贴近度方法在考试分类系统中的研究与应用

丛涌泉,管 婷,张春海,刘 超,刘晓东
(中国海洋大学 信息科学与工程学院,山东 青岛 266000)

摘 要:结合考试分析系统的具体情况对目前常用的贴近度方法进行统一的改进,在现有方法的基础上提出了一种新的贴近度方法,并给出了构造函数,根据新的贴近度方法能够灵活构造出不同精确度的贴近度函数以满足系统在不同条件下的各种需求。通过其后的例子可以看出通过改变文中提出的贴近度方法中相应的参数,可以构造出满足不同的精确度需求的新的贴近度函数来对目标进行有效分类处理,这对于下一步考试分析系统的编程和应用有着非常重要的理论指导意义。

关键词:贴近度;考试分析系统;模糊集
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2011)01-0250-04

Study and Application of Similarity Method in Examination Analysis System

CONG Yong-quan, GUAN Ting, ZHANG Chun-hai,
LIU Chao, LIU Xiao-dong
(Dept. of Information Science and Engineering, Ocean University of China, Qingdao 266000, China)

Abstract: This paper uniformly improves the presently common-used similarity methods by combining the details in the examination analysis system. Based on the presently used methods, a new similarity method with its structured function is proposed. This new method can structure flexibly the similarity function of different degrees of accuracy in order to meet various requirements under different conditions. From the examples in the paper, it is proved that by changing the corresponding parameters in the similarity method, the new similarity function can be structured to effectively categorize and deal with the object, which has an important significance of the theoretical guideline in the further program-designing and application of the examination analysis system.

Key words: similarity; examination analysis system; fuzzy sets

0 引 言

现在考试种类越来越多,作为考试的管理机构,如何准确高效地分析不同考试之间的相关性,将各类考试进行分类,对于考试的发展和决策具有重要意义。而目前在数据处理方面,已经有了许多以数据为中心的数据分类方法和查询技术[1]。文献[2]分析并提出了一种基于聚类和二进制 PSO 算法的特征选择方法。文献[3]介绍了一种基于本体的数据挖掘技术在商务信息智能分类中的应用,并从理论上阐述了该技术的可行性。而文献[4]结合模糊集理论研究了分布式处理海量数据中的分类方法问题。为了提高在考试分类过程中信息分类的准确率,节省考试资源,文中采用模糊贴近度方法对考试信息进行分析[5,6]。文中对现有的模糊贴近度方法进行了改进[7,8],给出了一种新的贴近度函数。通过其后的例子证明了基于这种贴近度方法能够构造出较现有经典贴近度方法更适用于大规模考试分类系统的具有更高精确性的贴近度函数,而且可以根据需求不断构造出符合用户需求的不同精确程度的贴近度函数。

以数据为中心的分类和存储系统现已广泛应用于无线传感器网络的数据处理机制中。而与大规模无线传感器网络中数据处理的特点相类似,在考试分类系统中,我们的数据处理方法也需要能够快速准确地在大量不同类的数据中分辨出到各类信息的相关性。但是由于考试分类中所处理数据的特点(属性少、不确定性强及分类模糊),现有经典的贴近度方法并不能很好的适用于本系统。而之前常用的许多模糊贴近度方法,如由 Li Dengfeng 等提出的贴近度方法[9]也没有达到本系统的要求。因此如何构造更精确的分类方

收稿日期:2010-05-23;**修回日期**:2010-08-20
基金项目:青岛市科技计划基金项目(08-1-3-2-jch)
作者简介:丛涌泉(1985-),男,山东德州人,硕士研究生,研究方向为数据库理论;张春海,教授,研究方向为数据库理论。

法来对数据进行精确的区分成为考试分类系统的核心问题。为了适应今后工作中对于不容类型数据快速准确分类的不同要求,文中提出了一种可以构造更高精确度的模糊贴近度方法。

1　区间值模糊集相关理论

目前,国内外对区间值模糊集的研究正在兴起,其中 Gehrke Mai, C. Walker, E. Walker. I. B. Turken, M. B. Gorzalczany 等学者对区间值模糊集作了研究[10],国内的李凡,邢汉承,裴道武等学者对此作了大量的工作[11]。在相似度量方面,已有了多种贴近度度量算法。

下面着重介绍区间值模糊集合相关的概念,首先介绍区间数的概念:

定义 1.1[2] 称 [0,1] 中的闭区间

$\underline{a} = [a^-, a^+]$ $(0 \le a^- \le a^+ \le 1)$

为[0,1]上的区间数,[0,1]上的区间数全体记为 L。在 L 中规定序关系(也用"$\le$"表示):

$\underline{a} \le \underline{b} \Leftrightarrow a^- \le b^-$ 且 $a^+ \le b^+$

易证"$\le$"是 L 中的偏序关系。$< L, "\le" >$ 是一个完备格,其中

$\underline{a} \vee \underline{b} = [a^- \vee b^-, a^+ \vee b^+]$

$\underline{a} \wedge \underline{b} = [a^- \wedge b^-, a^+ \wedge b^+]$

$\bigvee\limits_{i \in I} \underline{a}_i = [\bigvee\limits_{i \in I} a_i^-, \bigvee\limits_{i \in I} a_i^+]$

$\bigwedge\limits_{i \in I} \underline{a}_i = [\bigwedge\limits_{i \in I} a_i^-, \bigwedge\limits_{i \in I} a_i^+]$

且 $< L, "\le" >$ 中最大元为 $1 = [1,1]$,最小元为 $0 = [0,0]$。

定义 $< L, "\le" >$ 上运算:"C" 如下:

$\underline{a}^c = [1 - a^+, 1 - a^-]$

易证"C"具有如下性质:

① $(\underline{a}^c)^c = \underline{a}$

② $(\underline{a} \vee \underline{b})^c = \underline{a}^c \wedge \underline{b}^c$, $(\underline{a} \wedge \underline{b})^c = \underline{a}^c \vee \underline{b}^c$

③ $(\bigvee\limits_{i \in I} \underline{a}_i)^c = \bigwedge\limits_{i \in I} \underline{a}_i^c$, $(\bigwedge\limits_{i \in I} \underline{a}_i)^c = \bigvee\limits_{i \in I} \underline{a}_i^c$

④若 $\underline{a} \le \underline{b}$,则 $\underline{a}^c \ge \underline{b}^c$ $\forall \underline{a}, \underline{b} \in L$

即"C"是 $< L, "\le" >$ 上的余运算。

区间值模糊集的定义如下:

定义 1.2[12] 设:

$A = \{ < x, \mu_A(x) >, x \in X \}$

其中

$\mu_A: X \to L$

且满足条件:

$0 \le \mu_A(x) \le 1$ $(\forall x \in X)$

称 A 是 X 上的区间值模糊集。

简记作:$IVFS(X)$。

经典模糊集合的定义如下:

定义 1.3.[12] 设:

$A = \{ < x, \mu_A(x) >, x \in X \}$

其中

$\mu_A: X \to [0,1]$

且满足条件:

$0 \le \mu_A(x) \le 1$ $(\forall x \in X)$

称 A 是 X 上的模糊集。

简记作:$FS(X)$。

不难看出,通常所说的模糊集是区间值模糊集的一种特例,可以看作是区间值模糊集的子集。因此,在区间值模糊集中讨论的问题对模糊集和分明集都有非常重要的指导意义。

下面来构造一种基于距离的新的贴近度方法作为系统的理论支持。

2　贴近度方法

首先给出区间值模糊集中贴近度的定义:

定义 2.1 $S: IVFS(X) \times IVFS(X) \to [0,1]$ 称作贴近度当且仅当 $S(A,B)$ $A, B \in IVFS(X)$ 满足:

1) $0 \le S(A,B) \le 1$

2) 当 $A = B$ 时 $S(A,B) = 1$

3) $S(A,B) = S(B,A)$

4) 当 $A \le B \le C, A, B, C \in IVIFS(X)$ 时 $S(A,C) \le S(A,B)$ 且 $S(A,C) \le S(B,C)$

接下来在区间值模糊集中构造一类新的贴近度函数:

定义 2.2:函数 $S(A,B) = 1 - V(Q_1, Q_2, \ldots, Q_n))$ 其中 $A, B \in IVFS(X)$

$$Q_i = \{E_i^+ [\varphi_i^+ (|\mu_A^+ - \mu_B^+|)], E_i^- [\varphi_i^- (|\mu_A^- - \mu_B^-|)]\} \quad (1)$$

其中

(1) $V: [0,1]^n \to [0,1]$ 且满足

1) $x_1 = x_2 = \ldots = x_n = 0 \Leftrightarrow V(x_1, x_2, \ldots, x_n) = 0$

2) 对于每个变元 $x_i (i = 1,2,\ldots,n)$, $V(x_1, x_2, \ldots, x_n)$ 均为单增函数

(2) $Q_i: [0,1] \times [0,1] \to [0,1] (i = 1,2,\ldots,n)$ 且满足:

1) $Q_i(x,y) = 0 \Leftrightarrow x = y = 0$

2) 对于每个变元 x, y 均为单增函数

(3) $E_i^-, E_i^+: [0,1] \to [0,1] (i = 1,2,\ldots,n)$ 且满足

1) $E_i^-(x) = 0 \Leftrightarrow x = 0 \Leftrightarrow E_i^+(x) = 0$

2) 均为单增函数

(4) $\varphi_i^-, \varphi_i^+: [0,1] \to [0,1] (i = 1,2,\ldots,n)$ 且

满足

1) $\varphi_i^-(x)=0\Leftrightarrow\varphi_i^+(x)=0\Leftrightarrow x=0$

2) φ_i^-,φ_i^+ 均为单增函数

定理 1:定义 2.2 中的函数 $S(A,B)$ 是贴近度函数。

证明:

1) 由条件 1) 知 $0\leq 1-V(Q_1,Q_2,\ldots,Q_n)\leq 1$

即

$0\leq S(A,B)\leq 1$

2) 设 $A,B\in IVFS(X)$,若 $A=B$,则

$\mu_A^-=\mu_B^-,\mu_A^+=\mu_B^+$

由定义 2.2 知 $W,Q_i,E_i^-,E_i^+,\varphi_i^-,\varphi_i^+$ 均满足各自条件 1) 由此可证

$A=B\Leftrightarrow S(A,B)=1$

3) 由定义 2.2 简单可证 $S(A,B)=S(B,A)$。

4) 设 $A\leq B\leq C,A,B,C\in VIFS(X)$

$0\geq\mu_A^--\mu_B^-\geq\mu_A^--\mu_C^-$

$0\geq\mu_A^+-\mu_B^+\geq\mu_A^+-\mu_C^+$

因此

$|\mu_A^--\mu_C^-|\geq|\mu_A^--\mu_B^-|$

$|\mu_A^+-\mu_C^+|\geq|\mu_A^+-\mu_B^+|$

由定义 2.2 知 φ_i^-,φ_i^+ 均满足其条件 2),因此

$\varphi_i^+(|\mu_A^+-\mu_B^+|)\leq\varphi_i^+(|\mu_A^+-\mu_C^+|)$

$\varphi_i^-(|\mu_A^--\mu_B^-|)\leq\varphi_i^-(|\mu_A^--\mu_C^-|)$

由定义 2.2 知 E_i^-,E_i^+ 均满足其条件 2),因此

$E_i^-[\varphi_i^-(|\mu_A^--\mu_B^-|)]\leq E_i^-[\varphi_i^-(|\mu_A^--\mu_C^-|)]$

$E_i^+[\varphi_i^+(|\mu_A^+-\mu_B^+|)]\leq E_i^+[\varphi_i^+(|\mu_A^+-\mu_C^+|)]$

由定义 2.2 知 Q 和 V 均满足其条件 2),因此

$S(A,B)\geq S(A,C)$

综上可知,$S(A,B)$ 是贴近度函数。

下一节根据该贴近度函数构造出一些不同精确度的贴近度函数,并举例说明其有效性。

3 举例分析

本节通过举例来说明新的贴近度方法在考试分类系统中的作用:

例 3.1:

假设共有 3 个考试的考生信息集合:A1;A2;A3,每个集合主要包括考生的 3 种不同属性(a1,a2,a3)。

其中

a1:考生年龄(与平均年龄的差异程度,取值范围都为 0 到 1)

a2:考生学历(与平均学历的差异程度,取值范围都为 0 到 1)

a3:考生性质(与平均指标的差异程度,取值范围都为 0 到 1)

设 3 个集合的属性值如表 1 所示。

表 1 各集合的属性值

集合 \ 属性	a1	a2	a3
A1	0.2~0.8	0.5~0.9	0.1~0.1
A2	0.5~0.5	0.6~0.7	0~0.2
A3	0.7~0.8	0.1~0.2	0.4~0.6

为便于分析,写成如下形式:

$A_1=\{(0.2,0.8),(0.5,0.9),(0.1,0.1)\}$

$A_2=\{(0.5,0.5),(0.6,0.7),(0,0.2)\}$

$A_3=\{(0.7,0.8),(0.1,0.2),(0.4,0.6)\}$

现在来考虑本次考试中考生信息集合 $B=\{(0.4,0.6),(0.6,0.8),(0,0.2)\}$ 与三次考试的考生信息集合 A_1,A_2,A_3 的所属关系。

首先,以经典的贴近度函数,由 Li Dengfeng 提出的贴近度为例进行比较。其经典的贴近度函数如下:

$$s(A,B)=1-\frac{1}{\sqrt[p]{n}}\left\{\sum_{i=1}^{n}\left[\frac{[\mu_A(x_i)-\mu_B(x_i)]}{2}\right]^p\right\}^{\frac{1}{p}} \tag{2}$$

令 $p=1$,则用(2)式进行分析所获得分类结果为:

$s(A_1,B)=1,s(A_2,B)=1,s(A_3,B)=0.6$

所获得的分类结果是只有第三个集合和要分析的考试相关性不大。显然这种结果并不够理想,这需构造更精确的方法来进行区分。根据这一情况,根据(1)式再来构造一种更精确的贴近度函数再进行尝试。

令

1) $V=\left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]^{\frac{1}{p}}$

2) $Q_i(x,y)=\left(\frac{x+y}{2}\right)^p$

3) $E_i^+(x)=\frac{x}{2}=E_i^-(x)$

4) $\varphi_i^-(x)=\varphi_i^+(x)=x$

则

$$S(A,B)=1-\frac{1}{\sqrt[p]{n}}\left\{\sum_{i=1}^{n}\frac{1}{2^p}\left[\frac{|\mu_A^-(x_i)-\mu_B^-(x_i)|}{2}+\frac{|\mu_A^+(x_i)-\mu_B^+(x_i)|}{2}\right]^p\right\}^{\frac{1}{p}} \tag{3}$$

同样令 $p=1$,则用(3)式进行分类所获得结果为:

$s(A_1,B)=0.83,s(A_2,B)=0.93,s(A_3,B)=0.6$

可见使用新构造的这种方法使本次考试与其他三次考试的考生信息集合 A1,A2,A3 的不同相关得到

了反映。我们可以按这个分类结果对考试进行调整。

由以上例子可以看出，我们的贴近度方法对于在应用过程中构造不同精确的贴近度函数有重要的理论意义。可以以此为平台构造满足需要的不同精确度的贴近度函数。

例3.2：在分明集合中再举一例：

设之前来那个次考试中考生信息集合的不同属性值如下：

$A_1 = \{(0.2,0.2),(0.2,0.2),(0.2,0.2)\}$

$A_2 = \{(0.4,0.4),(0.4,0.4),(0.4,0.4)\}$

假设需要判断出本次考试的考生信息集合 $B = \{(0.3,0.3),(0.3,0.3),(0.1,0.3)\}$ 与已有的两个集合 A_1, A_2 的隶属关系。

首先，仍然先以经典的贴近度函数为例进行比较。令 $p=1$，则用(2)式进行分析所获得结果为：

$s(A_1,B) = 0.97$，$s(A_2,B) = 0.97$

显然，由此方法所得结果无效。

再考虑用式(3)的方法来进行分析。

当对于任意 $i \in \{1,2,\ldots,n\}$ 均有 $\mu_A^-(x_i) = \mu_A^+(x_i) = \mu_A(x_i)$ 时，

$$S(A,B) = 1 - \frac{1}{\sqrt[p]{n}}\left\{\sum_{i=1}^{n}\left[\frac{|\mu_A(x_i) - \mu_B(x_i)|}{2}\right]^p\right\}^{\frac{1}{p}}$$

同样令 $p=1$，则进行分类所获得结果为：

$s(A_1,B) = 0.9$，$s(A_2,B) = 0.876$

因此，得知本次考试与其他两次考试的关联程度为：与 A1 集合所属类别考试的关联程度较高（达0.9），与 A2 集合所属类别考试的关联程度较低（为0.83），据此可以区分开本次考试与其他两次考试的关系。可见，(3)式可以进行考生信息的区分。但是从结果不难看出，按(3)式进行分类的效果并不是非常理想，目标集合与两个类集的近似关系相差不到0.5，并没达到通常意义上的有效区分。

据此，根据新的情况，尝试着根据式(1)进一步构造一种精确度更高的贴近度函数。

令

1) $V = \frac{1}{n}\sum_{i=1}^{n} x_i$

2) $Q_i(x,y) = \frac{x+y}{2}$

3) $E_i^+(x) = \max(x) = E_i^-(x)$

4) $\varphi_i^-(x) = \varphi_i^+(x) = x$

则

$$S(A,B) = 1 - \frac{1}{2n}\sum_{i=1}^{n}[\max(|\mu_A^-(x_i) - \mu_B^-(x_i)|) + \max(|\mu_A^+(x_i) - \mu_B^+(x_i)|)] \quad (4)$$

用(4)式进行分类所获得结果为：

$s(A_1,B) = 0.9$，$s(A_2,B) = 0.833$

因此，得知本次考试与其他三次考试的关联程度为：与 A2 集合所属类别考试的关联程度最高（达0.9），与 A1 集合所属类别考试的关联程度较低（为0.833），这个结果达到了预期的目标，其精确度符合条件要求，可以按这个分类结果对目标进行相应的调整。

由本例不难看出，通过改变我们提出的贴近度方法中相应的参数，可以构造出满足不同的精确度需求的新的贴近度函数来对目标进行有效分类处理，这对于考试分析系统非常重要。

4 结束语

为了给考试分类系统设计一个高效的数据处理技术，作为今后系统设计的理论基础，着重提出了一类新的贴近度方法，并给出了构造函数。通过其后的例子证明了这种贴近度方法与现有方法相比，能够更精确地进行数据的分类处理。

参考文献：

[1] 刘大健. 模糊模式识别在模拟驾驶系统中的应用研究[D]. 杭州：浙江大学，2004：1-14.

[2] 张家柏，王小玲. 基于聚类和二进制 PSO 的特征选择[J]. 计算机技术与发展，2010，20(6)：31-35.

[3] 宋远芳. 基于本体的数据挖掘技术在商务智能中的应用[J]. 计算机技术与发展，2009，19(1)：184-186.

[4] 夏奇思，王汝传. 基于属性约简的粗糙集海量数据分割算法研究[J]. 计算机技术与发展，2010，20(4)：33-36.

[5] 柴造坡. 基于相似关系的变精度粗糙集的数据约简[J]. 哈尔滨师范大学自然科学学报 2009(4)：18-21.

[6] 陈　健，赵跃龙. 变精度粗糙集在手术诊断中的应用[J]. 闽江学院学报，2007，28(5)：39-42.

[7] Zadeh L A. Fuzzy sets[J]. Information and Control，1965，8：338-353.

[8] 秦克云，徐　扬. L 型直觉模糊集[J]. 兰州大学学报，模糊专集，1996，32：352-355.

[9] Li Deng-Feng. Some measures of dissimilarity in intuitionistic fuzzy structures[J]. Journal of Computer and System Sciences，2004(1)：115-122.

[10] Atanassov K T. New operations defined over the intuitionistic fuzzy sets[J]. Fuzzy Sets and Systems，1994，61：137-142.

[11] 段中兴，张德运. 基于误码率的模糊加权无线网络公平调度算法[J]. 西安交通大学学报，2005，39(12)：1303-1306.

[12] 罗承忠. 模糊集引论(上册)[M]. 北京：北京师大出版社，1989.