

一种基于机器学习的专家系统知识获取方法

马 昕 刘长龙 张贝克

(北京化工大学信息科学与技术学院, 北京 100029)

摘 要: 提出一种基于机器学习的混合知识获取方法,该方法结合了基于历史数据的规则提取方法和基于模型的规则提取方法。使用这两种方法提取规则,将其应用于对原油电脱盐系统的故障诊断中。实验结果表明,该方法能够有效的进行规则的提取,为故障诊断打下了良好的基础。其中基于历史数据的规则提取方法通过基于遗传算法的粗糙集约简来实现;基于模型的规则提取方法利用了符号有向图(SDG)的计算机自动推理结果,将因果图转化为规则。利用两种规则获取方法同时充实专家系统知识库,提供覆盖整个工艺流程的知识。

关键词: 知识获取;粗糙集约简;符号有向图;故障诊断

中图分类号: TP182

引 言

专家系统故障诊断^[1]是故障诊断中研究最多、应用最广的一类智能故障诊断技术。专家系统的建造过程主要分为知识获取、知识表示和系统实现三个部分^[2]。其中,知识获取是专家系统建造过程中最困难的一部分,也是最重要的一个阶段。专家系统的性能在很大程度上取决于所获知识的质量。

随着自动化技术、计算机和网络通信技术的飞速发展和广泛应用,人们面临的控制系统日益复杂,要求也越来越高,随之而来的对系统故障的快速诊断和处理也提出了更高的要求。成功及时的故障诊断将能够对系统实施良好的控制,从而保证控制系统能够高效稳定地运行。由于基于知识的故障诊断方法可以避免对精确数学模型的过分依赖,对于难以得到系统解析模型的复杂非线性系统的故障诊断问题是一种非常实用的方法。然而在实际的故障诊断问题中由于系统复杂的非线性特性,描述故障模式的信息常有某种程度的不完备、不确定性。对于故障特征的提取一直受到广泛的重视^[3]。有主元特征提取,基于人工神经网络的提取,模糊信息优化处理及基于互信息熵提取方法等。但主元特征提取

会因为输入变量的变化而改变主分量的特征值计算结果;当特征输入太多时,基于人工神经网络的方法存在耗时费工以及合适的网络结构选取问题;模糊信息优化处理和基于互信息熵的方法需要预先确定隶属函数或数据样本的概率分布。

本文提出一种基于机器学习的混合知识获取方法,利用符号有向图^[4]建立工艺流程机理模型,根据符号有向图计算机自动推理结果进行规则提取;利用粗糙集^[5]规则提取方法,从历史数据中挖掘出有用规则。将两种从不同途径获取的规则知识经过相容性检验后存入专家系统规则库,这样建立的规则库能够覆盖与工艺流程相关的深、浅层知识,帮助解决专家系统知识获取的“瓶颈”问题。

1 基本知识

1.1 粗糙集

一个信息系统 S 可以表达为四元组^[5-6]: $S = \langle U, C \cup D, V, f \rangle$ 。 U 是对象的非空有限集合; $C \cup D = A$ 是属性集合,子集 C 和 D 分别称为条件属性和决策属性; $V = \bigcup_{a \in A} V_a$ 是属性值的集合, V_a 表示了属性 $a \in A$ 的范围,信息函数 $f: U \times A \rightarrow V$ 指定 U 中每一个对象的属性值。

在粗糙集(RS)理论中,信息系统表示成决策表形式,决策表的列表示属性,行表示对象,每个单元格表示对象的属性值。容易得知,一个属性对应一个等价关系,一个决策表可以看作是对一族等价关系的定义。

对于每个 $R \subseteq A$,我们可以定义不可分辨关系

收稿日期: 2008-04-01

基金项目: 北京化工大学青年教师基金(QN0730)

第一作者: 女,1975年生,讲师,博士

E-mail: maxin@mail.buct.edu.cn

$IND(R) = \{(x, y) \in U^2, \forall a \in R, a(x) = a(y)\}$ 。
对每个子集 $X \subseteq U$, 把以下两个集合分别称为 X 的
 R 下近似和 R 上近似: $RX = \bigcup \{x \in U \mid [x]_R \subseteq X\}$; $\overline{R}X = \bigcup \{x \in U \mid [x]_R \cap X \neq \emptyset\}$ 。

R 是等价关系的一个族集, $r \in R$ 。若 $IND(R) = IND(R - r)$, 则称关系 r 在族集 R 中是可省略的, 否则就是不可省略的。若族集 R 中每个关系 r 都是不可省略的, 则称族集 R 是独立的, 否则就是依赖的或非独立的。若 $Q \subseteq R$ 是独立的, 并且 $IND(Q) = IND(R)$, 则称 Q 是关系族集 R 的一个约简 RED。

1.2 符号有向图

符号有向图 (SDG) 是一种定性模型表达方式。数学描述^[5]如下: SDG 模型 γ 是有向图 G 与函数 φ 的组合, 即 $\gamma = (G, \varphi)$ 。有向图 G 由 4 部分组成 $G = (V, E, \delta^+, \delta^-)$: 节点集合 $V = \{v_i\}$; 支路集合 $E = \{e_k\}$; 邻接关联符 δ^+ 表示 $E \rightarrow V$ (支路的起始节点) 和 δ^- 表示 $E \rightarrow V$ (支路的终止节点), 该“邻接关系”分别表示每一个支路的起始节点 $\delta^+ e_k$ 和终止节点 $\delta^- e_k$ 。函数 $\psi: E \rightarrow \{+, -, 0\}$, $\varphi(e_k) = \varphi(v_i, v_j)$ ($e_k = (v_i, v_j) \in E$) 称为支路 e_k 的符号。

SDG 模型 $\gamma = (G, \varphi)$ 的样本是节点状态值的函数 $\psi: V \rightarrow \{+, 0, -\}$, $\psi(v_i)$ ($v_i \in V$) 称为节点 v_i 的符号:

$$\begin{cases} \psi(v_i) = + & Xv_i - \overline{X}v_i \geq \epsilon v_i \\ \psi(v_i) = 0 & |Xv_i - \overline{X}v_i| < \epsilon v_i \\ \psi(v_i) = - & \overline{X}v_i - Xv_i \geq \epsilon v_i \end{cases}$$

其中 ϵv_i 代表节点 v_i 处于正常状态的域值。 $\psi(v_i)$ 为 0, 说明节点 v_i 的真实值在正常值范围内; $\psi(v_i)$ 为正, 说明节点 v_i 的真实值超出了上限阈值; $\psi(v_i)$ 为负, 则说明节点 v_i 的真实值超出了下限阈值。SDG 模型除了具有很强的系统状态表示能力外, 还能够进行有效的推理。在 SDG 模型样本中, 从初始节点的状态偏离开始, 导致其邻接下游节点的状态偏离, 并沿着相容路径, 一直影响到末端的节点, 导致其状态发生偏差。通过对相容路径的搜索即因果图的搜索, 就可以发掘出故障在复杂系统内部的发展演变过程。

2 基于机器学习的综合知识获取方法

由于专业领域知识的启发性难以捕捉和描述, 加之领域专家通常善于提供实例而不善于提供知

识, 所以知识获取被认为是专家系统研究开发中的“瓶颈”问题^[2]。

机器学习机制是知识获取的高级方式, 是人工智能领域的一个研究热点, 也是解决知识获取“瓶颈”问题的根本出路所在。机器学习可用图 1 所示的简化模型来说明。环境表示客观世界中获得的信息集合; 学习系统负责对所获取信息去粗取精、归纳总结; 知识库是知识存储的仓库; 执行环节利用知识库中的知识完成指定任务, 同时把情况反馈到学习系统。

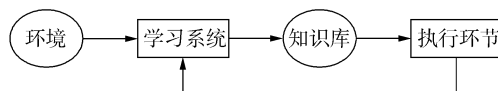


图 1 机器学习模型

Fig. 1 The model of machine learning

机器学习的结果一般都用产生式规则来表示, 因为产生式规则既可以表示过程性知识也可以表示说明性知识, 知识表达能力较强, 且直观、易于理解。

2.1 基于 RS 约简的规则提取

RS 主要用于增强系统自学习以及处理不完备信息的能力, 消除属性冗余值, 提取最小规则集合, 实现专家系统“由事例学习”的过程, 充实知识库。

下面采用基于遗传算法的近似计算方法, 来得到约简属性集合。虽然结果是次优的, 但大大节省了计算时间。基于遗传算法的约简算法步骤为:

输入: 决策表 $T = (U, C \cup D)$ 和遗传算法产生的变异率 τ 。

输出: 输出基于 τ 的一个约简 RED。

初始令冗余集 $L = \emptyset$ 和约简集 $S = \emptyset$, $n = \text{card}(C)$ 。

1) 由遗传算法计算变异率 τ , 根据 τ 对条件属性集合 $R = \{c_1, c_2, \dots, c_n\}$ 排序, 得集合 $R = \{b_1, b_2, \dots, b_n\}$;

2) 对 R 中每一个元素: 令 $R = R - \{b_i\}$; 调用子函数 $(\text{flag}, S, L) = \text{RED}(R, S, L, T)$, 如果 $\text{flag} = 1$ 则将元素 b_i 保留, 否则删除;

For $i = 1$ to n

$R = R - \{b_i\}$

$(\text{flag}, S, L) = \text{RED}(R, S, L, T)$

If $\text{flag} = 1$ then $R = R \cup b_i$

Next i

3) 集合 R 即为所求约简集。

子函数 $(\text{flag}, S, L) = \text{RED}(R, S, L, T)$ 负责

计算不可分辨关系,算法步骤如下:

1) 若集合 R 属于约简集 S , 令 $\text{flag}=1$, 退出子程序转至上面的步骤 2); 如果 R 属于冗余集 L , 令 $\text{flag}=0$, 退出子程序转至上面的步骤 2);

2) 对 U 中的对象重新排序, 将 R 中属性排列至前 $\text{card}(R)$ 位置;

3) For $i=2$ to $\text{card}(U)$

sign = 1

For $j=1$ to $\text{card}(R)$

If $a_j(x_i) \neq a_j(x_{i-1})$ then sign = 0

Next j

If sign = 1 then 将 R 加入集合 L 中, 令 $\text{flag}=0$, 退出子程序

Next i

4) 将 R 加入集合 S 中, $\text{flag}=1$, 退出子程序

该算法的计算结果总是获得一个约简, 重复使用该算法, 可以得到决策表的所有约简, 约简结果依赖于条件属性的排列顺序。

2.2 基于 SDG 模型的规则提取

除了通过对历史数据分析之外, 可以对工艺流程建立 SDG 模型, 利用 SDG 的推理能力, 得到描述流程机理的因果图集合。作者所在课题组开发了计算机辅助 SDG 推理软件^[7], 在对复杂工艺流程建模后, 可自动推理获得描述整个工艺流程的因果图集合, 如图 2 所示。

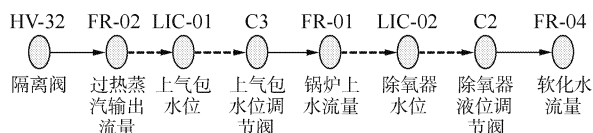


图 2 SDG 因果图示例

Fig.2 Cause-effect graph

图 2 中, 圆圈表示节点, 带箭头的线表示支路。实线表示支路上游节点对下游节点的影响关系为正, 虚线则表示支路上游节点对下游节点的影响关系为负。图 2 表示的内容转化为产生式规则如下:

IF FR-04 偏低 AND C2 偏低 AND LIC-02 偏高 AND FR-01 偏低 AND C3 偏低 AND LIC-01 偏高 AND FR-02 偏低 THEN HV-32 误差

因此可在各因果图基础上进行规则提取。假设有图 3 所示 SDG 模型, 其结论形式如下: A 变大 \rightarrow B 变小 \rightarrow C 变小。

2.3 规则相容性检验

为避免两种途径提供的规则可能存在非法性,

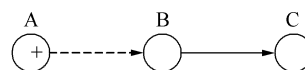


图 3 A 出现正偏差的解释

Fig.3 Explanation when the status of A is higher
系统中需引入对规则的检验功能:

1) 冗余性检验。若规则中结论相同, 只是所表示的故障征兆属性中, 有一些征兆互逆而其余征兆相同, 则合并该条规则;

2) 多义性检查。若出现故障条件相同而结论不同的规则, 则显示该条规则, 交由专家进行判别;

3) 完整性检验。若给出前提条件, 必然能得出结论。

经过以上规则的提取和检验后, 将各条规则加入到知识库中, 实现规则学习的全过程, 如图 4 所示。

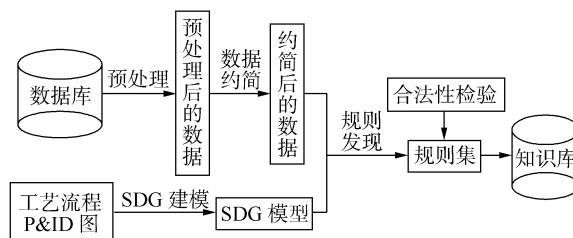


图 4 混合知识获取方法框图

Fig.4 Framework of knowledge acquisition

3 知识获取实例的研究

以常减压蒸馏装置的一个组成部分为例, 同时应用基于粗糙集和 SDG 模型两种方法来获取规则。此装置主要有以下几部分组成: 原油换热系统、原油电脱盐系统、初馏系统、常压蒸馏系统、减压蒸馏系统、加热炉及烟气余热回收系统、航煤精制系统、轻烃压缩回收系统、剂类加入系统。

这里以原油电脱盐系统为例, 其系统如图 5 所示。原油电脱盐就是在一定温度下, 在破乳剂、注水、混合、电场等因素综合作用下, 原油中小水滴结成水滴, 靠油水密度差而将原油中的水和溶解在其中的盐同时分离的过程。在原油电脱盐系统中涉及的操作参数有电脱盐温度、电脱罐压力、原油与破乳剂和水的混合压差、注水量、乳化层厚度、破乳剂、电脱盐罐的油水界面、电场强度(变压器输出电压)、罐底沉积物、电脱盐罐电流、电脱盐罐电压。这些参数都会影响原油的脱盐效果。

影响脱盐率的参数有很多, 但是在涉及到具体

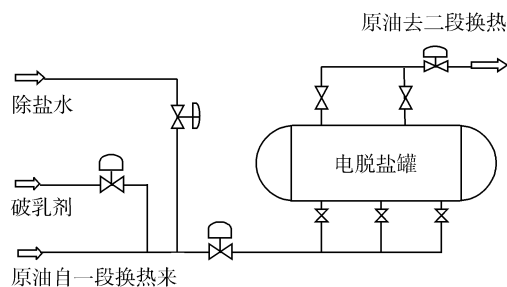


图 5 原油电脱盐系统图

Fig. 5 Electric desalting system

的故障诊断问题时,并不是每个参数都是同等重要的。我们在进行相关的数据分析时,其复杂度和这些参数的数目关系呈线性、平方,甚至更高。因此,去除一些无关的参数和相关性很小的参数是我们关心的问题。根据以上介绍的基于遗传算法的粗糙集属性约简方法对某石化厂一个月的数据进行处理。选择一个包含较少参数的原油脱盐系统数据源来说明属性约简问题。打开数据源后,可见此数据源共有 8 个字段(属性),如图 6 所示。共有 22720 条记录,如图 7 所示。与这些数据相对应的非正常工况有 18 种模式。

电脱盐系统 : 表		
字段名称	数据类型	
电脱盐温度	数字	
电脱盐操作压力	数字	
混合压差	备注	
乳化层厚度	数字	
油水界位	数字	
罐底沉积物	数字	
电脱盐罐的电流	数字	
电脱盐罐的电压	数字	

图 6 数据源字段

Fig. 6 The fields in the data source

电脱盐系统参数 : 表							
电脱盐温度	电脱盐操作压力	混合压差	乳化层厚度	油水界位	罐底沉积物	电脱盐罐的电流	电脱盐罐的电压
128.5	0.96	0.035	0.21	0.53	1	16	262
128.6	1.05	0.042	0.28	0.54	1	16	262
128.1	0.93	0.034	0.24	0.54	0	17	262
129.4	0.95	0.032	0.26	0.53	0	16	262
129.3	1.08	0.046	0.31	0.58	1	16	265
129.1	1.07	0.043	0.28	0.57	0	17	264
129.4	0.96	0.036	0.21	0.57	1	16	264
129.2	0.96	0.031	0.26	0.56	0	16	264
127.7	0.91	0.029	0.24	0.56	1	16	264
128.5	0.94	0.034	0.27	0.54	1	16	262
128.4	0.94	0.036	0.26	0.52	1	16	262
128.2	0.93	0.034	0.22	0.53	0	15	262
129.3	0.96	0.038	0.21	0.54	0	16	262

图 7 数据源记录

Fig. 7 The records in the data source

图 8 给出了一个电脱盐系统的信息系统,利用上面提到的基于遗传算法的属性约简方法对此信息系统进行数据分析。分析后的结果如图 8 所示。

电脱盐系统参数 : 表			
电脱盐温度	混合压差	油水界位	罐底沉积物
128.5	0.038	0.51	1
128.6	0.034	0.58	0
129.3	0.036	0.49	1
129.6	0.038	0.54	0
127.4	0.031	0.54	0
127.6	0.035	0.52	0
129.2	0.029	0.57	1
130.5	0.034	0.56	0
128.8	0.036	0.47	1
129.4	0.031	0.45	1
127.3	0.035	0.52	0

图 8 约简后结果

Fig. 8 The records after reduction

从图 8 中可以看到,对电脱盐系统的信息系统进行属性约简后,就只剩下了 4 个字段:电脱盐温度、混合压差、油水界位和罐底沉积物。利用这 4 个字段就可以表示原信息系统中的所有信息,而且这个过程是不损害原有信息表中的信息的,即根据原信息系统所获得的任何分类知识,通过约简后的字段集组成的信息系统同样可以得到。如果采用一般的属性约简算法,即去判断每个属性是不是必要的,如果不是就删除它,否则保留。此方法能够得到一个属性约简结果,但不一定能得到一个满意的属性约简结果。对于本信息系统若采用这种方法得到的一个约简则是字段电脱盐温度、电脱盐操作压力、混合压差、乳化层厚度、油水界位、罐底沉积物这 6 个字段。但像电脱盐操作压力对原油脱盐效果的影响其实是很小的。

从实际应用的角度来说,通过属性约简,减少信息系统的属性字段,可以大大减少信息系统的数量,从而提高数据分析的效率和速度。因为在化工系统监测采集的数据都是数以万计的,因此通过约简去除一个属性,其减少的数据量也是相当可观的,再加上很多数据分析的复杂度是字段数和记录数的平方,甚至更高的关系,所以通过约简后带来的数据分析效率的提高和速度的加快是非常显著的。

4 结束语

本文提出的混合知识获取方法,将基于粗糙集理论的规则提取方法和基于危险与可操作性分析的规则获取方法结合在一起,从知识库完备性、准确性出发,进行产生式规则知识的获取。并将其应用到原油电脱盐系统的故障诊断中,有较好的实用价值。专家系统的知识库采用 Access 数据库,规则提取算

法和人机交互界面采用 VC 编程实现。从不同途径获取的规则经过相容性检验后存入专家系统规则库,这样建立的规则库能够覆盖与工艺流程相关的深、浅层知识,帮助解决专家系统知识获取“瓶颈”问题。

参考文献:

- [1] 张萍,王贵增,周东华. 动态系统的故障诊断方法[J]. 控制理论与应用, 2000, 17(2): 153-158.
- [2] 石松芳,宋建萍. 专家系统中知识库维护的若干问题[J]. 湖北教育学院学报, 2006, 23(8): 24-26.
- [3] 郭小芸,马小平. 基于粗糙集故障诊断特征提取[J]. 计算机工程与应用, 2007, 43(1): 221-224.
- [4] 吴重光,夏涛,张贝克. 基于符号定向图(SDG)深层知识模型的定性仿真[J]. 系统仿真学报, 2003, 15(10): 1351-1355.
- [5] 任永功,王杨,闫德勤. 基于遗传算法的粗糙集属性约简算法[J]. 小型微型计算机系统, 2006, 27(5): 862-865.
- [6] Pawlak Z. Rough sets and intelligent data analysis[J]. Information Sciences, 2002, 147(1): 1-12.
- [7] 张贝克,夏涛,吴重光. 集成化 SDG 建模、推理与信息处理软件平台[J]. 系统仿真学报, 2003, 15(10): 1360-1363.

Knowledge acquisition methods for expert systems based on machine learning

MA Xin LIU ChangLong ZHANG BeiKe

(College of Information Science and Technology, Beijing University of Chemical and Technology, Beijing 100029, China)

Abstract: The objective of this paper is to present a hybrid method for knowledge acquisition which combines rule acquisition methods based on historical data with model-based methods. The former method derives rules through rough set reduction of a genetic algorithm, while the latter transforms the cause-effect graph to rules by using the automatic reasoning result of a signed directed graph (SDG). Using the rules obtained by the above hybrid method to enrich the rules base provides knowledge covering the whole flow process. An example of the use of the method is given for an electric desalting system.

Key words: knowledge acquisition; rough set reduction; signed directed graph (SDG); fault diagnosis