Vague集相似度量*)

朱振国[1,2]　王国胤[2]
(重庆交通大学计算机及信息学院　重庆 400074)[1]
(重庆邮电大学计算机科学与技术研究所　重庆 400065)[2]

摘　要　在不确定信息处理中,判定两个知识模式的相似度是知识划分、规则推理的前提。Vague集是处理模糊信息的一种有效工具,众多学者提出了基于Vague集的相似度量方法,但这些方法不足以准确描述Vague集的相似本质,并且度量的准确度较低。本文提出了一种度量Vague集相似度量方法的标准,为研究Vague集相似度量提供了参考;提出了一种新的Vague集相似度量的方法。实验结果表明,本文方法的区分能力高,并且有更好的度量效果。

关键词　Vague(值)集,相似度量,区分能力,模糊集

Similarity Measure of Vague Set

ZHU Zhen-guo[1,2]　WANG Guo-yin[2]
(School of Computer and Information, Chongqing Jiaotong University, Chongqing 400074, China)[1]
(Institute of Computer Science & Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[2]

Abstract　Vague set is a valid tool for processing uncertain information. The similarity measure of two uncertain patterns is important for intelligent reasoning. It is also a key problem to measure the similarity of vague values or vague sets in vague information processing systems. Many methods for similarity measure of vague sets have been proposed in recent years. However, these methods could not precisely describe the essences of the similarity between two vague sets. In this paper, some evaluation criterions for similarity measure of vague set is proposed by analyzing existed similarity measure methods and evaluation criterion. For measuring the similarity of vague sets, a new method is proposed. A new concept named differentiation ability is also developed. Simulation results show that our methods are better than existed methods.

Keywords　Vague value, Vague set, Similarity measure, Evaluation criterion, Distinguishing ability, Fuzzy sets

1　引言

现实世界中,人们所面对的信息往往是不精确、不确定的。如何描述不确定数据,以及利用不确定数据进行不确定性推理,一直是智能信息处理的关键问题。Zadeh于1965年提出的模糊集理论中[1],模糊集的单隶属度 $u_A(x)$ 描述对象属于某个集合的程度,它既包含了支持论域对象 x 的证据,也包含了反对 x 的证据,但它不能同时表示支持和反对 x 的证据。为了克服这个不足,Gau和Buehrer于1993年提出了Vague集理论[2]。Vague集提出了从真假(正反)两个方面对研究对象进行描述。在一个Vague集 V 中,用一个真隶属度函数 t_x 和一个假隶属度函数 f_x 表述其隶属度的界,它们分别表示对象 x 属于 V 的支持证据和反对证据的程度。这两个界构成在[0,1]中的一个子区间$[t_x, 1-f_x]$。例如,$[t_x, 1-f_x]=[0.4, 0.8]$,在Vague集中可解释为:对象 x 属于 V 的程度为0.4,不属于V的程度为0.2,不确定度为0.4。显然,用模糊集是无法表示和处理这类模糊信息的。比较之下,Vague集可以比Fuzzy集更好更准确地表达事物间的正反两方面的模糊性质,有利于对模糊信息的正确分析。

在不确定信息处理的过程中,经常要将模糊知识进行比较、匹配、分类、检索,以判定两个知识模式是否完全一致或近似一致。如果两者完全一致,或者虽不完全一致,但两者之间的相似程度落在给定的阈值内,就称这两个知识模式是匹配的,否则称为不匹配。因此,度量Vague集(值)之间的相似程度不仅有理论上的重要意义,而且有广泛的应用需求,是处理Vague集信息及其应用的基础。针对Vague集的知识相似度量这一问题,众多学者提出了多种Vague集的相似度量方法[3-19]。本文通过对这些方法的分析,提出Vague集相似度量区分能力的概念,并给出一种新的度量方法。实验结果表明,本文的方法具有更好的相似度量区分能力和度量效果。

2　Vague集基本概念

定义1[2]　设 U 是一个论域,对 U 的任一元素 x,U 中的一个Vague集 V 用一个真隶属函数 $t_V(x)$ 和一个假隶属函数 $f_V(x)$ 表示,$t_V(x)$ 是从支持 x 的证据所导出的 x 的隶属度下界,$f_V(x)$ 则是从反对 x 的证据所导出的 x 的否定隶属度下界,$t_V(x)$ 和 $f_V(x)$ 将区间[0,1]中的一个实数与 U 中的一个点联系起来,即:$t_V: U\to[0,1]$,$f_V: U\to[0,1]$,x 关于 V 的隶属度 $\mu_V(x)$ 记为:$[t_V(x), 1-f_V(x)]$,其中,$t_V(x)+f_V(x)\leqslant 1$。

*)基金项目:国家自然科学基金(No. 60573068, No: 60373111),重庆市教委科学技术研究项目(040505),新世纪优秀人才支持计划,重庆市自然科学基金重点项目(2005BA2003)。朱振国　硕士,讲师,主要从事模式识别、数据挖掘、网络安全;王国胤　博导,教授,主要从事Rough Set、神经网络、知识获取、数据挖掘、网络安全。

设 V 为一 Vague 集，$x\in U$，当 U 连续时，记 $V=\int[t_V(x),1-f_V(x)]/x,x\in U$；当 U 离散时，记 $V=\sum_{i=1}^{n}[t_V(x_i),1-f_V(x_i)]/x_i,x_i\in U$。

特别地，当 $t_V(x)=1-f_V(x)$ 时，Vague 集退化为模糊集；当 $t_V(x)$ 和 $1-f_V(x)$ 同时为 0 或 1 时，Vague 集退化为经典集合。为讨论方便，记 $t_V(x)$ 为 t_x，记 $f_V(x)$ 为 f_x。

定义 2[2] 设 A,B 是定义在论域 U 上的两个 Vague 集，A 和 B 是相等的，即 $A=B$，当且仅当 $\forall u\in U,t_A(u)=t_B(u),1-f_A(u)=1-f_B(u)$。

定义 3[2] 设 A,B 是定义在论域 U 上的两个 Vague 集，A 包含于 B，即 $A\subseteq B$，当且仅当 $\forall u\in U,t_A(u)\leqslant t_B(u),1-f_A(u)\leqslant 1-f_B(u)$。

定义 4[3] 假定 $0\leqslant t_x\leqslant 1-f_x\leqslant 1$，$x$ 关于 V 的隶属度 $\mu_V(x)$ 记为 $[t_x,1-f_x]$，则 Vague 值 x 可被分成三部分：真隶属度部分为 t_x，假隶属度部分为 f_x，未知部分即未知度为 $\Delta x=1-t_x-f_x$。

定义 5[3] 设 A 是定义在论域 U 上的一个 Vague 集，称 $s_A(x)=t_A(x)-f_A(x)$ 为 x 的 Vague 核。显然 $s_A(x)\in[-1,1]$，表征了现有证据对元素 x 的支持和反对两种趋势的对比。$s_A(x)$ 大于 0，说明 x 属于 A 的程度比 x 不属于 A 的程度大；$s_A(x)$ 为 0，说明 x 属于 A 的程度和 x 不属于 A 的程度相当；$s_A(x)$ 小于 0，说明 x 属于 A 的程度比 x 不属于 A 的程度小。

为了设计合适的相似性度量方法，本文提出如下的端点距离和相似度量区分能力的概念。

定义 6 设 x 是定义在论域 U 上的一个 Vague 集的某个 Vague 值，则其真隶属度函数 t_x 和假隶属度函数 f_x 在 $[0,1]$ 中构成子区间 $[t_x,1-f_x]$，称 $t_x,1-f_x$ 为 Vague 值 x 的端点。设 y 是定义在论域 U 上的一个 Vague 集的另一个 Vague 值，称 $|t_x-t_y|,|f_x-f_y|$ 为 Vague 值 x 和 y 的端点距离。

定义 7 设 V_1,V_2 是定义在论域 U 上的 Vague 集，$V_1=\{x_i\mid i=1,\cdots,n\},V_2=\{y_i\mid i=1,\cdots,m\}$，令数据集合 $S=\{(x,y)\mid\forall x\in V_1,\forall y\in V_2\}$。用相似度量方法 M 度量数据集 S 的相似度，根据度量结果把数据集 S 划分为 k 类（每类的值相同），则称相似度量方法 M 对于数据集合 S 的相似度量区分能力定义为：

$$Q_M(S)=k/(n\times m)\times 100\%$$

$Q_M(S)$ 的意义是显然的，在于相同的数据集，k 越大，$Q_M(S)$ 就越大，表明相似度量方法 M 在此数据集上相似度量区分能力越好。

3 Vague 值（集）相似度量的要素及度量方法的评价标准

3.1 Vague 值（集）相似度量的必要因素

Vague 值是在 $[0,1]$ 范围内的一个子区间，两个 Vague 值的相似性比较实质上是区间范围相近程度的度量。设 x,y 的 Vague 值分别为 $[t_x,1-f_x],[t_y,1-f_y]$，根据定义[2]，若 $t_x=t_y$ 且 $f_x=f_y$，则称 x 和 y 相等，即 x 和 y 完全相似。对于度量 Vague 值的相似性，我们认为应该综合考虑以下几点：

1. Vague 值区间两端的距离

从直观上看，Vague 值区间两端的距离越接近，说明它们的相似程度越大。例如对于 Vague 值 $a=[0.1,0.6],b=[0.1,0.7],c=[0.1,1.0]$，Vague 值 a 和 b 的两端距离分别是 0 和 0.1，而 a 和 c 的两端距离分别是 0 和 0.4，所以，我们肯定会得出 a 和 b 的相似度一定大于 a 和 c 的相似度的结论。

2. Vague 值的核的距离

根据定义[5]，$s(x)=t_x-f_x$ 为 Vague 值的核，它表征现有证据对元素 x 的支持和反对两种趋势的对比。如果两个 Vague 值的核相近，说明它们在支持和反对这两种趋势上是相近的。所以，Vague 值的核同样是 Vague 值相似度量的重要因素。

3. Vague 值未知度的距离

根据定义[4]，未知度 $\Delta x=1-t_x-f_x$ 说明了 Vague 值的不确定（未知）的信息。这部分的信息对表征两个 Vague 值的相似度也是很有意义的。如果一个数据的未知度大而另一个的小，则应该得出它们的相似度小的结论。

4. Vague 值的未知部分在支持、反对和中立这三个立场的比例

这个比例应该说是无法知道的数据，所以和以上 3 点不同，这个因素不是必备因素。比如 10 个人投票，1 人赞成 1 人反对 8 人中立，我们在没有征求本人意见的时候，是无法得知 8 个表示中立的人对于赞成或反对的倾向。常见的是用比例法分解 Vague 值的未知量。就比例法而言，根据不同的权重也有很多种划分方法。设某个 Vague 值 x，用 $t_x,f_x,\Delta x$ 分别表示支持、反对和未知的程度。文献[8]用三维表示法得到新的支持度，反对度和未知度，新的支持度为 $t_x+t_x\times\Delta x$，新的反对度为 $f_x+f_x\times\Delta x$，新的未知度为 $\Delta x\times\Delta x$。

根据以上分析，Vague 值区间两端的距离，核距离，未知度距离应该是度量 Vague 集相似度的必要因素。也就是说，如果一种 Vague 集相似度量的方法没有包含上述的必要因素，那么这种方法就不一定合理。而 Vague 值的未知部分在支持、反对和中立这三个立场的比例是个参考因素，关键是采用什么样的方法分离未知部分。

3.2 Vague 值（集）相似度量方法的评价标准

目前，很多学者对 Vague 值（集）间的相似度量进行了研究[3-19]。在这里，我们将这些度量方法按使用的参数是否固定，可以分为一般相似度量方法和可变参数度量方法两大类。一般相似度量方法参数固定或者不用人为设定参数，在 Vague 值的基础上直接计算。可变参数度量方法在使用时需要用户指定相应的权值，不同的权值得到的结果相差很大。所以，一般相似度量方法和加权度量方法不便于从相似度量值的角度比较。

不论哪一类度量方法，每一种相似性度量方法都可以用自己的标准来衡量 Vague 值（集）的相似性。对于两个 Vague 值（集）的相似度比较，不能单纯地用各种方法所得的值相比较，进而得出哪个度量方法好或哪个度量方法差的结论，因为每种度量方法的值都各成体系，不同方法的值并不能说明问题。比如 $x=[0.3,0.7],y=[0.4,0.6]$ 这两个 Vague 值，用 M_h[5] 方法得 0.9，用 M_{lf}[6] 方法得 0.95，我们不能说 M_h 的结果一定比 M_{lf} 的结果好。

文献[13,15,17,19]中给出了 Vague 值（集）的相似性度量的定义或度量方法必须具有的性质，但我们认为，Vague 值（集）的相似性度量方法应该满足如下标准：

(B1) 该方法必须包含衡量 Vague 值（集）相似的必要参数：Vague 值区间两端的距离 $|t_x-t_y|$ 和 $|f_x-f_y|$，核距离 $|S_x-S_y|$，未知度距离 $|\Delta x-\Delta y|$。

(B2)该方法必须具有良好的相似度量区分能力。

(B3)该方法必须满足如下定理1中Vague值(集)相似度量的基本性质。

定理1 设$VS_S(X)$表示X上的Vague集全体,X,Y,Z是论域U中的Vague值,则Vague值(集)相似度量公式M应满足如下性质。

(P1)$0\leqslant M(X,Y)\leqslant 1$(有界性);

(P2)$M(X,Y)=1$当且仅当$X=Y$(规一性);

(P3)$M(X,Y)=M(Y,X)$(对称性);

(P4)$M(X,Y)=0$当且仅当$V_X(x_i)=[1,1],V_Y(x_i)=[0,0]$或$V_X(x_i)=[0,0],V_Y(x_i)=[1,1]$;(特征点)

$$M_Z(x,y)=1-\frac{(2-t_x-t_y)|t_x-t_y|+(2-f_x-f_y)|f_x-f_y|+|S_x-S_y|+|\Delta x-\Delta y|}{(2-t_x-t_y)+(2-f_x-f_y)+2},$$

令$t_x+t_y=p,t_x-t_y=q,f_x+f_y=m,f_x-f_y=n$,则,

$$M_Z(x,y)=1-\frac{(2-p)|q|+(2-m)|n|+|q-n|+|-q-n|}{(2-p)+(2-m)+2}$$

$M_z(x,y)$包括了3.1节所讨论的度量Vague值(集)的必要参数,即Vague值区间两端的距离$|t_x-t_y|$和$|f_x-f_y|$,核距离$|S_x-S_y|$,未知度距离$|\Delta x-\Delta y|$。同时还考虑了确切的支持信息和反对信息所占的比重,并对未知信息给予了足够的重视。公式中系数$(2-t_x-t_y)$可以转化为$(\Delta x+\Delta y+f_x+f_y)$,即Vague值$x$和$y$的未知信息和确切的反对信息对$|t_x-t_y|$有重要的影响,当$(\Delta x+\Delta y+f_x+f_y)$比较大时,$|t_x-t_y|$的影响力就应该减小,此时Vague值$x$和$y$的相似度就小;当$(\Delta x+\Delta y+f_x+f_y)$比较小时,$|t_x-t_y|$的影响力就应该增大,此时Vague值$x$和$y$的相似度就大。对于系数$(2-f_x-f_y)$有同样类似的理解。

以下讨论$M_Z(x,y)$的性质,证明$M_Z(x,y)$满足定理1。

(P1) $0\leqslant M_z(X,Y)\leqslant 1$

由Vague集定义知:$0\leqslant|t_x-t_y|\leqslant 1,0\leqslant|f_x-f_y|\leqslant 1,0\leqslant|\Delta(X)-\Delta(Y)|\leqslant 1,0\leqslant|s_x-s_y|\leqslant 2$。当$|t_x-t_y|=0,|f_x-f_y|=0$时,核距离$|S_x-S_y|=0$,未知度距离

$$\frac{(2-t_x-t_y)|t_x-t_y|+(2-f_x-f_y)|f_x-f_y|+|S_x-S_y|+|\Delta x-\Delta y|}{(2-t_x-t_y)+(2-f_x-f_y)+2}=0$$

进而可得$|t_x-t_y|=0,|f_x-f_y|=0$,由此得出$t_x=t_y,f_x=f_y$,所以得到$x=y$。

当$x=y$时可得$t_x=t_y,f_x=f_y$,进而可得$|t_x-t_y|=0,|f_x-f_y|=0,|S_x-S_y|=0,|\Delta x-\Delta y|=0$,所以$M_Z(x,y)=1$。

$$\frac{(2-t_x-t_y)|t_x-t_y|+(2-f_x-f_y)|f_x-f_y|+|S_x-S_y|+|\Delta x-\Delta y|}{(2-t_x-t_y)+(2-f_x-f_y)+2}=1$$

根据(P1)的证明可知,只有当$|t_x-t_y|=1$且$|f_x-f_y|=1$时,才能出现这样的结果,由$|t_x-t_y|=1,|f_x-f_y|=1$可以得出Vague值x和y仅有的两种组合,即$x=[1,1],y=[0,0]$或$x=[0,0],y=[1,1]$。

当$x=[1,1],y=[0,0]$或$x=[0,0],y=[1,1]$时,根据(P1)的证明可知$M_Z(x,y)=0$。

(P5)若$X\subseteq Y\subseteq Z$,则$M(X,Z)\leqslant\min\{M(X,Y),M(Y,Z)\}$

由于$x\subseteq y\subseteq z$,由定义[2]有$t_x\leqslant t_y\leqslant t_z,f_x\leqslant f_y\leqslant f_x$,

而$|S_x-S_y|=|(t_x-f_x)-(t_y-f_y)|=t_y-t_x+f_x-f_y$,

$|\Delta x-\Delta y|=|(1-t_x-f_x)-(1-t_y-f_y)|=|(t_y-t_x)-(f_x-f_y)|$,

$|\Delta x-\Delta y|=0$,由此可得

$$M_Z(x,y)\leqslant 1-\frac{(2-t_x-t_y)\times 0+(2-f_x-f_y)\times 0+0+0}{(2-t_x-t_y)+(2-f_x-f_y)+2}=1$$

当$|t_x-t_y|=1$时可得$t_x=0,t_y=1$或者$t_x=1,t_y=0$。当$|f_x-f_y|=1$时可得$f_x=0,f_y=1$或者$f_x=1,f_y=0$。在这种情况下有4种组合方式,由Vague集定义可知,只有$t_x=0,t_y=1,f_x=1,f_y=0$和$t_x=1,t_y=0,f_x=0,f_y=1$这两种组合是有意义的,其他两种组合使得$t_x+f_x=2$或者$t_y+f_y=2$,这是不符合Vague集的性质的。当$t_x=0,t_y=1,f_x=1,f_y=0$或者$t_x=1,t_y=0,f_x=0,f_y=1$时,可知$|s_x-s_y|+|\Delta(X)-\Delta(Y)|=2$。由此可知当$|t_x-t_y|=1,|f_x-f_y|=1$时,$|s_x-s_y|+|\Delta(X)-\Delta(Y)|=2$,即它们同时达到最大值。所以,$0\leqslant|s_x-s_y|+|\Delta(X)-\Delta(Y)|\leqslant 2$,可得$M_Z(x,y)\geqslant 1-\frac{(2-t_x-t_y)+(2-f_x-f_y)+2}{(2-t_x-t_y)+(2-f_x-f_y)+2}=0$

$0\leqslant M_Z(x,y)\leqslant 1$证毕。

(P2) $M(X,Y)=1$当且仅当$X=Y$

当$M_Z(x,y)=1$时,可得

(P3) $M(X,Y)=M(Y,X)$是显然的。

(P4) $M(X,Y)=0$当且仅当$V_X(x_i)=[1,1],V_Y(x_i)=[0,0]$或

$V_X(x_i)=[0,0],V_Y(x_i)=[1,1]$;(特征点)

当$M_Z(x,y)=0$时,可得

因此$|S_x-S_y|+|\Delta x-\Delta y|=2\max\{(t_y-t_x),(f_x-f_y)\}$。

令$r_x=1-t_x,g_x=1-f_x$,由已知有$r_x\geqslant r_y\geqslant r_z,g_x\leqslant g_y\leqslant g_z$,

则$M_Z(x,y)=1-\frac{(r_x+r_y)(r_x-r_y)+(g_x+g_y)(g_y-g_x)}{r_x+r_y+g_x+g_y+2}$,

而$M(x,y)\geqslant M(x,z)$等价于:

$$\frac{(r_x^2-r_z^2)+(g_y^2-g_x^2)+2\max\{(r_x-r_y),(g_y-g_x)\}}{r_x+r_z+g_x+g_y+2}\leqslant\frac{(r_x^2-r_z^2)+(g_z^2-g_x^2)+2\max\{(r_x-r_z),(g_z-g_x)\}}{r_x+r_z+g_x+g_z+2}$$

令$m=r_x,n=r_y,z=g_y,w=g_x$,则有$1\geqslant m\geqslant n\geqslant 0,1\geqslant$

4 度量Vague值相似度的新方法

4.1 Vague值相似度量的新方法

根据第3节的分析,本节提出一种新的度量方法——多值算法:

定义8 设$x=[t_x,1-f_x],y=[t_y,1-f_y]$是Vague集$A$上的两个Vague值,$S_x,S_y$是$x$和$y$的Vague核,$\Delta x,\Delta y$是$x$和$y$的Vague未知度,则称$M_Z(x,y)$为Vague值$x$和$y$的多值算法相似度量。

$z \geqslant w \geqslant 0$。

设 $k(m,n,z,w)=$

$$\begin{cases} \dfrac{m^2-n^2+z^2-w^2+2(m-n)}{m+n+z+w+2} & m-n \geqslant z-w \\ \dfrac{m^2-n^2+z^2-w^2+2(z-w)}{m+n+z+w+2} & m-n < z-w \end{cases}$$

①当 $m-n \geqslant z-w$ 时，

$\dfrac{\partial k}{\partial m}=(m+n+z+w+2)^{-2}((2m+2)(m+n+z+w+2)-(m^2-y^2+z^2-w^2+2(m-n)))\geqslant$

$(2m+2-(m-n))(m+n+z+w+2)\geqslant 0$

②当 $m-n<z-w$ 时，

$\dfrac{\partial k}{\partial m}=2m(m+n+z+w+2)-(m^2-n^2+z^2-w^2+2(z-w))\geqslant(2m-(z-w))(m+n+z+w+2)\geqslant$

$(m+n)(m+n+z+w+2)\geqslant 0$

③当 $m-n \geqslant z-w$ 时，

$\dfrac{\partial k}{\partial n}=(m+n+z+w+2)^{-2}((-2m-2)(m+n+z+w+2)-(m^2-n^2+z^2-w^2+2(m-n)))\leqslant 0$

④当 $x-y<z-w$ 时，

$\dfrac{\partial k}{\partial n}=-2m(m+n+z+w+2)-(m^2-n^2+z^2-w^2+2(z-w))\leqslant 0$

由单调性可得 $M_z(x,y)\geqslant M_z(x,z)$，同理可证 $M_{z(y,z)\geqslant} M_z(x,z)$。

4.2 Vague集之间的相似度量

由 Vague 值的相似度量公式，我们可以进而得出相应的 Vague 集之间的相似度量公式。

定义 9 设 A 和 B 是论域 $U=\{x_1,x_2,x_3,\cdots,x_n\}$ 上的两个 Vague 集合，其中

$A=\sum_{i=1}^{n}[t_A(x_i),1-f_A(x_i)]/x_i$，$B=\sum_{i=1}^{n}[t_B(x_i),1-f_B(x_i)]/x_i$

假定 $V_A(x_i)=[t_A(x_i),1-f_A(x_i)]$ 表示 Vague 集 A 在 x_i 的隶属度，$i=1,2,\cdots,n$。

$V_B(x_i)=[t_B(x_i),1-f_B(x_i)]$ 表示 Vague 集 B 在 x_i 的隶属度，$i=1,2,\cdots,n$。

对应于定义 8，Vague 集 A 和 B 的相似程度可以由下面的公式计算：

$$S^*(A,B)=\frac{1}{n}\sum_{i=1}^{n}M_Z(V_A(x_i),V_B(x_i))$$

不难证明 $S^*(A,B)$ 满足 Vague 集之间相似度性质。

值得指出的是文献[6]中定理 6 认为 $T(A,B)=0$ 的充要条件是 $A=\sum_{i=1}^{n}[0,0]/u_i$，$B=\sum_{i=1}^{n}[1,1]/u_i$ 或者 $A=\sum_{i=1}^{n}[1,1]/u_i$，$B=\sum_{i=1}^{n}[0,0]/u_i$（$i=1,2,\cdots,n$）。这个定理有不完整之处。例如如果 $A=[1,1]/u_1+[0,0]/u_2$，$B=[0,0]/u_1+[1,1]/u_2$，根据文献[6]中的公式(5)可知，$T(A,B)=0$。因此 $A=\sum_{i=1}^{n}[0,0]/u_i$，$B=\sum_{i=1}^{n}[1,1]/u_i$ 或者 $A=\sum_{i=1}^{n}[1,1]/u_i$，$B=\sum_{i=1}^{n}[0,0]/u_i$（$i=1,2,\cdots,n$）是 $T(A,B)=0$ 充分非必要条件。通过分析，我们不难证明 $T_\cdot(A,B)=0$ 的充要条件是 $A=\sum_{i=1}^{n}[a_i,a_i]/u_i$，$B=\sum_{i=1}^{n}[b_i,b_i]/u_i$，满足 $a_i+b_i=1$，且 a_i,b_i 的值均是 0 或者 1（$i=1,2,\cdots,n$）。

5 与现有 Vague 集相似度量方法的对比研究

5.1 现有 Vague 集相似度量方法及分析

设 U 是一个论域，对 U 中的元素 x 和 y 以及 U 中的一个 Vague 集 V，x 和 y 的隶属度分别为 $\mu_V(x)=[t_x,1-f_x]$，$\mu_V(y)=[t_y,1-f_y]$；x 和 y 的 Vague 核分别为 $s(x)=t_x-f_x$，$s(y)=t_y-f_y$；未知度分别为 $\Delta x=1-t_x-f_x$，$\Delta y=1-t_y-f_y$。在以下论述中如不作特别说明，Vague 值 x 和 y 均按上述定义。

Chen 定义了 x,y 之间的相似度量 M_c[3,4]，$M_C(x,y)=1-\dfrac{|S(x)-S(y)|}{2}=1-\dfrac{|(t_x-t_y)-(f_x-f_y)|}{2}$。李凡和 An Lu 在文献[6,11]中均指出，M_c 度量 Vague 值 $x=[0,1]$ 和 $y=[0.5,0.5]$ 的相似度的结果为 1，表明 x 和 y 完全相等。[0,1]表示支持和反对的信息均为 0，而[0.5,0.5]表示支持和反对各占一半。用 M_c 度量的结果为 1，这不符合人的直觉看法。对于此方法中，Vague 值 $x=[0,1]$ 和 $y=[a,1-a]$，$(0<a\leqslant 0.5)$ 的相似性度量结果均为 1。

Hong D H 定义了 x,y 之间的相似度量 M_h[5]，$M_h(x,y)=1-\dfrac{|(t_x-t_y)|+|(f_x-f_y)|}{2}$。An Lu 指出[11]，$M_h$ 度量 Vague 值 $x=[0,1]$ 和 $y=[a,a]$，$(0<a\leqslant 1)$ 的相似度的结果为 0.5，比如[0,1]和[0.1,0.1]，[0.3,0.3]，[0.5,0.5]，[0.8,0.8]，[0.9,0.9]的相似度均为 0.5。这同样不符合人的直觉看法，并且区分能力很差。

李凡等定义了 x,y 之间的相似度量 M_{lf}[6]，$M_{lf}(x,y)=1-\dfrac{|S(x)-S(y)|}{4}-\dfrac{|(t_x-t_y)|+|(f_x-f_y)|}{4}$。由于 $M_{lf}=(M_c+M_h)/2$，因此 M_{lf} 也存在 M_c 和 M_h 相同的问题。

马志峰[7]等又定义了 x,y 之间的相似度量为 M_{mcf}，$M_{mcf}(x,y)=\left(1-\dfrac{|S(x)-S(y)|}{2}\right)\times(1-|g(x)-g(y)|)$。其中，$g(x)$ 和 $g(y)$ 是其定义的 Vague 值 x,y 的含糊度：$g(x)=\Delta_x+\delta_x$，$g(y)=\Delta_y+\delta_y$；而 δ_x 和 δ_y 是其定义的 Vague 值 x,y 的不确定度：$\delta_x=1-(|t_x-0.5|+|0.5-f_x|)$，$\delta_y=1-(|t_y-0.5|+|0.5-f_y|)$。$M_{mcf}$ 度量 Vague 值 $x=[1,1]$ 和 $y=[0.5,1]$ 的相似度的结果为 0，在度量 Vague 值 $x=[0.5,0.5]$ 和 $y=[0,1]$ 的结果为 1，这个结果显然也不能被接受和认同。

刘华文定义了 x,y 之间的相似度量 M_{lhw}[8]，$M_{lhw}(x,y)=1-\dfrac{|S'(x)-S'(y)|+2|(B'_x-B'_y)|}{4}$。其中，$S^l(x)=[t_x+t_x\times\Delta x]-[f_x+f_x\times\Delta x]$，$S^l(y)=[t_y+t_y\times\Delta y]-[f_y+f_y\times\Delta y]$ 为扩展核，考虑了未知信息倾向。$B^l_x=[t_x+t_x\times\Delta y]+[f_x+f_x\times\Delta y]$，$B^l_y=[t_y+t_y\times\Delta y]+[f_y+f_y\times\Delta y]$ 为扩展精确度，同样是考虑了未知信息倾向。M_{lhw} 度量 Vague 值 $x=[0,0]$ 和 $y=[1,1]$ 相似度的结果为 0.5，而当两个 Vague 值为[0,0]和[1,1]或为[1,1]和[0,0]时，即两个 Vague 值退化为经典集合时，其计算相似度量值应该为 0。因此，M_{lhw} 也存在缺陷。

阎德勤等定义了 x,y 之间的相似度量为 M_{ydq}[9-10]，$M_{ydq}(x,y)=1-\sqrt{\dfrac{(t_x-t_y)^2+(f_x-f_y)^2+(\Delta x-\Delta y)^2}{2}}$。$M_{ydq}$ 度量 Vague 值 $x1=[0,1]$ 与 $y1=[0,0]$，$x2=[0,1]$ 与 $y2=[1,1]$ 的相似度的结果为 0，但从直观上来看，$x1$ 和 $y1$ 的支持度

均为0，所以它们的相似度不应该为0，这显然不符合在投票模型中的直觉认识。

An Lu定义了 x,y 之间的相似度量 M_{al}[11]，$M_{al}(x,y)=\sqrt{(1-\Delta M_m)(1-\Delta M_i)}$。其中

$\Delta M_m=|(t_x+1-f_x)-(t_y+1-f_y)|/2$

$\Delta M_i=|(1-t_x-f_x)-(1-t_y-f_y)|$

M_{al} 度量 Vague 值 $x=[0,1]$ 和 $y=[0,0]$ 时的结果为0，在度量 Vague 值 $x=[0,1]$ 和 $y=[0.5,0.5]$ 的相似度的结果为0，这个结果同样不直观。

Lu J L 等定义了 x,y 之间的相似度量 M_{ljl}[12]，$M_{ljl}(x,y)=1-\frac{|pq|+|(2-m)n|+|q-n|}{p+(2-m)+2}$，其中 $p=t_x+t_y$，$q=t_x-t_y$，$m=f_x+f_y$，$n=f_x-f_y$。M_{ljl} 度量 Vague 值相似度的时候考虑了确切的支持信息和确切的反对信息占总体信息比例的影响，区分能力比较高，但是在度量 $x1=[0.1,0.9]$ 和 $y1=[0.2,0.9]$ 的相似度的结果为0.968，度量 $x2=[0.7,0.9]$ 和 $y2=[0.8,0.9]$ 的相似度的结果是0.953。在 $x2$ 和 $y2$ 的未知量远小于 $x1$ 和 $y1$ 的未知量，且在端点距离相等的情况下，$x2$ 和 $y2$ 的相似度应该大于 $x1$ 和 $y1$ 的相似度，而 M_{ljl} 计算的刚好相反，这样的结果表明该公式存在不足。

C C Lo 等定义了 x,y 之间的相似度量 M_{acl}[13]，$M_{acl}(x,y)=\frac{\min(\varphi_x,\varphi_y)}{\max(\varphi_x,\varphi_y)}$，其中 $\varphi_x=\frac{t_x+1-f_x}{2}$，$\varphi_y=\frac{t_y+1-f_y}{2}$。$M_{acl}$ 度量 Vague 值 $x=[0,0]$ 和 $y=[0,0.1]$ 的相似度的结果为0，但这两个的相似度应该很高。同样，度量 Vague 值 $x=[0,1]$ 和 $y=[0.5,0.5]$ 的相似度的结果为1，同样不是我们所能接受的。

李艳红等定义了 x,y 之间的相似度量 $M_{lyh}(x,y)$[14]，$M_{lyh}(x,y)=1-\sqrt{\frac{(t_x-t_y)^2+(f_x-f_y)^2}{2}}$。

裴振奎等定义了 x,y 之间的相似度量 M_{pzk}[15]，$M_{pzk}(x,y)=1-\frac{|t_x-t_y|+|f_x-f_y|}{4}-\frac{\max(|t_x-t_y|,|f_x-f_y|)}{2}$。

蔡立晶等定义了 Vague 集合 x,y 之间的相似度量 M_{clj}[16]，$M_{clj}(x,y)=1-\frac{|t_x-t_y|+|f_x-f_y|}{2}$。

黄国顺等定义了 x,y 之间的相似度量 M_{hgs}[17-18]，$M_{hgs}(x,y)=\frac{2-(|t_x-t_y|+|f_x-f_y|)}{2+(|t_x-t_y|+|f_x-f_y|)}$。

张诚一定义了 x,y 之间的相似度量 $M_{\alpha y}$[19]，$M_{\alpha y}(x,y)=1-\frac{|\delta(x)-\delta(y)|+|(\alpha_x-\alpha_y)|}{2}$，其中 $\delta(x)=[t_x+t_x\times\Delta x]$，$\delta(y)=[t_y+t_y\times\Delta y]$，$\alpha(x)=[f_x+f_x\times\Delta x]$，$\alpha(y)=[f_y+f_y\times\Delta y]$。

5.2 与现有 Vague 集相似度量方法的对比

1)表1的数据是从多个文献中选取的极具代表性的数据。从表1中可以明显看出带下划线的数据与其他数据的区别，说明了部分相似度量方法存在的问题，验证了5.1节所列出的反例和不足。

表1 各种相似度量方法的对比

No.	1	2	3	4	5	6	7	8
tx,1-fx	[0,1]	[1,1]	[0,0]	[0,1]	[0,1]	[0.1,0.9]	[0.7,0.9]	[0,0]
ty,1-fy	[0.5,0.5]	[0.5,1]	[1,1]	[0,0]	[1,1]	[0.2,0.9]	[0.8,0.9]	[0,0.1]
Mc	1	0.75	0	0.5	0.5	0.95	0.95	0.95
Mh	0.5	0.75	0	0.5	0.5	0.95	0.95	0.95
Mlf	0.75	0.75	0	0.5	0.5	0.95	0.95	0.95
Mmzf	1	0	0	0	0	0.95	0.76	0.76
Mlyh	0.5	0.65	0	0.29	0.29	0.93	0.93	0.93
Mlhw	0.5	0.81	0.5	0.25	0.25	0.88	0.97	0.99
Mpzk	0.5	0.63	0	0.25	0.25	0.93	0.93	0.93
Mydq	0.13	0.5	0	0	0	0.9	0.9	0.9
Mclj	0.5	0.75	0	0.5	0.5	0.95	0.95	0.95
Mhgs	0.33	0.6	0	0.33	0.33	0.90	0.90	0.90
Mzcy	0.5	0.88	0	0.5	0.5	0.92	0.96	0.99
Mal	0	0.61	0	0	0	0.92	0.92	0.92
Mljl	0.75	0.77	0	0.33	0.6	0.97	0.95	0.95
Mccl	1	0.75	0	0	0.5	0.91	0.94	0
Mz	0.5	0.72	0	0.4	0.4	0.93	0.94	0.94

2)文献[14-19]所定义的方法没有明显的直观不足，但是它们的共同特点是相似度量区分能力比较低，各种相似度量方法的相似度量区分能力见表2。

表2 各种相似度量方法的相似度量区分能力对比

相似度度量方法	Mc	Mh	Mlf	Mpzk	Mmzf	Mlyh	Mlhw	Mydq	Mclj	Mhgs	Mzcy	Mal	Mljl	Mccl	Mz
分类数	21	21	21	38	56	60	240	35	253	253	180	59	1367	129	638
相似度量区分能力(%)	0.48	0.48	0.48	0.87	1.29	1.38	5.51	0.80	5.80	5.80	4.13	1.35	31.39	2.96	14.65

表2数据经过如下几步得到：

第一步 在区间[0,1]内取出以0.1为步长的所有子区间，并且所有子区间端点之合小于等于1，由此得到的子区间共计有66个数据。

第二步 将得到的所有子区间作为 Vague 值(子区间的端点分别为 t_x 和 f_x)，分别组合成为内容相同的 Vague 集 V_1 和 V_2。

第三步 将 Vague 集 V_1 和 V_2 的数据两两组合，将此组合后的数据集合称为数据集 S(共计产生4356个组合数据)。用某种相似度量方法 M 对数据集 S 度量，若将相似度量结果相同的归为一类，由此得到某种相似度量方法 M 对数据集 S 度量的分类数。

第四步 再由定义5就可以得到某种相似度量方法 M 在数据集 S 上的相似度量区分能力，为便于数据对比，表1中

的相似度量区分能力扩大了100倍。

3)以上各种相似度量方法满足定理1的情况见表3,其中完全满足定理1的度量方法有 M_{pzk},M_{ydq},M_{zcy},M_{zcy},M_{zcl},M_z。

表3 各种相似度量方法满足定理1中相关性质的情况

相似度度量方法	Mc	Mh	Mlf	Mpzk	Mmzf	Mlyh	Mlhw	Mydq	Mclj	Mhgs	Mzcy	Mal	Mljl	Mccl	Mz
满足定理1的性质	P1-P4	P1-P4	P1-P4	P1-P5	P1-P4	P1-P4	P1-P3	P1-P5	P1-P4	P1-P5	P1-P5	P1-P4	P1-P4	P1-P5	P1-P5

4)对于 M_z 的进一步说明:虽然 M_z 的区分能力小于 M_{ljl} 的区分能力,但 M_z 在度量 $x1=[0.1,0.9]$ 和 $y1=[0.2,0.9]$ 时的结果为0.933,在度量 $x2=[0.7,0.9]$ 和 $y2=[0.8,0.9]$ 时的结果是0.942,这样的结果符合我们关注的已知的确切信息。因为 $x2$ 和 $y2$ 的未知量远小于 $x1$ 和 $y1$ 的未知量且端点距离相等,所以 $x2$ 和 $y2$ 的相似度应该大于 $x1$ 和 $y1$ 的相似度,这个结果比 M_{ljl} 的计算结果更合理更适用(用 M_{ljl} 计算时,$x2$ 和 $y2$ 的相似度小于 $x1$ 和 $y1$ 的相似度)。但要特别指出的是,不是说在端点距离相等且 $x2$ 和 $y2$ 的未知量小于 $x1$ 和 $y1$ 的未知量的情况下,就一定得出 $x2$ 和 $y2$ 的相似度大于 $x1$ 和 $y1$ 的相似度的结论。这是因为未知量大小不是影响相似度的唯一因素,在某些情况下可能不起作用。例如,当 $x1=[0.1,0.9]$,$y1=[0.1,0.9]$,$x2=[0.8,0.9]$,$y2=[0.8,0.9]$ 时,端点距离相等且 $x2$ 和 $y2$ 的未知量远小于 $x1$ 和 $y1$ 的未知量,但此时 $x2$ 和 $y2$ 的相似度等于 $x1$ 和 $y1$ 的相似度,均为1。

结束语 本文通过对现有的Vague集相似度量方法的分析,指出这些度量方法不足以准确描述Vague集的相似本质,给出了相似度量的一系列讨论,提出了度量Vague集相似性的必要参数和评价方法,定义了相似度量区分能力,为衡量Vague集相似度量方法提供了理论依据。在充分分析Vague集特点的基础上给出了一种新的相似度量的方法,经试验验证该方法有更好的度量效果,相似度量区分能力高,该方法有望在更广的领域得到应用。

参考文献

[1] Zadeh L A. Fuzzy sets. Information and control, 1965, 8(3): 338-353

[2] Gau Wen-Lung, Buehrer D J. Vague sets. IEEE Transactions on Systems, Man and Cybemetics, 1993, 23(2): 610-614

[3] Chen S M. Measures of similarity between vague sets. Fuzzy Sets Syst, 1995, 74(2): 217-223

[4] Chen S M. similarity measures between vague sets and between elements. IEEE Trans, Syst. Man, Cybern, 1997, 27(1): 153-158

[5] Hong D H, Kim C. A note on similarity measures between vague sets and between elements. Information Sciences, 1999, 115(1): 83-96

[6] 李凡,徐章艳. Vague集之间的相似度量. 软件学报,2001,12(6):922-927

[7] 马志锋,邢汉承. Vague决策表中的含糊规则获取. 计算机学报,2001,21(4):382-389

[8] 刘华文. 模糊模式识别的基础——相似度量. 模式识别与人工智能,2004,17(2):141-145

[9] 阎德勤,迟忠先. Vague集中的分解定理与相似度量. 计算机科学,2003,30(1):78-79

[10] 阎德勤,迟忠先,李艳红. 关于Vague集的相似度量. 模式识别与人工智能,2004,17(1):22-26

[11] Lu A, Ng W. Managing Merged Data by Vague Functional Dependencies// 23rd International Conference on Conceptual Modeling(ER2004). Shanghai, China, LNCS3288, 2004, 11: 259-272

[12] Lu Jingli, Yan Xiaowei, Yuan Dingrong, et al. A New Similarity Measure for Vague Sets. IEEE Intelligent Informatics Bulletin, 2005, 6(2): 14-18

[13] Lo C-C, Wang P, Chao K-M. A Fuzzy Analysis Method for New-Product Development // The 9th International Conference on Computer Supported Cooperative Work in Design. Coventry, UK, 2005, 5: 211-216

[14] 李艳红,迟忠先,阎德勤. Vague相似度量与Vague熵. 计算机科学,2002,29(12):129-132

[15] 裴振奎,徐九韵. Vague集之间相似度量的一种新方法. 广西师范大学学报,2003,21(1):138-143

[16] 蔡立晶,吕泽华,李凡. Vague集的三维表示及相似度量. 计算机科学,2003,30(5):76-77

[17] 黄国顺,刘云生. Vague集相似度量及其在模式识别中的应用. 复旦学报,2004,43(5):869-872

[18] 黄国顺,刘云生. 一类新Vague集相似度量. 计算机应用与软件,2005,22(7):24-26

[19] 张诚一,党平安. 关于Vague集之间的相似度量. 计算机工程与应用,2003,17(1):92-94

(上接第191页)

结束语 本文试图给出HMOS模型的一个理论规范。我们认为这种模型比较适合大规模分布系统的应用,但很多内容还有待展开。目前课题组正在开发和完善该理论模型的一个计算机软件实现。实现的细节就不在本文中赘述了。

致谢 首都经济贸易大学赵丹亚教授、王利教授、李宁副教授,以及项目组的全体成员。

参考文献

[1] 谢毅平,李书旺,张军. HMOS系统的规范介绍. 中国公安大学学报,2004(4)

[2] 张军,李书旺. HMOS分布模型的初步研究. 计算机科学与实践,2004(6)

[3] 张军,李书旺. 典型树状结构分布对象的复制、定位和联结服务. 计算机科学与实践,2005(3)

[4] Weider C, Reynolds J. Executive Introduction to Directory Services Using the X. 500 Protocol, RFC 1308. March 1992

[5] Yeong W, Howes T, Kille S. ISODE Consortium, X. 500 Lightweight Directory Access Protocol. RFC 1487, July 1993