文章编号：0253-374X(2015)07-1089-08 DOI:10.11908/j.issn.0253-374x.2015.07.020

实例检索中基于混合属性距离的相似性度量

朱芳来[1]，董志豪[1]，徐立云[2]
(1. 同济大学 电子与信息工程学院，上海 201804；2. 同济大学 机械与能源工程学院，上海 201804)

摘要：提出了一种基于混合属性距离的相似性度量方法. 利用各个属性间的距离及其组合权重求得实例间的总体距离，再用实例间的总体距离来刻画其相似程度. 基于区间数的距离计算公式和模糊集合理论，给出了属性值为模糊数、模糊区间数时的距离公式，并改进了属性值为隶属度函数时的距离公式. 同时考虑了属性权重问题，提出了一种基于距离离差信息的客观赋权方法，将主观权重和客观权重加以组合，以组合权重来计算实例的全局相似度. 以阀门的概念设计为例，验证了该方法在实例检索中的的可行性和合理性.

关键词：相似性度量；模糊数；模糊区间数；隶属度函数；客观权重
中图分类号：TB115 **文献标志码**：A

Similarity Measurement for Retrieval Based on Hybrid Attribute Distance

ZHU Fanglai[1], *DONG Zhihao*[1], *XU Liyun*[2]
(1. School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China; 2. School of Mechanical Engineering, Tongji University, Shanghai 201804, China)

Abstract: A new similarity measure method based on hybrid attribute distance was introduced. Overall distance between two cases was determined by distances between attributes and synthesis weights. First, a degree of similarity was characterized by overall distance. Based on the distance formula of interval numbers and the theory of fuzzy set, a distance formula of fuzzy numbers and fuzzy interval numbers was given and the distance formula of membership function was improved. Meanwhile, the deviation information of distance values was used to calculate objective weights regarding weights assignment, then objective weights and subjective weights were integrated into synthesis weights to calculate the overall similarity of the cases. Lastly, the method was applied to the conceptual design of valve. The effectiveness and feasibility of the method in case retrieval were demonstrated.

Key words: similarity measurement; fuzzy numbers; fuzzy interval numbers; membership function; objective weights

基于实例推理(CBR)的核心思想是利用人们过去解决问题的方案和经验来解决新问题，其中实例检索是 CBR 系统的关键. 实例检索就是在实例库中找到一个与当前问题实例最相似的实例，因此实例间的相似度计算决定了实例检索的质量和可靠性.

目前，基于实例属性的相似性度量方法已有很多研究. 文献[1]提出了语义距离的概念，用以描述两个对象语义上的相近程度. 在此基础上，文献[2]采用模糊相似优先比来描述问题实例与各个实例间的相似顺序. 文献[3]对文献[2]中的方法进行了改进，采用海明距离的模型描述实例间距离. 另外，曼哈顿距离[4]、最近相邻[5]、特征区间相似度[6]等算法也相继被提出. 但传统的实例间相似度计算一般要求有确定的属性值. 然而，现实问题求解过程中，由于客观环境的动态不确定性和决策者主观思维结构的复杂性，导致信息感知的不完全性和不精确性，所以问题的描述中包含有与不精确信息对应的不精确特征属性. 若采用传统的相似性计算模型，往往是无效的. 文献[7-8]利用区间数描述案例的不精确属性，但有些属性的精确属性值本身就是区间数，其模糊属性值应为模糊区间数. 所以，这些方法并不适用于含有模糊数、模糊区间数等模糊属性值的复杂问题. 另一方面，针对实例权重的计算问题，目前大多数检索模型只考虑了属性的主观权重，但主观权重只能反映决策者对各属性的偏好或者属性本身的重要程度，并不能反映属性本身信息对决策结果的贡献.

本文提出了一种基于属性总体距离的相似性度

收稿日期：2014-07-21
基金项目：上海市科委同济大学-青浦区科技合作项目(08002370095)
第一作者：朱芳来(1965—)，男，教授，博士生导师，工学博士，主要研究方向为未知输入观测器、故障检测.
E-mail：zhufanglai@tongji.edu.cn

量模型,用模糊数、模糊区间数表示模糊属性值,在精确属性间距离公式的基础上给出了新的模糊属性间距离公式,并改进了属性值为隶属度函数时的距离公式.其次,本文提出一种新的基于属性距离方差的客观赋权法,在此基础上将主客观权重加以组合,利用组合权重计算实例间的总体距离.

1 基于属性距离的相似性度量模型

1.1 实例描述

设实例库中有 K 个实例,$B=\{C_0,C_1,\cdots,C_k,\cdots,C_K\}$.不妨设一个实例有 n 个属性,则实例 C_k 可以表达为 $C_k=\{a_{k1},a_{k2},\cdots,a_{kj},\cdots,a_{kn}\}$. n 个属性的权重分配表达为 $\boldsymbol{W}=(w_0,w_1,\cdots,w_j,\cdots,w_n)$.其中,$0\leqslant w_j\leqslant 1$,且 $\sum_{j=1}^{n} w_j=1$.又设需求解的问题实例为 C_0,则有 $C_0=\{a_{01},a_{02},\cdots,a_{0j},\cdots,a_{0n}\}$.

1.2 C_0 与各实例的相似度

C_0 与实例库 B 中所有实例的属性距离构成距离矩阵 $\boldsymbol{M}$,对矩阵 $\boldsymbol{M}$ 的每一列进行量纲一处理得到矩阵 $\boldsymbol{M}_n$.然后由属性主观权重和客观权重得到组合权重 $\boldsymbol{W}$,则根据下式可得问题实例 C_0 与其他各实例间的总体距离 D_k:

$$\boldsymbol{M}_n\boldsymbol{W}^{\mathrm{T}}=(D_1,D_2,\cdots,D_k,\cdots,D_K) \tag{1}$$

式中:$\boldsymbol{W}^{\mathrm{T}}$ 为组合权重向量 $\boldsymbol{W}$ 的转置.那么,C_0 与 C_k 间的相似度 $S_k=1-D_k$,相似度最大者即为问题实例 C_0 的解.

2 属性距离计算方法

传统相似度分析方法并没有考虑属性值为模糊数、模糊区间数的情形,对此本节提出了新的距离公式.属性值为隶属度函数时一般采用 Euclid 距离公式,但这种方法存在不足,本节提出了改进后的距离公式.

2.1 属性分类

要对实例属性进行分类,首先对属性通常所具有的属性值进行分类.现以阀门主要属性为例进行说明,如表1所示.

属性值分类:

(1)确定符号属性值.这种属性值通常用明确的术语表示,如表1中所有实例的属性"型式"用字母表示.

(2)确定数属性值.这种属性值是一个确定的数,如表1中所有实例的属性"公称通径 D_n"的值就是一个确定数.

表1 阀门主要属性

Tab.1 Main attributes of valve

实例	属性					
	型式	公称压力 P_n/MPa	公称通径 D_n/mm	升程 H/mm	适用温度 T/℃	阀杆拉压强度 P 语义描述
C_1	A	1.0	55	19	[300,400]	很差
C_2	B	1.6	125	44	[350,500]	差
C_3	A	2.5	50	13	[250,400]	好
C_4	A	2.5	40	10	[400,500]	较好
C_5	BS	1.6	65	14	[100,200]	较好
C_6	A	4.0	55	17	[350,500]	很好
C_7	A	2.5	50	24	[350,450]	差
C_0	A	约3.0	45	约[11,13]	约[300,400]	好

(3)确定区间数属性值.这种属性值是一个边界确定的区间数.对属性"适用温度 T"而言,除去 C_0 外,其他属性值均为区间数,如[300,400].

(4)模糊数或模糊区间数属性值.这种属性值是一个不确定的数或一个没有确定边界的区间,如表1中的"约3.0"、"约[300,400]".

(5)模糊概念属性值.这种属性值通常是一概念变量,每一个概念变量都对应一个模糊集和隶属度函数 u,如表1中阀杆拉压强度 P 语义描述可以分为很好、好、较好、差、很差共五类.每类依次对应模糊集 A_1,A_2,A_3,A_4,A_5.它们各自的隶属度函数表示如图1所示.

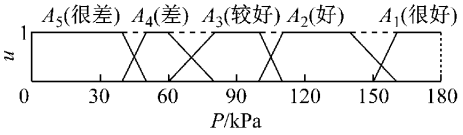

图1 模糊集的隶属度函数

Fig.1 Membership function of fuzzy set

2.2 属性间距离计算

2.2.1 确定符号间距离

对于确定符号属性,当实例集中相应的属性与问题属性完全匹配时,其距离赋为零,否则赋为无穷大.通过确定符号属性可以筛选出一个实例子集.

2.2.2 确定数间距离

对两个确定数属性值 x_1 和 x_2,其距离定义为

$$f_{\mathrm{DN}}(x_1,x_2)=|x_1-x_2|$$

2.2.3 确定区间数间距离

为了解决实际问题的需要,诸多学者提出了多种区间数距离公式[9],如 P-距离和 Hausdorff 距离等.但这些距离都只考虑区间数左右端点的偏差,丢

失了一些有用的信息. 故本文采用基于期望值与宽度的 EW 型距离公式[10]，这种算法不仅包含的信息量多，而且计算简单.

定义 1　设 $\mathbf{R}$ 为实数域，$x^{(\mathrm{L})}, x^{(\mathrm{U})} \in \mathbf{R}$，且 $x^{(\mathrm{L})} \leqslant x^{(\mathrm{U})}$，若 $I=(x^{(\mathrm{L})}, x^{(\mathrm{U})})$，称 I 为区间数；$E(I)=\dfrac{x^{(\mathrm{L})}+x^{(\mathrm{U})}}{2}$ 为 I 的期望值；$B(I)=\dfrac{x^{(\mathrm{U})}-x^{(\mathrm{L})}}{2}$ 为 I 的宽度.

定义 2　设两区间数 $I_1=[x_1{}^{(\mathrm{L})}, x_1{}^{(\mathrm{U})}]$，$I_2=[x_2{}^{(\mathrm{L})}, x_2{}^{(\mathrm{U})}]$，则 I_1 与 I_2 的距离

$$f_{\mathrm{DI}}(I_1, I_2)=\sqrt[p]{|E(I_1)-E(I_2)|^p+\frac{1}{3}|B(I_1)-B(I_2)|^p},\quad p\geqslant 1 \tag{2}$$

2.2.4　模糊数或模糊区间数间距离

目前，并未见到“约 3.0”或“约[300,400]”这样的模糊数或模糊区间数间距离的算法，对此，本文基于式(2)提出了一种新的距离公式.

(1) 模糊数的处理

对于“约 3.0”这样的属性值可以用三角模糊数表示.

定义 3[11]　若 $\widetilde{S}$ 可用如下隶属度函数定义：

$$\alpha=u_{\widetilde{S}}(x)=\begin{cases}\dfrac{x-s_1}{s_M-s_1}, & s_1\leqslant x\leqslant s_M\\ \dfrac{x-s_2}{s_M-s_2}, & s_M<x\leqslant s_2\\ 0, & \text{其他}\end{cases} \tag{3}$$

则称 $\widetilde{S}$ 为三角模糊数，其中$[s_1, s_2]$为支撑区间，点$(s_M, 1)$是峰值，如图 2 所示. 实际应用中，通常用 s_1，s_2，s_M 三个值来构造三角模糊数. 因此，用 $\widetilde{S}=(s_1, s_2, s_M)$表示.

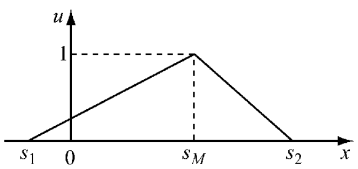

图 2　三角模糊数

Fig.2　Triangular fuzzy number

假设实例 C_p 的第 j 个属性值为模糊数 $F_{pj}^{(n)}=$“约 x_{pj}”，则 $F_{pj}^{(n)}$ 可以用三角模糊数 $\widetilde{S}=(s_1, s_2, s_M)$表示，其中：

$$s_M=x_{pj}$$

$$s_1=\min(x_{1j}, x_{2j}, \cdots, x_{kj}, \cdots, x_{Kj})$$

$$s_2=\max(x_{1j}, x_{2j}, \cdots, x_{kj}, \cdots, x_{Kj})$$

(2) 三角模糊数 $\widetilde{S}$ 的 α 级区间 $I(\alpha)$

令 $\widetilde{S}$ 的 α 级区间 $I(\alpha)=[x^{(\mathrm{L})}(\alpha), x^{(\mathrm{U})}(\alpha)]$，其中 $s_1\leqslant x^{(\mathrm{L})}(\alpha)\leqslant s_M$，$s_M<x^{(\mathrm{U})}(\alpha)\leqslant s_2$，由式(3)得

$$\alpha=\frac{x^{(\mathrm{L})}(\alpha)-s_1}{s_M-s_1}, \alpha=\frac{x^{(\mathrm{U})}(\alpha)-s_2}{s_M-s_2}$$

从而求得 $x^{(\mathrm{L})}(\alpha)=s_1+\alpha(s_M-s_1)$，$x^{(\mathrm{U})}(\alpha)=s_2-\alpha(s_2-s_M)$，所以三角模糊数 $\widetilde{S}=(s_1, s_2, s_M)$的 α 级区间

$$I(\alpha)=[x^{(\mathrm{L})}(\alpha), x^{(\mathrm{U})}(\alpha)]=[s_1+\alpha(s_M-s_1), s_2-\alpha(s_2-s_M)]$$

即模糊数最终可以用其对应的三角模糊数的 α 级区间 $I(\alpha)$表示. 模糊数 $F_{pj}^{(n)}=$“约 x_{pj}”的 α 级区间

$$I_{pj}(\alpha)=[x_{pj}^{(\mathrm{L})}(\alpha), x_{pj}^{(\mathrm{U})}(\alpha)]=[s_{pj1}+\alpha(s_{pjM}-s_{pj1}), s_{pj2}-\alpha(s_{pj2}-s_{pjM})]$$

例如，求取表 1 中的模糊数“约 3.0”的 α 级区间. 易知 $s_1=1$，$s_2=4$，$s_M=3$，则

$$x^{(\mathrm{L})}(\alpha)=s_1+\alpha(s_M-s_1)=1+2\alpha$$

$$x^{(\mathrm{U})}(\alpha)=s_2-\alpha(s_2-s_M)=4-\alpha$$

模糊数“约 3.0”的 α 级区间为$[1+2\alpha, 4-\alpha]$.

(3) 模糊区间数的处理

对于“约[300,400]”这样的模糊数属性值可以用梯形模糊数表示.

定义 4[12]　若 $\widetilde{T}$ 可用如下隶属度函数定义：

$$\alpha=u_{\widetilde{T}}(x)=\begin{cases}\dfrac{x-t_1}{t_M^{(\mathrm{L})}-t_1}, & t_1\leqslant x<t_M^{(\mathrm{L})}\\ 1, & t_M^{(\mathrm{L})}\leqslant x\leqslant t_M^{(\mathrm{U})}\\ \dfrac{x-t_2}{t_M^{(\mathrm{U})}-t_2}, & t_M^{(\mathrm{U})}<x\leqslant t_2\\ 0, & \text{其他}\end{cases} \tag{4}$$

则称 $\widetilde{T}$ 为梯形模糊数，其中$[t_1, t_2]$为支撑区间，在 $x\in[t_M^{(\mathrm{L})}, t_M^{(\mathrm{U})}]$时其隶属度为 1，如图 3 所示.

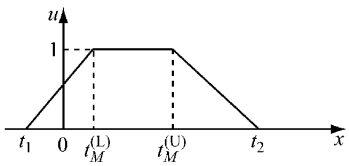

图 3　梯形模糊数

Fig.3　Trapezoidal fuzzy number

实际应用中，通常用 $t_1, t_M^{(\mathrm{L})}, t_M^{(\mathrm{U})}, t_2$ 四个值来构造梯形模糊数. 因此用 $\widetilde{T}=(t_1, t_M^{(\mathrm{L})}, t_M^{(\mathrm{U})}, t_2)$表示.

假设实例 C_p 的第 j 个属性值为模糊区间数 $F_{pj}^{(i)}=$“约$[x_{pj}^{(\mathrm{L})}, x_{pj}^{(\mathrm{U})}]$”，则 $F_{pj}^{(i)}$ 可以用梯形模糊数 $\widetilde{T}=(t_1, t_M^{(\mathrm{L})}, t_M^{(\mathrm{U})}, t_2)$ 表示，其中，$t_M^{(\mathrm{L})}=x_{pj}^{(\mathrm{L})}$，$t_M^{(\mathrm{U})}=x_{pj}^{(\mathrm{U})}$，$t_1=\min(x_{1j}^{(\mathrm{L})}, x_{2j}^{(\mathrm{L})}, \cdots, x_{kj}^{(\mathrm{L})}, \cdots, x_{Kj}^{(\mathrm{L})})$，$t_2=\max(x_{1j}^{(\mathrm{U})}, x_{2j}^{(\mathrm{U})}, \cdots, x_{kj}^{(\mathrm{U})}, \cdots, x_{Kj}^{(\mathrm{U})})$. 同样，可以给出梯形模糊数的 α 级区间

$$I(\alpha)=[x^{(\mathrm{L})}(\alpha),x^{(\mathrm{U})}(\alpha)]=[t_1+\alpha(t_M^{(\mathrm{L})}-t_1),t_2-\alpha(t_2-t_M^{(\mathrm{U})})]$$

即模糊区间数最终可以用其对应的梯形模糊数的 α 级区间 $I(\alpha)$ 表示. 模糊区间数 $F_{pj}^{(i)}$ ="约$[x_{pj}^{(\mathrm{L})},x_{pj}^{(\mathrm{U})}]$"的 α 级区间

$$I_{pj}(\alpha)=[x_{pj}^{(\mathrm{L})}(\alpha),x_{pj}^{(\mathrm{U})}(\alpha)]=[t_{pj1}+\alpha(t_{pjM}^{(\mathrm{L})}-t_{pj1}),t_{pj2}-\alpha(t_{pj2}-t_{pjM}^{(\mathrm{U})})]$$

例如,对于表1中的属性"适用温度",求模糊区间数"约[300,400]"的 α 级区间. 易知 $t_1=100$, $t_2=500$, $t_M^{(\mathrm{L})}=300$, $t_M^{(\mathrm{U})}=400$,则

$$x^{(\mathrm{L})}(\alpha)=t_1+\alpha(t_M^{(\mathrm{L})}-t_1)=100+200\alpha$$

$$x^{(\mathrm{U})}(\alpha)=t_2-\alpha(t_2-t_M^{(\mathrm{U})})=500-100\alpha$$

模糊区间数"约[300,400]" α 级区间为$[100+200\alpha,500-100\alpha]$.

(4) 模糊数及模糊区间数间基于 α 级区间的距离公式

基于上述对模糊数与模糊区间数的处理及区间数距离式(2),提出了新的模糊数及模糊区间数(统一用 F_{pj} 表示,下标 p 表示实例 C_p, j 表示第 j 个属性)间的距离公式为

$$f_{\mathrm{F}}(F_{pj},F_{qj})=\frac{\int_0^1 g(I_{pj}(\alpha),I_{qj}(\alpha))h(\alpha)\mathrm{d}\alpha}{\int_0^1 h(\alpha)\mathrm{d}\alpha} \quad (5)$$

其中,

$$g(I_{pj}(\alpha),I_{qj}(\alpha))=[|E(I_{pj}(\alpha))-E(I_{qj}(\alpha))|^p+\frac{1}{3}|B(I_{pj}(\alpha)-B(I_{qj}(\alpha))|^p]^{\frac{1}{p}},\ p\geqslant 1$$

$h(\alpha)$作为加权函数,是定义在[0,1]上的正值连续函数,由于决策者往往更偏好于隶属水平较高的点,所以这里选择 $h(\alpha)$为增函数.

2.2.5 模糊概念属性间距离

模糊概念属性值一般用隶属度函数表示,如表1中的属性"阀杆拉压强度",而隶属度函数型式较多且有些型式的隶属度函数表达式复杂,如正态分布、岭形分布等,不易用其 α 级区间表示,所以在计算模糊概念属性间的距离时多采用 Euclid 距离公式[1].

定义5 当模糊集 A_i 与 A_j 表示为论域 $X=[a,b]$上的隶属度函数 u_{A_i} 与 u_{A_j},且 u_{A_i} 与 u_{A_j} 在$[a,b]$上可积,则 A_i 与 A_j 的 Euclid 距离

$$f_{\mathrm{E}}(A_i,A_j)=\left[\frac{1}{b-a}\int_a^b(u_{A_i}(x)-u_{A_j}(x))^2\mathrm{d}x\right]^{\frac{1}{2}}$$

但 Euclid 距离公式存在不足,如对图4中的隶属度函数 $u_{A_6}(x)$, $u_{A_7}(x)$, $u_{A_8}(x)$,根据 Euclid 距离公式,会有 $f(A_6,A_8)=f(A_7,A_8)$,但这显然不符合实际情况.

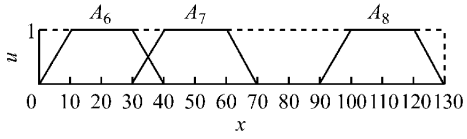

图4 A_6, A_7, A_8 的隶属度函数

Fig.4 Membership function of A_6, A_7 and A_8

对此,本文给出一种新的距离公式.

设模糊集 A_i 与 A_j 的隶属度函数如下:

$$u_{A_i}(x)=\begin{cases}l_{A_i}(x), & t_1\leqslant x<t_M^{(\mathrm{L})}\\ 1, & t_M^{(\mathrm{L})}\leqslant x\leqslant t_M^{(\mathrm{U})}\\ r_{A_i}(x), & t_M^{(\mathrm{U})}<x\leqslant t_2\\ 0, & \text{其他}\end{cases}$$

$$u_{A_j}(x)=\begin{cases}l_{A_j}(x), & h_1\leqslant x<h_M^{(\mathrm{L})}\\ 1, & h_M^{(\mathrm{L})}\leqslant x\leqslant h_M^{(\mathrm{U})}\\ r_{A_j}(x), & h_M^{(\mathrm{U})}<x\leqslant h_2\\ 0, & \text{其他}\end{cases}$$

其中:$[h_1,h_2]$为 $u_{A_j}(x)$的支撑区间;l 和 r 分别为相应区间的隶属度函数. A_i 与 A_j 的隶属度函数有三种关系,分别对应三种不同的距离公式,如下所示:

(1) 当 $t_2\leqslant h_1$ 时,模糊集 A_i 与 A_j 的隶属度函数如图5所示,则

$$f_{\mathrm{L1}}(A_i,A_j)=\frac{1}{2}\left[\int_{t_1}^{h_2}|u_{A_i}(x)-u_{A_j}(x)|\mathrm{d}x+2\int_{t_M^{(\mathrm{L})}}^{h_M^{(\mathrm{U})}}1-r_{A_i}(x)-l_{A_j}(x)\mathrm{d}x\right] \quad (6)$$

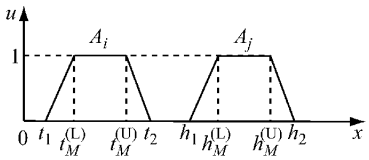

图5 A_i 与 A_j 的隶属度函数($t_2\leqslant h_1$)

Fig.5 Membership function of A_i and A_j ($t_2\leqslant h_1$)

(2) 当 $t_2>h_1$ 且 $t_M^{(\mathrm{U})}\leqslant h_M^{(\mathrm{L})}$ 时,模糊集 A_i 与 A_j 的隶属度函数如图6所示,则

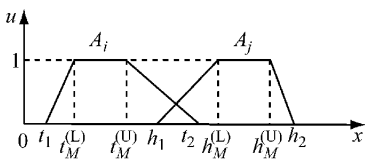

图6 A_i 与 A_j 的隶属度函数($t_2>h_1$ 且 $t_M^{(\mathrm{U})}\leqslant t_M^{(\mathrm{L})}$)

Fig.6 Membership function of A_i and A_j ($t_2>h_1$ and $t_M^{(\mathrm{U})}\leqslant t_M^{(\mathrm{L})}$)

$$f_{L2}(A_i,A_j)=\frac{1}{2}\left[\int_{t_1}^{h_2}|u_{A_i}(x)-u_{A_j}(x)|\,\mathrm{d}x+2\int_{t_M^{(U)}}^{h_M^{(L)}}1-\max(r_{A_i}(x)-l_{A_j}(x))\,\mathrm{d}x\right] \quad (7)$$

(3) 当 $t_M^{(U)}>h_M^{(L)}$ 时，模糊集 A_i 与 A_j 的隶属度函数如图 7 所示，则

$$f_{L3}(A_i,A_j)=\frac{1}{2}\int_{t_1}^{h_2}|u_{A_i}(x)-u_{A_j}(x)|\,\mathrm{d}x \quad (8)$$

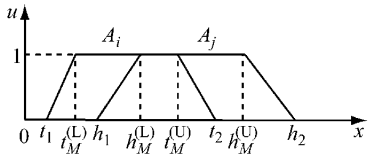

图 7　A_i 与 A_j 的隶属度函数（$t_M^{(U)}>h_M^{(L)}$）

Fig.7　Membership function of A_i and A_j ($t_M^{(U)}>h_M^{(L)}$)

现在利用新的距离公式求解图 4 中 A_6 与 A_8，A_7 与 A_8 的距离，显然应使用式(6)，则

$$f_{L1}(A_6,A_8)=90,\ f_{L1}(A_7,A_8)=60$$

即

$$f_{L1}(A_6,A_8)>f_{L1}(A_7,A_8)$$

与实际情况相符合.

2.3　小结

利用上述提出的距离公式，可以求得各属性间距离. C_0 与 C_k 的第 j 个属性间距离记为 d_{kj}，则 C_0 与实例库 B 中所有实例的属性距离构成距离矩阵 $\boldsymbol{M}$，并对 $\boldsymbol{M}$ 的每一列进行量纲一化处理，得到新的矩阵 $\boldsymbol{M}_n$，如下所示：

$$\boldsymbol{M}=\begin{bmatrix}d_{11}&d_{12}&\cdots&d_{1n}\\d_{21}&d_{22}&\cdots&d_{2n}\\\vdots&&&\vdots\\d_{K1}&d_{K2}&\cdots&d_{Kn}\end{bmatrix}$$

$$\boldsymbol{M}_n=\begin{bmatrix}d'_{11}&d'_{12}&\cdots&d'_{1n}\\d'_{21}&d'_{22}&\cdots&d'_{2n}\\\vdots&&&\vdots\\d'_{K1}&d'_{K2}&\cdots&d'_{Kn}\end{bmatrix}$$

其中，$d'_{kj}=\dfrac{d_{kj}}{\sqrt{\sum_{i=1}^{K}d_{ij}^2}}$ 为距离矩阵中 d_{kj} 经过量纲一化后的值.

3　属性的权重计算

在进行实例检索时，各属性权重的合理性和准确性直接影响结果的可靠性. 属性权重按性质主要分为：①主观权重. 反映决策者对各属性的偏好或者属性本身的重要程度，用 $\boldsymbol{W}^{(1)}=(w_1^{(1)},w_2^{(1)},\cdots,w_j^{(1)},\cdots,w_n^{(1)})$ 表示；②客观权重. 反映属性所含的信息量对检索结果的贡献，此类权重与其属性值对实例的区分能力成正比，用 $\boldsymbol{W}^{(2)}=(w_1^{(2)},w_2^{(2)},\cdots,w_j^{(2)},\cdots,w_n^{(2)})$ 表示.

目前大多数检索模型只考虑了属性的主观权重，为此本文提出一种新的基于属性距离方差的客观赋权法，在此基础上得到属性的组合权重 $\boldsymbol{W}=\sigma(\boldsymbol{W}^{(1)},\boldsymbol{W}^{(2)})=(w_1,w_2,\cdots,w_i,\cdots,w_n)$.

3.1　基于层次分析法的主观赋权法

层次分析法(AHP)[12]是一种定性和定量相结合的决策方法，可以有效解决属性个数为 3 时权重值难以确定的问题，现在简略地给出 AHP 方法确定属性权重的实现步骤.

(1) 根据属性重要程度获得判断矩阵 $\boldsymbol{E}$

决策者根据对各属性的偏好或者属性本身的重要程度进行两两比较，可以采用插入法[13]对属性进行排序，然后采用 1～9 标度法[12]得到判断矩阵

$$\boldsymbol{E}=(e_{ij})_{n\times n}=\begin{bmatrix}e_{11}&e_{12}&\cdots&e_{1n}\\e_{21}&e_{22}&\cdots&e_{2n}\\\vdots&&&\vdots\\e_{n1}&e_{n2}&\cdots&e_{nn}\end{bmatrix}$$

其中，e_{ij} 表示决策者确定的第 i 个属性相对于第 j 个属性的重要程度系数，并且满足 $e_{ij}=1/e_{ji}$.

从属性值分类来说，属性值为确定数或确定区间的属性一般较为重要，模糊数或模糊区间数次之. 从具体实例来说，需根据其实际应用来确定哪个属性较为重要. 如：对船用阀门来说，“抗腐蚀性”比“适用温度”重要；对锅炉用阀门来说，“适用温度”比“抗腐蚀性”重要.

(2) 单个属性主观权重. 用和法确定单个属性的主观权重，其计算公式为

$$w_i^{(1)}=\frac{1}{n}\sum_{j=1}^{n}\frac{a_{ij}}{\sum_{k=1}^{n}a_{kj}},\ i=1,2,\cdots,n$$

根据计算所得实例各属性主观权重 $w_i^{(1)}$，构建量纲一化后的主观权重向量

$$\boldsymbol{W}^{(1)}=(w_1^{(1)},w_2^{(1)},\cdots,w_j^{(1)},\cdots,w_n^{(1)})$$

3.2　基于属性距离离差信息的客观赋权法

对于第 j 个属性，若问题实例 C_0 与实例库 B 中所有实例的距离差异很小，则该属性对实例检索的贡献小，应赋予较小的权重；反之则应赋予相对较大的权重. 本文给出一种新的基于属性距离方差信息的客观赋权法，具体步骤如下：

(1) 对量纲一化后的距离矩阵 $\boldsymbol{M}_n$ 中的每一列

求方差,得到方差向量

$$\boldsymbol{\delta}=(\delta_1,\delta_2,\cdots,\delta_j,\cdots,\delta_n)$$

式中:δ_j 为矩阵 $\boldsymbol{M}_{\mathrm{n}}$ 中第 j 列向量的方差. δ_j 越小,说明实例间关于第 j 个属性的距离差异较小.

(2) 对方差向量进行量纲一化处理,以此作为基于属性距离方差信息的权重向量,如下所示:

$$\boldsymbol{W}^{(2)}=(w_1^{(2)},w_2^{(2)},\cdots,w_j^{(2)},\cdots,w_n^{(2)})$$

其中,$w_j^{(2)}=\dfrac{\delta_j}{\sum_{i=1}^{n}\delta_i}$ 为第 j 属性的客观权重.

3.3 组合权重

本文采用乘法合成计算组合权重,如下所示:

$$w_i=w_i^{(1)}w_i^{(2)}/\sum_{j=1}^{n}(w_j^{(1)}w_j^{(2)}),\ i=1,2,\cdots,n \tag{9}$$

从式(9)可以看出,组合权重综合考虑了决策者偏好和实例自身的属性信息.

4 案例分析

表1为一个简化了的阀门设计案例库,从表1可知案例中选用了阀门设计的六个属性,$C_1\sim C_7$ 为案例库中已有的阀门设计方案,C_0 为需要设计的问题实例,其中既有模糊属性值也有确定属性值. 现以阀门概念设计中实例检索为例,来说明该相似性度量方法的应用.

4.1 计算距离矩阵

属性"型式"为确定符号属性,所以首先就属性"型式"选出与问题实例 C_0 距离为0的实例:C_1,C_3,C_4,C_6,C_7. 以下计算均针对这五个实例及实例的后五个属性:P_{n},D_{n},H,T,P. 式(5)中取 $h(\alpha)=\alpha$.

首先计算问题实例 C_0 与实例 C_1 的各个属性间距离 d_{1j},$j=1,2,\cdots,5$.

(1) 属性 P_{n} 间距离

由表1可知,实例 C_0 的 P_{n} 属性值 $a_{01}=$"约3.0",为模糊数,其 α 级区间 $I_{01}(\alpha)=[1+2\alpha,\ 4-\alpha]$;实例 C_1 的 P_{n} 属性值 $a_{11}=1.0$,其对应的 α 级区间 $I_{11}(\alpha)=[1,1]$.

由式(5)(p 取2)得 C_0 与 C_1 间关于属性 P_{n} 的距离

$$d_{11}=f_{\mathrm{F}}(a_{01},a_{11})=\frac{\int_0^1 g(I_{01}(\alpha),I_{11}(\alpha))h(\alpha)\mathrm{d}\alpha}{\int_0^1 h(\alpha)\mathrm{d}\alpha}=$$

$$\int_0^1\frac{\sqrt{2}}{2}\sqrt{\alpha^2+2\alpha+6}\alpha\mathrm{d}\alpha\Big/\int_0^1\alpha\mathrm{d}\alpha=1.869$$

(2) 属性 D_{n} 间距离

由表1可知,实例 C_0 与 C_1 的 D_{n} 属性值均为确定数:$a_{02}=45$,$a_{12}=55$. 易得 C_0 与 C_1 间关于属性 D_{n} 的距离

$$d_{12}=f_{\mathrm{DN}}(a_{02},a_{12})=|a_{02}-a_{12}|=10$$

(3) 属性 H 间距离

由表1可知,实例 C_0 的 H 属性值 $a_{03}=$"约[11,13]"为模糊区间数,易知 $t_M^{(\mathrm{L})}=11$,$t_M^{(\mathrm{U})}=13$,$t_1=10$,$t_2=24$,则模糊区间数"约[11,13]"的 α 级区间 $I_{03}(\alpha)=[10+\alpha,\ 24-11\alpha]$. 实例 C_1 的 H 属性值 $a_{13}=19$,其对应的 α 级区间 $I_{13}(\alpha)=[19,19]$.

由式(5)(p 取2)得 C_0 与 C_1 间关于属性 H 的距离

$$d_{13}=f_{\mathrm{F}}(a_{03},a_{13})=\frac{\int_0^1 g(I_{03}(\alpha),I_{13}(\alpha))h(\alpha)\mathrm{d}\alpha}{\int_0^1 h(\alpha)\mathrm{d}\alpha}=$$

$$\frac{\sqrt{3}}{6}\int_0^1\sqrt{147\alpha^2-212\alpha+244}\alpha\mathrm{d}\alpha\Big/\int_0^1\alpha\mathrm{d}\alpha=5.736$$

(4) 属性 T 间距离

由表1可知,实例 C_0 的 T 属性值 $a_{04}=$"约[300,400]"为模糊区间数,易知 $t_M^{(\mathrm{L})}=300$,$t_M^{(\mathrm{U})}=400$,$t_1=100$,$t_2=500$,则模糊区间数"约[300,400]"的 α 级区间 $I_{04}(\alpha)=[250+50\alpha,\ 500-100\alpha]$. 实例 C_1 的 T 属性值 $a_{14}=[300,400]$,为确定区间数,其对应的 α 级区间 $I_{14}(\alpha)=[300,400]$.

由式(5)(p 取2)得 C_0 与 C_1 间关于属性 T 的距离

$$d_{14}=f_{\mathrm{F}}(a_{04},a_{14})=\frac{\int_0^1 g(I_{04}(\alpha),I_{14}(\alpha))h(\alpha)\mathrm{d}\alpha}{\int_0^1 h(\alpha)\mathrm{d}\alpha}$$

$$=25\int_0^1\sqrt{4\alpha^2-12\alpha+\frac{28}{3}}\alpha\mathrm{d}\alpha\Big/\int_0^1\alpha\mathrm{d}\alpha=72.793$$

(5) 属性 P 间距离

由表1可知,实例 C_0 与 C_1 的 P 属性值均为模糊概念:$a_{05}=$"好",$a_{15}=$"很差",对应的模糊集分别为 A_2 和 A_5,其隶属度函数如图1所示. 根据第2.2.5节中的分析,由式(6)得 C_0 与 C_1 间关于属性 P 的距离

$$d_{15}=f_{\mathrm{L1}}(a_{05},a_{15})=$$

$$\frac{1}{2}\left[\int_0^{160}|u_{A_2}(x)-u_{A_5}(x)|\mathrm{d}x+\right.$$

$$\left.2\int_{40}^{110}1-r_{A_2}(x)-l_{A_5}(x)\mathrm{d}x\right]=210$$

(6) 距离矩阵 $\boldsymbol{M}$

实例 C_0 与 C_1 各属性间距离为

$$(d_{11},d_{12},d_{13},d_{14},d_{15})=$$

(1.869,10,5.736,72.793,210)

依次计算实例 C_0 与其他实例各属性间的距离，得到距离矩阵 $\boldsymbol{M}$，并对其进行量纲一化处理得 $\boldsymbol{M}_n$，如下所示：

$$\boldsymbol{M}=\begin{bmatrix}1.869 & 10 & 5.736 & 72.793 & 210\\ 0.494 & 5 & 2.069 & 92.814 & 0\\ 0.494 & 5 & 4.062 & 117.242 & 80\\ 1.213 & 10 & 3.984 & 109.820 & 75\\ 0.494 & 5 & 10.528 & 83.936 & 140\end{bmatrix}$$

$$\boldsymbol{M}_n=\begin{bmatrix}0.783 & 0.603 & 0.427 & 0.336 & 0.763\\ 0.207 & 0.302 & 0.154 & 0.429 & 0.000\\ 0.207 & 0.302 & 0.302 & 0.542 & 0.291\\ 0.508 & 0.603 & 0.297 & 0.508 & 0.273\\ 0.207 & 0.302 & 0.784 & 0.388 & 0.509\end{bmatrix}$$

4.2　计算组合权重

(1) 主观权重

设属性 P_n 与 D_n 同等重要，P_n 和 D_n 比 H 十分重要，P_n 和 D_n 比 T 的重要性介于稍微重要和明显重要之间，P_n 和 D_n 比 P 的重要性介于同等重要和明显重要之间，T 比 H 明显重要，T 比 P 和 P 比 H 都是稍微重要. 根据 1～9 标度法可得判断矩阵

$$\boldsymbol{E}=\begin{bmatrix}1 & 1 & 7 & 4 & 2\\ 1 & 1 & 7 & 4 & 2\\ 1/7 & 1/7 & 1 & 1/5 & 1/3\\ 1/4 & 1/4 & 5 & 1 & 3\\ 1/2 & 1/2 & 3 & 1/3 & 1\end{bmatrix}$$

用和法求得量纲一后的权重向量

$\boldsymbol{W}^{(1)}=(0.331,0.331,0.041,0.171,0.126)$

(2) 客观权重

对量纲一化后的距离矩阵 $\boldsymbol{M}_n$ 中的每一列求方差，得到方差向量

$\boldsymbol{\delta}=(0.067,0.027,0.057,0.007,0.082)$

对 $\boldsymbol{\delta}$ 进行量纲一处理，得客观权向量

$\boldsymbol{W}^{(2)}=(0.280,0.113,0.238,0.030,0.340)$

(3) 组合权重

利用乘法合成计算组合权向量

$\boldsymbol{W}=(0.493,0.200,0.052,0.027,0.228)$

可以看出，只考虑主观权重时，P_n 与 D_n 的权重相等，但加入客观权重后，D_n 的组合权重要小于 P_n 的组合权重，这显然更为合理. 这是由于组合权重包含了属性本身信息量对决策结果的贡献. 通过矩阵 $\boldsymbol{M}_n$ 可以看出，实例间属性 D_n 的距离之间的差异相对于属性 P_n 要小，因此属性 D_n 对检索的贡献相对较小.

4.3　C_0 与各实例的相似度

由式(1)得 C_0 与各实例间的总体距离为

$$(D_1,D_3,D_4,D_6,D_7)=\boldsymbol{M}_n\boldsymbol{W}^{\mathrm{T}}=$$
$$(0.712,0.182,0.259,0.463,0.329)\quad(10)$$

C_0 与各实例的相似度为

$$(S_1,S_3,S_4,S_6,S_7)=$$
$$(0.288,0.818,0.741,0.537,0.671)\quad(11)$$

可见，五个实例与问题实例的相似度顺序为 C_3，C_4，C_7，C_6，C_1，即实例 C_3 与新问题 C_0 最为相似，而实例 C_1 与新问题 C_0 最不相似，运算结果与通过表 1 得到的直观判断相符.

4.4　结果分析

如果没有模糊属性间距离公式，模糊数或模糊区间数属性间的距离计算只能转换为确定数及确定区间数属性间的距离计算，此时计算结果为(计算过程省略)：

$$(D_1',D_3',D_4',D_6',D_7')=$$
$$(0.580,0.180,0.348,0.446,0.341)\quad(12)$$
$$(S_1',S_3',S_4',S_6',S_7')=$$
$$(0.420,0.820,0.652,0.554,0.659)\quad(13)$$

可以看出，与问题实例 C_0 最为相似的依然是实例 C_3，但将模糊属性值转换为精确属性值意味着筛选条件更为严苛，这会因为部分信息的丢失而导致结果的不准确.

一方面，在结果(13)中 $S_4'<S_7'$，即 C_7 比 C_4 更相似于 C_0，但通过表 1 可以看出，这与人们的直观判断是不相符的，而且 S_4' 与 S_7' 只相差 0.007，分辨率不高. 而通过新方法得到结果(11)则不存在这个问题.

另一方面，当通过设定阈值来筛选相似实例时，如果阈值取 0.7，从结果(11) 可以看出，会输出 C_3 和 C_4，如果依照结果(13)，则只输出 C_3. 但通过表 1 可以看出，C_4 也有一定的参考价值，所以通过新方法得到的结果更合理准确.

5　结语

为了解决当前 CBR 系统对模糊概念无能为力、求解不精确等弊端，并使其获得更广泛应用，本文在分析已存在的相似性度量方法基础上，提出一种基于混合属性距离的相似性度量方法. 针对案例检索过程中所存在的不精确信息，基于区间数的距离计算模型和模糊集合理论，提出了模糊数及模糊区间数基于 α 级区间的距离公式，并改进了属性值为隶

属度函数时的距离公式.另外,研究了基于属性距离方差的客观赋权方法,在计算属性权重时综合考虑了主观权重和客观权重,使得组合权重不仅反映属性的特点和决策者的偏好,而且反映了属性本身信息量对实例检索的贡献.通过举例表明,采用所述模型能有效处理复杂环境下存在不精确信息的设计问题求解,且具有计算简单和结果准确的优点,从而提高了CBR系统实例检索的有效性和可靠性,是对相似度计算模型的有益探索,有着广泛的用途.

参考文献:

[1] 何新贵. 模糊数据库中的语义距离及模糊视图[J]. 计算机学报, 1989, 12(10): 757.
HE Xingui. Semantic distance and fuzzy users' view in fuzzy databases [J]. Chinese Journal of Computers, 1989, 12(10): 757.

[2] 钟诗胜, 王知行, 何新贵. 一个混合属性的实例检索模型[J]. 软件学报, 1999, 10(5):521.
ZHONG Shisheng, WANG Zhixing, HE Xingui. A model for mixed attributions cases indexing [J]. Journal of Software, 1999, 10(5):521.

[3] 王晓亮, 刘西拉. 基于事例推理系统中的模糊检索[J]. 上海交通大学学报, 2008, 41(11): 1783.
WANG Xiaoliang, LIU Xila. Fuzzy retrieval in case-based reasoning systems [J]. Journal of Shanghai Jiaotong University, 2008, 41(11): 1783.

[4] Gupta K M, Montezemi A R. Empirical evaluation of retrieval in case-based reasoning systems using modified cosine matching function [J]. IEEE Transactions on Systems, Man and Cybernetics, Part A:Systems and Humans, 1997, 27(5):601.

[5] 蒋占四, 陈立平, 罗年猛. 最近邻实例检索相似度分析[J]. 计算机集成制造系统, 2007, 13(6): 1165.
JIANG Zhansi, CHEN Liping, LUO Nianmeng. Similarity analysis in nearest-neighbor case retrieval [J]. Computer Integrated Manufacturing Systems, 2007, 13(6): 1165.

[6] 任凯, 浦金云. 基于案例属性特征区间相似度的改进算法研究[J]. 控制与决策, 2010, 25(2): 307.
REN Kai, PU Jinyun. Research on mended range attributes similarity calculation models of case-based reasoning [J]. Control and Decision, 2010, 25(2): 307.

[7] Slonim T Y, Schneider M. Design issues in fuzzy case based reasoning [J]. Fuzzy Sets and Systems, 2001, 117(2):251.

[8] 周凯波, 冯珊, 李锋. 基于案例属性特征的相似度计算模型研究 [J]. 武汉理工大学学报:信息与管理工程版, 2003, 25(1): 24.
ZHOU Kaibo, FENG Shan, LI Feng. Research of the similarity calculation models based on the features of case properties [J]. Journal of Wuhan University of Technology: Information and Management Engineering Edition, 2003, 25(1): 24.

[9] 胡启洲,张卫华.区间数理论的研究及其应用 [M].北京:科学出版社,2010.
HU Qizhou, ZHANG Weihua. Research and application of interval number theory [M]. Beijing: Science Press, 2010.

[10] 包玉娥, 彭晓芹, 赵博. 基于期望值与宽度的区间数距离及其完备性[J]. 模糊系统与数学, 2013, 27(6): 133.
BAO Yu'e, PENG Xiaoqin, ZHAO Bo. The interval number distance and completeness based on the expectation and width [J]. Fuzzy Systems and Mathematics, 2013, 27(6): 133.

[11] 曹炳元.应用模糊数学与系统 [M].北京:科学出版社,2005.
CAO Bingyuan. Fuzzy mathematics and system [M]. Beijing: Science Press, 2005.

[12] 彭祖赠,孙韫玉.模糊数学及其应用 [M]. 武汉:武汉大学出版社,2007.
PENG Zuzeng, SUN Wenyu. Fuzzy mathematics and applications [M]. Wuhan: Wuhan University Press, 2007.

[13] 严蔚敏,吴伟民.数据结构(C语言版)[M]. 北京:清华大学出版,2007.
YAN Weimin, WU Weimin. Data structure (C language version) [M]. Beijing: Tsinghua University Press, 2007.