计算机科学1998Vol. 25 No. 5

模糊数据库中近似相等的研究*)

A Study on Approximate Equality in Fuzzy Database

张师超　蒋运承

（广西师范大学数学与计算机科学系　桂林541004）

摘　要　Based on semantic distance of fuzzy attribute and the grey relational algebra theroy, the concept of grey dependency is presented in this paper. By using the concept in fuzzy database, we may obtain a method to judge whether two fuzzy data are approximately the same in their semantics. And it plays an important role in the retrieval of fuzzy database. We also discuss its applications on the fuzzy selection and fuzzy natural join.

关键词　Fuzzy database, Grey dependency, Approximate equality

近几年，国外对模糊数据库进行了研究，在能源决策及信息检索方面已有所应用。如日本和美国已把模糊数据库用于能源决策、医疗会诊等许多领域，显示了其重要的实用价值。在国内，许多学者也对模糊数据库进行了有效的研究。例如，何新贵提出了模糊关系型数据库的数据模型[1]，并引入了语义距离及模糊视图的概念[2]；刘凤玉等研究了模糊数据库的设计与实现[3]；张晏青论述了分布式模糊数据库对模糊信息检索的支持[4]等等。

特别，模糊数据库中的数据有各种不同的表示方法，从而在模糊数据库中数据之间不能用精确的相等来衡量，我们有必要对模糊数据库中数据之间的近似相等进行研究。文[2]提出了语义距离（属性之间或元组之间）的概念，用它来判断两属性或元组之间是否近似相等。其主要特点是：(1)有较好的理论基础，如切比雪夫距离、欧几里得距离等都满足距离的公理；(2)根据属性或元组之间的语义距离提出了"ε-相等"的概念，可用以构成各种选择条件。其主要的不足体现在：判断两属性或两元组是否近似相等时只单纯考虑这两个属性或元组之间的语义距离，未考虑关系中其它对应属性或元组对它们的影响。本文在模糊值语义距离的基础上，根据灰色关系代数理论提出了"灰色依赖"的概念，用它可以提供判断两属性或元组在语义上是否近似相等的方法。该方法能克服文[2]中的不足，并且在模糊数据库的检索中有很重要的作用。下面，首先叙述一下有关模糊数据库的检索问题。

一、模糊数据库中的检索

在模糊数据库中，如何进行模糊数据的检索是一个重要的研究课题，近几年国内外通过对模糊数据库的研究表明，模糊数据库在模糊信息检索、决策等方面有广阔的应用前景，它有助于提高信息检索的效率和缩小自然语言与数据库检索语言的差距，可望对智能检索给予有益的支持。下面仅举一简例。

在侦破案件中对罪犯嫌疑情况登记中可能包括"年龄"、"身高"、"体重"、"健康状况"等等栏目，其值不太可能确切填写，都是一些模糊值。例如对"年龄"的估计只能给出"老"、"中"、"青"、"少"、"大约多少岁"或"多少岁左右"等等。同样，"身高"也只能估出"高个子"、"矮个子"或"多少米左右"等。假设有一模糊数据库存放了这些模糊数据，现要对它进行检索，如查找"身高在1.62米左右，体重大约50公斤的青年人很可能是谁?"，显然这不能使用通常数据库中的检索方法。本文所提出的"灰色依赖"能有效地解决模糊数据库中的检索问题。下面将进行详细讨论。首先介绍一些模糊值间的语义距离的有关概念，详细情况请参考文[2]。

*) 本文得到国家自然科学基金和国家863计划的资助。

二、模糊值间的语义距离

2.1 模糊数的表示

模糊数有各种各样的表示方法，下面仅举几例。

1)模糊区间数表示。若 D 是一个可排序的论域，则 D 中的一个模糊子集可用模糊区间数[a,b]/p 表示。其中 a∈D,b∈D,[a,b]表示一个闭区间，称为一个区间数，表示 D 中介于 a 和 b 之间(包括 a 和 b)的所有元素，p 是一个可信度(0<p≤1)。

2)模糊中心数表示。若能在论域 D 的元素之间定义一个广义距离，则 D 上的一个模糊子集可用一个模糊中心数(c,r,p)表示，即该模糊子集落在以 c 为中心，r 为半径的"超球"之中的可信度为 p。其中 c 是 D 中的一个元素，r 是一个实数小量，0<p≤1。

3)隶属函数表示。隶属函数是由 Zadeh 提出的一种表示模糊集的方法，在论域 D 为有限集时，可用列表表示，否则可用一个函数过程来计算隶属函数。

4)模糊集合表示。论域 D 上的一个模糊子集通常可用一个序偶 S(D)/p 表示，即该模糊子集落在子集 S(D)中的可信度为 p。其中 S(D)是 D 的一个子集，p 是可信度(0<p≤1)。

2.2 模糊值间的语义距离

根据以上给出的几种模糊数表示方法，下面来讨论模糊值间的语义距离。

1. 当模糊数用隶属函数表示时，两个模糊值 $f_1(x)$与 $f_2(x)$之间的语义距离可定义为两函数之差的某种范数，即：$\|f_1(x)-f_2(x)\|$。

2. 当模糊数用模糊区间数表示时，两个模糊值 $f_1=[a_1,b_1]/p_1$，$f_2=[a_2,b_2]/p_2$之间的语义距离可定义为：$u_1*[\rho(a_1,a_2)+\rho(b_1,b_2)]+u_2*|p_1-p_2|$。其中 $u_1,u_2>0$，$u_1+u_2=1$为两个权系数，并假设论域的元素之间已定义了一种语义距离 $\rho(x,y)$。

3. 当模糊数用模糊中心数表示时，求两个模糊值(c_1,r_1,p_1)和(c_2,r_2,p_2)之间的语义距离可先把它们化为区间数$[c_1-r_1,c_1+r_1]/p_1$和$[c_2-r_2,c_2+r_2]/p_2$，再利用2求距离。

4. 求精确值 d 与区间数、中心数或隶属函数之间的语义距离时，先把精确值化为区间数[d,d,1]、中心数[d,0,1]或视为隶属函数 f(x)，其中当 x=d 时 f(x)=1，当 x≠d 时 f(x)=0，然后再计算。

三、模糊数据库中的近似相等

在通常数据库中，根据灰色系统原理而建立起来的灰色关系代数具有很重要的实用价值[3]。它的主要思想是：在关系中所有元组的影响下，通过计算两个元组之间的依赖程度来分析它们之间的不确定性关系。将灰色关系代数理论引入模糊数据库，我们可以提出"灰色依赖"的概念，用它可以提供判断两属性或两元组是否近似相等的方法。为了介绍该方法，我们分别引进属性值(模糊值)间的语义距离、灰色属性依赖值和灰色元组依赖值等概念。

设模糊关系 $R=(F(A_1),F(A_2),\cdots,F(A_n))$，$t_i$，$t_j$ 是 R 上的两个元组，它们的属性依次分别表示为：

$(t_i.a_1,t_i.a_2,\cdots,t_i.a_n)$

$(t_j.a_1,t_j.a_2,\cdots,t_j.a_n)$

下面详细地介绍如何利用灰色依赖值来判断两属性或两元组是否近似相等的具体算法。

1. 求属性值(模糊值)的语义距离。给定元组 t_i 和 t_j，则 $t_i.a_k$ 与 $t_j.a_k$ 的语义距离为：

$\Delta_j^i(k)=|t_i.a_k-t_j.a_k|\quad 1\leqslant k\leqslant n$

其中"-"运算表示第二节中所介绍的求模糊值之间的语义距离的运算。

2. 为了说明整个关系中所有元组(设有 m 个)对元组 t_i 的影响，引进符号 Δ_{max}^i及 Δ_{min}^i，分别定义如下：

$$\Delta_{max}^i=\max_{1\leqslant j\leqslant m}\ \max_{1\leqslant k\leqslant n}\ (\Delta_j^i(k))$$

$$\Delta_{min}^i=\min_{1\leqslant j\leqslant m}\ \min_{1\leqslant k\leqslant n}\ (\Delta_j^i(k))$$

显然，Δ_{max}^i表示关系中所有元组的属性对元组 t_i 的对应属性的最大语义距离；Δ_{min}^i表示关系中所有元组的属性对元组 t_i 的对应属性的最小语义距离。

3. 求灰色属性依赖值。给定元组 t_i 和 t_j，则属性 a_k 的灰色属性依赖值 $\zeta_j(k)$由下式给出：

$$\zeta_j(k)=(\Delta_{min}^i+\xi\Delta_{max}^i)/(\Delta_j^i(k)+\xi\Delta_{max}^i)$$

其中ξ是0到1之间的一个系数，它的作用是当 Δ_{max}^i太大时，可以减弱 Δ_{max}^i的影响，从而可以提高 $\Delta_j^i(k)$的语义距离的意义。

显然，$\zeta_j(k)$是一个0到1之间的数，用它可以衡量关系中在其它元组的属性的影响下，属性 $t_i.a_k$ 对属性 $t_j.a_k$ 的依赖程度。很明显，可以认为元组 t_i 的属性 a_k 与它对关系中所有元组中的属性 a_k 的依赖值的最大者近似相等。

4. 求灰色元组依赖值。对元组中所有属性的灰色属性依赖值 $\zeta_j(k)$取平均就可得到灰色元组依赖值 ζ_j，即：

$$\zeta_j=\left(\sum_{k=1}^{n}\zeta_j(k)\right)/n$$

显然，ζ_i 描述了在关系中所有元组的影响下，元组 t_i 对元组 t_j 的依赖程度，我们可以认为元组 t_i 与它对关系中所有元组的依赖值的最大者近似相等。

特别注意：一定要适当地选用各属性值（模糊值）的语义距离，使它们具有相同的度量单位，否则上述提供的判断两元组是否近似相等的公式就失去了意义。

下面举例说明该方法在模糊数据库中的具体应用。首先讨论它在模糊选择中的应用。设有一存放嫌疑罪犯有关情况的模糊关系 R_1：

	姓名	身高	体重	年龄	FS1
r_1	张三	1.70左右	大约60	年青	0.91
r_2	李四	高个子	52左右	28左右	0.93
r_3	王二	大约1.57	45左右	大约20	0.85

现又获得有关某嫌疑罪犯的记录 r 如下：

身高	体重	年龄	FS1
1.62左右	大约50	年青	0.90

问一：该嫌疑罪犯最可能是谁？

假设经计算得到：

$\zeta_1=0.8963$，$\zeta_2=0.9782$，$\zeta_3=0.6419$

因此，结果是：姓名　李四

问二：对该嫌疑罪犯的嫌疑次序如何？

由上面给出的灰色元组依赖值很容易得到结果，次序为：李四、张三、王二。

下面再讨论一下它在模糊自然连接中的应用。设有另一存放嫌疑罪犯有关情况的模糊关系 R_2：

	身高	体重	健康状况	FS2
r1′	1.68左右	50左右	一般	0.90
r2′	1.71左右	70左右	良好	0.88
r3′	高个子	大约65	良好	0.95

问：模糊关系 R_1 与模糊关系 R_2 的自然连接的结果？

假设经计算对于属性身高和体重来讲，R_1 中的 r_1 对 R_2 中的 r_2' 有最高的灰色元组依赖值；R_1 中的 r_2 对 R_2 中的 r_1' 有最高的灰色元组依赖值；R_1 中的 r_3 对 R_2 中的 r_3' 有最高的灰色元组依赖值，因此，结果是：

姓名	身高	体重	年龄	健康状况	FS
张三	1.70左右	大约60	年青	良好	0.88
李四	高个子	52左右	28左右	一般	0.90
王二	大约1.57	45左右	大约20	一般	0.85

说明：对于 FS 的取值有很多方法，在这里采用 $FS=\min\{FS1,FS2\}$。

结语　对于模糊数据库中的数据，相等应是一个模糊概念。我们在模糊值语义距离的基础上，根据灰色关系代数理论提出了"灰色依赖"的概念，并研究了用灰色依赖值（属性之间或元组之间）判断模糊数据库中两属性或两元组是否近似相等的方法。该方法的特点是：(1)不需要定义 ε-相等的概念；(2)判断两对象之间是否近似相等时，考虑了整个关系中所有对象对这两个对象的影响；(3)该算法有坚实的理论基础，即：灰色关系代数理论；(4)该方法可望有广泛的应用，而且特别适用在过去的工具无能为力的复杂对象或事物的检索中。

参 考 文 献

[1] 何新贵，模糊关系数据库的数据模型，计算机学报，1989，2

[2] 何新贵，模糊数据库中的语义距离及模糊视图，计算机学报，1989，10

[3] 刘凤玉等，模糊关系数据库 FRDBMS 的设计与实现，计算机研究与发展，1991，4

[4] 张晏青，模糊信息检索，计算机研究与发展，1988，10

[5] Ka-Wing Wong，Extension Relational Algebra and Grey Relational Algebra，ACM SIGICE Bulletin，22(4)1997