粗糙集与 Vague 集*)

闫德勤[1] 迟忠先[2]
(辽宁师范大学计算机系 大连116029)[1] (大连理工大学计算机系 大连116024)[2]

摘 要 本文研究了粗糙集和 Vague 集及它们的关系,这两类集都是人工智能、知识挖掘和知识发现的重要工具。从集合基数表示的角度说,粗糙集也是 Vague 集的一种。通过分析这两种集的关系,提出了相关定理。同时,我们还提出了粗糙 Vague 集的概念,并初步研究了其性质。

关键词 粗糙集,Vague 集,粗糙 Vague 集

Rough Sets and Vague Sets

YAN De-Qin[1] CHI Zhong-Xian[2]
(Department of Computer Science, Liaoning Normal University, Dalian 116029)[1]
(Department of Computer Science, Dalian University of Technology, Dalian 116024)[2]

Abstract In this paper, rough sets and Vague sets are approached. Both rough sets and Vague sets are important tools for AI and KDD. This paper explores the relationship of them. In the point view of cardinal number expression, rough sets are a kind of Vague sets. By analyzing relationship of the two sets, a theorem is given. Besides, a new concept of RV sets is proposed, about which some properties are studied.

Keywords Rough sets, vague sets, RV sets

1 引言

目前,在计算机科学及应用的多种领域特别是在人工智能(AI)、知识挖掘知识发现(KDD)中,粗糙集(rough sets)理论有着重要的实际应用,其研究成为一个热点领域[2~6]。粗糙集的成功应用在于其提供了一种有效表示和处理知识的一种工具,随着应用的广泛和深入对这个理论发展的要求也在不断增强。Vague 集[1]拓广了模糊集(Fuzzy sets)对事物表达的范围,同时也提供了一种对知识表示的新工具。Vague 集从一定意义上讲在对事物属性的描述上较模糊集提供了更多的选择方式,因而在学术界和工程技术界引起了广泛关注。本文研究了粗糙集与 Vague 集的关系,建立了粗糙集与 Vague 集的联系,给出了粗糙集对应的 Vague 集形式及相关定理。同时,提出了粗糙 Vague 集(RV sets)的概念,并初步研究了其性质。

2 粗糙集与 Vague 集的基本概念

设 $U=\{x_1,x_2,\cdots,x_n\}$ 是一有限集,称为论域,R 是 U 上的一个等价关系,U/R 表示在 U 上导出的所有等价类;$[x]_R$ 表示包含元素 x 的 R 等价类,$x\in U$。对任集合 $X\subseteq U$

$R_-(X)=\{x\in U|[x]_R\subseteq X\}$

$R^-(X)=\{x\in U|[x]_R\cap X\neq\varnothing\}$($\varnothing$ 为空集)

分别称 $R_-(X)$ 与 $R^-(X)$ 为 X 的 R 下近似和 X 的 R 上近似。上下近似不相等时,称对于 R,X 为粗集。在省略 X 时由(R_-,R^-)表示粗糙集。

$Bn_R(X)=R^-(X)-R_-(X)$ 称为 X 的 R 边界,也称为粗糙集的边界。$Pos_R(X)=R_-(X)$ 称为 X 的 R 正域,也称为粗糙集的正域。$Neg_R(X)U-R_-(X)$ 为粗糙集的负域。粗糙集的精度由下式表示:

$$d_R(X)=\frac{|R_-(X)|}{|R^-(X)|}$$

其中,$|\cdot|$ 表示集合的基数。

定义1[1] 设 X 是一个对象空间,其中的任意一个元素用 x 表示,X 中的一个 Vague 集 V 用一个真隶属函数 t_v 和一个假隶属函数 f_v 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 X 中的每一个点联系起来,即

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

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

x 关于 V 的隶属度 $V(x)$ 表示为:

$[t_v(x),1-f_v(x)]$。

其中,$t_v(x)+f_v(x)\leqslant 1$。

当 X 为连续空间时,一个 Vague 集 V 表示为:

$$V=\int_X[t_v(x),1-f_v(x)]/x\mathrm{d}x,x\in X。$$

当 X 为离散空间时,一个 Vague 集 V 表示为:

$$V=\sum_{i=1}^{n}[t_v(x_i),1-f_v(x_i)]/x_i,x_i\in X$$

3 粗糙集对应的 Vague 集

为使讨论方便,这里把 Vague 集定义中的是一个对象空间 X 记为论域空间 U,其中的任意一个元素用 x 表示。设 V 为一 Vague 集,对于 $x\in U$ 集 V 把区间[0,1]分为三部分:$t_v(x)$、$f_v(x)$ 和中间部分 $m_v(x)=1-t_v(x)-f_v(x)$。为了简便分别记 $t_v=t_v(x)$,$f_v=f_v(x)$,$m_v=m_v(x)$。这三部分反映着 Vague 集 V 在一维空间上的粗糙性。

*)辽宁师范大学校基金资助。**闫德勤** 博士,教授,主要研究领域为人工智能、图像处理等。**迟忠先** 教授,博士生导师,主要研究领域为知识发现、数据仓库、数据挖掘等。

对于粗糙集，对任集合 $X\subseteq U$ 令：

$$t_R(X)=\frac{|R_-(X)|}{|U|} \tag{1}$$

$$1-f_R(X)=\frac{|R^-(X)|}{|U|} \tag{2}$$

(1)(2)式构成了论域 U 到实数区间[0,1]的映射，t_R 由完全属于 X 的等价类(的基数)构成真隶属函数，f_R 由完全无关于 X 的等价类(的基数)构成假隶属函数，从而形成 Vague 集：

$[t_R(X),1-f_R(X)]$。

因此，(1)(2)式也是对粗糙集进行 Vague 集描述的定义式。由定义式(1)(2)可得出：

$$f_R(X)=\frac{|U-R^-(X)|}{|U|}=\frac{|neg(X)-Bn_R(X)|}{|U|},$$

$$m_R(X)=1-t_R(X)-f_R(X)=\frac{|Bn_R(X)|}{|U|}。$$

例1 设 R 为论域 U 上的一个等价关系，$U/R=\{E_1,E_2,E_3\}$。其中，

$E_1=\{x_1,x_2,x_3,x_4\}$，

$E_2=\{x_5,x_6\}$，

$E_3=\{x_7,x_8\}$。

对于集合 $X=\{x_1,x_2,x_3,x_4\}$ 的 R 粗糙集为：

$R_-=\{x_1,x_2,x_3,x_4\}$，

$R^-=\{x_1,x_2,x_3,x_4,x_5,x_6\}$。

由(1)(2)式，该粗糙集对应的 Vague 集为：$[t_R(X),1-f_R(X)]=[0.5,0.75]$。

如果把集合 $X\subseteq U$ 看作一个对象，其相应的(由粗糙集导出的)Vague 集则是建立在等价机理上的肯定度与否定度表示。

由式(1)、(2)，把粗糙集转化为 Vague 集具有两方面的意义：一是提供一种对于以集合为整体的对象进行 Vague 集表示的方法，这样的方法为集合群体性研究提供理论工具；二是建立粗糙集与 Vague 集的理论联系，这对两方面的理论研究具有意义。

为叙述简便，记 V_R 为由粗糙集形成的 Vague 集的全体。V_R 中集合的相等、交、并、补等运算的定义完全与 Vague 集[1]中相应的运算一致，这里就不重新定义。一般情况下记 $t_R=t_R(X)$，$f_R=f_R(X)$，以及 $m_R=m_R(X)$。

定理1 设 A、B 为论域 U 上的两个等价关系，对于集合 $X\subseteq U$ 两个粗糙集 (A_-,A^-)、(B_-,B^-) 的精度分别为 d_A、d_B，相应的 Vague 集为 $\hat{A}$、$\hat{B}$。由 Vague 集 $\hat{C}=\hat{A}\cap\hat{B}$ 和 $\hat{D}=\hat{A}\cup\hat{B}$ 所对应的粗糙集 (C_-,C^-) 和 (D_-,D^-) 的精度分别为 d_C、d_D。则有 $d_Ad_B=d_Cd_D$ 成立。

证明：设 A、B 为论域 U 上的两个等价关系，其相应的 Vague 集为 $\hat{A}$、$\hat{B}\in V_R$。由 $\hat{C}=\hat{A}\cap\hat{B}$ 知，$t_{\hat{C}}=\min(t_{\hat{A}},t_{\hat{B}})$

$1-f_C=\min(1-f_{\hat{A}},1-f_{\hat{B}})=1-\max(f_{\hat{A}},f_{\hat{B}})$

对于 Vague 集 $\hat{C}$ 相应的粗糙集有：

$C_-=\{A_-$或 $B_-\mid$相应的真隶属度为 $t_C\}$，

$C^-=\{A^-$或 $B^-\mid$基数最小$(\neq f_C)\}$。

因此，$d_C=\dfrac{|\min(A_-,B_-)|}{|\min(A^-,B^-)|}$。

同样，由 $\hat{C}=\hat{A}\cup\hat{B}$ 知，$t_C=\max(t_{\hat{A}},t_{\hat{B}})$

$1-f_C=\max(1-f_{\hat{A}},1-f_{\hat{B}})=1-\min(f_{\hat{A}},f_{\hat{B}})$

因此，对于 Vague 集 $\hat{C}$ 相应的粗糙集有：

$C_-=\{A_-$或 $B_-\mid$相应的真隶属度为 $t_C\}$，

$C^-=\{a^-$或 $B^-\mid$基数最大$(\neq f_C)\}$。

有 $d_C=\dfrac{|\max(A_-,B_-)|}{|\max(A^-,B^-)|}$。

$$d_Cd_D=\frac{|\min(A_-,B_-)\max(A_-,B_-)|}{|\min(A^-,B^-)\max(A^-,B^-)|}=\frac{|A||B_-|}{|A^-||B^-|}=d_Ad_B。$$

证毕。

4 粗糙 Vague 集(RV sets)

为解决用粗糙集理论方法处理用模糊集表示的问题，人们引入粗糙模糊集[5,6](RF)的概念。同样当用粗糙集解决用 Vague 集表示的问题时也要有一定的工具。为此，这里引入粗糙 Vague 集(RV)的概念。

定义2 在论域 U(又称对象空间)上，设 R 是一个等价类，V 是一个 Vague 集。由 R 和 V 构成的粗糙 Vague 集(RV sets)定义如下：

$Rt_-(V)=\inf\{t_v(x)\mid x\in[x]_R\}$

$Rt^-(V)=\sup\{t_v(x)\mid x\in[x]_R\}$

$Rf_-(V)=\sup\{f_v(x)\mid x\in[x]_R\}$

$Rf^-(V)=\inf\{f_v(x)\mid x\in[x]_R\}$

其中 Rt_-、Rt^- 表示在同一等价类上真隶属度的最小和最大值，Rf_-、Rf^- 表示在同一等价类上假隶属度的最小和最大值。上下近似 Vague 集表示为：

$V^-=[Rt^-(V),1-Rf^-(V)]$，

$V_-=[Rt_-(V),1-Rf_-(V)]$。

定理2 设 (V_-,V^-)、(W_-,W^-) 为定义2给出的粗糙 Vague 集，有下式成立：

(1)$V_-\subseteq V\subseteq V^-$

(2)$(V\cup W)^-=V^-\cup W^-$

$(V\cap W)_-=V_-\cap W_-$

(3)$V_-\cup W_-\subseteq(V\cup W)_-$

$(V\cap W)^-\subseteq V^-\cap W^-$

(4)若 $V\subseteq W$，则 $V_-\subseteq W$ 且 $V^-\subseteq W^-$

证明：(1)：由定义2知 $Rt_-\leqslant t_v\leqslant Rt^-$，$1-Rf_-\leqslant 1-f_v\leqslant 1-Rf^-$ 所以(1)式成立。

(2)：任给 x 和等价关系 R，有：

$$\begin{aligned}Rt(V\cup W)^-&=\sup\{\max(t_v(x),t_w(x))\mid x\in[x]_R\}\\&=\max(\sup\{t_v(x),t_w(x)\mid x\in[x]_R\})\\&=Rt(V)^-\cup Rt(W)^-,Rf(V\cup W)^-\\&=\inf\{\max\}(f_v(x),f_w(x))\mid x\in[x]_R\}\\&=\max(\inf\{f_v(x),f_w(x)\mid x\in[x]_R\})\\&=Rf(V)^-\cup Rf(W)^-。\end{aligned}$$

所以有 $(V\cup W)^-=V^-\cup W^-$。

同理可证 $(V\cap W)_-=V_-\cap W_-$，

所以(2)式成立。

(3)：任给 x 和等价关系 R：

$t_{V_-}=Rt_-(V)=\inf\{t_v(x)\mid x\in[x]_R\}$，

$t_{W_-}=Rt_-(W)=\inf\{t_w(x)\mid x\in[x]_R\}$，

$t_{(V\cup W)_-}=Rt_-(V\cup W)=\inf\{\max(t_v(x),t_w(x))\mid x\in[x]_R\}$。

设 $\alpha=\max(t_{V_-},t_{W_-})$ 不妨设 $\alpha=t_{V_-}$，则存在一个元素 $x_0\in[x]_R$，使得 $\alpha=t_v(x_0)\geqslant t_w(x_0)$。由 t_{V_-} 的定义，$\alpha=t_{V_-}\leqslant\{\max(t_v(x),t_w(x))\mid x\in[x]_R\}$。

因此有，

$\alpha\leqslant\inf\{\max(t_v(x),t_w(x))\mid x\in[x]_R$，

即 $\max(t_{V_-},t_{W_-})\leqslant t_{(V\cup W)_-}$。

另一方面：

$1-f_{V_-}=1-Rf_-(V)=\inf\{f_v(x)\mid x\in[x]_R\}$，

$1-f_{W_-}=1-Rf_-(W)=\inf\{f_w(x)\mid x\in[x]_R\}$，

$1-f_{t(V\cup W)-}=1-Rf_-(V\cup W)=\inf\{\max(f_v(x),f_w(x))|x\in[x]_R\}$。

利用上面的证明方式可得：

$\max(1-f_{V-},1-f_{W-})\leqslant 1-f_{t(V\cup W)-}$，综上所述，得到 $V_-\cup W_-\subseteq(V\cup W)_-$，同理可证$(V\cap W)_-=V_-\cap W_-$，所以(3)式成立。

(4)：由 $V\subseteq W$ 得 $t_v\leqslant t_w$，$1-f_v\leqslant 1-f_w$。

$\inf\{t_v(x)|x\in[x]_R\}\leqslant\inf\{t_v(x)|x\in[x]_R\}$，

$\sup\{f_v(x)|x\in[x]_R\}\geqslant\sup\{f_w(x)|x\in[x]_R\}$。

$1-\sup\{f_v(x)|x\in[x]_R\}\leqslant 1-\sup\{f_w(x)|x\in[x]_R\}$。

因此，$t_{V-}\leqslant t_{W-}$，$1-f_{V-}\leqslant 1-f_{W-}$，即 $V_-\subseteq W_-$ 成立，

同样，由 $t_v\leqslant t_w$，$1-f_v\leqslant 1-f_w$，得：

$\sup\{t_v(x)|x\in[x]_R\}\leqslant\sup\{t_v(x)|x\in[x]_R\}$，

$\inf\{f_v(x)|x\in[x]_R\}\geqslant\inf\{f_w(x)|x\in[x]_R\}$。

$1-\inf\{f_v(x)|x\in[x]_R\}\leqslant 1-\inf\{f_w(x)|x\in[x]_R\}$。

因此，$t_{V^-}\leqslant t_{W^-}$，$1-f_{V^-}\leqslant 1-f_{W^-}$，即 $V^-\subseteq W^-$ 成立。从而(4)式成立。证毕。

结论 本文以粗糙集基数表示的形式建立了产生粗糙集的 Vague 集形式，提出了粗糙 Vague 集的概念，给出了相关定理，对于处理一定形式的数据或知识信息具有工具性的使用价值，对于粗糙集与 Vague 集结合的理论与应用研究具有一定意义。

参考文献

1 Gau W L, Buehrer D J. Vague sets. IEEE Trans. Systems Man Cybernet, 1993, 23(2): 610～614

2 Pawlk Z, et al. Rough Sets. Communications of the ACM, 1995, 38(11): 89～95

3 Pawlk Z. Rough set theory and its application to data analysis. Cybernetics and Systems, 1998, 29(9): 661～668

4 Ziarko W. Introduction to the special issue on rough sets and knowledge discovery. International Journal of Computational Intelligence, 1995, 11(2): 223～226

5 曾黄麟. 粗集理论及其应用. 重庆：重庆大学出版社，1996

6 张文修，等. 粗糙集理论与方法. 北京：科学出版社，2001

(上接第131页)

较差。它的优点是当处理机分配完之后没有通信代价(除了动态负载平衡)。横向划分的缺点是通信代价随着树的增长变得很严重。它的优点是所有的处理机从始至终都参与任务计算，而且负载平衡比较好。纵向划分的缺点是随着树的层数的增加一些处理机将处于闲置状态。负载平衡需要额外的数据移动。而且每个处理机的负载不很均匀。这种方法比横向划分方法需要比较少的通信量，比动态数据划分方法的负载平衡度高。混合并行方法采用上述两种方法的优点。在各个处理机独立工作之前采用静态数据划分的并行方法，将各个子任务由单独的处理机负责，这样可以达到减少通信代价。这种方法同动态处理机组负责一个子任务，在处理机组内部采用静态数据分片的并行方法，一个处理机组保存了对应子任务的全部数据。各个处理机组之间需要数据交换，处理机组内部需要传送消息，当所有的处理机组只包含一个处理机时通信代价为零。各个处理机组的处理机的数量是和任务的大小成比例的，因此混合并行方法的负载平衡程度比较好。

表1 5种并行决策树的比较

方法	数据移动	传送统计图表	数据倾斜	负载平衡	精度
动态数据分片方法	多	没有	严重	差	高
横向数据分片方法	没有	多	中	好	高
纵向数据分片方法	没有	没有	中	中	高
混合并行方法	中	中	少	好	高
决策树合并	没有	没有	很少	很好	略低

总结和讨论 随着生物信息领域的飞速发展，出现越来越多的生物数据库，这些数据库的最大特点就是数据量庞大。传统的内存算法不适合庞大的生物数据，这样的执行时间有时是无法接受的，所以并行挖掘算法就变得很重要。前面对现阶段的并行建立决策树的算法进行了综述，并且对各个算法进行了性能评价。混合并行的方法在这些方法中是比较好的，它结合了各种方法的优点，但是文[8]没有考虑纵向划分的混合方法，纵向划分的方法在通信代价方面优于横向划分方法，而且生物数据会形成很多属性，属性值一般比较少，更适合纵向划分，因此这是值得研究的一个方面。

参考文献

1 Wang D, Wang X, Honavar V, Dobbs D. Data-Driven Generation of Decision Trees for Motif-Based Assignment of protein Sequences to Functional families. In: Proc. of the Atlantic Symposium on Computational Biology, Genome Information Systems&Technology. 2001

2 Quinlan J R. C4. 5: programs for Machine Learning. Morgan Kaufmann, San Matoo, CA, 1993

3 Agawal R, et al. An interval classifier for database mining applications. In: proc. of the VLDB Conf. Vancouver, British Columbia, Canada, Aug. 1992. 560～573

4 Michie D, spiegelhalter D J, Taylor C C. Machine Learning, Neural and Statistical Classification. Ellis Horwood, 1994

5 Hunt E B, Marin J, Stone P T. Experiments in Induction. Academic press, 1996

6 Quinlan J R. Discovering rules from large collections of examples: A cass study. In: Michie D, ed. Expert Systems in the Micro Electronic Age, Edinburgh University press, 1979

7 Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Eng., 1993, 5(6): 914～925

8 Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining. In: Proc. of the Fifth Int'l Conf. on Extending Database Technology, Avignon, france, 1996

9 Shafer J, Agrawal R, Mehta M. SPRINT: A scalable classifier for data mining. In: Proc. of the 22nd VLDB Conf. 1996

10 Chattratichat J, et al. Large scale data mining: Challenges and responses. In: Proc. of the Third Int7 Conf. on Knowledge discovery and Data Mining, 1997

11 Joshi M V, Karypis G, Kumar V. ScalparC: a new scalable and efficient parallel classification algorithm for mining large datasets. In: Proc. of the Intl. parallel processing Symposium, 1998