维普资讯 http://www.cqvip.com

基于 Rough 集的交叉覆盖算法设计*)

吴 涛 张 铃

(安徽大学人工智能研究所 智能计算与信号处理重点实验室 合肥230039)

Covering Design Algorithm Based on Rough Set Theory

WU Tao ZHANG Ling

(Institute of Artificial Intelligence, Anhui University; Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Hefei 230039)

Abstract This paer gives an artificial network(RCSN)combining rough set theory and covering design algorithm, which reduces condition attribute using rough set theory and designs the structure of neural network with covering design algorithm. An instance shows this kind of network has the advantages of fast computation and high accuracy; the method also can cut down the occupying of memory and the cost of data collecting.

Keywords Rough set, Attribute reduction, Covering algorithm

1. 引言

人工神经网络因其特有的信息处理能力和独到的解算能力取得了长足的发展，引起各方面的广泛关注，当前已经有许多人工神经网络模型在视觉、语言、信号处理、自动控制、函数逼近、金融证券等领域获得成功的应用，但诸如 BP 网络等模型也存在一些带有共性的缺点，比如：学习和训练时间很长、学习收敛性很难保证[1]。文[2]根据神经元的几何意义提出的多层前向神经网络的交叉覆盖算法，针对学习样本的特征构造神经网络，可以有效地处理海量数据，在一定意义上解决了多年来一直未解决的作为分类器的多层前向网络的设计问题。由波兰数学家 Z. Pawlak 提出的粗糙集理论[3]无需提供除问题所需处理的数据集合之外的任何先验信息，仅根据观测数据删除冗余信息，通过比较知识不完整的程度——粗糙度、属性间的依赖性与重要性，发现数据间的关系，在不损失信息的前提下提取有用特征，简化信息处理，为研究不精确、不确定知识的表达、学习、归纳方法等提供了一个有力的数学工具，为智能信息处理提供了有效的处理技术，成为当前机器学习、知识发现等领域的研究热点，且已经在信息系统分析、人工智能、决策支持系统、知识与数据发现、模式识别与分类、故障检测等方面取得了较为成功的应用[4,5]。本文将粗糙集与覆盖算法相结合，应用粗糙集的方法提取保持信息完整的最小属性集，在此基础上建立交叉覆盖网络(RCSN 网络)，实验证明这种网络运行速度快、准确性高，并可大幅度地降低数据的收集与存储成本。

2. 属性的约简与覆盖算法

2.1 Rough 集与属性约简

粗糙集理论将知识理解为使用等价关系集 R 将有限非空论域 U 进行划分，其主要思想是利用已知的知识库，将不精确或不确定的知识用已知的知识库中的知识来(近似)刻画，在不损失信息的前提下，约简信息属性和属性值，获得最小化的决策规则。现简要介绍其基本概念。

近似空间 $A=(U,R)$ 称为一个知识库，若 $r\in R$，$x,y\in U$，$(x,y)\in r$，则称 x，y 是 r 不可分辨的，$[x]_r=\{y|(x,y)\in r\}$ 是 $x\in U$ 的 r 等价类，$X_1,X_2,\cdots,X_n$ 是 U 上由 r 生成的等价类全体，记为 U|r，它构成 U 的一个划分，即：

$$U|r=\{X_1,X_2,\cdots,X_n\},\ \bigcup_{i=1}^{n}X_i=U$$

显然，$IND(R)=\bigcap R=\bigcap\{r|r\in R\}$ 也是论域 U 上的一个等价关系。

设 $X\subset U$，若 X 可以表示成 r 的一些等价类的并，则称 X 是 r 可定义的，否则，称 X 是 r-rough 的。记：

$$r_*(X)=\bigcup\{Y|Y\in U|R:Y\subseteq X\},$$

$$r^*(X)=U\{Y|Y\in U|R:Y\cap X\neq\phi\}$$

分别称为 X 的 r-下近似与 r-上近似。下近似 $r_*(X)$ 也称作 X 的 r-正区域，记作 $POS_r(X)$，它是由根据现有知识判断出肯定属于 X 的对象所组成的最大集合；上近似 $r^*(X)$ 为由那些根据现有知识判断出可能属于 X 的对象所组成的最小集合。$U-r^*(X)$ 称作 X 关于 r 的负域，记作 $NEG_r(X)$，可以解释为由那些根据现有知识判断出肯定不属于 X 的对象所组成的集合。$r^*(X)-r_*(X)$ 称作 X 的边界，记作 $BN_r(X)$，它为由那些根据现有知识判断出可能属于 X 但不能完全肯定是否一定属于 X 的对象中所组成的集合。比值：$a_r=\frac{card(r_*(X))}{card(r^*(X))}$ 称为 $X\subseteq U$ 的近似精度，其中，card(·)表示集合的基数。

$r\in R$，若 $IND(R-\{r\})=IND(R)$，则称 r 是等价关系集 R 中是可约去的，否则是不可约去的。若族集 R 中的每个关系 r 都是不可约去的，则称 R 是独立的，否则就称 R 是相关的。若 $P\subset R$，且 $IND(P)=IND(R)$，则称 P 是 R 中的一个约简，可见，等价关系集 P 与 R 包含相同的信息量。R 的所有约简的交称为 R 的核，记为 CORE(R)，它是 R 中所有不可约去的关系的集合。

若 P，Q 是 U 上的两个等价关系的集合，设 P 和 Q 是全域 U 上的等价关系的族集，称 $POS_P(Q)=\bigcup_{X\in U|Q}P_*(X)$ 为 Q 的 P-正区域，记作 $POS_P(Q)$，它是全域 U 上所有使用分类 U|P 所表达的知识能够正确地分类于 U|Q 的等价类的个体的

*)本文得到国家自然科学基金科研基金(60175018)资助。吴 涛 博士生，讲师，主要进行人工神经网络、数据挖掘及其在经济领域中的应用研究。张 铃 教授，博士生导师，主要研究领域为人工智能理论、人工神经网络理论。

维普资讯 http://www.cqvip.com

集合。若 $r\in P$，且：$POS_{IND(P-\{r\})}(IND(Q))=POS_{IND(P)}(IND(Q))$，则称关系 r 在族集 P 中是 Q-可约去的，否则称为 Q-不可约去的，如果在族集 P 中的每个关系 r 都是 Q-不可缺的，则称 P 关于 Q 是独立的，否则就称为 P 关于 Q 是相关的。若 $S\subset P$，且 $POS_S(Q)=POS_P(Q)$，称 S 为 P 的 Q-约简，族集 P 中的所有 Q-约简的交，称为族集 P 的 Q-核，记作 $CORE_Q(P)$，它是 P 中所有 Q-不可约去的关系的集合。

一个信息系统可以用四元组 $K=(U,A,V,f)$ 表示，其中：U 是对象的全体，即论域，A 是属性全体，它又可分为条件属性 C 和决策属性 D，即：$A=C\cup D$；$V=\bigcup_{a\in A}V_a$，V_a 是属性 a 的值域；$f:U\times A\rightarrow V$ 是一个信息函数，$f_x:A\rightarrow V$，$x\in U$，反映了对象 x 在 K 中的完全信息，其中 $f_x(a)=f(x,a)$。

这样，信息系统的一个属性就是 U 上的一个等价关系，可以按照前面的定义，约简条件属性。

2.2 覆盖算法

覆盖算法是根据样本数据自身的结构，构造性地建立神经网络模型。

设样本空间中的向量的长度都相等，即样本空间中的样本点都位于 n+1维空间中某个中心在原点的球面 S^n 上（若不然，可通过变换：$T:D\rightarrow S^n$，$T(x)=(x,\sqrt{R^2-|x|^2})$，将样本点映射到球面 S^n 上，其中，$R\geqslant\max\{|x_i|\}$）。则 $\omega' x-\theta>0$ 表示球面上由超平面 $\omega' x-\theta=0$ 所分割的正半空间的部分，称为球面上的"球形领域"，若 ω 与 x 等长，则 ω 就是这个球形领域的中心。每个球形领域作为一个神经元，取其激励函数为 $\sigma(\omega' x-\theta)$，其中，$\sigma(x)=\begin{cases}1 & x>1\\ 0 & 其它\end{cases}$，则激励函数就是"球形领域"的特征函数。适当地选择样本学习的起始点，构造一组覆盖所有学习样本的"球形领域"，并按"距离就近原则"建立神经网络。通过求球形领域的重心和平移球心位置，使之可以覆盖更多的样本点。

3. RCSN 网络的结构

我们将 Rough 集理论和覆盖算法结合起来，用 Rough 集理论在不减少信息的前提下，约简属性，降低输入向量的维数，用覆盖算法构造"球形领域"覆盖学习样本，这样可以减少覆盖的个数，降低运算量，提高学习的效率。其网络结构如下：

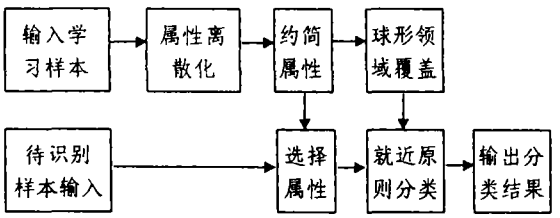

图1 RCSN 网络结构图

4. RCSN 网络的算法步骤

(1)求条件属性的约简

①离散化连续的条件属性；

②求出决策属性 D 的条件属性 C 的正区域 $POS_C(D)$；

③对每一个条件属性 c_i，计算 $POS_{C-\{c_i\}}(D)$，若 $POS_{C-\{c_i\}}(D)=POS_C(D)$，则 c_i 是可约去的，否则，c_i 不可约去，从而得条件属性的核 CORE(C)；

④求包含 CORE(C)的最小集合 P，使得 $POS_P(D)=POS_C(D)$，得条件属性 C 的最小约简 P。

(2)删除样本中不包含于 P 中的属性所在的列，得简化的学习样本 X。

(3)设学习样本分为$\{U_1,U_2,\cdots,U_k\}$k 类，构造样本 X 的覆盖：

①求简化的学习样本 X 的最大模 R，并将 X 投影到中心在原点、半径为 R 的球面上；

②取类别号 i=1，构造覆盖 C(i)；

③若 U_i 中没有尚未覆盖的点，则转⑧，否则，任取尚未被覆盖的一点 $a_i\in U_i$；

④求以 a_i 为中心的覆盖 $C(a_i)$，其对应的权值 $\omega=a_i$，阈值 $\theta=d(\omega)$，其中：

$$d^1(\omega)=\max_{x\in U_i}\{\langle a_i,x\rangle\},$$

$$d^2(\omega)=\min_{x\notin U_i}\{\langle a_i,x\rangle\mid\langle a_i,x\rangle>d^1(\omega)\},$$

$$d(\omega)=\frac{1}{2}(d^1(\omega)+d^2(\omega));\qquad(1)$$

⑤求 $C(a_i)$所覆盖的点的重心，并将其映射到球面上，设投影点为 a_i'，按(1)式计算其阈值 θ'，得球形领域 $C(a_i')$；

⑥若 $C(a_i')$覆盖的点数大于 $C(a_i)$所覆盖的点数，则令 $a_i'\rightarrow a_i$，$\theta'\rightarrow\theta$，返回⑤，否则，转⑦；

⑦求 ω 的平移点 ω''，并求对应的球形领域 $C(a_i'')$，若 $C(a_i'')$覆盖的点数大于 $C(a_i)$所覆盖的点数，转⑤，否则，得 C(i)的一个覆盖；

⑧若 i<k，则 i+1→i，转③，否则，转(4)

(4)根据(1)求出的属性约简 P，简化待识别样本，并将其投影到球面上；

(5)用(3)中所得的覆盖，按就近原则，对待识别样本进行分类。所谓的就近原则即：若存在一个 i，使得待识别的点位于 C(i)的某个球形领域内，则该点的类别号为 i，否则，找出中心与该点距离最近的球形领域，该领域所属的类别就是此待识别样本的类别。

5. RCSN 网络的应用

我们使用 UCI 数据库中由 Johns Hopkins 大学的 Vince Sigillito 提供的电离层数据库(ionosphere database)进行试验。该数据库是收集于加拿大的 Goose Bay，Labrador 的雷达信号，它包含17脉冲信号，每个脉冲有两个属性，信号分为"好"和"坏"两种，"好"意指可反映电离层的结构，而"坏"是指信号穿过电离层，不能反映电离层的结构，即数据库共有35个属性，该数据库的前200个样本中"好"和"坏"各占50%，后150个样本中，类别为"好"的有123个。Sigillito，V.G.，Wing，S.P.，Hutton，L.V. 和 Baker，K.B. 以此数据库的前200个样本作学习样本、后150个作测试样本，用感知机和 BP 网络对该数据库进行分类(1989)。其结果是：线性感知机的正确率为90.7%，非线性感知机的正确率为92%，BP 网络的正确率为96%。我们应用 rough 理论对34个属性进行约简，得最小独立属性集为8个属性，然后，以前200个样本作学习样本、后150个作测试样本，用 CSN 和 RCSN 对该数据库进行分类，结果如下：

表1 CSN 和 RCSN 的参数比较

项目	CSN	RCSN
学习样本数	200	200
测试样本数	150	150
正确率(%)	92.053	95.364
覆盖数	74	52
运行时间(s)	1.05	0.65

维普资讯 http://www.cqvip.com

众所周知，运行时间长一直是感知机与 BP 神经网络在实际应用中难以克服的困难，覆盖算法利用神经元的几何意义，开辟了一条新的研究神经网络的路径，克服了网络运行速度瓶颈问题。而 Rough 集理论在保证信息量不变的前提下，提取最少个数的特征属性，降低样本空间的维数，加大了样本点的分布密度，从而，可以减少覆盖的个数，提高系统的精度。同时，也减少了数据的收集与存储成本。但是，CSN 与 RCSN 的运行时间和精度与学习中初始点的选取密切相关，如何选择样本的学习顺序是我们正着手研究的问题。

参 考 文 献

1 张钹，张铃. 人工神经网络的设计方法. 清华大学学报(自然科学版)，1998，38(S1)：1～4

2 张铃，张钹，殷海风，多层前向网络的交叉覆盖算法. 软件学报，1999，10(7)：737～742

3 Pawlak A. Rough Sets. Theoretical Aspects of Reasoning about Data [M]. Kluwer Academic Pub.，1992

4 张文修，吴伟志. 粗糙集理论介绍和研究综述. 模糊系统与数学，2000，14(4)：1～12

5 王珏，苗夺谦，周育健. 关于 Rough Set 理论与应用的综述[J]. 模式识别与人工智能，1996，9：337～344

6 张铃，张钹. M-P 神经元模型的几何意义及其应用. 软件学报，1998，9(5)：334～338

7 Pawlak Z. Rough set theory and its applications to data analysis [J]. Cybernetics and Systems：An International Journal，1998，29：661～688

8 吴鸣锐，大规模模式识别问题的分类器设计研究：[工学博士学位论文]. 北京：清华大学计算机系，2000

9 刘清. Rough 集及 Rough 推理. 科学出版社，2001. 8

(上接第56页)

5 算法示例

数据表1可以描述为一个知识表达系统 KS＝(Universe，Attr)，其中 Universe＝{对象1，对象2，对象3，对象4，对象5，对象6，对象7，对象8}，Attr＝{C_1，C_2，C_3，D}。当表1被描述为一个知识表达系统后，就可以用对知识表达系统操作的简化运算对它进行操作。比如 Rule(对象1，{C_1，C_2，C_3}，{D})是它的一条函数规则，可以根据定义直接判断，也可以根据定理1来判断，我们要用知识表达系统函数规则的简化算法对这条规则进行简化。

数据表1　可以描述为一个知识表达系统

Universe 对象	C_1高度	C_2头发	C_3眼睛	D 分类
对象1	矮	黑	蓝	—
对象2	高	黑	蓝	—
对象3	高	黑	棕	—
对象4	高	红	蓝	+
对象5	矮	黄	蓝	+
对象6	高	黄	棕	—
对象7	高	黄	蓝	+
对象8	矮	黄	棕	—

生成 $[对象1]_{\cap ind\{C_1,C_2,C_3\}}$ 和 $[对象1]_{\cap ind\{D\}}$，$[对象1]_{\cap ind\{C_1,C_2,C_3\}}$＝{对象1}，$[对象1]_{ind\{D\}}$＝{对象1，对象2，对象3，对象6，对象8}；

$[对象1]_{\cap ind\{C_1,C_2,C_3\}} \subseteq [对象1]_{\cap ind\{D\}}$，所以 *Rule*(对象1，{$C_1$，$C_2$，$C_3$}，{*D*})是函数规则；Family＝{$C_1$，$C_2$，$C_3$}；

生成 $[对象1]_{\cap ind(Family-\{C_1\})}$＝{对象1，对象2}，因为 $[对象1]_{\cap ind(Family-\{C_1\})} \subseteq [对象1]_{\cap ind\{D\}}$，所以 Family＝{$C_2$，$C_3$}；

生成 $[对象1]_{\cap ind(Family-\{C_2\})}$＝{对象1，对象2，对象4，对象5，对象7}，因为 $[对象1]_{\cap ind(Family-\{C_2\})} \not\subset [对象1]_{\cap ind\{D\}}$，所以 Family＝{$C_2$，$C_3$}；

生成＝$[对象1]_{\cap ind(Family-\{C_3\})}$＝{对象1，对象2，对象3}，因为 $[对象1]_{\cap ind(Family-\{C_3\})} \not\subset [对象1]_{\cap ind\{D\}}$，所以 Family＝{$C_2$，$C_3$}；

输出 Family＝{C_2，C_3}，算法结束。

结束语　本文旨在研究粗集理论中函数规则的简化和集族的相对交简化之间的相互转化关系。这项研究在粗集理论中的位置及其在粗集理论与其它研究领域中的位置表示在图1中。从图1可以看到，对于知识表达系统，粗集理论中已经涉及到了它的一些简化问题，但缺乏形式语言描述；对于等价关系族和集族的简化，粗集理论已经用数学语言描述了几种具体的形式，但缺乏对两类简化问题关系的探讨；粗集理论与其它的一些应用领域的接口主要是决策表的简化。我们认为：对知识表达系统、等价关系族、集族的简化应该可以提出更多的具体形式，相应地，两类转化问题之间会存在更为复杂的转化关系；应用领域与粗集理论的接口也不应该限制到决策表的简化上。本文的工作是在这个大思路下产生的。沿着这个思路，还可以产生其它的结果，这些结果将使得粗集理论更加丰富和严谨，使得其它领域对粗集理论的应用更加广阔和灵活。

附录：本文经常使用的符号说明

-：如果符号的左右两边是两个子集合，那么符号表示两个集合的差运算；

$\subseteq$：左边的子集包含在右边的子集中；

$\subset$：左边的子集真包含在右边的子集中；

$\not\subset$：左边的子集不包含在右边的子集中；

$\in$：左边的元素属于右边的集合；

$\varnothing$：空子集；$\forall$：量词任意；$\exists$：量词存在；

Equ[Universe]：集合 Universe 上所有的等价关系组成的集合；

IFF：当且仅当；

$[x]_r$：x 是集合中的一个元素，r 是这个集合上的一个等价关系，$[x]_r$ 表示 x 所在的 r 的等价类；

$P \xrightarrow{T} Q$：P，T，Q 是三个命题，$P \xrightarrow{T} Q$ 表示根据命题 T，由命题 P 可推出命题 Q；

$P \xrightarrow{T} Q$：P，T，Q 是三个命题，$P \xrightarrow{T} Q$ 表示根据命题 T，命题 P 成立当且仅当命题 Q 成立；

P()：如果括号中是一个集合，则符号表示这个集合的幂集。

参 考 文 献

1 Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning about Data，Kluwer Academic Publications，Dordrecht，1991

2 刘清. Rough 集及 Rough 推理. 科学出版社，2001

3 曾黄麟. 粗集理论及其应用——关于数据推理的新方法(修订版)，重庆大学出版社，1998