文章编号:1002-0411(2003)01-032-04

复杂工业过程中数据挖掘模型研究

罗印升[1,2]　李人厚[1]　梅时春[1]

(1. 西安交通大学系统工程研究所　西安　710049;　2. 陕西理工学院电气工程与自动化系　汉中　723003)

摘　要:针对复杂工业过程中产生积累的大量数据,以新的视角分析了应用数据挖掘的基础和数据特点,讨论了数据挖掘的基本思想,提出了集成数据挖掘的模型结构,为复杂工业过程的监控开辟了一条新的途径并指出了要进一步研究的问题.

关键词:数据;数据挖掘;复杂工业过程

中图分类号:TP274　　**文献标识码**:B

DATA MINING MODEL RESEARCH ON COMPLEX INDUSTRY PROCESS

LUO Yin-sheng[1,2]　LI Rin-hou[1]　MEI Shi-chun[1]

(1. *Institute of System Engineering, Xi'an Jiaotong University, Xi'an* 710049,

2. *Department of Electrical Engineering & Automation, Shanxi Institute of Technology, Hanzhong*　723003)

Abstract: According to a large number of data in complex industry process, This paper analyses the applied basis of Data Mining (DM) and data characteristics with new views, discusses DM idea, puts forward integrated DM model, gives the future topics, provides a new path for complex industry process monitoring and control.

Keywords: data, DM, complex industry process

1　引言(Introduction)

随数字技术、信息技术的飞速发展和计算机技术广泛应用,各个行业已产生积累了大量的数据.面对如此庞大的数据,人们要么置之不理,要么通过数据库管理系统与统计分析方法相结合,实现查询、检索及表报功能,进行联机分析处理 OLAP(或 OLTP 联机事务处理).从而得出可供决策参考的统计分析数据,但数据中隐藏的内在有用知识无法得到,我们处在数据丰富而知识贫乏(Data Rich but Information Poor)的矛盾之中.另一方面,就人工智能中的知识获取而言,它主要依靠用户或者领域专家手工将知识输入到知识库中,耗时费力,且因主观性易于出错.因此迫切需要能够自动发现获取知识的新技术,这便是数据挖掘产生的直接推动力.于是内容涉及人工智能、数据库技术、模式识别、数据可视化、统计学等的一个新兴交叉学科领域数据挖掘技术诞生了.

复杂工业过程领域更是如此,本文就数据挖掘在该领域的应用问题进行讨论.在第 2 部分 DM 概述的基础上,第 3 部分介绍复杂工业过程中 DM 的基础与现状,第 4 部分总结复杂工业过程中数据的特点,第 5 部分提出复杂工业过程中 DM 的基本思想,第 6 部分提出了复杂工业过程中 DM 的模型结构,第 7 部分进一步的研究工作,结束语为第 8 部分.

2　DM 简介(DM summary)

2.1　DM 的定义

文[1]指出数据挖掘是指从大量数据中提取或"采掘"知识.进而形成了广义的 DM 定义,数据挖掘是从存在于大量数据的数据库、数据仓库或信息储存体中发现有趣知识的过程.DM 的对象不仅是数据库,也可以是文件系统或其它任何组织在一起的数据集合.文[2]定义了知识发现(KDD, Knowledge Discovery in Database)的概念.它是指识别出存在于数据库中可信的、新颖的、具有潜在应用价值和最终可理解模式的非平凡过程.而数据挖掘是此过程的一个特定关键步骤.由此可知这两个定义的

收稿日期:2002－02－25

最终目标都是从大量数据集合中挖掘出各种有价值、被人理解的可用于指导实践的知识. KDD 的定义完整准确但表面上仅局限于数据库,而[1]中的 DM 定义对其进行了扩充指组织在一起的数据集合. 但这里的数据集合应该是来源于实践中有用系统的,而不是杂乱无章的数据堆积. 通常对 *KDD* 和 *DM* 不加区别混合使用.

2.2 DM 的过程与任务

DM 的过程可归纳为:数据准备(数据选择、清洗、变换)、数据采掘(各种挖掘算法的运用)和知识的表达、解释与验证三个阶段,是一个循环往复的过程. 主要任务有分类、回归分析、聚类、预测、关联性、变化和偏差分析、模式发现和路径发现等.

3 复杂工业过程中 DM 的基础与现状(Base and state of DM in complex industry process)

由于工业生产过程常伴随着物理化学反应、生化反应、相变过程及物质与能量的转换和传递,又面临效益、品种、质量和环境保护、安全等多方面的挑战,因此它已经成为一个十分复杂的大系统.

3.1 基础

现代工业生产过程中,以计算机为核心的各种控制系统(如 DCS、FCS、CIPS)的广泛应用,使得能够采集、存储大量关于工业生产过程的过去(历史)和当前运行状态(动态)的丰富有价值的数据(包括,产品质量变量、过程变量、设备状态);也为能够应用先进控制技术奠定了坚实的基础. 以往人们往往过分重视控制算法的开发,而忽视了对系统运行中有关操作人员的行为特征、经验,设备状态及生产过程状态的大量数据的全面分析利用,更难从中发现知识. 部分原因由于系统的复杂性、处理数据量的庞大和缺乏相应的技术手段,而 DM 的产生为解决这些问题提供了强有力的工具. 它可以充分利用积累的历史数据和当前数据,提取潜在的模式、规则,以新的视角为复杂工业过程系统的监控开辟新的途径. 这使得在复杂工业过程中展开数据挖掘应用研究有了坚实的基础.

3.2 现状

DM 理论与技术的研究已经较为广泛,国外已广泛应用于商业、金融、电信和企业管理中. 在工业过程中的应用也有成功的报道,英国已应用 DM 于化工过程的监控中,取得了显著的效益[3]. 美国钢铁公司和神户钢铁公司运用 DM 技术研究分析产品性能规律进行质量控制. CASSIOPEE 故障发现系统被欧洲三大航空公司用来诊断和预测波音 737 客机的故障,其应用获得了"欧洲创新应用"的一等奖. 国内有关工业方面 DM 的理论和应用研究才开始,除宝钢有初步的应用成果报道外,成功的应用报道很少[4].

4 复杂工业过程中数据的特点(Data characteristics in complex industry process)

要进行复杂工业过程中的数据挖掘,在遵循 DM 一般方法的基础上,必须认真分析其数据特点并结合工艺上的要求进行. 其数据特点可概括如下[3,5]:

(1) 数据量巨大、高维且有较强的耦合性

不论是计算机监督还是计算机控制的工业系统,都要定时采集系统的变量和设备状态,以供显示、控制之用,另外还有重复测量和冗余测量的数据,日积月累这些数据量是非常巨大的;同时由于工业系统的行为状态是许多变量因素共同作用的结果,它们之间有较强的耦合及非线性关系.

(2) 工业噪声和过程中的不确定性

工业过程系统工作环境复杂,电、磁、噪声干扰较强,加之系统存在的不确定性,因而数据易受污染.

(3) 动态性与数据类型的多样性

各种变量的值是不断变化的,这反映了系统的动平衡过程,是系统本质的反映. 在过程的监控中通过辨识、观察诸变量的数值,预测系统的状态及变化趋势. 其数据的类型也是多样的如,数值型(整型、实型)、非数值型、逻辑型等.

(4) 多时标性与不完整性

因系统的复杂性,众多变量的变化快慢各异,所以采集信号的频率不同,导致时间上的不同步. 在数据的记录上也可能丢失数据,造成数据的不完整.

(5) 多模态性

数据是系统状态变化的反应. 既有正常工作状态,又有各种异常状态(包括未知状态、不确定状态)和故障态的数据. 前者是主体,后者的数据量相对较少. 但它们都是 DM 中所不可缺少的.

5 复杂工业过程中数据挖掘的基本思想(DM idea in complex industry process)

由于工业过程的复杂性及特点,其数据挖掘必须将传统数据挖掘方法与现代以计算智能 CI

(Computational Intelligence)为核心的众多理论方法融合集成起来. 以工业过程积累的大量数据为研究对象,发挥现行监控平台的作用,结合生产工艺要求,研究适用于复杂工业系统数据特点的鲁棒、有效的新挖掘方法,解决数据丰富、知识贫乏的问题. 并运用新方法获取知识,补充、完善和更新知识系统和监控手段. 指导操作,优化生产,改进工艺,提高产品质量.

5.1 融合集成各种 DM 方法

结合数据先验知识、工程背景知识,将基于统计学的传统 DM 方法、机器学习方法及计算智能方法融合集成起来,研究能从大量数据中提取反映系统状态的模式、规则的算法与策略. 该方法对数据的不确定性、不完整性,噪声有较强的鲁棒性.

5.2 建立系统的状态监控与故障诊断模型

基于数据挖掘的新方法,对复杂工业过程的历史数据包括正常状态、异常状态及各种故障状态数据(系统中各种变量在故障发生后,它们中的一部分或全部其数据或状态变化都会与正常情况下有所不同,其中包含了丰富的故障信息)进行聚类、分类分析,建立系统状态评估和故障预测模型. 在此基础上,对系统过程,设备状态的当前数据进行在线挖掘分析. 对已建立的模型进行检验和修正,形成系统动态分析模型. 并进行可视化处理,提供操作指导与分析,使操作员得到充分的信息,进行生产过程监视和操作,从而实现更高水平的监控,防止事故隐患发生,保证安全生产.

5.3 构建产品质量控制模型,改进工艺、提高产品质量

所谓生产过程最优化可以概括为:在满足必要的约束条件下,改变生产过程的工艺参数,使某种与经济效益有关的目标函数达到极值. 在生产优化实施中,对大量生产数据进行挖掘找到产品质量与工艺参数的模型关系. 分析诸多变量作用下的产品质量规律,帮助质检人员、工艺人员弄清影响产品质量的主次因素,提出相应的对策,进一步调整工艺参数,进行质量控制,为实现生产过程操作最优化提供指导.

5.4 丰富知识库和决策支持系统,为先进控制的实现创造更好的条件

工业过程对象已变为一个十分复杂的系统,产生了更为困难的过程控制问题以及对高性能控制器的要求. 这样经典的控制方法难以胜任,以计算机为工具的高等控制方法已经应用于这些系统中. 但结构复杂,计算量大. 通常必须有知识库作为支持,而知识的获取是关键. 数据挖掘技术可以将提取的潜在模式、规则评估检验后归入知识库,使得高等控制充分发挥作用,提高生产过程的控制水平.

6 复杂工业过程中数据挖掘的模型(DM model in complex industry process)

在前述分析复杂工业过程中数据挖掘的基础、数据特点及讨论基本思想的基础上,我们把计算机技术、控制技术和信号处理技术结合起来,建立起智能代理(Agent)和面向对象的递阶式、分布式工业过程数据挖掘模型结构,如图 1 所示. 为了发挥现行 DCS/FCS/CIPS 系统的作用,我们提出了集成数据挖掘系统和典型的 CIPS 系统的融合结构如图 2 所示.

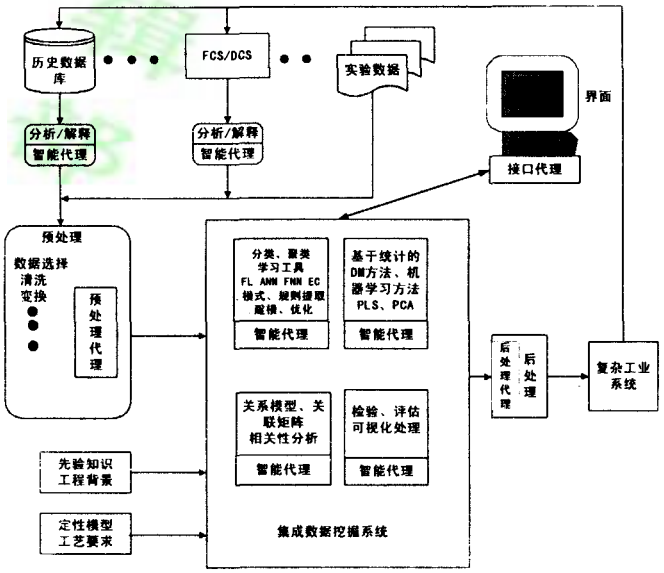

图 1 集成数据挖掘模型

Fig. 1 Model of integrated data mining system

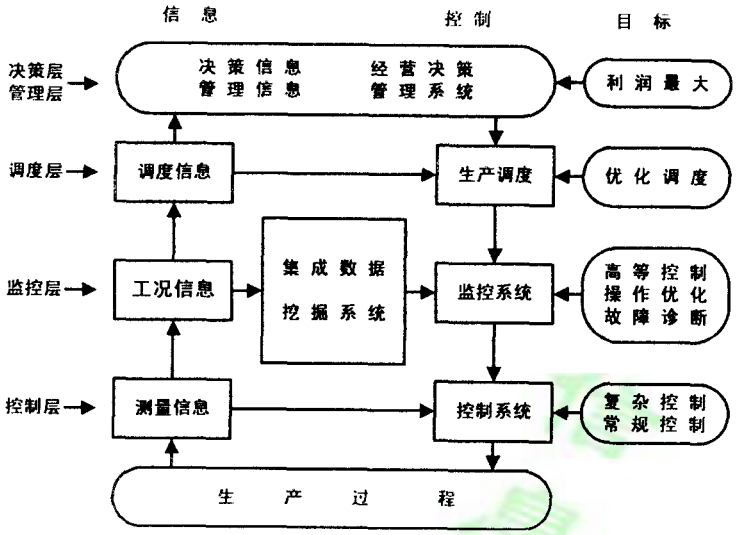

图 2　集成数据挖掘系统和典型的 CIPS 系统的融合结构

Fig. 2　Fusion structure of integrated data mining system & CIPS

以 DCS/FCS/CIPS 中采集积累的数据、人工记录日志及实验仿真数据为数据源,进行数据挖掘的前端处理,以解决数据的不完整性、噪声及重复、冗余问题等提高数据质量;之后通过集成的 DM 系统运用各种 DM 新算法进行知识、规则提取,状态辨识及模式的建立;并以领域操作、监督人员熟悉的方式可视化,对所获取的知识、规则及模式进行评估检验,确定其可信度;应用发掘的有用知识、规则及模式对工业系统实施监控、诊断或丰富知识库. 数据挖掘的过程并非一次就能成功或结束,而是一个不断的、反复的过程,从而逐渐获得有用的新知识.

7　进一步的研究工作(Future works)

数据挖掘应用于复杂工业过程系统中是一个崭新的领域,有许多问题有待进一步探索.

(1) 多智能体的递阶、分布式 DM 集成模型中,多智能体的协同机制问题.

(2) 以计算智能为核心,使得各种挖掘算法适合于工业连续过程的挖掘问题.

(3) 数据挖掘系统如何和 OLAP 技术及现有的监控系统相融合而协调工作,减少工作量.

(4) 适用于复杂工业过程监控的不同挖掘阶段的结果可视化问题.

(5) 解决原始数据间时间不一致问题和挖掘算法的可扩充性问题.

8　结束语(Conclusion)

数据挖掘技术是一个新的交叉学科领域. 本文以新的视角,在分析复杂工业过程中数据挖掘的基础、数据特点的基础上,讨论了工业过程中数据挖掘的思想,提出了集成的 DM 系统模型. 为进一步实现复杂工业过程的监控、优化与故障诊断开辟了新的途径.

参 考 文 献(References)

1　jiawei Han & Micheline Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. 2000

2　U M Fayyad, G piatetsky-shapiro, P Smyth, and Ruthurusamy, editors. Advanced in Knowledge Discovery and Data Mining. Cambridge, MA: AAAI/MIT Press, 1996

3　Xue ZWang. Data Mining and Knowledge discovery for process monitoring and control. London: springer, 1999

4　吴少敏,冯建生. 数据挖掘技术及其应用. 冶金自动化,2001,**6**:5～8

5　梅时春,李人厚,罗印升. 过程监控中数据挖掘与知识发现理论及应用. 微机算计信息,2001,**91**(2):1～3

作者简介

罗印升(1964—),男,陕西理工学院电气工程与自动化系副教授,硕士学位,现为西安交通大学系统工程研究所博士生. 研究领域为工业过程数据挖掘、智能控制理论与方法等.

李人厚 (1935—),男,西安交通大学系统工程研究所教授,博士生导师. 研究领域为智能控制理论与方法、知识发现、CSCW 理论与应用等.

梅时春 (1965—),男,二炮工程学院讲师,现为西安交通大学系统工程研究所博士生. 研究领域为数字图像水印、智能控制理论与方法.