研究论文

基于基因表达式编程的化工过程故障诊断知识抽取

李秀喜，熊海霞，杨国军
（华南理工大学化学与化工学院，广东 广州 510640）

摘要：专家系统是化工过程故障诊断最常用的技术之一。专家系统的基础是专家知识，而知识获取一直是专家系统的“瓶颈”问题，所以知识提炼是开发化工过程故障诊断专家系统的关键技术。本文提出了一种基于基因表达式编程（GEP）的化工过程故障诊断知识的提取技术，通过模糊函数对数据进行模糊化处理，利用 GEP 演化特性从数据库中找出异常以及产生这些异常的原因，从而获得用于故障诊断的知识规则。实际案例研究结果显示，该技术与领域专家结合能有效提取故障诊断知识，可作为化工过程故障诊断专家系统的知识获取手段。

关键词：专家系统；故障诊断；基因表达式编程；知识提取
中图分类号：TQ 021.8 **文献标识码**：A **文章编号**：0438－1157（2010）02－0392－06

Extracting knowledge based on GEP for fault diagnosis of chemical processes

LI Xiuxi, XIONG Haixia, YANG Guojun
(*School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou* 510640, *Guangdong, China*)

Abstract: Expert system based on expert knowledge is one of the most common technologies in chemical fault diagnosis. Knowledge acquisition is the bottleneck of expert system, so knowledge extraction is the key technology of expert system. In this paper, an extracting knowledge technology based on gene expression programming (GEP) for fault diagnosis of chemical processes was presented. Fuzzy processing of the data was performed with the ambiguity function, and then from the database the anomalies and the reasons of these anomalies were identified by using GEP evolution properties. Therefore the knowledge rules were obtained which could be used in fault diagnosis. Practical case study showed that the technology combined with experts in the field could effectively extract fault diagnosis knowledge, and could be used as a knowledge acquisition tool in expert system for fault diagnosis of chemical processes.

Key words: expert system; fault diagnosis; gene expression programming; extracting knowledge

引 言

专家系统诊断利用专家积累的丰富实践经验，模仿专家分析问题和解决问题的思路，而且能够解释自己的推理过程，解释结论是如何获得的，无论是在理论上还是在工程上应用都很广泛[1-4]。专家

2009－10－21 收到初稿，2009－10－30 收到修改稿。
联系人及第一作者：李秀喜（1966—），男，副研究员。
基金项目：国家自然科学基金项目（20536020，20876056）。
Received date: 2009－10－21.
Corresponding author: LI Xiuxi, cexxli@scut.edu.cn
Foundation item: supported by the National Natural Science Foundation of China (20536020, 20876056).

系统的基础是知识，然而知识获取一直是专家系统应用的“瓶颈”问题。

数据库技术极大地加快了人类积累数据的速度，尤其在化工生产领域，由于自动化程度的提高和数据库技术的应用，积累了大量的生产历史数据，问题就在于如何从这些数据中挖掘出有用的知识。决策树算法是一种从训练集中发现分类知识的数据挖掘方法。由于决策树学习得到的模型是树状结构，且易于表示为多个 IF-THEN 规则，已成为常用的知识挖掘工具。目前典型的决策树算法是 C4.5[5-6]等，然而，C4.5 构造决策树的核心思想是贪心算法，使得决策树对历史数据非常准确，但应用到新数据时，准确性急剧下降。在构造树的过程中，需要对分类算法数据集进行多次顺序扫描和排序，降低了算法的学习效率。另外，C4.5 还不能进行模糊知识挖掘。

本文利用基因表达式编程（gene expression programming，GEP）技术提出了一种用于故障诊断规则知识的自动提取算法，其动机为：①从异常的系统运行状态中找出未知的故障现象，获取新的知识；②在所获取的故障知识中，一个故障现象对应多个故障原因，而故障原因的表达式描述即故障原因症兆往往十分复杂，利用数据挖掘工具可以对这些表达式做出不断的修正；③为知识工程师进行知识库维护时提供决策支持，知识工程师可以通过数据挖掘技术所提供的规则，对知识库中的知识进行不断修正。

1　GEP 算法

受生物遗传基因表达特点的启发，Ferreira[7-8]于 1999 年首先提出了 GEP 的知识发现新技术。与遗传算法（genetic algorithm，GA）、遗传编程（genetic programming，GP）类似，GEP 属于进化算法的一种。GEP 用固定长字符串来代表计算机程序。当进化这些程序的适应度时，它们随后以不同大小和形状的表达树（expression trees，ETs）表达。重组时，通过修正和遗传到新代来生成染色体个体而不是表达树。GEP 融合了 GA 和 GP 的优点，从而提供更大潜力来解决复杂问题。Ferreira 指出 GEP 比 GA 和 GP 的效率高 2～4 倍。

基因表达式编程遗传操作的基本单位是染色体（chromosome），遗传信息载体的最小单位是基因（gene）。染色体由一个或多个基因组成，基因由线性的定长字符串编码而成。基因由头（head）和尾（tail）组成，头部包含了变量集（varible set）中的变量和函数集（function set）中的函数，而尾部只包含变量。头部和尾部的长度有以下关系

$$|\text{tail}|=|\text{head}|\times(n-1)+1 \tag{1}$$

其中，n 是此基因含有的函数集中所有函数参数的最大数目[9-10]。文献［11］证明了满足该约束关系的染色体集合在遗传操作下封闭。GEP 算法流程如图 1 所示。

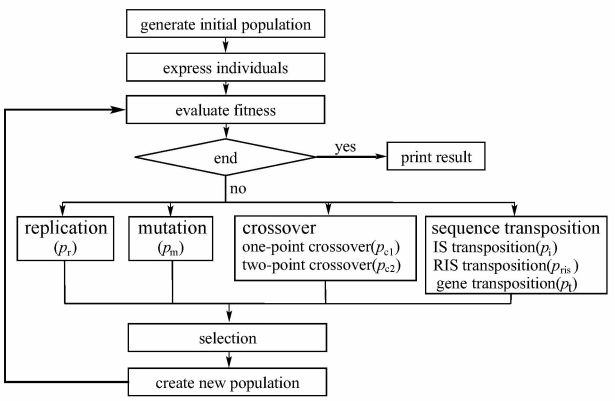

图 1　GEP 算法流程

Fig. 1　Flow chart of GEP algorithm

2　基于 GEP 的模糊知识规则提取

利用数据挖掘技术提取知识，就是从已经发生的问题中找出规则，以避免类似的问题再次发生。本文的目的就是从历史异常数据库中找出曾经发生过的异常以及产生这些异常的原因，以获得用于故障诊断的知识规则。知识规则采用产生式规则表述形式，即“IF 条件表达式 THEN 结论”的形式。GEP 模糊知识规则提取算法的基本步骤为：

（1）数据筛选，包括异常数据标识、n 类分类问题转化为 n 个 2 分类问题；

（2）变量模糊化；

（3）对每一个 2 分类问题，运行 GEP 算法，获得最终树表达式；

（4）对每一个最终树表达式剪枝修整，改写成模糊 IF-THEN 规则；

（5）规则确认与调整，并添加到知识库中。

2.1　数据筛选

首先从历史数据库中找出有问题的数据，即故障记录数据，并将其分类标记。为了获得较理想的知识规则，在数据筛选时，应使正常数据记录和异常数据记录的个数大致相当。若有多个异常，也要

使各个异常记录个数匹配，不能相差太大，以免出现模型偏向数据记录个数多的一方。若有多个设备异常，则将多分类问题转化为多个 2 分类问题。本文采用的多分类问题分解策略如图 2 所示。

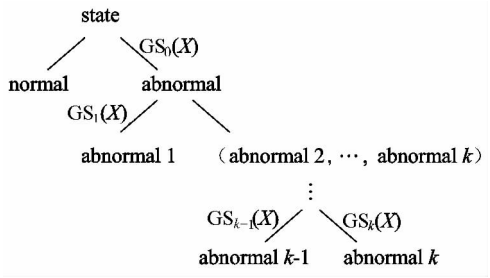

图 2 多分类问题分解策略

Fig. 2 Decomposition strategy of multi-classification problem

2.2 变量模糊化与模糊运算

一般来说，专家们在描述一个过程的状态或采样特征时，经常使用“好/坏”、“高/低”或“高、中、低”等模糊性词语，如塔釜压力过高、进料流量太低等。本文采用梯形模糊隶属度函数对数据模糊化，梯形隶属度函数 $u(x)$ 描述如图 3 所示，这种形式的函数需要 a_1、a_2、a_3、a_4（$a_1 < a_2 < a_3 < a_4$）4 个参数来描述。

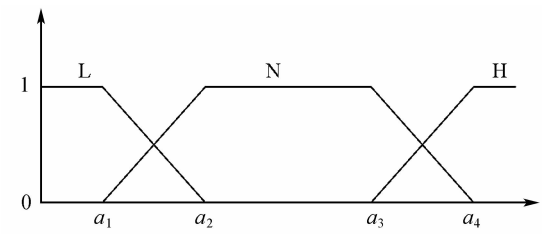

图 3 梯形隶属度函数 $u(x)$

Fig. 3 Trapezoidal membership function $u(x)$

在化工生产过程中，各个变量的特性如变化快慢、高低限、对过程安全的影响等是不同的。对 a_i（$i=1$，2，3，4）的设定必须要求对过程知识有充分的了解，这里由现场工程师设定。

本文采用标准模糊运算。设 A、B 是两个模糊集合，$u_A(x)$ 是元素 x 属于模糊集合 A 的隶属度，$u_B(x)$ 是元素 x 属于模糊集合 B 的隶属度，则标准模糊运算为

AND：$u_{A\wedge B}(x) = \mathrm{MIN}(u_A(x), u_B(x))$；

OR：$u_{A\vee B}(x) = \mathrm{MAX}(u_A(x), u_B(x))$；

OR：$u_{\neg A}(x) = 1 - u_A(x)$

图 4 是一个模糊逻辑运算实例，其代表的规则为：IF x_1 = low OR x_2 = low AND x_7 = high THEN 状态 = 异常 WITH $CF = 0.7$。

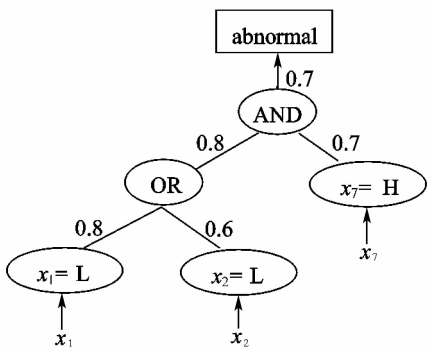

图 4 模糊逻辑运算实例

Fig. 4 Example of fuzzy logic operations

L—low；N—normal；H—high

2.3 两层 GEP 算法

本文采用两层 GEP 编码算法，共同进化，第二层调用第一层的个体。第一层用于获得函数“x_i = value”的隶属度值，其中 value 表示“低、中、高”，因此，编码简单。函数集只有一个函数{“=”}。终端集包括所有变量和 3 个逻辑变量值{低、中、高}。遗传算子只有单点交叉和变异，且规定函数节点不参与变异。第二层通过对第一层表达式的调用和模糊逻辑运算获得完整的规则表达式树。两层 GEP 调用如图 5 所示。

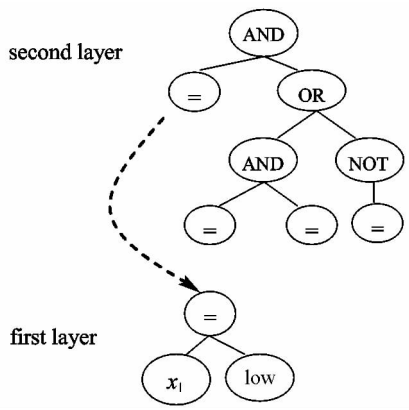

图 5 两层 GEP 调用示意图

Fig. 5 Schematic diagram of two-layer GEP call

适应值是度量群体中的个体在进化过程中解决问题所表现出的优良程度，本文采用的适应值函数 f 定义为

$$f_i = w_1 \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + w_2 \times \mathrm{Simp} \qquad (2)$$

其中，Simp 表示个体编码树的复杂度，定义为

$$\mathrm{Simp} = \frac{\mathrm{MaxNode} - 0.5 \times \mathrm{NumNode} - 0.5}{\mathrm{MaxNode} - 1} \qquad (3)$$

因此，$0.5 \leqslant \mathrm{Simp} \leqslant 1$。其中，$w_1$、$w_2$ 是权重；

TP 为正确分类样本数量；FP 为满足规则条件，但错误分类样本数量；FN 为不满足条件，但分入规则结果类的样本数量。MaxNode 为表达式树的最大节点数；NumNode 为个体树的实际节点数。为了增加全局搜索能力，适应值较低（<50%）的个体有 20% 的选择概率，好的适应值个体（>50%）有 80% 的选择概率。同时，规定最好的个体（或其中之一）直接复制到后代中。

3　案　例

本文以某石化厂的重质糠醛精制生产流程为例说明算法的有效性。润滑油糠醛精制的目的是除去润滑油油料中的非理想组分，降低润滑油的酸值、残炭等，提高润滑油的黏温性能。来自减压蒸馏装置侧线的馏分润滑油和来自丙烷脱沥青装置（轻脱）的残渣润滑油原料，含有大量的润滑油非理想组分——多环短侧链的芳香烃和环烷烃，硫、氮、氧化合物及少量的胶质。这些组分使润滑油颜色深、酸值高、残炭值大，黏温特性和抗氧化性变坏，腐蚀金属设备。因此，润滑油原料必须经过精制，去除所含的非理想组分，才能达到润滑油基础油的标准[12]。

重质糠醛精制装置由原料脱气（塔 T408）、糠醛抽提（塔 T401）、精制液汽提（炉 F401，塔 T402）、抽出液汽提（塔 T403/1，塔 T403/2，炉 F402，塔 T404，塔 T405）、水溶液回收（塔 T406，容 C405，塔 T407）等子系统组成。其中糠醛抽提是装置的核心部分，其关键设备是萃取塔 T401，其操作直接影响精制油的收率和质量。糠醛在萃取塔内与原料油逆流接触并分层，溶解原料油的非理想组分，从而达到精制的目的。重质糠醛精制流程如图 6 所示。

根据现场工程师的经验和历史故障记录，本文从糠醛精制系统中选出 40 个关键变量，其中流量变量 10 个、液位变量 9 个、温度变量 13 个、压力变量 8 个。从数据库中筛选出正常记录 1000 个、异常记录 1125 个。异常状态分为 2 类：一类是 T403/1 异常，包含 459 条记录；另外一类是 T406 异常，有 666 条记录。对流量的监测，本文不直接使用流量变量，而是监测流量的增量变化，即当前时刻的前 5 min 流量均值与前 30 min 的均值之差

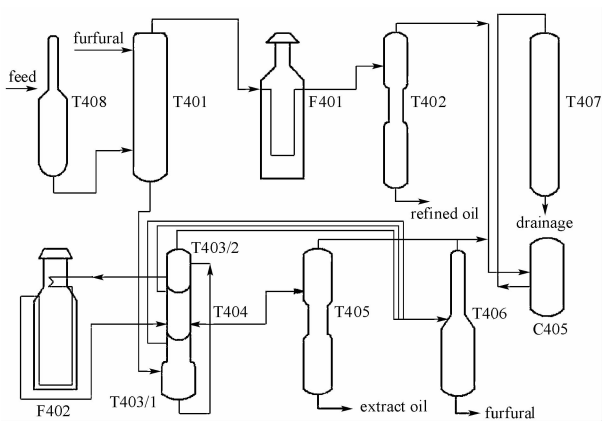

图 6　重质糠醛精制装置流程

Fig. 6　Flowsheet of heavy furfural refining unit

$$\Delta F = \mathrm{mean}(F,5) - \mathrm{mean}(F,30) \tag{3}$$

按照上述多故障分类策略，需要运行 3 次 GEP 算法。本文的两层 GEP 运行参数见表 1。

（1）第 1 次运行将正常与异常分开，$GS_0(X)$ 表达式树如图 7 所示。

整理后得到如下规则

IF　TI4045＝low OR LIA4023＝low OR
　　NOT（LIA4022＝normal）OR
　　NOT（PI2027＝normal）
THEN　系统异常

（2）第 2 次运行提取出 T403/1 的异常规则，得到运行结果 $GS_1(X)$ 为

IF　LIA4022＝low AND PI4027＝low
THEN　T403/1 异常

结果分析：LIA4022（T403/1 的液位）是抽出液汽提过程的关键监测变量，PI2027 是 T403/1

表 1　GEP 参数

Table 1　GEP parameters

Layer number	F	T	Pop	G	h	p_{c1}	p_{c2}	p_m	p_r
first	=	all 40 measured variables, 3 logic values(low, normal, high)	200	300	1	0.8	0	0.2	0.1
second	AND, OR, NOT	all gene expression of first layer	200	300	7	0.4	0.4	0.2	0.1

Note: F—function set; T—terminal set; Pop—size of population; G—maximum generation; h—head; p_{c1}—probability of one-point crossover; p_{c2}—probability of two-point crossover; p_m—probability of mutation; p_r—probability of replication.

表 2 故障诊断知识

Table 2 Fault diagnosis knowledge

Number	Rule	Operational suggestion
1	IF LIA4022=low AND PI4027=low THEN T403/1 liquid level is low	increase top pressure of T403/1
2	IF LIA4023=low AND ΔFIC4008=low THEN T406 liquid level is low	increase feed flow rate of return wet furfural at top seven of T406
3	IF LIA4023=low AND ΔFIC4002=high THEN T406 liquid level is low	check T401 furfural flow control valve , then replace it by vice-line valve
4	IF TI4045=low THEN top temperature of T406 is low	increase top temperature of T406

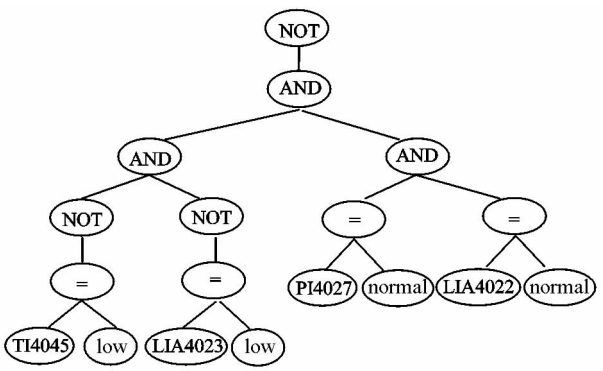

图 7 $GS_0(X)$ 表达式树

Fig. 7 $GS_0(X)$ expression trees

的塔顶压力测量变量；抽出液自 T401 底部出来之后，进入三效蒸发塔（T403/1、T403/2、T404）；抽出液中含有大量糠醛（约 85%）和少量非理想组分，在三效蒸发塔中，糠醛由于沸点比非理想组分沸点低，分别从 T403/1、T403/2、T404 的顶部蒸发出来，进入糠醛干燥塔 T406。从上述分析可知，当相关流量参数并没有明显变化时，塔压过低（即 PI4027=low），T403/1 的糠醛蒸发量会增大，造成液位降低（即 LI4022=low）。

（3）第 3 次运行提取 T406 的异常规则，其结果 $GS_2(X)$ 为

IF （LIA4023=low AND ΔFIC4008=low）

OR（LIA4023=low AND ΔFIC4002=high）

OR（TI4045=low）

THEN T406 异常

结果分析：T406 是糠醛干燥塔，水和糠醛从 T406 顶部进入，由于水和糠醛可以形成共沸物（常压沸点 97.5℃），因此 T406 采用共沸精馏使糠醛和水分离；经干燥后的糠醛从 T406 底部出来，从 T406 顶部出来的是糠醛和水的共沸物。从运行结果来看，T406 出现了两个异常，一个液位低（LIA4023 = low），另一个是塔顶温度低（TI4045=low）。造成液位低的原因是塔顶部 7 层回流湿醛进料量减少（ΔFIC4008=low）或塔底出口流量过大（ΔFIC4002=high）。

上述提取出的异常规则经过现场工程师的分析确认，整理出了 4 条故障诊断知识（表 2），并添加到知识库中。

4 结 论

本文利用 GEP 构建了从历史数据中提取出用于故障诊断的知识规则的算法步骤。通过实例说明了所提出的算法能较好地对数据库中数据进行规则的提炼。将此技术应用于对未知故障知识规则的提炼，所得结果对知识工程师和领域专家进行知识库更新和维护都具有很好的意义。当然，知识的自动提取并不能代替专家的工作，而应视为对专家工作的一个补充。将专家经验收集手段与知识自动提取技术相结合，可获得更加完备的知识，并进一步增强系统的自我学习能力，有助于故障诊断专家系统的进一步应用。

References

[1] Liao S H. Expert system methodologies and applications—a decade review from 1995—2004. *Expert System with Applications*, 2005, **28**: 93-103

[2] Becraft W R, Lee P L. An integrated neural network/expert system approach for fault diagnosis. *Computers & Chemical Engineering*, 1993, **17** (10): 1001-1014

[3] El-Shal S M, Morris A S. A fuzzy expert system for fault detection in statistical process control of industrial processes. *IEEE Transactions on Systems*, 2000, **30** (2): 281-289

[4] Qian Yu, Xu Liang, Li Xiuxi, Li Lin, Andrzej Kraslawski. LUBRES: an expert system development and implementation for real-time fault diagnosis of a lubricating oil refining process. *Expert Systems with Applications*, 2008, **35** (3): 1252-1266

[5] Quinlan J R. Induction of decision trees. *Machine Learning*, 1986, **1** (1): 81-106

[6] Quinlan J R. C4.5: Programs for Machine Learnings. Canada: Morgan Kaufmann, 1993

[7] Ferreira C. Gene expression programming in problem solving//Invited Tutorial of the 6th Online World Conference on Soft Computing in Industrial Applications. Berlin, 2001: 10-24

[8] Ferreira C. Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. Angra do Heroismo, Portugal: Springer, 2002

[9] Vassilios K K, Andreas S. Data mining based on gene expression programming and clonal selection//IEEE Congress on Evolutionary Computation. Vancouver, BC, Canada, 2006: 16-21

[10] Wilson S W. Classifier Conditions Using Gene Expression Programming. Berlin: Springer-Verlag, 2008

[11] Zuo Jie (左劼). Key technology research of gene expression programming. Chengdu: Sichuan University, 2004

[12] Shui Tiande (水天德). Modern Oil Production Processes (现代润滑油生产工艺). Beijing: China Petrochemical Press, 1997: 6-15