

基于粗糙集理论的多属性评估方法的研究

张 磊,侯太平,甄维坤,李 涛,向 高

(解放军理工大学气象学院,南京 211101)

摘 要:传统的模糊综合评价利用模糊数学理论来描述评价过程中涉及的一些模糊因素,取得了一定效果。但这种方法存在如隶属度参数的调整的主观性,网络结构不易确定,模糊规则的选取等问题不易解决。文章提出结合粗糙集理论(RS)和模糊神经网络(FNN)的一种网络评估模型,使得网络的结构容易确定,并利用 Rosetta 软件轻松提取模糊规则,是一种具有通用性的网络评估模型。

关键词:粗糙集;综合评估;模糊神经网络;Rosetta

中图分类号:TP18 **文献标识码:**A

Study of Multi-attribute Evaluation Method Based on Rough Set Theory

ZHANG Lei, HOU Tai-ping, ZHEN Wei-kun, LI Tao, XIANG Gao

(Institute of meteorology, PLA University of Science and Technology, Nanjing 211101, China)

Abstract: The original fuzzy comprehensive evaluation use fuzzy math theory to describe the process of evaluation and some fuzzy factors, this evaluation model is feasible. but some problem consist in the model, for example, how to adjust the parameters of membership functions. How to design the structure of the model, how to select the fuzzy rules. In this paper, it designs a new model that combines rough set (RS) and fuzzy neural network (FNN). It is easy to design the structure of the model and pick up the fuzzy rules with Rosetta. It is a evaluation model that can used in many cases.

Key words: rough set; comprehensive evaluation; fuzzy neural network; Rosetta

在传统的模糊综合评价中,存在评价变量较多,使网络结构复杂化,并且会出现模糊规则组合爆炸的情况,严重影响了网络的学习速率。由于模糊神经网络的结构确定缺乏理论指导,所以网络结构主要靠人为的给定,基于人为给定的网络结构可能效果并不理想。

本文结合粗糙集理论和模糊神经网络方法确定一种评估方法。由于 RS 理论具有强大的数值分析能力,而 FNN 网络具有准确的逼近收敛能力和较高的精度,二者结合可以得到一种可理解性好、收敛速度快、评估结果可靠的神经网络。我们可以用 RS 理论对模糊神经网络的结构进行设计,运用粗糙集理论对属性进行约简,对规则进行筛选,简化网络结构。

1 粗糙集理论

粗糙集(RS)理论^[1]是 Z.pawlak 教授在 1982 年提出的,它是一种刻画含糊、不完整性、不确定性的数学工具,它能够利用数据本身提供的信息,不需要任何的

先验知识。它的基本思想是在保持分类能力不变的前提下,通过知识约简,导出概念的分类规则,删除不相关或不重要的属性,只利用数据集上的等价关系对知识的不确定程度进行度量,从而避免了对知识的主观评价。

1.1 粗糙集相关概念

决策表:决策表可以表示为 $S=(U,A,V,f)$, U 为对象的非空有限集合,成为论域; $A=C\cup D$ 是属性集合,子集 C 和 D 分别称为条件属性和结果属性集; $V=\bigcup_{a\in A}V_a$ 是属性集的集合, V_a 表示属性 $a\in A$ 的属性值范围。 $f:U\times A\rightarrow V$ 是一个信息函数。

1.2 区分表

设有决策表 $S=(U,A,V,f)$, $A=C\cup D$ 是属性集合, $C=\{a_i|i=1,\dots,m\}$, $a_i(x_j)$ 是样本在属性 a_i 上的取值, $b(k,(i,j))$ 表示区分表中对应样本对 (x_i,x_j) 和属性 a_k 的元素,则区分表 $b(k,(i,j))$ 的定义是:

收稿日期:2010-01-10

作者简介:张 磊(1982—),男,辽宁盘锦人,硕士研究生,研究方向:智能决策;侯太平(1971—),男,副教授,硕士生导师,研究方向:数据挖掘,人工智能。

$$b(k, (i, j)) = \begin{cases} 1 & a_k \in C \text{ 且 } a_k(x_i) \neq a_k(x_j) \text{ 且 } d(x_i) \neq d(x_j) \\ 0 & d(x_i) = d(x_j) \end{cases}$$

其中 $i, j=1, \dots, n; k=1, \dots, m$.

区分表实质上是可辨识矩阵的另一种描述,但这种二进描述法特别适用于计算机的实现。

1.3 基于粗糙集理论的规则获取^[2]

主要是通过通过对原始决策系统进行简约处理,即在保持决策系统结论属性和条件属性之间的依从关系不变的前提下,对决策系统进行简约,包括属性简约和值简约。

不失一般性,用粗糙集来形成规则集的过程可描述如下:

(1)数据预处理 由于粗糙集理论只能处理离散属性,所以必须在属性约简之前对连续属性离散化。对原始数据集进行连续属性离散化及消噪处理,将不完备信息系统转化为完备信息系统。

(2)构建决策表 根据样本数据建立决策表。在评估系统中,决策表以二维表格的形式表示,每一行描述一个属性,每一列为论域 U 的一个样本。

(3)属性约简 是从决策表中消去一些冗余的条件属性而不会改变其决策能力,通常采用可辨识矩阵进行属性约简计算。属性约简的目标就是要从条件属性集合中发现部分必要的条件属性,用尽可能少的属性产生决策规则。

(4)属性值约简 是在属性约简的基础上对决策表的进一步约简,属性约简有利于从决策表中分析发现对决策分类起作用的属性,但是属性约简只是在一定程度上去掉了决策表中的冗余属性,还没有充分去掉决策表中的冗余信息,值约简可以进一步去除起作用属性中的冗余信息。

(5)规则提取 经过属性约简与属性值约简,可以得到精简的决策信息系统,根据它可得到具有一定决策概率的不精确决策规则,但是其中大量的决策规则是冗余的。利用分类可信度和覆盖率定义生成决策规则。

$$\alpha_R(D) = \frac{|[x_k]_R \cap D_k|}{|[x_k]_R|}$$

$$\beta_R(D) = \frac{|[x_k]_R \cap D_k|}{|D_k|}$$

式中: $\alpha_R(D)$ 为可信度; $\beta_R(D)$ 为覆盖率; x_k 代表第 k 条规则的决策属性类; R 代表针对规则的条件属性所做的分类。

粗糙模糊模型是靠大量的样本数据来驱动的,即从大量的样本数据中挖掘出模糊规则。

规则知识获取可以看作是一个知识系统,提取有用属性,简化属性表达和推理规则的一个过程。

2 建立基于粗糙集理论的模糊神经网络模型

粗糙模糊神经网络的结构^[3]如图 1 所示,网络可分为五层:输入层、隶属函数生成层、规则推理层、决策层、去模糊化层。

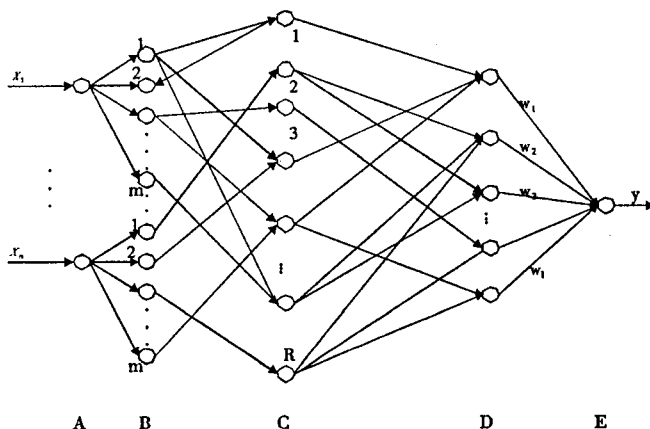


图 1 基于粗糙集的模糊神经网络结构图

设网络的第 L 层的第 j 个节点的输入为 I_j^L , 输出为 O_j^L 。

A 层为网络的输入层, n 为输入变量的个数。输入的变量个数为要评估的对象的条件属性个数。在这里去经过粗糙集约简过的变量个数。

B 层为模糊化层, B 层对 A 层的输入起模糊化作用。B 层中的 m_1, \dots, m_n 分别为各输入变量的语言变量数。B 层为隶属函数生成层, 可以计算输入变量对各个语言变量的隶属度。节点的作用函数通常取正态型函数。

$$u_i(x_j) = \exp\left[-\frac{(\omega_i x_j - \omega_c)^2}{1/\omega_d}\right]$$

I 为模糊化等级; ω_i 为输入变量的量化因子; ω_c 为隶属函数的中心元素; ω_d 为隶属函数的尺度因子, 初始值为随机赋值。

C 层为规则推理层。C 层的每一个节点代表一条规则。这些规则的获取和描述的理论依据是粗糙集理论。通过 Rosetta^[4] 软件生成并筛选规则, 得到模糊规则数。网络的节点数等于网络的规则数。执行模糊“与”运算。假设有 R 条规则, 则 C 层的节点数为 R 。

$I_k^c = O_{i_1}^b \cdot O_{i_2}^b \cdots O_{i_n}^b; i_n = m(n-1)+1, \dots, m \times n, O_k^c = I_k^c, K=1, 2, \dots, R$

D 层为去模糊化层, 它的节点数为评估系统的决策属性个数来决定的。执行模糊“或”运算。

假设 D 层的节点数为 L

$$I_k^0 = O_k^0 / \sum_{i=1}^R O_i^0, O_k^0 = I_k^0; k=1, 2, \dots, L$$

E 层为输出层,起去模糊化作用,输出清晰值。

$$I_k^1 = \sum_{i=1}^L (O_i^0 \cdot \omega_{ik}), O_i^1 = I_i^1; k=1, 2, \dots, L$$

网络的最终输出 $y = I_k^1$;

连接权值 ω_{ik} 的初始值设为规则的可信度值。

对于建立的模糊神经网络来说,定义误差函数为:

$$E = \frac{1}{2} \sum_{t=1}^m (y_0(t) - y(t))^2$$

其中 m 是学习样本数, $y_0(t)$ 是期望输出, $y(t)$ 是实际输出。反向传播思想被用来监督学习,通过调整网络各参数,使误差函数 E 达到最小。

网络各参数的调整公式可表示为:

$$W_{ij}(n+1) = W_{ij}(n) - \eta \partial E / \partial W_{ij} + \alpha (W_{ij}(n) - W_{ij}(n-1))$$

$$c_{ij}(n+1) = c_{ij}(n) - \eta \partial E / \partial c_{ij} + \alpha (c_{ij}(n) - c_{ij}(n-1))$$

$$\sigma_{ij}(n+1) = \sigma_{ij}(n) - \eta \partial E / \partial \sigma_{ij} + \alpha (\sigma_{ij}(n) - \sigma_{ij}(n-1))$$

其中 η 为学习率, α 为惯性系数 ($0 < \alpha < 1$)。

3 实验仿真

本文采用 UCI 数据集中的 IRIS 数据进行试验。IRIS 共有 150 组样本。有 4 个条件属性 (X_1, X_2, X_3, X_4) 和 1 个决策属性 (d)。使用 Rosetta 软件对样本进行分类,分为训练子表和测试子表,样本数分别为 100, 50。对样本数据进行离散化和约简并生成规则,可得到规则 30 条。通过可信度大于 0.75 和支持度大于 4 的约定,得到 10 条规则。

(1) $X_1 \in ([*, 5.5])$ AND $X_2 \in ([2.9, 3.2])$ AND $X_3 \in ([*, 3.2])$ AND $X_4 \in ([*, 0.8]) \Rightarrow d(1)$

(2) $X_1 \in ([*, 5.5])$ AND $X_2 \in ([3.2, *])$ AND $X_3 \in ([*, 3.2])$ AND $X_4 \in ([*, 0.8]) \Rightarrow d(1)$

(3) $X_1 \in ([5.5, 6.3])$ AND $X_2 \in ([3.2, *])$ AND $X_3 \in ([*, 3.2])$ AND $X_4 \in ([*, 0.8]) \Rightarrow d(1)$

(4) $X_1 \in ([5.5, 6.3])$ AND $X_2 \in ([*, 2.9])$ AND $X_3 \in ([3.2, 4.9])$ AND $X_4 \in ([0.8, 1.7]) \Rightarrow d(2)$

(5) $X_1 \in ([6.3, *])$ AND $X_2 \in ([2.9, 3.2])$ AND $X_3 \in ([3.2, 4.9])$ AND $X_4 \in ([0.8, 1.7]) \Rightarrow d(2)$

(6) $X_1 \in ([5.5, 6.3])$ AND $X_2 \in ([2.9, 3.2])$ AND $X_3 \in ([3.2, 4.9])$ AND $X_4 \in ([0.8, 1.7]) \Rightarrow d(2)$

(7) $X_1 \in ([6.3, *])$ AND $X_2 \in ([2.9, 3.2])$ AND $X_3 \in ([4.9, *])$ AND $X_4 \in ([1.7, *]) \Rightarrow d(3)$

(8) $X_1 \in ([6.3, *])$ AND $X_2 \in ([3.2, *])$ AND $X_3 \in ([4.9, *])$ AND $X_4 \in ([1.7, *]) \Rightarrow d(3)$

(9) $X_1 \in ([5.5, 6.3])$ AND $X_2 \in ([*, 2.9])$ AND $X_3 \in ([4.9, *])$ AND $X_4 \in ([1.7, *]) \Rightarrow d(3)$

(10) $X_1 \in ([6.3, *])$ AND $X_2 \in ([*, 2.9])$ AND $X_3 \in ([4.9, *])$ AND $X_4 \in ([1.7, *]) \Rightarrow d(3)$

模糊神经网络的结构为 4-12-10-3-1。设定迭代次数为 1 000 次,得到总误差为 0.00732。对测试子表进行测试,50 样本只有 2 个未能识别,识别率为 96%。由此可见,使用此方法进行多属性决策是可行的。

模糊神经网络的结构为 4-12-10-3-1。设定迭代次数为 1 000 次,得到总误差为 0.00732。对测试子表进行测试,50 样本只有 2 个未能识别,识别率为 96%。由此可见,使用此方法进行多属性决策是可行的。

4 结束语

本文结合粗糙集理论和模糊神经网络的理论建立了一种网络模型,这种模型具有粗糙集和模糊理论的优点。网络模型的结构容易确定,网络的各参数有明确的意义。利用粗糙集实验软件 Rosetta 可以轻松的对评价属性进行约简,并生成规则,达到简化网络结构的效果。此模型在模糊评价中具有一般性,能够适用于多数的多属性模糊综合评价。

参考文献:

- [1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science, 1982(11):341-356.
- [2] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2001.
- [3] 易继锴, 侯媛彬. 智能控制技术 [M]. 北京: 北京工业大学出版社, 2007.
- [4] 程玉胜. Rosetta 实验系统在机器学习中的应用[J]. 安庆师范学院学报(自然科学版), 2005, 11(2): 89-92.