面向专家的知识库优化[1]

盛昭瀚，赵卫东，陈国华
（南京大学管理科学与工程研究院，南京 210093）

摘要：知识库的质量是影响智能系统性能的主要因素，而知识获取一直是设计智能系统的瓶颈问题，这是由于目前人类认识的局限性，导致知识工程师和专家之间的不协调关系造成的. 为克服上述不利局面，本文利用粗糙集等理论，得到含有噪声的初始知识库，然后采用遗传算法、可视化技术和知识校验等技术对规则库和案例库进行了优化，从而在知识获取过程中建立了知识工程师和专家之间的新型的关系，其中专家处于中心地位，知识工程师只是起辅助作用，即整个知识获取过程是面向专家的.

关键词：面向专家；知识库；遗传算法；优化
中图分类号：TP18　　**文献标识码**：A　　**文章编号**：1007-9807(2001)03-0040-06

0 引 言

知识库是专家系统和智能决策支持系统的重要组成部分，其质量直接决定了系统的性能高低. 通常，知识库中的知识主要以规则和案例的形式存在. 传统上知识的获取是由知识工程师和专家的多次交互完成，他们之间的交流管道存在障碍，专家能够有效地解决问题，但要其整理出知识是有一定难度的，尤其是他们拥有的只可意会、难以言传的经验知识，它们是专家长期实践的结晶，就目前人类的认知水平，难以直接得到. 这样知识获取就成为设计智能系统的瓶颈；而且由于主客观的因素，知识库难免出现矛盾、蕴涵、冗余和循环等问题，影响了系统的推理能力.

机器学习是获取知识的有效途径. 为避免传统知识获取的困难局面，本文结合故障诊断专家系统的一些实践，提出由专家收集大量的案例，然后采用粗糙集等机器学习工具获得初始规则集. 由于机器学习的局限性（实质上是归纳学习的非单调性），难以获得完全、无噪声的知识库，所以有必要对知识库进行补充、优化和校验，以得到高质量的知识库，增加系统的实用性. 上述过程，实质上是面向专家的，知识工程师的主要任务是引导和协助专家对知识获取的一些问题进行技术处理，从而在他们之间形成了新的合作关系，有利于构造高质量的知识库.

知识库的优化，已引起人们注意[1-4,10]，出现了许多有效的方法，如状态空间图[1]、可视化技术[2]和杂交算法等. 遗传算法是模仿自然界生物进化的一种仿生技术[5]，它因能达到最优解而在函数优化、机器学习等领域得到广泛的应用. 本文采用它优化知识库. 在优化过程中，遗传算法具有下述特点：(1)知识编码直接选择规则本身作为染色体，规则前提和结论中的事实和命题为基因；(2)规则适应值由专家确定；(3)规则库中的矛盾、冗余和蕴涵等问题采用可视化技术协调完成. 此外，还研究了案例库的校验，讨论了相应的算法.

1 知识的获取

目前机器学习的主要方法有基于信息论的决策树、基于集合论的学习等方法. 其中粗糙集理论

1 收稿日期：1999-08-14；修订日期：2000-03-?
基金项目：国家自然科学基金资助项目（79970070）.
作者简介：盛昭瀚（1944-），男，江苏人，教授，博士生导师.

已逐渐引起人们的注意.

粗糙集理论是由 Pawlak 提出的[6]，它在机器学习中用于发现分类规则，其基本思想是：把专家的决策实例经过预处理（如连续属性离散化、空值的处理等），整理成决策表形式，然后通过一系列简化（包括属性和值简化），得到最小决策规则规则集．这种方法对于完全信息或相容的决策表，能够得到比较满意的结果，所以在控制领域、医疗诊断、图象处理、模式识别等领域受到了重视[7]．

尽管粗糙集理论也有简（优）化规则的措施，但对案例数量较大的场合计算量很大，且专家一时难以收集全各种情况下的故障案例，知识的获取还应结合其他方法．这样便得到初始知识库．由于主客观的原因，初始规则库可能存在不完全、矛盾、蕴涵、冗余和循环等问题，有待于进一步优化处理．

2　知识编码

规则库是由大量的规则组成，它是专家经验的结晶．为简便起见，假设规则为 Horn 子句，且其不确定性用可信度 C 表示：R_i：if a then b_i $C(R_i)$．其中 a_i 为前提，由简单事实、命题或其组合，即 $a=\{<a_{i1},f(a_{i1})>,\cdots,<a_{in},f(a_{in})>\}$，式中 f 表示事实、命题的可信度，$f(a_{ij})=\min\{f(a_{ij}),j=1,\cdots,m\}$；b 为结论，由一个简单事实或命题组成，其可信度 $f(b_i)=f(a)*C(R_i)$，$C(R_i)$表示规则的可信度．用 Horn 子句表示为：R_i：$b_i\leftarrow a_i$ $C(R)$．

规则库精简的目的，是为了消除其中的矛盾、冗余和蕴涵等缺陷．在保持知识库同样功能的前提下，减少库的规模．用遗传算法时，首先考虑知识的编码问题．不同的码长和码制，对问题的求解精度和算法的效率都会有影响．常用的码制采用二进制．规则库中的规则数量大，规则前提的属性（事实）多，有些规则比较复杂，前提和结论部分包含有语法结构．采用二进制编码会带来编码过长和固定编码长度的问题．本文采用规则本身作为编码，即染色体为规则，基因为规则前提和结论的逻辑命题和事实．这样的编码方法具有方便直观、实用性强和易于理解的优点．

3　用遗传算法优化规则库

文[4]提出一种用遗传算法精化知识库的方法．但此方法存在下述主要问题：（1）忽略了规则的重要参数—置信度；（2）没有考虑专家的监督指导作用，当原始规则库存在不一致或在规则库演化过程中不能及时排除不符合实际的规则；（3）因采用规则本身作为编码，算法的复杂性较高，仅适用一些小型知识库的精化，对于大型知识库，缺乏简化措施，算法的效率会降低，难以得到最优解．针对上述问题，本文基于可视化等技术，以专家为中心，给出了一些改进措施．

规则编码的特殊性，导致遗传算法的许多步骤不同于以往的许多做法．用遗传算法优化规则库的原理如图1（假设推理机是满足要求的），其过程如下：

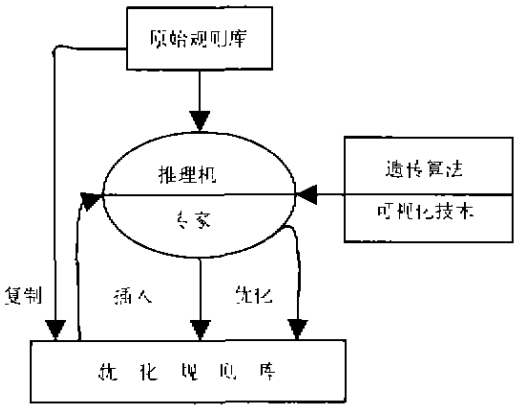

图1　用遗传算法优化规则库

（1）优化知识库（相对原知识库而言）置空．

（2）选择原知识库简短、常用和可信度大的规则若干．

（3）复制（2）中的规则到优化知识库．

（4）对原规则库中的每一规则，实施操作：把规则的前件作为已知事实，由推理机用优化规则库中的规则前向推理．若推理结果与规则的结论不一样，或推理结果空（无规则匹配），说明优化知识库中的规则不足，则需插入新的规则．插入的新规则不一定有意义，需要专家的评判．只有合乎事实的规则才能加入规则库，以免引入噪声．规则的可信度由专家确定．最后得到用于优化的初始规则库．

（5）规则库优化：把原规则库的每一规则前

提与优化知识库匹配，再根据优化规则库各规则的响应情况，计算其适应值，然后进行遗传操作。

规则的适应值按下列原则计算：

·一般情况下，被激活的规则适应值增加；矛盾规则和结论错误者除外，未激活的规则适应值不变；

·对于得到同一正确结论的规则，前提所含命题越少者的适应值增加较多；

·矛盾规则（即对于相同前提，有不同结论的规则）的处理，首先由专家确定正确响应的规则，并增加其适应值，错误响应的规则其适应值减少；

·对于激活后又能激活其他规则的规则，适应值应增加较大，例如有下列规则 R_1—R_5：

R_1：e←a∧b∧c；R_2：d←a∧c∧e；R_3：c←d∧e；R_4：f←c∧e；R_5：g←a∧b∧f。

当规则 R_1 和 R_2 被激活时，将随之激活规则 R_3 和 R_4，故规则 R_1 和 R_2 的适应值增加较大。遗传算法的基本算子包括：

1 选择：选择适应值较高的规则，删除适应值较低的规则，保持规则库的种群数小于原始规则的容量。

2 复制：选择适应值较高的规则，按一定概率复制；

3 交叉：这里指互换两规则的前提有关命题，互换后的规则有效性由专家评判，符合实际的规则的可靠度由专家给出。如有规则：

R_1：e←a∧b∧c；R_2：d←f∧g∧e

互换后得到的有效规则可能为：

R_3：e←f∧g∧e；R_4：d←a∧b∧c

因要校验互换后规则的有效性，这一步的工作量可能较大，需要专家的经验启发和解释。

4 变异：因规则变异会带来许多无实际意义的规则，增加专家的工作量，所以要算作慎重考虑。

(6) 重复步骤(3)直至最优规则库各规则的适应值不再发生改变为止。

为减小算法的复杂性，在步骤(1)前可以用模糊演绎图（Fuzzy Deduction Graph，FDG）和 L 系统等技术来实现推理过程的可视化，来检查规则中的矛盾和循环。模糊演绎图是规则的一种图形表示方式，其画法简单直观。例如有模糊规则集 R，其模糊演绎图如图2。图中圆圈结点表示规则前提或结论的事实，简单命题，而箭头的边表示规则。这样整个知识库就构成一个有向网络图，用四元组表示为：FDG=（N，E，f，C），式中各符号的含义：N：结点集合；E：有向边集合；函数 f：N→[0，1]，表示结点的可信度；函数 C：E→[0，1]，表示规则的可信度。

R={ r：p2←p1 0.9
r：p3←p2 0.8
r：p4←p1 and p2 0.9
r：p5←p3 0.77
r：p6←p2 and p3 0.8
r：p1←p8 0.9
r：p8←p6 0.85 }

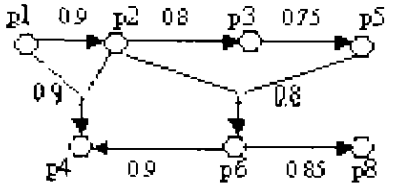

图2 规则的模糊演绎图

为实现规则推理的可视化，模糊规则须增加可视化算子，用 BNF 范式表示：

模糊规则::=〈前件〉〈后件〉〈可信度〉；

前件::=〈前提〉

前提::=〈对象名，关系算子，状态特征值，可视化算子〉

后件::=〈对象名，状态特征值，可视化算子〉

为使可视化算子伴随推理过程，引入了形式语言系统 L 系统[7]，它通过符号串的解释可以转化为可视化的工具。模糊演绎树可以由 L 系统描述，推理路径的描述也就可以采用有限符号集。L 系统包含许多子系统，分别生成各类形式语言，采用 0L 系统，这里 0 是指每一个符号在重写时与相邻符号无关[7]。

整个推理过程用 0L 语言描述，且被解释执行。这样就使规则推理和可视化过程并被执行。0L 系统的使用，简化了可视化的过程。推理开始时，已知事实用可视化结点表示，显示在屏幕上，结点的内容附着在表示在结点的图标上，双击鼠标即可显示其内容。这些事实会激活相关规则，使规则也显示出来，直至得到结论。同样，规则的内容也附着在表示规则的有向边上，双击鼠标也显示规则内容。为使各规则被激活（其前件的状态为

激活状态时，其后件结点动作，它将传递消息给相关规则前件，如此反复直至得到结果。

推理过程的可视化（模糊演绎图）不仅增强了系统的解释能力，而且可以帮助知识工程师发现规则库中的错误，例如若图中含有环时，则规则库中可能含有环，导致系统进入死循环；当图中的两个结点有两条或两条以上的有向边相连时，规则库中可能存在矛盾等等。在设计故障诊断的专家系统中，首先收集大量的典型案例，由专家利用经验或有粗集等机器学习方法从中抽取出初始规则库，然后用上述的算法优化。实践证明这种做法不仅可以得到简洁的规则库，规则的矛盾、遗漏、冗余等现象大大减小；而且知识库的构造可以脱离知识工程师，一定程度上避免了传统知识的瓶颈问题。

4 案例库的校验

案例推理是人工智能的一种问求解方法，案例作为一种知识表达方式，表达专家的难以形式化的经验知识。实践证明，案例推理可以弥补规则推理的不足，而且它与推理规则的结合，可以提高系统的推理能力，其效率大于两者的能力之和[1]。而且对系统案例推理能力的校验，对系统的性能的提高有很大关系。案例推理的效果，与案例的检索、案例的调整等模块的功能有关，所以案例库的校验主要集中在案例的检索、案例的调整。类似规则库的优化，案例库的校验需要专家的积极参与。文[1]虽然对案例库的校验作了深入的研究，但其淡化了专家的作用，离开了专家的知识库的校验和验证，结果可能会有局限性，因为事物的精确表示在其本身，只有知识拥有者才真正了解知识的意义。在文[1]的基础上，讨论了面向专家的案例库校验。

案例库的校验通常分为二步：

（1）选择校验准则，以决定案例库和案例推理机能否接受。这个问题由领域专家掌握，包括案例能否覆盖领域，是否有"病态"案例，案例集容度是否合适等问题。这一步是案例校验的基础。

（2）准备测试案例。案例库的校验不仅是对案例库质量的评价，更重要的是对案例推理的检索能力和调整功能的测试。案例推理系统应对案例库中的案例作出正确的反映，而且也应有一定的"外推能力"，即对案例库外案例能在案例库中检索到相似的案例，并经过适当的调整得到可以接受的解。所以，测试案例应有两部分组成：案例集本身和一定数量并能覆盖领域范围的外部案例，后者有专家收集。

（3）案例推理校验，主要包括相似案例检索和案例的调整。

4.1 案例库的检索

案例库的检索校验算法如下：

（1）案例库中的案例检索

1）把案例库中所有案例拷贝到测试案例库；

2）对测试案例库的每一个案例，做如下操作：

a. 将案例作为当前问题，等待案例库系统检索；b. 案例库系统对测试案例检索，并把检索出的最相似案例的结果与测试的结果做比较；若结果相同，则标记此测试案例为"正确"，否则标记为"错误"。

3）统计测试案例集的检索结果，若发现有标记为"错误"的测试案例，说明案例检索功能有问题，修改之，重复2）直至测试案例集的检索基本上没有错误；

4）置测试案例集为空。

（2）案例库的"外推"能力（"插值"）

1）由专家再采集适当数量并均匀分布在案例集的案例组成测试案例集（不在案例库中），此案例库也在案例的调整使用；

2）对测试案例库的每一案例，做如下操作：

a. 将案例作为当前问题，等待案例库系统检索；b. 案例库系统对测试案例检索，并把检索出的最相似案例的结果与测试案例的结果做比较，由专家判断结果是否可以接受。若结果可以接受，则标记此测试案例为"正确"，否则标记为"错误"。

3）统计测试案例集中标记为"正确"的案例数量，计算其比例；

4）由专家决定此比例是否可以接受。若可以接受，说明案例的检索模块正常；否则应考虑案例的检索功能修改，回到2），直至专家认为可以接受为止。

4.2 案例的调整

案例的调整检验算法如下：

（1）案例库中的案例调整

1) 把案例库中所有案例拷贝到测试案例库；

2) 对测试案例库每一个案例，做如下操作：

a. 从案例库中删除当前测试案例；b. 将案例作为当前问题，等待案例库系统检索；

c 案例库系统对测试案例检索，并把检索出的最相似案例的结果按案例的调整算法调整；

d. 测试案例的调整结果与测试案例的实际解做比较：由专家判断是否可以接受. 若结果可以接受，则标记此测试案例为"正确"，否则标记为"错误"；e. 恢复当前测试案例.

3) 统计测试案例集的检索结果，计算标记为"错误"的测试案例比例，由专家决定是否可以接受. 若不能接受，说明案例调整功能有问题. 修改之，重复2)，直至测试案例集的调整基本上达到要求；

1) 置测试案例集为空.

2) 案例库的"外推"能力. 将测试案例库换成 4.1.2) 1) 中的案例集，重复 1) 中的各步.

5 结束语

结合医疗诊断知识库的设计，面向专家对规则库用遗传算法进行优化. 实验证明对于较小规模的知识库是可行的. 同样对案例库的校验，也能提高系统案例推理的能力. 目前工程上已出现规则推理和案例推理复合的实例，证明两种推理方法复合的性能大于两者的性能之和，所以提高案例库与规则库组成的知识库的质量对于复合系统的性能有很大影响.

参 考 文 献：

[1] Zupan B, Albert Mo Kim Cheng. Optimization of rule based systems using state space graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 1998,10(2):278-293

[2] Lee J, Liu K F R, Chiang Weiling. A fuzzy petri net-based expert system and its application to damage assessment of bridges[J]. IEEE Transactions on Systems, Man and Cybernetics, 1999,29(3):350-369

[3] Lopez-Suarez A, Kamel M. Reorganization knowledge to improving performance[J]. IEEE Transactions on Knowledge and Data Engineering, 1998,10(1):193-191

[4] Blas Parym. Knowledge base improvement through genetic algorithms[J]. Information Systems, 1998,111:65-79

[5] Goldberg D. Genetic algorithms in search, optimization and machine learning[M]. MA: Addison Wesley, 1989

[6] Pawlak Z. Rough sets-theoretical aspects of reasoning about data[M]. Kluwer Academic Pub., 1991

[7] Chandwani M, Chaudhat N S. Knowledge representation using fuzzy deduction graphs[J]. IEEE Transactions on Systems, Man and Cerbernetics, 1996,26(6):848-856

[8] Rozenberg G. The mathmetical theory of L systems[M]. New York: Academic Press, 1980

[9] Golding A R, Rosenbloom P S. Improving accuracy by combining rule-based and case-based reasoning[J]. Artificial Intelligent, 1996,87:215-254

[10] Gonzalez A J, Xu Lingh, Gupta U M. Validation techniques for case-based reasoning systems[J]. IEEE Transactions on Systems, Man and Cybernetics, 1998,28(4):467-477

Expert-oriented optimization of knowledge base

SHENG Zhao-han, ZHAO Wei-dong, CHEN Guo-hua

Graduate School of Management Science & Engineering, Nanjing University, Nanjing 210093, China

Abstract: The quality of knowledge base is important for knowledge intensive systems (KIS), such as expert systems and intelligent decision support systems. But over a long time, knowledge acquisition is a bottleneck in designing KIS due to the bad coordination between knowledge engineers and experts, which is limited by human epistemology level nowadays. To solve the problem, this paper utilizes some machine learning theories (mainly rough set theory) to get a rough knowledge base and then choose genetic algorithms, visualization technology, knowledge verification and validation technology to optimize it. We, in

this way, build a new relation between knowledge engineers and experts in which experts play a decisive role while knowledge engineers are accessory, i.e., it is expert-oriented during the knowledge acquisition course.

Key words: expert-oriented; knowledge base; genetic algorithm; optimization

(上接第24页)

[21] 袁安照，余光胜. 现代企业组织创新[M]. 太原：山西经济出版社，1998
[22] 汪丁丁. 回家的路：经济学家的思想轨迹[M]. 北京：中国社会科学出版社，1998
[23] 斯蒂芬·P·罗宾斯. 管理学[M]. 第4版. 北京：中国人民大学出版社，Englewood Cliffs, NJ: Prentice Hall，1997
[24] 何清涟. 现代化的陷阱——当代中国的经济社会问题[M]. 北京：今日中国出版社，1998

Mountains-forming: new logic of knowledge and learning based enterprise

RUI Ming-jie, FAN Sheng-jun
School of Management, Fudan University, Shanghai 200433, China

Abstract: For the coming knowledge economy, we need new enterprise logic different from any others before. We summarize here the new logic of enterprise as the knowledge and learning based, compare it with the mountains-forming campaign appeared in the geological history, and make a frame study of it. The theoretical base of the new view of enterprise is knowledge and learning theories, its typical organization pattern is the team-based organization, the essence of management is to accomplish competence promoting through transforming and combination of knowledges, marketing theme is to pursue the knowledge transcending. By all these things, we have portrayed the outline of the new enterprise logic.

Key words: knowledge and learning based enterprise; team-based organization; competence promoting; knowledge transcending

(上接第39页)

Nonlinear time-varying systems identification by feedforward neural networks

GU Cheng-kui, WANG Zheng-ou
Institute of System Engineering, Tianjin University, Tianjin 300072, China

Abstract: A new identification method based on feed-forward networks is presented for nonlinear time-varying systems. We apply local extended Kalman Algorithm to train feed-forward networks, this algorithm needs no matrix inversion computation and has the higher convergence speed and the smaller storage required in comparison to the global extended Kalman Algorithm. Simulation results show the present method has better effect on nonlinear time-varying systems identification.

Key words: nonlinear time-varying systems; feed-forward neural networks; local algorithm; Kalman filtering