

# Tactic of Case Retrieval Based on Rough Set Theory and Artificial Neural Network

Wu Deng<sup>1,2,3</sup>, Fajun Guo<sup>1</sup>, Chao Feng<sup>1</sup>

<sup>1</sup>Software Institute, Dalian Jiaotong University, Dalian, China, 116028

<sup>2</sup>Key Laboratory of Intelligent Manufacture of Hunan Province, Xiangtan University Xiangtan China, 411105

<sup>3</sup>Key Laboratory of Advanced Design and Intelligent Computing (Dalian University), Ministry of Education Dalian China, 116622

Email: dw7689@163.com

**Abstract:** In order to improve efficiency and quality of case retrieval in case-based reasoning system, a case retrieval model based on the integration of rough set theory and artificial neural network is presented. The paper analyzes and deals with case database using rough set theory in order to obtain the minimum case feature subset which is regarded as the inputs of RBF neural network for training the network. The pretreated requirement problem is inputted the optimized network model to find out the nearest case that is the design plan for design reference.

**Keywords:** rough set theory; artificial neural network; case retrieval; case-based reasoning; feature vector

## 基于粗糙集理论和人工神经网络的实例检索策略

邓 武<sup>1,2,3</sup>, 郭发军<sup>1</sup>, 丰超<sup>1</sup>

<sup>1</sup>大连交通大学 软件学院, 大连, 中国, 116028

<sup>2</sup>智能制造湖南省普通高等学校重点实验室, 湘潭大学 湘潭, 中国, 411105

<sup>3</sup>先进设计与智能计算省部共建教育部重点实验室, 大连大学, 大连, 中国, 116622

Email: dw7689@163.com

**摘 要:** 为了提高实例推理过程中实例检索的效率和质量, 将粗糙集理论与神经网络相结合, 提出了基于粗糙集理论和人工神经网络的实例检索模型。该方法在实例层次组织的基础上, 利用粗糙集理论对实例库进行分析和处理, 求得能够覆盖原始实例特征的具有最大完备度的最小属性集, 作为人工神经网络输入, 对网络进行训练, 得到优化网络模型, 将预处理的待求问题向量输入优化网络模型, 提取最为接近的相似实例, 作为设计参考。

**关键词:** 粗糙集理论; 人工神经网络; 实例检索; 实例推理; 特征向量

## 1 引言

在传统人工设计中, 方案设计主要采用类比设计或者经验设计, 工业中 60% 的产品设计表现为在原产品基础上的改型或者改进。基于实例的推理是利用以往求解类似问题的以实例形式表示的经验知识进行推理, 从而获得当前问题求解结果的一种推理模式。一般来说, 实例推理的推理过程包括问题特征的抽取、相应实例检索、实例的修改与调整、方案的评价和新实例的存储等基本步骤。

资助信息: 国家自然科学基金资助项目(60870009); 大连大学先进设计与智能计算省部共建教育部重点实验室开放课题资助项目; 智能制造湖南省高校重点实验室开放课题资助项目。

实例检索作为实例推理的关键技术, 对求解效率和推理准确性都有较大的影响, 而设计实例的特征属性众多, 重要程度不同, 权重一般在实际中很难判断, 通常是在经验知识的基础上通过假设或平均权重的办法来处理, 因而无法证明其正确性, 难以保证实例检索确实有效。为了解决以上问题, 本文提出将粗糙集理论和人工神经网络相结合, 实现优势互补, 用于实例检索。该方法是粗糙集理论来对实例库进行分析和处理, 通过知识约简, 导出问题的决策或分类规则集。利用人工神经网络的连接权来确定各特征属性的权重, 从而减少权重确定过程中主观因素的影响。采用人工神经网络进行相似实例匹配, 不需要组织大量产生式规则, 也不需要树搜索, 通过学习机制可以自动调整其内

部连接强度,具有自组织和自学习能力和很强的自适应能力,同时在神经网络的检索过程中不需要定义各特征属性的重要度,从而避免了在特征属性权值确定过程中的主观因素影响,保证实例检索的客观性,并且具有记忆功能的神经网络可减少在实例检索过程中对数据库的频繁操作,提高检索效率。

## 2 粗糙集理论

粗糙集理论 (Rough Sets Theory, 简称 RST) 是波兰数学家 Z.Pawlak 于 1982 年提出的一种分析数据的数学理论,该理论能有效地处理不精确、不确定与不完全的各种数据信息。其主要思想是在保持系统分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则集。目前粗糙集理论在机器学习、知识获取、决策分析、专家系统、模式识别等领域得到了广泛的应用。

**定义1.** 设一个信息系统  $S = \{U, A, V, F\}$ , 其中  $U$  是对象集合 (论域),  $C \cup D = A$  是属性集合, 子集  $C$  和  $D$  分别称为条件属性和决策属性。  $V = \bigcup_{r \in R} V_r$  是属性的集合,  $V_r$  表示属性  $r \in A$  的属性值范围,  $F: U \times A \rightarrow V$  是一个信息函数, 它指定  $U$  中每一对象  $x$  的属性值。具有条件属性和决策属性的信息系统称为决策表。

**定义 2.** 设一个信息系统  $S = \{U, A, V, F\}$ , 对于每个属性子集  $B \subseteq A$ , 定义一个不可分辨关系  $\text{Ind}(B)$  为:  $\text{Ind}(B) = \{(x, y) \in U^2: f_a(x) = f_a(y), \text{任意 } a \in B\}$ 。有序对  $B_S = (U, \text{IND}(B))$  被称为近似空间。

**定义 3.** 设一个信息系统  $S = \{U, A, V, F\}$ ,  $B \subseteq A$ ,  $X \subseteq U$ , 定义集合  $R_-(X) = \{x \in U: [x]_R \subseteq X\}$  为  $X$  的  $B$  下近似集,  $R_+(X) = \{x \in U: [x]_R \cap X \neq \emptyset\}$  为  $X$  的  $B$  上近似集, 其中  $[x]_R$  表示  $x$  在等价关系  $R$  下与其等价的元素构成的集合;  $Bn_R(X) = R_+(X) - R_-(X)$  称为  $X$  的  $R$  边界域;  $\text{Pos}_R(X) = R_-(X)$  称为  $X$  的  $R$  正域,  $\text{Neg}_R(X) = U - R_+(X)$  称为  $X$  的  $R$  负域; 如果  $Bn_R(X) = \emptyset$ , 则称  $X$  为可由  $R$  定义的精确集,  $Bn_R(X) \neq \emptyset$ , 称  $X$  为不可由  $R$  精确定义, 称  $X$  为粗糙集。

**定义 4.** 不可分辨关系作为一种二元等价关系, 将数据空间  $U$  划分为一系列等价类, 根据属性集  $B$  提供的信息, 包含数据实体  $x$  的等价数  $[X]_B$  可定义为:  $[X]_B = \{y \in U | x \text{ IND}(B) y\}$ 。

**定义 5.** 设  $R \subseteq A$  是一族等价关系,  $r \in R$ , 如果  $\text{Ind}(R) = \text{Ind}(R - r)$ , 则称  $r$  是  $R$  中不必要的; 否则称  $r$  是  $R$  中必不可少的。如果所有  $r \in R$  都是必不可少的, 我们称  $R$  为独立的; 否则称  $R$  为依赖的。

**定义 6.** 设一个信息系统  $S = \{U, A, V, F\}$ ,  $A = C \cup D$  且  $B \subseteq C$ , 定义  $D$  的  $B$  正域  $\text{Pos}_B(D) = \bigcup \{B(X): X \in \text{IND}(D)\}$ 。  $\text{Pos}_B(D)$  是论域中所有通过用分类  $\text{IND}(B)$  表达知识能够确定地划入  $\text{IND}(D)$  类的对象集合。

**定义 7.** 设一个信息系统  $S = \{U, A, V, F\}$ ,  $A = C \cup D$ ,  $D$  和  $C$  间的依赖度定义为:  $K = \gamma_C(D) = |\text{Pos}_C(D)| / |U|$  ( $0 \leq K \leq 1$ )。对于属性  $c \in C$ , 如果  $\gamma_C(D) = \gamma_{C-c}(D)$ , 则称属性  $c$  相对于决策属性  $D$  是冗余的, 否则是不可缺少的, 属性  $c$  在  $C$  相对于  $D$  的重要性可以定义为  $S_{SGF}(c, C, D) = \gamma_C(D) - \gamma_{C-c}(D)$ , 如果  $C$  中的任意属性相对于  $D$  都是不可缺少, 那么  $C$  相对于  $D$  是独立的。

**定义 8.** 设一个信息系统  $S = \{U, A, V, F\}$   $A = C \cup D$  且  $B \subseteq C$ , 如果  $B$  是  $D$  独立的, 且  $\gamma_C(D) = \gamma_B(D)$ , 则称  $B$  是  $C$  的一个  $D$  相对约简, 记为  $R_{\text{red}}(C)$ 。一般而言, 约简不只一个, 各约简集合的交集称为核, 记作  $C_{\text{core}}(C)$ 。核是属性集合中不可能去掉的知识特征集合, 去掉核中任意属性都将导致整个知识系统分类能力的下降。

## 3 人工神经网络

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, 用神经网络本身结构表达输入与输出关联知识的隐函数编码, 输入空间与输出空间的映射关系是通过网络结构不断学习、调整、最后以网络的特定结构表达。人们利用 ANN 系统的学习功能、联想记忆功能、分布式并行信息处理, 已在模式识别、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。ANN 的知识是以固定联结结构存储的, 不易修改和补充。神经网络一般不能处理具有语义形式的输入, 可以实现无导师聚类学习, 但不能确定哪些知识是冗余的, 哪些有用的等。

RBF 神经网络是由输入层、隐藏层和输出层构成, 其结构如图 1 所示。从图可以看出网络从输入到输出的映射为

$$f(x) = \sum_{i=1}^m w_i \phi(x, c_i) \quad (1)$$

其中:  $x = (x_1, x_2, \dots, x_n)^T \in R_n$  为输入矢量,  $W = (w_1, w_2, \dots, w_m)^T \in R_m$  为输出权值矩阵,  $f$  为输出矢量,  $\phi$  为径向基函数,  $c_i$  为径向基函数的第  $i$  个聚类中心。

径向基函数采取常用的高斯函数, 即

$$\phi(x, c_i) = \exp\left(-\frac{\|x - c_i\|^2}{\sigma^2}\right) \quad (2)$$

RBF神经网络的核心问题是确定隐层节点数及相应中心节点位置和宽度，当他们被确定后，RBF神经网络从输入到输出就形成一个线性方程组，此时输出权矢量可用最小二乘法获得。

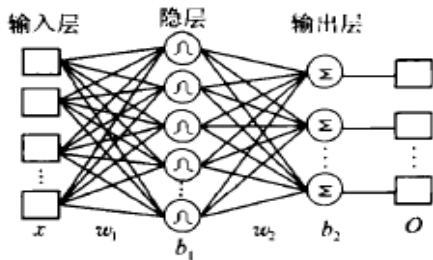


Figure 1. RBF neural network architecture  
图 1. RBF 神经网络结构

#### 4 实例检索策略

实例检索是基于实例推理的重要支撑环节，目的是根据给定的问题描述，利用实例的相关信息，从实例库中寻找适合当前问题的最相似实例，作为满足设计要求的参考实例。检索就是搜索的过程，实例检索和修改模型如图 2 所示。

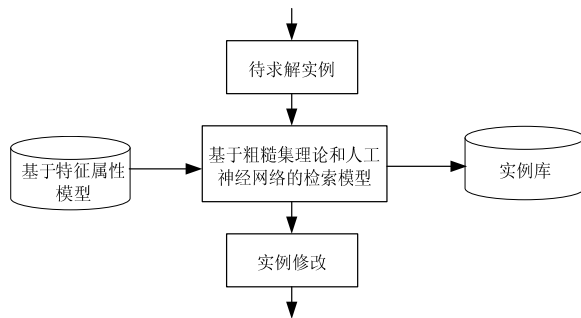


Figure2. Case retrieval and modification model  
图 2.实例检索与修改模型

##### 4.1 冲模实例表达的方法和策略

实例表达作为整个冲模设计系统的基础，其目的在于对一个已经存在的成功设计实例进行记录，使之成为计算机可以识别和利用的数据结构。一个冲模设计实例表达了一系列复杂的内容，其中可以包括产品的功能、结构等产品属性的信息，产品的材料、工艺信息，还可以包括设计规范、设计经验等归纳性知识，以及设计的过程性知识等。对于这样一个复杂的结构，目前的设计实例主要有以下三种策略：

###### (1) 分层次的表达

把一个设计实例分解成为不同的等级层次分别进行表达，是设计类型的 CBR 系统使用最多的一种策略。该方法将一个实例分解成为不同等级的子实例，在设计

当中可以根据当前问题的特点，只针对某些子实例进行推理设计。

###### (2) 不同角度同时表达

一个产品的设计不仅可以从不同角度来共同描述，而且可以映射为不止一个设计实例。

###### (3) 采用多媒体

多媒体技术使一个实例包含不同数据记录格式成为可能，从而可以在一个设计实例当中同时包含不同的角度、类型的信息。

设计实例表达所采用的具体方法主要有框架、语义网络、面向对象的表达等。其中的框架、面向对象方法体现了一定的对知识的层次性描述，比较适用于分层次的设计实例的表达。考虑到设计实例的分解主要体现了对象和属性、一般和特殊、整体和局部这三种层次关系。面向对象表达方法中类的继承体现了一般到特殊的层次关系，类中的属性定义则体现了对象与属性的关系，但不能体现整体和局部的关系。针对面向对象方法的这种不足，本文的研究采用了具有分解层次的面向对象的方法表达设计实例。

#### 4.2 基于粗糙集理论和人工神经网络的实例检索策略

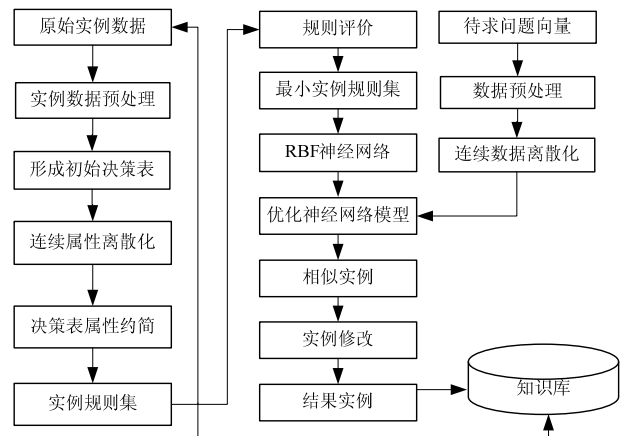


Figure3. Case retrieval model based on RST and ANN  
图 3.基于粗糙集理论和人工神经网络的实例检索模型

冲模设计实例包含了丰富的信息，由于实例库庞大和复杂，需要采用一定的检索策略，对实例库进行有序的组织和管理。从具体方法上，实例检索策略主要有最邻近策略、归纳推理策略、知识导引策略和模板检索策略。但并不是每一个设计信息都对实例的性质具有等同的影响，因而只需考虑那些有助于实例间彼此区分的关键特征属性。本文采用三级检索的策略，第一级实例检索采用归纳推理的策略，利用粗糙集理论对实例库进行

分析和处理,求得能够覆盖原始实例特征的具有最大完备度的最小规则集。第二级将最小规则集作为神经网络的输入,对网络进行训练,得到优化神经网络模型;第三级是将待求问题预处理后,输入优化神经网络模型,检索出最为接近的设计问题的实例,如图3所示。

#### 4.2.1 特征属性决策表

在粗糙集理论中,决策表是一种特殊而重要的知识表达系统,多数决策问题都可以用决策表形式来表达。决策表也是一种分类,其中决策属性相当于分类中的类属性。对决策表的处理主要是对其进行约简和最小化,约简既包括对属性集的约简,也包括对属性值的约简和冗余规则的删除。本文把实例信息表示为属性决策表,根据公式(1)得到粗糙集理论的知识表达系统为  $S=(U, A)$ 。由于实例属性通常包含定性和定量的属性,而粗糙集理论不能直接处理定量属性,因此对于定量的问题先要采用一定的方法进行离散化处理,变成定性问题后,再用粗糙集理论进行处理。

#### 4.2.2 连续属性离散化

离散化的关键就是要把属性的值域范围划成合适的离散区间。连续属性离散化就是在特定的连续属性的值域范围内设定若干个离散化划分点(断点),将属性的值域范围划分成一些离散的区间,再用不同的符号或整数值代表属于每个区间的属性值。本文使用一种混合聚类算法——基于分层聚类的 k-means 算法来离散化决策表中连续数据。

#### 4.2.3 属性约简算法

常用的基于粗糙集理论属性约简算法主要有:基于可辨识矩阵和逻辑运算的属性约简算法,归纳属性约简算法,基于互信息的属性约简算法,基于特征选择的属性约简算法等。这些算法没有考虑到数据领域知识的特殊性以及用户需求的灵活性,因而可以利用一些启发式知识,计算各属性重要度的评价来简化约简过程,本文采用一种基于粗糙集的属性重要度约简算法约简。

#### 4.2.4 RBF 神经网络学习

RBF 神经网络学习把训练过程分为前后两个阶段:首先确定基函数的中心点  $c_i$  和标准方差  $\delta_j$ ; 然后确定隐层到输出层间的权值。

(1) 采用减聚类算法得到RBF神经网络的中心参数  
在 RBF 网络学习中,如何确定  $c_i$  和  $\delta_j$  是一个关键问题。本文采用减聚类算法指导聚类学习,通过一个聚类自动终止判据控制聚类个数,得到合理的中心参数。

(2) 用最小二乘法进行RBF 网络的权值调整  
设有n组输入样本  $x_i$ ,  $d_i$  是每组样本的期望输出值。定义目标误差函数为

$$E = \frac{1}{2} \sum_{i=1}^n \|d_i - y_i\|^2 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m (d_{ij} - y_{ij})^2 \quad (3)$$

$y_i$  是在  $x_i$  输入下的网络输出向量。 $c_i$  和  $\delta_j$  被确定后,我们开始训练隐层至输出层之间的权值。由于隐层到输出层间的关系是一个线性方程组,这时求它们的权值就成为线性优化问题,所以可采用最小二乘法。由于每个输出神经元对应一个线性组合,因而函数值最大的输出神经元就是分类结果。

## 5 结论

本文根据产品设计的特点,将粗糙集理论与神经网络相结合,构建了基于粗糙集理论和 RBF 神经网络集成系统的实例检索模型。该技术有效提高了冲模实例检索的正确性,降低了系统对领域专家的依赖性,提高了冲模实例检索的效率。

## References (参考文献)

- [1] Pawlak Z. Rough sets. International Journal of Computer and Information Sciences [J]. 1982, Vol.11, No.5: 341-356.
- [2] WANG Delu, SONG Xuefeng. Identification of Industry Clusters Life Cycle Based on Integration of Rough Sets and Neural Network[J].Journal of China University of Mining & Technology,2010,Vol.39, No.2: 284-289.
- [3] JIN Wei, CHEN Huiping. A Hybrid Hierarchical kmeans Clustering Algorithm [J]. Journal of Hohai University Changzhou, 2007, Vol.21, No.3:7-10.
- [4] YANG Zhi, HU Jinzhu, HU Longjiang. Component Retrieval Based on Knowledge Base and Case Reasoning[J].Computer Engineering, 2005, Vol.21, No.11: 159-162.
- [5] Albert Fornells, Elisabet Golobardes, Josep Maria Martorell, etc. A Methodology for Analyzing Case Retrieval from a Clustered Case Memory[J]. Case-Based Reasoning Research and Development. 2007, No.8: 122-136.