第 32 卷 第 20 期 计 算 机 工 程 2006 年 10 月
Vol.32 № 20 Computer Engineering October 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)20—0191—02 文献标识码：A 中图分类号：TP183

基于神经网络和粗糙集规则的提取方法

庄传礼[1,2]，杨 萍[1,2]，李道亮[2]，傅泽田[1,2]
(1. 中国农业大学经济管理学院，北京 100083；2. 中国农业大学教育部精细农业系统集成研究重点实验室，北京 100083)

摘 要：在利用粗糙集对连续性数据进行分类规则挖掘时，需要对数据进行离散化处理，但是离散结果往往会破坏原有数据的隐含信息，提取的分类规则质量难以保证。该文设计了一种基于自组织人工神经网络与粗糙集理论的分类规则提取方法，利用神经网络自动分类的功能，对离散前后的数据进行分类，比较两次分类结果是否一致，当达到一致性结果后，再利用粗糙集理论对数据约简，进行规则提取，有效地解决了原始数据信息丢失的问题，通过实例证明了该方法的合理性。

关键词：规则挖掘；粗糙集；自组织人工神经网络；离散化

Extracting Rules Based on Artificial Neural Networks and Rough Sets Theory

ZHUANG Chuanli[1,2], YANG Ping[1,2], LI Daoliang[2], FU Zetian[1,2]
(1. College of Economics & Management, China Agricultural University, Beijing 100083; 2. Key Laboratory of Modern Precision Agriculture System Integration, China Agricultural University, Ministry of Education, Beijing 100083)

【Abstract】 The continuous value is discretized before using the rough set method to mine the classification rules. But more information concealed in the original data is lost after the discretization, the quality of the extracted classification rules is very poor. A new method based on self-organizing artificial neural networks and rough set theory is designed to extract classification rules of continuous value. Because self-organizing artificial neural networks can train themselves and make an auto-classification on the input mode, it is used twice to classify the data before and after discretization. It extracts the rules by rough sets reduction until the results of two classifications are consistent. Through the analysis of case studies, the rationality of the extraction rule is testified.

【Key words】 Rules extraction; Rough sets; Self-organizing artificial neural networks; Discretization

1 概述

粗糙集方法(RST)是一个重要的分类规则挖掘工具，在不完整性、不确定性问题方面有着明显的优势。该方法是由波兰科学家Pawlak提出的，它以不可分辨的关系划分知识，形成新的知识表达系统，通过知识约简，利用上、下近似集描述对象获得新方法[1]。但是粗糙集方法是一种结构化的、数值化的信息处理方法，适合处理离散数据，对于连续数据的处理能力有限[2,3]，而研究对象的一些属性经常包含一些连续性的数值，这就需要对这些数值进行离散化处理，可是对数据进行离散，势必会破坏数据原有的隐含信息，在这种数据上产生的分类规则，显然不能保证分类规则的可靠性。

人工神经网络(ANN)是 20 世纪 80 年代中期兴起的一门非线性科学，它模拟人脑的一些基本特征，利用非线性映射和并行处理方法，通过网络权值的不断学习、调整、完成输入与输出空间的映射关系。在模式识别、机器学习、决策支持、知识发现和数据挖掘等领域得到广泛的应用。尽管对规则挖掘方面不如粗糙集方法直观清晰，但是它具有较强的数值逼近能力，能够处理定量的、数值化的信息，得出较为精细的结果[2]。

把人工神经网络方法与粗糙集方法结合在一起挖掘分类规则能够起到优势互补的作用。文献[4]就是采用了这个思路，该文先用粗糙集方法约简属性，再利用人工神经网络挖掘规则，取得了较为明显的改进，但是由于人工神经网络是一个“黑箱”，挖掘的规则不易表达，同时该文并未解决连续性数据离散所存在的问题。

2 规则提取方法

本文设计的规则提取方法共分为如下 4 个步骤：

(1)利用自组织人工神经网络对数据进行分类。把这一步分类的结果记录下来，作为离散后分类的对比分类。一般情况下，自组织人工神经网络可以通过自身学习，对输入属性进行自动分类，分类的结果较为合理，也可以作为以后挖掘规则使用的决策表中的决策属性。现在常用自组织人工神经网络的方法有竞争学习网络、自组织特征映射网络、学习矢量量化网络等 3 类，这些网络的算法可以参考人工神经网络的书籍或文献[5]。也已经有许多自组织人工神经网络的软件(如 Matlab ANN toolbox、神经网络 Office 插件等)，借助这些程序设计和实现自组织人工神经网络，可以提高工作效率。

(2)对数据的条件属性进行离散。这一步的目的是为了以后应用粗糙集方法挖掘分类规则。不同的离散方法对提取的规则会有着重要的影响，适当的离散方法不仅提高规则挖掘

基金项目：科技部国际重点合作基金资助项目(2003DF000004)；AsiaIT&C基金资助项目(117839/C/G-41-15)
作者简介：庄传礼(1972－)，男，博士生、讲师，主研方向：人工智能，数据结构；杨 萍，博士生；李道亮，教授；傅泽田，教授、博导
收稿日期：2006-01-16 **E-mail：**zhuangchuanli@163.com

的效率，还能够有效地反映数据的原始信息，使挖掘的规则准确表达条件属性与决策属性之间的关系。

常用的离散化方法有等距离划分法、等频率划分法、Naive Scalar 法和 Semi-naive Scalar 法等。这些方法并没有一个统一的优劣判断方法，必须通过不断的尝试和积累才能获得满意的结果。

(3)利用自组织人工神经网络对离散后的数据再进行分类。分类的类别应该同前一次相同，可以适当地调整相应的学习率。如果分类结果同前一次一致或者达到预设的准确度，则可以运行到下一步，否则，应该考虑离散化的过程是否存在问题，调整离散方法或者相关参数，再进行离散，直到再分类的结果与第 1 次分类结果一致或者达到预设的准确度，进入下一步。

(4)把满足一致性要求的离散后数据整理成决策表，可以把条件属性和分类结果合并成决策表，也可以指定某个属性作为决策属性。对决策表用粗糙集理论进行约简，然后根据预先设置的覆盖度和置信度水平提取规则。一般而言，由于求最小约简是 NP 问题，因此目前大都采用启发式算法，主要有最大覆盖删除算法、格点搜索算法、SAV 遗传算法等。方法流程见图 1。

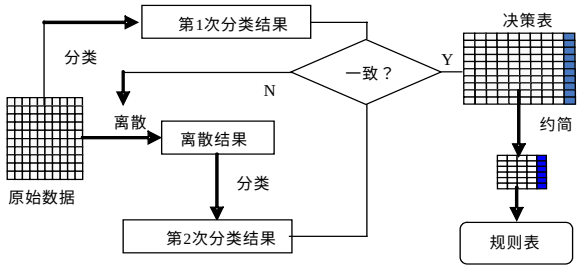

图 1 方法流程

3 应用举例

表 1 给出了我国 12 个煤矿废弃地理气候特征，对其分类有助于借鉴煤矿废弃地植被恢复的经验，但是表 1 给出的属性较多，类型不一，目前尚无明确的分类方案。本文按照所提供的方法寻找分类规则。

表 1 原始数据

地点	a	b	c	d	e	f	g	h
山西朔州	6.9	2 779	440	2 476	30	1 400	1 570	3.6
山西阳泉	10.9	2 767	576	3 325	72	1 500	1 362.5	2.8
辽宁铁法	6.8	2 350	650	3 400	18	250	1 667	3.6
山西垣曲	13.3	2 307	631	4 391	56	480	2 072	3.5
内蒙通辽	0.1	2 850	383	2 525	68	1 200	1 784	3.7
河北迁安	10.1	2 782	735	3 200	44	400	1 100	2.5
黑龙江桦南	3.1	2 427	475	2 750	39	260	900	3.6
辽宁抚顺	6.5	2 386	800	2 800	12	230	1 600	3.5
内蒙伊克昭	7.3	3 100	385	3 000	35	1 100	2 000	3.7
河南焦作	14.9	2 300	604	4 700	76	500	8 00	2.9
内蒙赤峰	5.0	2 950	368	2 400	40	1 050	2 100	3.7
山西陵川	7.3	2 518	663	3 600	62	1 058	1 614	4.0

表 1 中，用 a、b、c、d、e、f、g 分别代表平均气温(℃)、平均日照(h/a)、平均降雨(mm)，平均积温(℃)、沙粒含量(%)、海拔(m)、蒸发量(mm)、风速(m/s)，对表 1 的分析如下：

(1)对表 1 中各属性进行归一化。归一化处理的目的是为了在应用 Matlab 工具箱时，方便地定义取值区域。本文离散后的分类，由于取值范围一致，因此没有进行归一化处理。

$$x_{ij}=\frac{x_{ij}^{'}-x_j^{\min}}{x_j^{\max}-x_j^{\min}}\quad i=1,...,8,\ j=1,...,12$$

其中，$x_{ij}^{'}$ 是表 1 中未处理的数据，$x_j^{\max}$、$x_j^{\min}$ 是第 j 行的最大值和最小值。

经处理后，应用 Matlab6.5 人工神经网络工具箱中的自组织竞争学习网络模型，建立网络模型。

输入层把属性个数设置 8 个，竞争层根据聚类数目设定，本文设置 3 个神经元。网络竞争层权值学习速率设定为 0.01，阈值学习速率为 0.001，权值学习函数为 Kohonen 函数，阈值学习函数为公平阈值学习函数，分类结果如表 2 所示。

表 2 分类的结果

类别	废弃地
1	山西朔州、内蒙通辽、内蒙伊克昭、内蒙赤峰、
2	山西阳泉、山西垣曲、河南焦作、山西陵川
3	辽宁铁法、河北迁安、黑龙江桦南、辽宁抚顺

从地理位置来看，3 个类别如下：

1)地点基本上在内蒙古高原的边界，沿大兴安岭末端延长线分布；

2)主要位于吕梁山以东，太行山以西；

3)基本上处于东北平原；各类别具有相同的地理分布特征，分类结果有一定的合理性。

(2)对表 1 中属性数据离散化，本文采用的等距离划分法进行离散，间距为每个属性值的极差的 1/3，区间取值左闭右开。

(3)再次应用自组织竞争学习网络模型进行分类，参数设置同上，分类结果验证与前面一致。

(4)把离散表加入分类结果，生成决策表 3。

表 3 决策表

地点	a	b	c	d	e	f	g	h	分类
山西朔州	2	2	1	1	1	3	2	3	1
山西阳泉	3	2	2	2	3	3	2	1	2
辽宁铁法	2	1	2	2	1	1	3	3	3
山西垣曲	3	1	2	3	3	1	3	3	2
内蒙通辽	1	3	1	1	3	3	3	3	1
河北迁安	3	2	3	2	2	1	1	1	3
黑龙江桦南	1	1	1	1	2	1	1	3	3
辽宁抚顺	2	1	3	1	1	1	2	3	3
内蒙伊克昭	2	3	1	1	2	3	3	3	1
河南焦作	3	1	2	3	3	1	1	1	2
内蒙赤峰	1	3	1	1	2	3	3	3	1
山西陵川	2	1	3	2	3	3	2	3	2

本文利用波兰波兹南科技大学智能决策支持系统实验室开发的粗糙集软件 ROSE2.2 中生成的 satisfactory description 法，提取该决策表的分类规则。当最小覆盖度为 50%，最小置信度为 100%时，系统产生 33 条规则。

规则验证：本文从网上收集到另外 3 处煤矿废弃地的信息，由于网上信息发布的目的不同，有些信息是不完整的。但是可以看出利用生成的规则能够进行分类，并且分类的结果显示这 3 个煤矿从地理位置上和以上的分类基本吻合。

规则与验证结果如表 4、表 5 所示。表 5 中括号内数字表示离散结果。 (下转第 209 页)