构造性覆盖算法的知识发现方法研究

张 旻[①②] 张 铃[②]
[①](解放军电子工程学院 702 研究室 合肥 230037)
[②](安徽大学人工智能与信号处理国家教育部重点实验室 合肥 230039)

摘 要 该文提出一种新的基于构造性覆盖算法的知识发现方法。由于覆盖网络构造方法的特殊性，使得形成的每个覆盖领域都很有价值，对覆盖领域内样本分析能挖掘出数据内在的知识，且可以根据需求构造不同的覆盖网络，形成对数据的多侧面的分析；实验结果表明覆盖算法用于知识发现的方法是有效可行的。

关键词 覆盖算法, 覆盖领域, 等价关系, 知识发现
中图分类号：TP18，TP311.13 **文献标识码**：A **文章编号**：1009-5896(2006)07-1322-05

Study on the Method of Knowledge Discover Based on the Structured Covering Algorithm

Zhang Min[①②] Zhang Lin[②]
[①](702 *Research Room*, *Electronic Engineering Institute of PLA*, *Hefei* 230037, *China*)
[②](*Ministry of Education*, *Key Lab of IC&SP at Anhui University*, *Hefei* 230039, *China*)

Abstract This paper proposes a new method of knowledge discovery based on the structured covering algorithm. Since the network of covering domains is constructed through a special method, it makes each covering domain a valuable pattern. Through analyzing the samples covered by the covering domain, certain valuable pattern will be found, which includes clustering information of samples, association rules among the data, the outlier analysis, etc. And in order to meet different requirement of users, different covering networks can be structured, therefore, multi-sides of data can be analysed. The experiments show that using the covering algorithm to discover knowledge of data is effective and feasible.

Key words Covering algorithm, Covering domain, Equipollent relationship, Knowledge discover

1 引言

数据挖掘是指从数据中识别合理的、新颖的、有潜在价值的、以及最终可理解模式(pattern)的非常规过程。数据挖掘面临的一个主要问题是数据中潜在可能的关系模式数量太大，因此要想搜索到有用的模式，必须采用一定的人工智能技术，特别是机器学习领域的方法。文献[1,2]提出一种M-P模型的几何表示，在此基础上形成了神经网络覆盖算法，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，在样本的分类应用中取得了很好的效果[3－5]。算法在构造神经覆盖网络时充分利用了样本的类别标记以及样本相似性等样本自身特征，形成覆盖领域既包含了样本的聚类信息，又包含了样本的类别信息，因而是一个很有价值的模式。本文在研究覆盖算法的基础上，发现覆盖算法除了适合分类研究外，还非常适合进行数据挖掘，因此首次提出了利用覆盖算法进行数据挖掘。

为了保证一定的完整性，本文首先简单介绍文献[1]中给出的几何表示，然后研究如何在此基础上进行知识发现，最后给出了实验结果和结论。

2004-12-16 收到, 2005-10-17 改回
国家自然科学基金(60175018)资助课题

2 构造性覆盖神经网络

1943 年McCulloch和Pitts[6]根据神经元传递中的“0,1 率”和神经传递中信号不但有不同的强度，而且有兴奋和抑制两种情况，第一次提出神经元的数学模型(M-P模型)，此模型一直沿用至今。

将神经元看作是一个有 n 个输入和一个输出的元件，此元件的激励函数可表示为 $y=\sigma(Wx-\theta)$，其中 σ 是符号函数，即 $\sigma(x)=\begin{cases}1, & x>0\\ -1, & x\le 0\end{cases}$，现在分析一下这个模型的几何意义。

(1) 超平面表示 由神经元激励函数的定义可知，它由两个函数复合而成，其中第一个函数为 $(Wx-\theta)$，若令其等于零，得 $Wx-\theta=0$。这个方程在 n 维空间中表示一个超平面 P；当 $(Wx-\theta)>0$ 时，表示点落在超平面的正半空间内，此时，$\sigma(Wx-\theta)=1$；当 $(Wx-\theta)<0$ 时，表示点落在超平面的负半空间内，此时 $\sigma(Wx-\theta)=-1$。于是，一个M-P 神经元的功能可看成是一个由超平面划分的空间位置的识别器。这给神经元一个相当直观的解释。然而当 n, m 较大时，即 n 维空间中 m 个超平面的相交情况时，就非常复杂、很不直观。故此后很少有人再用神经元超平面表示的几何意义来帮助研究神经网络学习了。

(2) 超球面上的“领域”(neighborhood)表示　几何直观往往是研究的很好向导，但是超平面的方法遇到困难，张铃教授对此进行了改变[1]，将会给我们提供新的直观帮助。

如图1(a)，若我们限定输入向量的长度相等，即输入向量限定在n维空间的某个球面上，那么这时 $(Wx-\theta)>0$ 就表示球面上落在p正半空间的部分，这个部分恰好是球面上的某个“球形领域”，若取W与x等长，则这个“球形领域”的中心恰好是W，其半径为 $r(\theta)$,它是 θ 的单调下降函数。若我们取 $\sigma'(x)=\begin{cases}1, & x>0\\ 0, & x\le 0\end{cases}$，且取神经元的激励函数为 $\sigma'(Wx-\theta)$，则一个神经元的激励函数正好是它所代表的球面上“球形领域”的特征函数。这样，我们就能非常直观地进行神经网络的各项研究。

通过这种变换，我们就将神经网络的最优设计问题转化成某种最优覆盖问题。当给定的输入向量的长度不相等时，可用下面给出的方法将它变换成长度相等的情况。设输入的定义域为 n 维空间中的有界集合 D，令 S^n 是 n+1 维空间中的 n 维的超球面，作变换

$T: D\to S^n$，$T(x)=\left(x,\sqrt{r^2-\|x\|^2}\right)$将样本点映射到球面 S^n 上，其中，$r\ge \max\{|x_i|\}$。

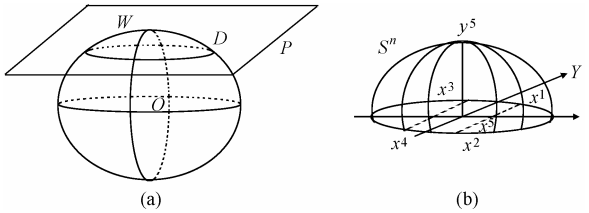
这个变换几何直观上可以理解为，将D看成是位于n+1维空间中过原点的一个n维超平面上，而且D位于 S^n 的内部，则变换T就是将D上的点垂直投射到 S^n 的上半球面上(图1(b))。这种变换显然是一一对应的。如上面所述，这时每一个神经元 (W,θ) ，就是在超球面 S^n 上以W为中心，以 $r(\theta)$ 为半径的一个“球形领域”的特征函数(其中 $r(\theta)=r(\cos(\theta/R))$。

(3) 覆盖领域的构造　设样本集 $S=\{r^t=(x^t,y^t),$ $t=0,1,\cdots,p-1\}$，集合中不同的 y^t (类标记)只有 k 个，不妨设 y^t 的前 k 个的值均不相同。令样本输出为 y^t 样本标记的集合 $I(t)$ (即 $I(t)=\{i\mid y^i=y^t\}$) ,其对应的输入集合记为 $P(t)$, t=0,1,2,,$\cdots$, k−1。并将输入样本的上标按 $I(0), I(1),\cdots, I(k-1)$ 的顺序排列，于是，我们能够取一批“球形领域”$\left\{C_j^t, t=0,1,\cdots,k-1; j=1,2,\cdots,k_t\right\}$。令 $C=\cup C_j^k$，使得 C^t 只覆盖 j 属于 $I(t)$ 的 x^j，而不覆盖 j 不属于 $I(t)$ 的 x^j，且 C^t 互不相交。这样球形领域就能达到学习分类的目的。固定 t，令 $P(t)=\{x^i\mid i\in I(t)\}$，对样本输入 $x^i\in P(t)$,覆盖领域按式(1)构造。

$$\left.\begin{aligned} d^1(i)&=\max_{j\notin I(t)}\{<x^i,x^j>\},\\ d^2(i)&=\min_{j\in I(t)}\left\{\begin{matrix}<x^i,x^j>|<x^i,x^j>\\ >d^1(i)\end{matrix}\right\}\\ d(i)&=\frac{1}{2}\left(d^1(i)+d^2(i)\right)\end{aligned}\right\} \tag{1}$$

其中<x, y>表示内积。

计算 $d(i)$，作以 x^i 为中心、阈值 θ =$d(i)$的覆盖 C_j^i，并通过求球形领域的重心和平移领域中心位置，使之可以覆盖更多的样本点，并按此方法求出样本的全部覆盖领域。



(a)　(b)

图 1 超平面“球形领域”和 $D\to S^n$ 变换图

(a) 超平面P与超球面相交，形成“球形领域”的示意图

(b) 从 $D\to S^n$ 变换图

Fig.1 Schematic diagrams of superplane of a sphere neighborhood and its transformation

(a) Schematic diagram of a sphere neighborhood

(b) Transform diagram of $D\to S^n$

3　覆盖算法在数据挖掘中应用

3.1　覆盖算法分析

下面对覆盖算法做进一步分析：设样本集 $S=\{r^t=(x^t,y^t),t=0,1,\cdots,k-1\}$，那么，当输入为$x^i$时网络的输出为$y^i$，记住$x^i$与$y^i$之间的对应关系；至于“联想”就是当输入为$x^i$+$\Delta_i$时，其输出仍为$y^i$。满足上面两种功能要求的网络就可以认为是一个“联想记忆器”了。另一方面，若将“当输入为x^i+Δ_i时，其输出仍为y^i”理解为输入落在x^i的附近时，其输出应为y^i，否则其输出就不为y^i；将在“x^i附近”看作是x^i的一个球形领域，由神经元的几何意义得知，若取以x^i为中心，以Δ_i为半径的“球形领域”，则其对应的神经元就能完成上述联想记忆的任务。若将多层前向神经网络作为“分类器”来进行设计，那么其设计就相当于用若干“球形领域”将输入x^i按其所属的类别把它们划分开来。这是覆盖算法在分类领域成功应用的基本思想。

式(1)中建立的以x^i为中心，Δ_i为半径的“球形领域”之所以具有联想记忆功能，是因为领域中样本的类标记都一致(y^i)，样本投影在超球面的相互距离较近，样本间具有很强的相似性，因此将在“x^i附近”球形覆盖领域的样本“联想”为具有等价关系的不可分辨样本；而不同覆盖领域在超球面的位置不同，反映出它们覆盖的样本具有一定的差异；对于远离同类点的离群样本，在超球平面上形成了孤点，只能由新的覆盖领域进行覆盖，这就为离群点的分析创造了条件。

为了形象说明覆盖算法问题，图 2 从二维覆盖图进行说明。

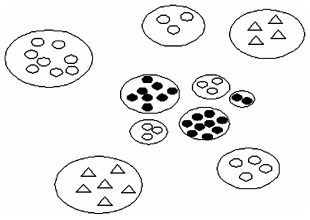


图 2　原样本覆盖图

Fig.2 Covering picture of the samples

样本共分为黑实心点、空心圆点、空心三角点 3 类，分别用C_1, C_2, C_3表示，分别有 3、5、2 个覆盖领域，每个领域的覆盖样本数不同。覆盖领域内的样本有以下特点：一是同一覆盖领域内的样本具有相同的类标记，样本间位置相近，因此样本具有相似性；二是同一类标记的样本，如果差异较大，不会在同一覆盖领域内，会形成多个覆盖领域；三是投影位置相近的样本，如果类别不同，也不可能聚集在同一覆盖领域。这些独特的特征使得覆盖网络中的每个覆盖领域就是一个很好的模式，值得挖掘。

3.2 覆盖领域的不可分辨关系

从上面的分析可知同一覆盖领域的样本具有很强的相似性，因此我们可以定义覆盖领域等价关系的概念，即将同一覆盖领域样本看成是不可分辨的等价关系。

设x, y是覆盖网络U的两个样本, 若 $(x, y)\in$ 同一覆盖领域C_j^t, 则称 x 与 y 在整个覆盖网络中是不可分辨的。U/R 是 U 上由 R 生成的等价类全体, 它构成了 U 的一个划分。即

$$U/R=\{X_1, X_2, \cdots, X_n\}$$

其中 $X_i \subseteq U$ 是等价类，对应的是第i个覆盖领域中的样本，$X_i \neq \Phi$，$X_i \cap X_j=\Phi$，且$i\neq j, i, j=0,1,\cdots,n$，$\bigcup X_i = U$ 。这样有多少覆盖领域，就能形成相同数量的等价类。

3.3 覆盖领域的近似集

有了以上的覆盖的等价关系，就可以参照粗糙集的相关知识定义覆盖领域样本的近似集。

若论域中存在子集 $X \subseteq U$ ，我们根据关系 R 定义的基本集合($Y_i, i=1,\cdots,n$)的描述来划分 X。为了准确地说明某些 Y_i 在 X 中对象的隶属度的情况，这里考虑两个子集：

覆盖领域 X 下近似集： $R\_(X)=U\{Y_i \in U \mid R: Y_i \subseteq X\}$

覆盖领域 X 上近似集： $R^-(X)=U\{Y_i \in U \mid R: Y_i \cap X \neq \phi\}$

有了覆盖领域等价关系和近似集的概念，就可以参照粗糙集的一些运算规则，完成对数据的各种运算，获取数据的多种知识了。

3.4 覆盖样本的知识发现

覆盖网络中有多少个覆盖领域，就会形成多少个等价的样本集，且可以进行各种运算，形成多种模式的样本集。是不是所有形成的模式都是有价值的？答案是否定的，因此我们有必要定义覆盖模式支持度的概念。

下面我们研究对覆盖领域样本的知识分析。

定义 1 C_i^j 为样本类标记为 i 的第 j 个覆盖领域，$\mathrm{card}(C_i^j)$ 为该覆盖领域的样本数量，$\mathrm{card}(X)$ 为整个学习样本数据数量，则该覆盖领域的支持度为： $s=\dfrac{\mathrm{card}(C_i^j)}{\mathrm{card}(X)}$ 。

支持度说明该覆盖领域模式在所有事务中有多大的代表性，显然，支持度越大，该模式越重要，应用越广泛。所以可以预先根据需要，给出 s，我们只关心大于阈值 s 的覆盖领域，这些覆盖领域的关联规则被认为是新颖、实用的，而且能大大提高数据挖掘的效率。

按式(1)构造某一覆盖领域时，该覆盖领域的样本为同一类别，有时在构造覆盖领域时允许少量的其它类别的样本进入覆盖领域，这样就有必要定义下面确信度的概念。

定义 2 C_i^j 为样本类标记为 i 的第 j 个覆盖领域，$\mathrm{card}(C_i^j)$ 为该覆盖领域的样本数量，$\mathrm{card}(C_{i,i}^j)$ 为该覆盖领域样本类别为 i 的样本数量，该覆盖领域类别为 i 的确信度 c 表示为： $c_k=\dfrac{\mathrm{card}(C_{i,i}^j)}{\mathrm{card}(C_i^j)}$ 。

如果给出一定的确信度，对覆盖领域的条件稍加改变，允许类标记不为 i 的样本进入覆盖领域 C_i^j，能够克服由于噪声干扰引起的个别其它类别的样本投影位置偏差对构造覆盖网络的影响，从而得到满足确信度的覆盖模式。

覆盖算法还有以下特点：

(1) 多侧面覆盖样本规则分析　在覆盖学习过程中，样本投影是利用样本的属性特征，对于一个信息系统(如为决策系统将决策属性视为条件属性即可转变为信息系统)选择不同的属性特征为条件和决策特征，就能形成不同的决策系统，而不同决策系统的样本投影到球面的位置就会发生变化，构造的覆盖网络也会随之发生变化，也就是等价关系R发生了变化，形成的等价类和划分也不同，可以满足对数据多侧面的规则分析，方便、灵活的挖掘各种有价值的规则[8]。

(2) 侧面覆盖知识的合成　对于多侧面获得的覆盖信息，我们可以进行如下合成处理。

设学习样本按 $T=(U, A, C, D)$进行构造，得到满足兴趣度的覆盖领域形成划分：

$$U/R=\{X_1, X_2, \cdots, X_n\}$$

同理如果按 $T'=(U, A, C', D')$ 进行构造，得出满足兴趣度的覆盖领域形成划分：

$$U/R'=\{Y_1, Y_2, \cdots, Y_m\}$$

定义两个覆盖领域的合成划分为

$$U/RR'=\{X_i \cap Y_j \mid X_i \subset U/R, Y_j \subset U/R'\}$$

这样就能得到不同侧面样本的覆盖合成规则。

(3) 大规模数据的处理　一般来说，没有一个标准的步骤来指导神经网络的开发工作，神经网络模型需要花大量的时间进行设计。但是神经覆盖算法将空间搜索策略改变为构造覆盖方法，可有效防止组合爆炸，使得构造的难度大大减少，时间大大缩短，因此适合对特征空间维高、样本数据大的问题进行构造[7]。从几个应用领域实践结果[4,6,7]来看,效果很好，如成功完成了对 700 类手写体汉字构成的 49000 个 256 维海量数据的构造[4,6]，反映出覆盖算法满足对海量数据的处理。

综上所述覆盖算法适合数据的知识发现。

4 算法实现

(1) 覆盖领域的构造　设学习样本为 X，构造样本 X 的

覆盖步骤如下：

(a) 求学习样本 X 中样本的最大模 r，并将 X 中的点投影到中心在原点、半径为 r 的球面上；

(b) 取类别号 i=1，构造覆盖 $C(i)$;

(c) 若 X_i 中没有尚未覆盖的点，转(h)，否则，任取 X_i 中尚未被覆盖的一点 a_i；

(d) 按式(1)求以 a_i 为中心，阈值 θ=$d(\omega)$ 的覆盖 $C(a_i)$；

(e) 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'，按步骤(d)中的公式计算其阈值 θ'，得球形领域 $C(a_i')$；

(f) 若 $C(a_i')$覆盖的点数大于 $C(a_i)$ 所覆盖的点数，则令 $a_i' \to a_i$，$\theta' \to \theta$，返回(e)，否则，转(g)；

(g) 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(e)，否则，得 $C(i)$的一个覆盖，保存覆盖样本数 $n(i)$；若 $i<m$，则 $i+1 \to i$，转(c)，否则，转(h)；

(h) 构造结束。

训练结束得到n个超球形覆盖领域，即训练样本被分为n类不同覆盖集合$C=\{C_1, C_2, \cdots, C_n\}$。

(2) 数据挖掘　设置支持度的门限，保留高于支持度的覆盖领域，并根据数据挖掘需要，对这些覆盖领域数据进行相应处理，挖掘感兴趣的规则。

5　实例分析

我们开发了覆盖算法的应用程序，对来源于 http://www.ics.uci.edu/mlearn/MLRepository.html 的国旗数据库进行数据分析。

数据库有 194 个例子(国家或地区)，从提供的信息来看，数据文件包含各个国家以及国旗的具体内容，除去国名，涉及 29 个属性，包括洲、地域、面积、人口、语言、宗教、竖条数、横条数、不同颜色数、红、绿、蓝、黄、白、黑、橙、支配颜色、圆圈数、直立十字、X 形十字、四分部分数量、太阳或星数量、月牙、三角、图标、文字、左顶部、左底部，　该数据库没有决策属性，是一个典型的信息系统。直接从 29 个属性直接分析，很难分析出有用的知识，我们利用覆盖算法从以下几个方面对数据进行分析：

(1) 29 个属性的覆盖学习　选择 support=2%，确信度为 100%, 对 194 个例子来说，覆盖领域的样本数应大于 4, 同一覆盖领域的样本标记一致。

(a) 按语种覆盖分类　首先，指定语言为决策属性，其余 28 个为条件属性，将信息系统变成决策系统。语言的类别有英语、西班牙语、法语、德语、斯拉夫语、其它印欧语系、汉语、阿拉伯语、日语/土耳其语/芬兰语/马儿扎语、其它语种等共 10 类，经过学习后，得出满足支持度的覆盖领域样本有：

$C_{英语\,1}$={百慕大群岛，斐济，蒙特塞拉特岛，St.-Helena, British-Virgin-Isles，开曼群岛，福克兰群岛，Turks-Cooks-Islands}

$C_{英语\,2}$ ={安提瓜岛-巴布达岛, 巴哈马群岛, 圭亚那, St-Kitts-Nevis, St.-Lucia, 特立尼达岛-多巴哥岛}

$C_{英语\,3}$ ={澳大利亚, Cook-Islands，新西兰, 纽埃岛，图瓦卢，英国(UK)，西萨摩亚群岛}

$C_{西班牙\,1}$ ={阿根廷, 萨尔瓦多, 危地马拉, 洪都拉斯, 尼加拉瓜}

$C_{西班牙\,2}$ ={智利，哥斯达黎加，古巴，波多黎各}

$C_{法语}$ ={Burkina，几内亚，马里，塞内加尔}

$C_{阿拉伯语}$ ={伊拉克，约旦，科威特，北也门，南也门，苏丹，叙利亚，阿拉伯联合酋长国}

$C_{其它语种}$={安哥拉，刚果，埃塞俄比亚，柬埔寨，越南，扎伊尔}

语言属性相同的相似国旗情况如表 1 所示。

表 1　语言属性相同的相似国旗情况表

Tab.1 Similarity of flag with same language attribute

语言	国家数	相似国旗数	比例(%)
英语	43	8	18.6
		6	13.9
		7	16.3
阿拉伯语	19	8	42
西班牙语	21	5	23.8
		4	19.1
其他语种	46	6	13

(b) 按宗教覆盖分类　指定宗教为决策属性，其余 28 个为条件属性，将信息系统变成决策系统。宗教的类别有天主教、其它基督教(Other Christian)、穆斯林、佛教、印度教、异教徒(Ethnic), 其他类型等共 8 类，经过学习后，得出满足支持度的覆盖领域样本有：

$D_{天主教}$={阿根廷，智利，哥斯达黎加，多米尼加共和国，萨尔瓦多, 危地马拉, 洪都拉斯, 尼加拉瓜，圣马力诺，乌拉圭}

$D_{其它基督教\,1}$={{安提瓜岛-巴布达岛, 巴哈马群岛, 巴巴多斯岛, St.-Lucia}

$D_{其它基督教\,2}$={British-Virgin-Isles，开曼群岛，斐济，St-Helena，Turks-Cooks-Islands}

$D_{其它基督教3}$={澳大利亚, Cook-Islands，法罗群岛，冰岛，新西兰, 纽埃岛，图瓦卢，英国(UK)，西萨摩亚群岛}

$D_{穆斯林\,1}$ ={伊拉克，约旦，科威特，北也门，南也门，苏丹，叙利亚，阿拉伯联合酋长国}

$D_{穆斯林2}$={阿尔及利亚，Comorro-Islands，马尔代夫，尼日利亚，巴基斯坦}

宗教属性相同的相似国旗情况如表 2 所示。

表 2　宗教属性相同的相似国旗情况表

Tab.2 Similarity of flag with same religion attribute

宗教	国家数	相似国旗数	比例(%)
天主教	40	10	25
其它基督教	60	5	8.3
		4	6.7
		9	15.0
穆斯林	19	8	42
		5	26.3

(c) 进行宗教覆盖和语言覆盖的样本合成　以分别从宗

教和语言角度对数据进行了覆盖聚类分析，现对覆盖结果进行合成, 设定覆盖样本数≥4，得出以下结论：

Y_1={澳大利亚, Cook-Islands，新西兰, 纽埃岛，图瓦卢，英国(UK)，西萨摩亚群岛}

Y_1. 语言＝英语 and Y_1. 宗教＝其它基督教

Y_2 ={阿根廷, 萨尔瓦多, 危地马拉, 洪都拉斯, 尼加拉瓜}

Y_2.语言＝西班牙 and Y_2.宗教＝天主教

Y_3={伊拉克，约旦，科威特，北也门，南也门，苏丹，叙利亚，阿拉伯联合酋长国}

Y_3. 语言＝阿拉伯语 and Y_3.宗教＝伊斯兰教

Y_4={ British-Virgin-Isles，开曼群岛, 斐济，St-Helena，Turks-Cooks-Islands}

Y_4.语言＝英语 and Y_4.宗教＝其它基督教

语言、宗教属性相同的相似国旗情况如表 3 所示。

表 3 语言、宗教属性相同的相似国旗情况表

Tab.3 Similarity of flag with combination features of religion and language

语言	宗教	国家数	相似国旗数	比例(%)
英语	其它基督教	36	7	19.4
			5	13.8
阿拉伯语	穆斯林	19	8	42.0
西班牙语	天主教	20	5	25.0

(2) 去除数据中洲、地域、面积、人口等属性重新进行覆盖学习 用剩下的 23 个属性进行覆盖学习，同样分别从宗教、语言两个侧面进行覆盖分析，然后进行合成处理，最终结果如下：

Y_1={澳大利亚, Cook-Islands，新西兰, 纽埃岛，图瓦卢，英国(UK)}

Y_1. 语言＝英语 and Y_1. 宗教＝其它基督教

Y_2={阿根廷, 萨尔瓦多, 洪都拉斯, 尼加拉瓜}

Y_2.语言＝西班牙 and Y_2.宗教＝天主教

Y_3={伊拉克，约旦，科威特，北也门，苏丹，叙利亚，阿拉伯联合酋长国}

Y_3. 语言＝阿拉伯语 and Y_3. 宗教＝伊斯兰教

Y_4={British-Virgin-Isles，开曼群岛, 斐济，Turks- Cooks-Islands}

Y_4.语言＝英语 and Y_4.宗教＝其它基督教

语言、宗教属性相同的相似国旗情况如表 4 所示。

表 4 语言、宗教属性相同的相似国旗情况表

Tab.4 Similarity of flag with combination features of religion and language

语言	宗教	国家数	相似国旗数	支持度(%)
英语	其它基督教	36	6	16.7
			4	11.1
阿拉伯语	穆斯林	19	7	36.8
西班牙语	天主教	20	4	20.0

从国旗的 29 个属性进行覆盖学习后得到覆盖样本聚类的情况看，国旗确实在某种程度上反映了国家的或地区的文化、宗教信仰、历史等因素。从语言方面分析的结论：英语、西班牙语、法语、阿拉伯语、其他语种的个别国家的国旗有相似性，而剩下语种的国旗差别很大，没有什么相似性；从宗教覆盖领域样本聚类的情况看，天主教、其他基督教、伊斯兰教等国家国旗具有一定的相似性，其他的宗教国旗没有这种相似性。结合宗教和语言的综合结果可以得出说阿拉伯语，信仰伊斯兰教的 19 个国家中有 8 个国家的国旗基本一致，占 42%；而说英语，信仰其他基督教的 36 个国家中分别有 7 个和 5 个两组国家的国旗基本一致，分别占 19.4%和 13.8%；说西班牙语， 信仰天主教的 20 个国家有 5 个国家的国旗基本一致，占 25%，其他语种和宗教信仰的国家的国旗没有相似性。去除洲、地域、面积、人口等属性，得出基本同样的结论。

6 结束语

由于覆盖网络构造的特殊性和直观性，形成的每个覆盖领域就是一个有价值的模式；将同一领域的样本定义为不可分辨的等价类，这样就可以利用已知的数据工具完成对数据的各种分析；覆盖算法直观，具有广阔的应用前景。

本文虽然取得了一定的成果，但仍有很大的改进的空间，特别是如何利用数学工具完成对等价的覆盖数据深入处理，获取数据内在更多的有价值知识，目前研究的还不够，现正在进行这方面的工作，以期待覆盖算法成为数据挖掘领域真正有效可行算法。

参 考 文 献

[1] Zhang Ling, Zhang Bo. A geometrical representation of McCulloch-Pitts neural model and its applications. *IEEE Trans. on Neural Networks*, 1999, 10(4): 925−929.

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

[4] 陶品, 张钹等. 构造型神经网络双交叉覆盖增量学习算法. 软件学报, 2003, 14(2): 194−201.

[5] 叶少珍, 张钹等. 一种基于神经网络覆盖构造算法的模糊分类器. 软件学报, 2003, 14(3): 429−434.

[6] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*,1943, 5: 115−133.

[7] 张燕平, 张铃等. 基于覆盖的构造学习算法SLA及其在股票预测中的应用. 计算机研究与发展，2004, 41(6)：979−984.

[8] 张燕平, 张铃, 吴涛. 机器学习中的多侧面递进算法MIDA. 电子学报, 2005, 33(2): 327−331.

张 旻： 男, 1966 年生, 博士，副教授, 从事数据挖掘、计算智能等方向研究.

张 铃： 男, 1937 年生，教授，博士生导师，研究方向为智能计算、人工智能等.