11}^{\max}, \cdots, v_{cn}^{\max}, \sigma_{11}^{\max}, \cdots, \sigma_{cn}^{\max}, 1, \cdots, 1) \quad (28)$$

其中 $v_{ij}^{\max}$、$v_{ij}^{\min}$、$\sigma_{ij}^{\min}$、$\sigma_{ij}^{\max}$为对应隶属函数的中心和方差的最大最小值. 以染色体 H_1 为中心,在搜索空间内随机生成其余 $L-1$的个染色体,从而形成初始种群.

3.4.2 **适应度函数** 由于采用了实数编码方式,因此直接采用优化目标作为适应度函数,即采用式(6)做为适应度函数:

$$Fit = \frac{1}{N}\sum_{k=1}^{N} e_k \quad (29)$$

3.4.3 **遗传操作** 遗传算法中主要包括以下三种遗传操作:选择、交叉和变异. 为保证群体多样性和算法的有效性,每一种遗传操作均给出几种具体的实现方法,在实际的操作过程中,由算法随机选择.

选择操作采用轮盘赌选择法. 为防止最优个体在选择操作时被忽略,本文同时采用最优个体保存法. 交叉操作随机采用离散交叉和算术交叉. 变异操作随机采用均匀变异和高斯变异.

3.4.4 **遗传算法步骤** 给定种群大小 L,进化代数 T,交叉概率 P_c,变异概率 P_m,标记第 t代的种群为 P_t,则遗传算法的详细步骤如下:

(1)将模糊分类系统编码为染色体,并在搜索空间内形成初始种群 P_t, $t=0$

(2)计算染色体的适应度值

(3)选择 $L/4$对染色体进行交叉操作,形成 $L/2$个新的染色体,记为 P_t'

(4)选择 $L/2$个染色体进行变异操作,形成 $L/2$个新的染色体,记为 P_t''

(5)在 $2L$个染色体 (P_t, P_t', P_t'')中选择 L个染色体,并保留最优个体,形成 $t+1$代种群

(6) $t=t+1$,如果 $t>T$,则停止,并将最优个体作为优化解,并解码为模糊分类系统,否则转(2).

3.4.5 流程图

图 1给出了算法的流程图.

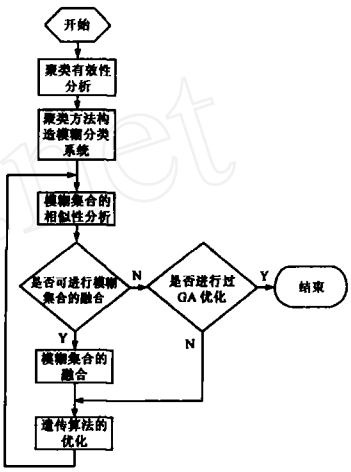


图 1 算法流程图

4 仿真

Iris是一个典型的分类问题,被众多学者用来作为各种分类算法的评估标准. 本文采用 Iris数据构造模糊分类系统,验证本文提出算法的有效性.

采用 3.1节的模糊聚类有效性函数,来确定最优的模糊规则数目,结果如表 1所示,为直观起见,图 2给出了 PC和 PE的图形表示. 除 XB外,其他的聚类有效性函数均给出最优模糊规则数目为 3的结果,而 XB给出最优模糊规则数目为 2,次优模糊规则数目为 3,也符合 Iris的分布特性. 本文采用"从众"策略,选择模糊规则数目为 3.

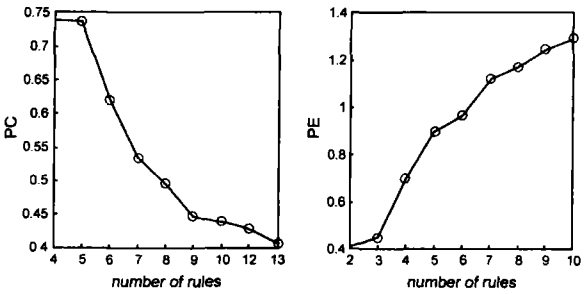


图 2 聚类有效性分析

利用 3.2节基于模糊聚类的方法,构造初始的模糊分类系统. 图 3给出初始模糊分类系统的隶属函数,可见模糊集合具有较大的重叠,解释性差. 错误划分样本数目为 17,精确性指标为 88.67%,精确性差.

表 1　聚类有效性分析

c	PC	PE	NFI	MinHT	MeHT	XB	FHV	DPA	PD	FP
2	0.7403	0.4074	0.4807	0.8468	0.6476	0.1008	0.0464	0.4236	0.4356	0.0002
3	0.7363	0.4529	0.6044	1.4058	0.7710	0.2278	0.0414	0.8695	0.5999	-0.0009
4	0.6204	0.6984	0.4938	1.0839	0.5859	0.4248	0.0678	0.7043	0.3312	0.0212
5	0.5351	0.8987	0.4189	0.8731	0.4275	0.8546	0.0875	0.6833	0.2960	0.0350
6	0.4959	0.9676	0.3951	0.4770	0.4172	0.6892	0.0751	0.6279	0.3669	0.0034
7	0.4479	1.1201	0.3559	0.4557	0.3669	1.3662	0.1131	0.4814	0.2557	0.0120
8	0.4396	1.1707	0.3595	0.5300	0.3686	1.9045	0.1546	0.3811	0.1638	0.0171
9	0.4282	1.2449	0.3567	0.6608	0.4162	0.6574	0.1275	0.6284	0.3004	0.0212
10	0.4068	1.2977	0.3409	0.6881	0.3954	0.6688	0.1178	0.5301	0.3380	0.0047

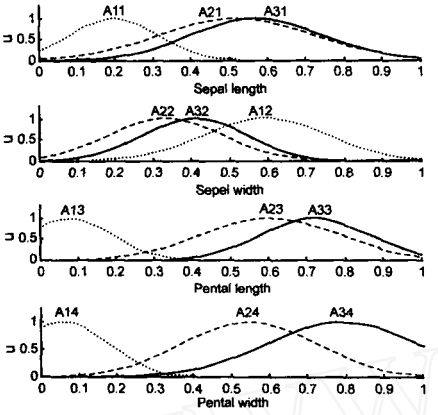


图 3　初始模糊分类系统的隶属函数

利用模糊集合的相似性测度,对初始的模糊分类系统进行模糊集合的相似性分析和融合,取相似性融合的测度阈值为 0.4. *Sepal length* 的隶属函数 A_{21} 和 A_{31} 的相似性测度为 0.7907, *Sepal width* 的隶属函数 A_{22} 和 A_{32} 的相似性测度为 0.6587, *Pental length* 的隶属函数 A_{23} 和 A_{33} 的相似性测度为 0.3825, *Pental width* 的隶属函数 A_{24} 和 A_{34} 的相似性测度为 0.3007. 可见需要对集合 $(A_{21}、A_{31})$ 和 $(A_{21}、A_{31})$ 进行融合. 融合后 *Sepal length* 的隶属函数为 2 个,其相似性测度为 0.1491, *Sepal width* 的隶属函数为 2 个,其相似性测度为 0.2825,均不再需要进行相似性融合.

为提高模糊分类系统的精确性,采用遗传算法优化约简后的系统. 遗传算法参数如下:种群大小 $L=40$,交叉概率 $Pc=0.8$,变异概率 $Pm=0.05$,进化代数 $T=100$. 经过遗传算法优化后,错误划分样本数目为 12,精确性指标为 92%.

上述的模糊集合相似性分析和遗传算法优化反复迭代进行,直至满足以下终止条件:所有隶属函数的模糊集合的相似性测度均小于其阈值 0.4;或者存在相似性测度大于阈值 0.4 的模糊集合,但是融合后经过遗传算法优化,增加的错误划分样本数目大于 2

在约简和优化过程中, *Sepal length* 和 *sepal width* 变量由于其隶属函数最终融合为一个模糊集合,并且剔除该模糊集合后,模糊分类系统的精确性指标基本不变,因此可以删除该两个变量,最终得到如下的模糊分类系统:

R^1: If *pental length* is medium and *pental width* is medium
Then pattern belongs to Class 1 with $CF=0.9197$
R^2: if *pental length* is short and *pental width* is narrow
Then pattern belongs to Class 2 with $CF=0.3594$　　(30)
R^3: If *pental length* is long and *pental width* is wide
Then pattern belongs to Class 3 with $CF=0.4657$

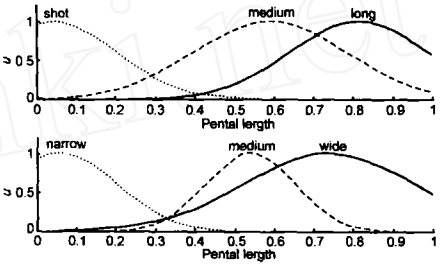


图 4　最优模糊分类系统的隶属函数

其隶属函数如图 4所示,隶属函数能够容易赋予语义值,具有较高的解释性. 模糊分类系统错误划分样本数目为 3,精确性指标为 98%.

为说明本方法的有效性,表 2给出了本文与其他文献的结果比较. 可见,本文提出的方法能够用较少的模糊规则和模糊集合数目,达到较精确的分类目的. 文献 [21] 虽然实现了样本的完全分类,但其模糊规则和模糊集合数目较大,解释性差,且泛化能力未知.

表 2　不同分类方法的性能比较

	模糊集合数目	模糊规则数目	精确性 (%)
文献 [18]	11	3	97.5
文献 [19]	9	3	96.2
文献 [3]	12	4	98
文献 [7]	7	5	98
文献 [20]	12	3	98
本文	6	3	98
文献 [21]	18	5	100

为验证本文提出方法对未知数据的泛化能力,采用 5 次折叠交叉验证法 (5-fold cross validation),验证结果为:错误划分的验证样本为 5 个,精确性为 96.67%. 参考文献 [20] 在相同的建模精度情况下,其泛化能力为 95.33%,本文略有提高,究其原因是参考文献 [21] 采用了四个特征变

量,降低了系统的泛化能力.

5 结论

本文提出一种基于数据的模糊分类系统的设计方法,同时考虑精确性和解释性.首先采用聚类有效性函数确定规则数目,利用基于模糊聚类的算法辨识初始的模糊分类系统.然后利用模糊集合的相似性融合对初始模糊分类系统进行约简,提高其解释性;采用遗传算法对约简后的系统进行优化,提高其精确性,该过程迭代进行直至满足终止条件.Iris数据分类的仿真,验证了该方法的有效性.

对于高维高噪声分类系统,聚类有效性分析可靠性降低,为保证精确性应采用较大规则数目,另外特征变量过多,均使得分类系统的解释性较差,从而需要进行特征变量选择和规则约简,这将是本文下一步的工作.

参考文献:

[1] Kuncheva L I. Fuzzy Classifier Design (Studies in Fuzziness and Soft Computing) [M]. New York: Heidelberg, 2000.

[2] Abe S, Thawonmas R. A fuzzy classifier with ellipsoidal regions [J]. IEEE Trans. Fuzzy Systems, 1997, 5(3): 358 - 368.

[3] Shi Y, Eberhart R, Chen Y. Implementation of evolutionary fuzzy system [J]. IEEE Trans. Fuzzy Systems, 1999, 7(2): 109 - 119.

[4] Castellano G, Fanelli1 A M. Modeling fuzzy classification systems with compact rule base [A]. 1999 International Conference on Computational Intelligence for Modeling, Control and Automation [C]. Vienna, Austria: IOS Press, 1999: 287 - 292.

[5] Jin Y. Advanced Fuzzy Systems Design and Applications [M]. New York: Physical-Verl, 2003.

[6] Nauck D D. Fuzzy data analysis with NEFCLASS [J]. International Journal of Approximate Reasoning, 2003, 32(2-3): 103 - 130.

[7] Ishibuchi H, Nakashima T, Murata T. Three-objective genetics-based machine learning for linguistic rule extraction [J]. Information Sciences, 2001, 136(1-4): 109 - 133.

[8] Abonyi J, Roubos H, Szeifert F. Data-Driven generation of compact, Accurate, and linguistically sound fuzzy classifiers based on a decision tree initialization [J]. International Journal of Approximate Reasoning, 2003, 32(1): 1 - 21.

[9] Chang X, Lilly J H. Evolutionary design of a fuzzy classifier from data [J]. IEEE Trans System Man Cybernetic Part B, 2004, 34(4): 1894 - 906.

[10] Gustafson D, Kessel W. Fuzzy clustering with a fuzzy covariance matrix [A]. Proc of IEEE Conf on Decision and Control [C]. San Diego, USA: IEEE Press, 1979. 761 - 766.

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

[12] Xie X L, G Beni A. validity measure for fuzzy clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847.

[13] Roubens M. Pattern classification problems and fuzzy sets [J]. Fuzzy Sets and Systems. 1978, 1(4): 239 - 253.

[14] Rivera F F, Zapata E L, Carazo J M. Cluster validity based on the hard tendency of the fuzzy classification [J]. Pattern Recognition Letters. 1990, 11(1): 7 - 12.

[15] Gath I, Geva A B. Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions [J]. Pattern Recognition Letters. 1989, 9(2): 77 - 86.

[16] 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.
Gao X-B. Fuzzy cluster analysis and its applications [M]. Xi'an: Xidian University Press, 2004. (in Chinese)

[17] Setnes M, Babuska R, Kaymak U, Lemke H R N. Similarity measures in fuzzy rule base simplification [J]. IEEE Trans on Systems Man and Cybernetics Part B. 1998, 28(3): 376 - 386.

[18] Wang J S, Lee G. C S. Self-adaptive neuro-fuzzy inference system for classification application [J]. IEEE Trans Fuzzy System. 2002, 10(6): 790 - 802.

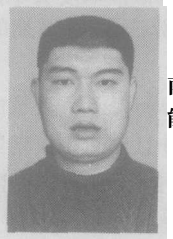
[19] Wu T P, Chen S M. A new method for constructing membership functions and fuzzy rules from training examples [J]. IEEE Trans System Man Cybernetic Part B. 1999, 29(1): 25 - 40.

[20] 童树鸿,沈毅,刘言志. 基于聚类分析的模糊分类系统构造方法 [J]. 控制与决策. 2001, 16(SUPP1): 737-740, 744.
Tong S, Shen Y, Liu Z. Approach to construct fuzzy classification system with clustering [J]. Control and Decision. 2001, 16: 737-740. (in Chinese)

[21] Russo M. Genetic fuzzy learning [J]. IEEE Trans. Evolutionary Computation. 2000, 4(3): 259 - 273.

作者简介:

邢宗义 男,1974年2月出生于山东临沂,南京理工大学机械学院副教授,主要从事模糊建模与工业过程智能控制伺服系统等研究.



张 永 男,1969年8月出生于江苏连云港,南京理工大学博士研究生,主要从事智能控制与智能系统、模糊建模等研究.