文章编号：1672-2892(2010)05-0607-06

基于粗糙集和决策树法的认知无线电知识挖掘

余晓航[1a]，李磊民[1b]，黄玉清[1b]

(1.西南科技大学 a.信息工程学院；b.国防科技学院，四川 绵阳 621010)

摘 要：对粗糙集、决策树C4.5算法进行了研究，提出用粗糙集和决策树相结合的方法设计CR知识挖掘模型，并通过案例研究其可行性；利用基于MATLAB 802.11a物理层仿真平台收集的数据作为CR感知样值，通过样本值训练决策树序列，构建决策树提取知识，并用混淆矩阵法对设计模型的准确性及性能进行评价。实验结果表明，该方法设计模型的分类准确率高，增强了知识的可解释性，能够初步达到认知无线电知识挖掘和对以往经验学习的目的。

关键词：认知无线电；知识挖掘；决策树；粗糙集；C4.5算法

中图分类号：TN914；TP274 **文献标识码：**A

Knowledge discovery for cognitive radio based on rough set and decision tree method

YU Xiao-hang[1a], LI Lei-min[1a], HUANG Yu-qing[1b]

(1a.School of Information Engineering; 1b.School of Manufacturing Science and Engineering, Southwest University of Science, Mianyang Sichuan 621010, China)

Abstract: It is one of the key issues that making knowledge discovery effectively in a Cognitive Radio(CR) engine design. Basing on the research about Rough Set Theory and C4.5 algorithm of decision tree, this study presented a model of CR knowledge discovery designed by combination of rough set and decision methods and studied its feasibility through a case. Using data based on simulation platform of MATLAB 802.11a physical layer as CR perception sample, decision tree sequence was trained, and decision tree was built for knowledge extraction. Then the accuracy and performance of the design model was evaluated by confusion matrix. The simulation results show that the proposed design model gets high classification accuracy rate, can enhance the interpretability of knowledge, and therefore has preliminarily achieved the purpose of knowledge discovery for cognitive radio and learning from the experiences.

Key words: Cognitive Radio; knowledge discovery; decision tree; rough set; C4.5 algorithm

认知无线电(CR)的概念[1]最初由Joseph Mitola 博士提出，它是一种智能的无线电通信系统，其核心思想是使无线通信设备具有发现“频谱空穴”并合理利用的能力，包含如频谱共享[2]、系统性能优化[3]等技术。其中知识是认知无线电引擎智能推理与学习的基础[4]。目前，知识挖掘技术常见的应用案例多发生在零售业、制造业、财务金融保险、通信及医疗服务行业[5]。国内外将CR与知识挖掘技术相结合的研究很少，这是一探索性课题。在现有研究中比较典型的有：文献[6]介绍基于支持向量机(Support Vector Machine, SVM)[7]的CR知识学习引擎设计，并没有给出规则的表示方式；文献[8]主要介绍基于本体的知识表示和基于知识挖掘技术的智能工程决策平台的设计与实现，但没有与CR研究结合起来。

粗糙集和决策树是知识挖掘和学习的重要方法，通常用来分析数据和形成预测模型。本文利用基于MATLAB 802.11a物理层仿真平台收集的数据作为CR感知样值，在对粗糙集和决策树C4.5算法进行研究的基础上，提出用一种基于粗糙集理论和信息熵概念决策树的改进方法[9]来设计CR知识挖掘模型，从而获取知识。认知引擎通过对获取知识的学习推理进行知识库的积累和更新，随着不断的学习，认知引擎存入知识库中的知识又作为推理引擎后续工作的基础。通过案例研究结果显示：该模型既能够保留原始信息特点，又能够保持较高的知识约简效率，准确地对模拟用户的需求信息进行知识性描述。此外，文中采用混淆矩阵方法对分类方法进行了评估，评估结果

收稿日期：2010-01-21；**修回日期：**2010-03-12

基金项目：西安电子科技大学综合业务网理论及关键技术国家重点实验室资助项目(ISN10-09)

显示该模型算法的分类准确度和稳定度能够初步达到对认知无线电知识挖掘和对以往经验学习的要求。

1 基于粗糙集和决策树的数据挖掘方法设计

1.1 粗糙集和决策树方法概述

粗糙集[9-10]是用来研究不完整数据和不精确知识的表达、学习、归纳的一套理论，由波兰理工大学 Pawlak Z 教授提出。

粗糙集理论的知识表达方式通常采用信息系统(Information System)的形式，它可以表现为四元有序组 $IS=(U,A,V,f)$，其中 U 为有限对象的全体集合，即论域；A 为全体属性的非空有限集合，$A=C\cup D$ 且 $C\cap D=\varnothing$，C 为全体条件属性集合，D 为全体决策属性集合；V 是属性 A 的值域；f 为一个信息函数，反映了对象 x 在 IS 中的全部信息。假设 $x\in U,\forall q\in A$，U 中的每个对象 x 都可以由属性集合 A 的值表示，该值就是从对象 x 中提取出的知识信息或称为规则。信息系统中，在保证划分决策表原始的分类能力不变的情况下，应该保持较高的约简效率即以最简单的决策属性对条件属性集合进行分类[11]，最终达到利用 C 相对于 D 的任一约简来代替 C 作为一条知识或规则的目的。

决策树的主要方法有Quinlan J R提出的ID3算法和C4.5算法[12-13]等；其基本思想为：不断地选择最优的属性，并据此来划分数据样本，建立相应的节点，直至把具有最高信息增益的属性作为当前节点的测试属性。C4.5是从ID3演变而来的，它用信息增益率来选择属性，解决了数值缺失、属性值的范围连续、决策树的修剪及规则导出等问题。另外，利用C4.5 算法建立决策树的速度较ID3 算法快，而且决策树结构也较ID3 算法合理，同时也找到了较好的规则信息。

信息增益率(GainRatio)定义为：

$$GainRatio=\frac{Gain(s,a)}{SplitInfo(s,a)} \tag{1}$$

式中：$Gain(s,a)$ 为信息增益；$SplitInfo(s,a)$ 为分裂信息，代表了按照属性 a 把样本 s 分裂成 n 部分而生成的潜在广度信息和均匀性信息。

$$\begin{cases} Gain(s,a)=Info(s_i)-\sum_{i=1}^{k}\frac{|s_i|}{|s|}Info(s_i); \quad SplitInfo(s,a)=-\sum_{i=1}^{n}\frac{|s_i|}{|s|}\lg 2(\frac{|s_i|}{|s|}) \\ Info(s)=-\sum_{j=1}^{m}freq(c_j,s)\lg 2\left[freq(c_j,s)\right] \end{cases} \tag{2}$$

式中：$|s|$为数据集合 s 的样本个数；$|s_i|$为子集 s_i的样本个数(i=1,2,⋯,n)；$freq(c_j,s)$ 为 s 中的样本属于 c_j 类别的频率(j=1,2,⋯,m)，m 是 s 中样本的类别数量。

1.2 CR 知识挖掘模型设计

在无线电知识挖掘中，复杂、庞大、不规则的采集数据量是其遇到的首要问题，它涉及到电磁环境、无线信道特征以及用户需求等方面。对此，需要设计一种无线电知识挖掘方法。该方法的设计应该注意 3 个方面的目的：

1) 其能处理浩瀚的数据，消除冗余信息；2) 其分类速度要快，描述应尽可能简单并易于转换为数据库查询语言；3) 能及时准确地提取认知引擎优化控制无线通信系统所需的决策知识。

通常，数据挖掘的方法有神经网络方法、遗传算法、决策树方法、粗集方法、覆盖正例排斥反例方法、统计分析方法和模糊集方法等。基于上述设计要求以及粗糙集和决策树强大的优势互补性，本文将两种方法有机结合，即采用粗糙集进行数据约简，去除冗余属性，然后利用决策树方法来产生分类规则提取知识。设计框图如图 1 所示。图 1 中，数据样本模块利用基于 MATLAB 802.11a 物理层仿真平台收集的数据作为 CR 感知样值。粗糙集方法预处理模块采用粗糙集属性约简原理[14]，在不减少数据有效信息前提下，对冗余属性进行合并或删除。采用 C4.5 算法对经过预处理后的低维样本数据进行决策建树，与此同时选用 k 次迭代交叉验证，将最小误分代价对应的决策树作为最终构建的决策树。混淆矩阵法评估模块用来对决策树模型的准确性和性能的优劣进行评价。

主要步骤如下：

1) 数据的采集，利用基于 MATLAB 802.11a 物理层仿真平台收集的数据作为 CR 感知样值，该平台采用正交频分数字复用技术，调制方式有 BPSK,QPSK,16-QAM 和 64-QAM，对应星座点数为 2,4,16,64；编码效率有 1/2,2/3 和 3/4。该平台可模拟的数据率为 6 Mbps,9 Mbps,12 Mbps,18 Mbps,24 Mbps,36 Mbps,48 Mbps,54 Mbps，共包括 8

个调制参数：BPSK-1/2,BPSK-3/4,QPSK-1/2,QPSK-3/4,16QAM-1/2,16QAM-3/4,64QAM-2/3,64QAM-3/4。

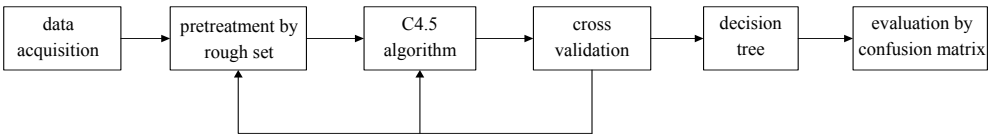

Fig.1 Knowledge discovery for CR framework
图 1 CR 知识挖掘模型框图

在该仿真实验中，数据被定义为 4 个不同的节点，每个节点有 8 种不同状态，其中调制方式与编码效率定义为调制参数节点，该节点包括 8 种状态，对应上述的 8 个调制参数；信噪比节点的每种状态是一个区间，共包括 8 个区间：-inf~10,10~11,11~14,14~18,18~22,22~26,26~28,28~inf(单位为 dB)；数据率节点的每种状态是一个特定的数据率，与该平台可模拟的 8 种不同的数据率对应；误包率节点的每种状态也是一个区间，误包率平均分为 8 个区间。

2) 采集到的数据样本经过预处理后得到低维样本数据。采用粗糙集属性约简原理，通过属性约简和值约简删除其中不相关或不重要的属性，约去过剩的条件属性值，使知识简化。将源于 MATLAB802.11a 仿真平台的 *CR* 感知样值进行预处理后得到的数据包含信噪比、调制方式、误包率、误码率和数据率等设置参数。

3) 选择 C4.5 算法构建决策树，根据每个条件属性信息增益的大小，确定父节点和子节点的生成次序，在由这些节点构成的树序列中，用 *K* 次迭代交叉验证的方法选择最优树作为最终的决策树。

其基本思路为：选取最能分辨不同类别样本的属性作为根节点，同时把数据样本分为对应的若干块；紧接着再从每一块数据样本中选出分辨力最大的属性作为连接根层的节点。如此不断循环，直到每一子节点当且仅当只包含一类样本时为止，这样就构建起决策树序列。交叉验证方法从构建起来的树序列中计算最小的平均误分代价，并将最小误分代价对应的决策树序列作为最终构建的决策树。

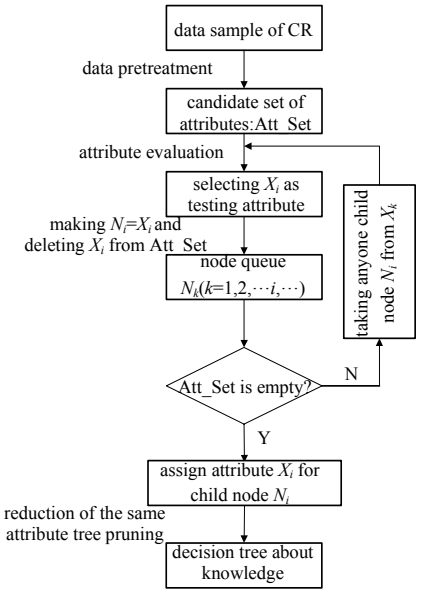

Fig.2 Flowchart of knowledge discovery for CR
图2 CR知识挖掘流程图

4) 采用混淆矩阵法来评价最终构建的决策树模型的准确性和性能。它既能识别出误差的性质，也能识别出误差的数量，用户可以根据错误分类的相对严重程度来评价模型的表现。评估的过程是将模型应用到测试数据集上，并将预测结果与数据集中的观察结果进行比较。

1.3 模型算法实现流程

基于粗糙集和决策树方法(C4.5)的 CR 知识挖掘模型算法流程如图 2 所示：用 Data 表示当前输入 CR 样本数据集，当前候选属性集用 Attribute_Set 的缩写 Att_Set 表示，$X_k \in Att\_Set$，(k=1,2,⋯,i,⋯)；记 X_i 为当前属性，N_i 为当前结点；定义 $\overline{\mathrm{Gain}(X_k)}$ 为所有属性的信息增益平均值，计算公式为：$\overline{\mathrm{Gain}(X_k)}=\frac{1}{i}\sum_{k=1}^{i}\mathrm{Gain}(X_k)$。

主要步骤如下：

1) 对 CR 数据样本集 Data 各项属性数据进行预处理，对属性约简和值约简删除其中不相关或不重要的属性、约去过剩的条件属性值，得到候选属性集合 Att_Set；

2) 初始化树的根节点 N，并确定 Attribute_Set 叶节点属性；

3) 计算候选属性 Attribute_Set 中每个属性 X_k，选取信息增益率最大的 GainRatio(X_i)且同时获取的信息增益 Gain(X_i)属性又不低于所有属性平均值 $\overline{\mathrm{Gain}(X_k)}$ 的属性 X_i 作为测试属性；

4) 将当前属性 X_i 赋值给当前结点 N_i，将该属性的属性值作为该属性的分叉节点，并将分叉节点插入队列中；

5) 从候选属性 Attribute_Set 中将当前使用属性 X_i 删除；

6) 从队列中取出 1 个节点，递归进行 3)~5)步骤，直到候选属性 Attribute_Set 为空；

7) 为每个叶子节点分配类别属性，对相同的类别属性进行合并，进行树剪枝，构建决策树。

2 仿真及结果分析

实验借助一款基于开源项目 Weka 和 Xelopes 的开源数据挖掘软件AlphaMiner，采用 C4.5 算法分类器，利用基于 MATLAB 802.11a 物理层仿真平台收集的数据作为 CR 感知样值。

2.1 决策树构建的实现

剪枝后可以直接从生成的决策树中提取相应的决策规则。决策树具有直观性，易于理解，本文所指规则由条件属性和目标属性组成，采用 IF THEN 形式，如图 3 中所示。每条规则都是从根节点到叶节点的路径。叶节点表示具体的结论，而叶节点以上的节点及其边表示相应条件的条件取值。

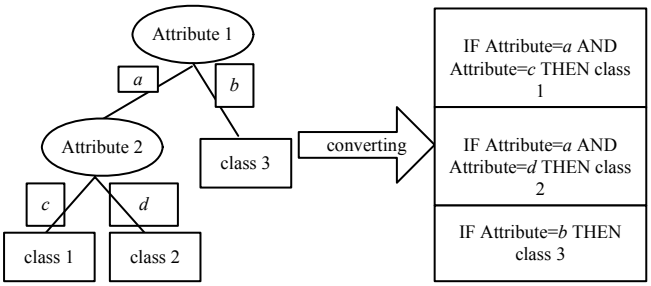

Fig.3 Conversion between decision tree and decision rule

图3 决策树与决策规则之间的转换

由此生成的决策树如图 4 所示。在该树状图中知识容易理解，规则获得路径及逻辑表达直观，节点信息清晰明确，从根到叶的每条路径创建一个规则。

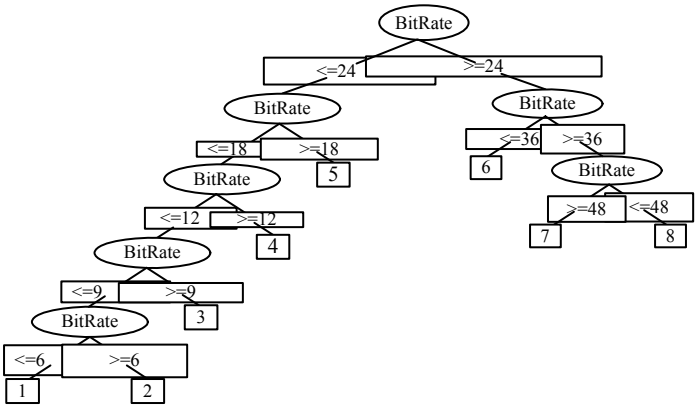

Fig.4 Arborescence of decision rule

图4 决策规则树型图

该分类规则沿着给定路径上的每个属性和属性相关联值形成规则前件(“IF”)的一个合取项，叶节点包含类预测，形成后件(“THEN”)部分，得出最后结论。由规则转换示意图(图 3)可从图 4 中提取出对应的以 IF THEN 形式表示的规则，如表 1 所示。其中表 2 中规则序列号 $N(N=1,2,\cdots,8)$表示 C4.5 算法在测试样本数据集中提取出的规则个数，共有 8 条规则；BitRate 单位为 Mbps；类名 $C_i(i=1,2,\cdots,8)$表示在相应的规则下的无线通信调制方式与编码效率，共包括 8 种状态：BPSK-1/2,BPSK-3/4,QPSK-1/2,QPSK-3/4,16QAM-1/2,16QAM-3/4,64QAM-2/3,64QAM-3/4。划分后的每一类数据采用的传输比特率(BitRate)以及采用的调制方式与编码效率是确定的，三者之间的一一对应关系如表 2 所示。以表 2 的第 2 列信息为例，它表示划分为类“1”的待处理数据采用的调制方式为二相相移键控 BPSK(Binary Phase Shift Keying)，编码效率为 50%，信息传输比特率(BitRate)为 6 Mbps。

表 1 决策树生成决策规则

Table1 Decision rule from decision tree

N	rules	C_i	confidence/(%)	record
1	BitRate <= 6	1	100	532
2	BitRate <= 9 AND BitRate >6	2	100	543
3	BitRate <= 12 AND BitRate >9	3	100	932
4	BitRate <= 18 AND BitRate >12	4	100	2 701
5	BitRate <= 24 AND BitRate >18	5	100	6 115
6	BitRate > 24 AND BitRate <=36	6	100	11 139
7	BitRate > 36 AND BitRate <=48	7	100	5 114
8	BitRate > 48	8	100	5 691

表 2 类名、调制方式与比特率对应关系

Table2 Corresponding relation among class name,modulation and bit rate

class name	modulation and coding efficiency	bit rate/Mbps
C_1	BPSK-1/2	6
C_2	BPSK-3/4	9
C_3	QPSK-1/2	12
C_4	QPSK-3/4	18
C_5	16QAM-1/2	24
C_6	16QAM-3/4	36
C_7	64QAM-2/3	48
C_8	64QAM-3/4	54

2.2 规则知识描述

由图 4 决策树及表 1、表 2 可将所提取规则知识用 IF-THEN 形式描述，例如规则 1,2,8 可描述为：

1) IF 比特率(BitRate)在区间(0,6 Mbps]，THEN 系统采用编码效率为 50%的 BPSK 调制方式；

2) IF 比特率(BitRate)在区间(6,9 Mbps]，THEN 系统采用编码效率为 75%的 BPSK 调制方式；

……

8) IF 比特率(BitRate)在区间(48,+∞ Mbps]，THEN 系统采用编码效率为 50%的 QAM 调制方式。

采用粗糙集和决策树方法对 CR 感知样本进行分析处理获得的决策规则,就是数据挖掘所要获取的 CR 知识，通过这些规则可以对 CR 用户通信需求进行分类。为了适应特定的频谱环境，CR 引擎通过规则知识适时调整诸多通信参数(如：频率、功率、调制方式、星座大小、编码方式、编码速率等)，满足用户需求。该决策树概括了在不同误码率、信噪比和比特率条件下，系统选择匹配的调制方式和编码效率，以期满足误比特率 BER(Bit Error Ratio)、信号带宽、频谱效率和数据速率等多个指标的要求。

2.3 模型分析评价

采用混淆矩阵法对决策树的准确度和性能进行评估。混淆矩阵表示的是真实值和预测值的交叉表，它能够识别错误的种类和个数。通过识别出误差的性质从而识别出误差的数量，然后依据相对错误分类的程度来评估 CR 知识挖掘模型的性能。表 3 为基于 C4.5 算法的分类误差矩阵表，表 4 为基于粗糙集和 C4.5 算法的分类误差矩阵表(pred 表示预测分类；real 表示实际分类)。

表 3 基于 C4.5 算法的分类误差矩阵表

Table3 Matrix table of classification error based on C4.5

real	pred								
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	∑
C_1	38	1	0	0	0	0	0	0	38
C_2	4	26	2	2	1	5	0	1	41
C_3	3	5	63	4	0	2	2	1	80
C_4	1	4	8	176	7	8	2	0	206
C_5	2	0	7	8	451	10	5	3	486
C_6	0	5	7	2	20	1 032	19	2	1 087
C_7	0	0	2	1	0	5	333	3	344
C_8	0	0	0	3	4	4	8	424	443
∑	48	41	89	196	483	1 066	369	434	2 726
note:accuracy rate=87.4%;*Kappa*=0.86									

表 4 基于粗糙集和 C4.5 算法的分类误差矩阵表

Table4 Matrix table of classification error based on rough set and C4.5

real	pred								
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	∑
C_1	38	0	0	0	0	0	0	0	38
C_2	3	31	2	1	0	0	0	0	37
C_3	2	3	65	2	1	1	1	0	75
C_4	0	5	7	195	6	6	0	1	220
C_5	2	0	10	1	451	9	6	0	479
C_6	0	5	5	0	17	1 044	20	0	1 091
C_7	0	0	2	0	1	5	333	2	343
C_8	0	0	0	3	2	5	7	426	443
∑	45	44	91	202	478	1 070	367	429	2 726
note:accuracy rate=94.7%;*Kappa*=0.93									

比较表3和表4可知：经过粗糙集预处理后，测试样本的知识挖掘分类正确率提高7.3%。结果表明，经粗糙集方法处理和决策树方法挖掘过后获得的知识分类准确度达到90%以上，能有效实现对主用户(频谱初始授权用户)和认知用户(未授权用户)信息特征的提取。在实验室条件下，能够初步达到认知无线电知识挖掘和对以往经验学习的目的。

3　结论

知识是认知无线电中认知引擎推理学习的基础，本文探索性提出用粗糙集和决策树相结合的方法设计CR知识挖掘模型，该模型既能够保留原始信息特点，又能够保持较高的知识约简效率，准确地对模拟用户的需求信息进行知识性描述。可以获取CR的学习和推理所需的长期经验知识以及暂时状态和环境知识。利用802.11a仿真平台模拟无线通信系统，收集样值作为CR感知信息来进行相关知识挖掘，并通过混淆矩阵方法对设计模型进行评价，满足预期设计的目的要求。下一步研究工作的重点是如何将提取出的CR知识准确、灵活地表示以及如何实现CR知识的共享与重用。

参考文献：

[1] Mitola J. Cognitive radio:making software radios more personal[J]. IEEE Pers. Commun., 1996,6(4):139-153.

[2] WANG Jiao,HUANG Yu-qing,JIANG Hong. Improved Algorithm of Spectrum Allocation Based on Graph Coloring Model in Cognitive Radio[C]// Proc. 2009 International Conference on Communication and Mobile Computing(CMC). Los Alamitos:IEEE Computer Society, 2009:53-357.

[3] ZHANG Xiao-qin,HUANG Yu-qing,JIANG Hong,et al. Design of Cognitive Radio Node Engine Based on Genetic Algorithm[C]// 2009 WASE International Conference on Information Engineering. Taiyuan:[s.n.], 2009:22-25.

[4] Joseph Mitola III. Cognitive Radio An Integrated Agent Architecture for Software Defined Radio[D]. Ph.D dissertation of Sweden:Royal Institute of Technology(KTH), 2000.

[5] 于　勇. 数据挖掘研究及其应用[J]. 信息与电子工程, 2003,1(1):30. (YU Yong. Data Mining Research and Application[J]. Information and Electronic Engineering, 2003,1(1):30.)

（下转第624页）