

# 基于数据仓库的钢铁企业过程数据分析系统

赵相东<sup>1</sup>, 陆剑峰<sup>1</sup>, 张 浩<sup>1,2</sup>

(1. 同济大学 CIMS 研究中心, 上海 200092; 2. 上海电力学院电力与自动化工程学院, 上海 200090)

**摘 要:** 结合数据仓库 (DW) 和统计过程控制 (SPC) 相关理论, 依据钢铁企业生产过程的特点, 提出了适合钢铁企业的过程数据分析系统的通用架构和功能. 之后举例说明了过程数据分析中相关参数选择的过程. 详细介绍了利用改进的支持向量机方法进行性能参数预测以及采用粗糙集方法进行规则生成的过程. 实验证明, 生成的预测模型以及规则能够为企业决策提供支持, 满足了企业需求.

**关键词:** 数据仓库; 过程数据; K 近邻算法; 主成分分析; 支持向量机; 粗糙集

**中图分类号:** TP311

**文献标识码:** A

## Process Data Analysis System Based on Data Warehouse in Iron & Steel Enterprise

ZHAO Xiangdong<sup>1</sup>, LU Jianfeng<sup>1</sup>, ZHANG Hao<sup>1,2</sup>

(1. CIMS Research Center of Tongji University, Shanghai 200092, China;

2. School of Electric Power and Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China)

**Abstract:** The general framework and function of process data analysis system are introduced by integrating theories of data warehouse (DW) and statistical process control (SPC) with characteristics of the production process in iron & steel enterprise. And then an example of selecting relative parameters in process data analysis is given. Lastly the process of performance parameters prediction by improved support vector machine (SVM) and the process of generating rules set by rough set are explained in detail. Experiments show that the generated prediction model and rules can support enterprise's decision and meet enterprise's requirement.

**Keywords:** data warehouse; process data; K-nearest neighbor (KNN); principal component analysis (PCA); support vector machine (SVM); rough set

## 1 引言 (Introduction)

当前企业的规模越来越大, 生产工艺、生产设备和生产过程越来越复杂, 依据物理化学机理建立精确数学模型 (机理建模) 的方法已越来越困难. 基于数据的控制、决策分析系统成为当前信息系统的研究热点. SPC 和 DW 已在工业和服务行业得到广泛推广和应用. SPC 是指用统计方式来分析过程及输出, 通过适当的措施来达到并保持过程稳定, 从而实现保证产品质量的目的<sup>[1]</sup>. DW 从企业信息系统数据库中抽取并保存历史数据, 通过相关数据的整合和分析, 以支持企业决策<sup>[2]</sup>. DW 为质量持续改进提供了有力的工具. 它集成了企业质量标准数据, 生产过程数据、产品质量成本数据, 最大限度地利用企业的现有信息资源, 充分发挥了信息集成、数据处理效率高, 分析手段多样的优点. 统计过程控制分析一般只针对局部生产过程某些参数进行控制. 基于数据仓库的过程数据分析 (DWPA) 系统

结合了常用统计、人工智能和数据挖掘等方法, 可以从历史数据中发现某些参数的变化趋势, 并从大量相关因素分析中挖掘出某些参数不符合标准的原因, 深入揭示了变量之间的关系. 本文提出的 DWPA 系统结合 DW 及 SPC 相关理论, 已在国内多家钢铁企业实施<sup>[3]</sup>, 取得了显著的效益.

## 2 基于钢铁企业的 DWPA 系统的基本架构 (General framework of DWPA in iron & steel enterprise)

DWPA 系统以数据仓库为核心<sup>[4]</sup>. 按照数据仓库通用的设计思想, 分为 5 层结构, 即数据源层、操作数据层 (ODS)、主题数据层、数据分析层以及分析结论应用部分. 数据源层主要包括过程控制系统中的实时数据库, 此外还包括制造执行系统 (MES) 中的生产实绩数据、检化验数据, 企业资源计划 (ERP) 系统中的制造标准数据、成本数据、销售数据以及质量异议数据, 如图 1 所示.

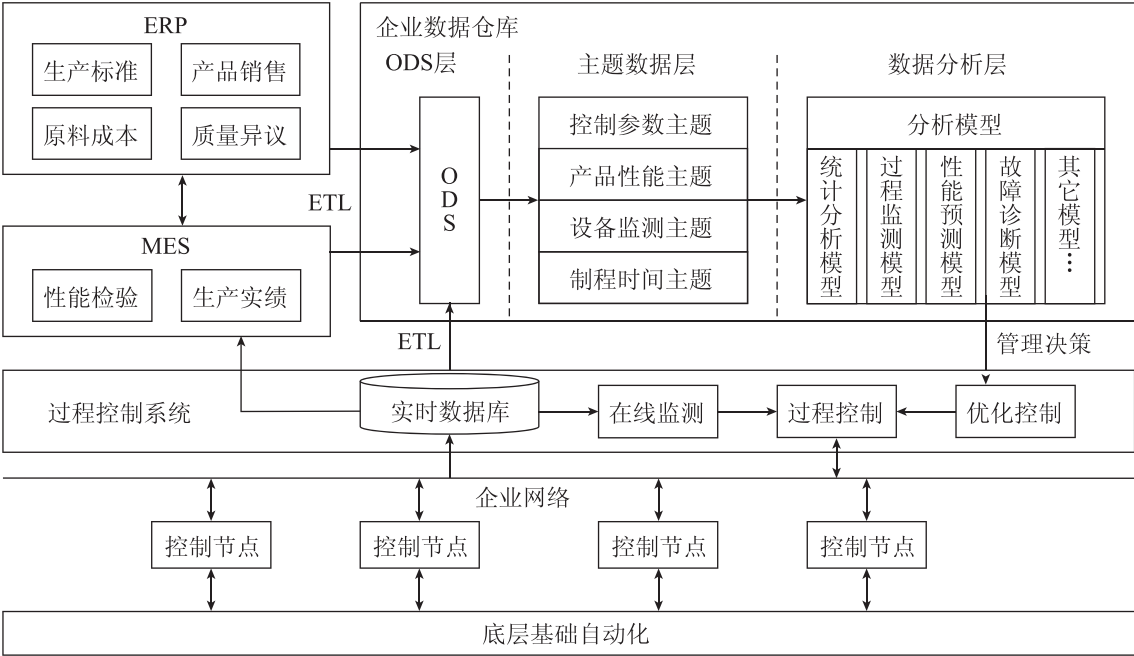


图 1 钢铁企业 DWPA 系统基本架构

Fig.1 General framework of DWPA in iron & steel enterprise

各数据源的数据经过抽取、转化、装载（ETL）过程以统一的数据格式存储在数据仓库的 ODS 层。ODS 层之后是主题数据层，主题数据层中包含了不同功能的主题数据集。主题数据集是按照特定的业务需求，将 ODS 中的相关的数据按照特定的键值组合在一起形成新的数据集用以支持进一步的分析。主题数据集构成整个数据仓库平台数据分析的数据基础。接下来是数据分析层，包括各种分析模型，利用主题数据层的数据，输出相应的知识，用于支持管理决策，优化生产控制。

3 DWPA 系统在钢铁企业的应用（Application of DWPA in iron & steel enterprise）

数据仓库可以保存并集成大量相关的生产历史数据，通过建立不同的分析模型，利用合适的查询和分析工具，可以实现针对生产过程全方位的预测、故障原因分析，辅助管理者进行控制优化，提升产品质量。下面以牌号 SS400 热轧带钢的力学性能分析为例，说明参数选择、预测建模以及判定规则生成的过程。

3.1 相关参数的选择

带钢的力学性能主要包括屈服强度、拉伸强度等，其决定因素主要包括 PCS 的控制参数，MES 中的生产实绩及钢卷自身的化学成分数据。企业希望通过历史数据的分析，找到各种控制因素与性能之间的影响关系，进行性能优化和预测。目前，力学性能以及相关过程控制历史数据已整合在主题数据集中，下面的工作首先是进行参数的筛选。

钢卷的相关数据包括生产班组、班次、板坯连铸机号等离散型数据以及钢卷的宽度、厚度、成分、过程温度等连续型数据。对于离散型变量的筛选可以采用方差分析，一般的，性能指标符合正态分布，假设某分类变量对其无影响，即在该分类变量下不同组的性能参数仍然符合原分布，则组间方差与组内方差的比值符合  $F$  分布，对应  $\alpha$  值如果大于 0.1 则原假设成立，如果小于 0.05 则原假设不成立。不失一般性，这里以钢种 SS400 的拉伸强度作为分析变量，选取一定时间段内的 2 099 条记录作为分析样本。选择生产班组作为分类变量，得到统计值如表 1。

表 1 生产班组单因素方差分析  
Tab.1 Production class of one-way ANOVA

	Sum of Squares	df	Mean Square	$F$	Sig.
Between Groups	102.840	3	34.280	0.071	0.976
Within Groups	1013916	2095	483.969		
Total	1014019	2098			

由表 1 可见, 得到的  $F$  值为 0.071, 对应  $\alpha = 0.967 \gg 0.1$ , 则可认为生产班组对拉伸强度无影响. 按此方法, 可知连铸机号等离散型变量对拉伸强度也无影响.

对于连续型变量, 利用 K 近邻 (KNN) 算法<sup>[5]</sup> 的“弱点”来解决. 已知, KNN 算法如果包括了与输出变量不相关的因素, 会导致预测精度急剧下降. 受此启发, 可以首先引入全部因素, 依次尝试删除 KNN 模型中的每个因素, 如果预测精度提高则彻底删除该因素, 精度降低则保留该因素. 经过一次所有因素的遍历, 保留下来的因素是下步所需要的. 具体算法如下:

采用 Shepard 提出的全局 KNN 算法, 首先选取全部  $n$  个参数,  $k$  个训练样本, 分析计算每个样本的预测值  $f_1(x_q)$ .  $f_1(x_q) = \sum_{i=1}^k \omega_i f(x_i) / \sum_{i=1}^k \omega_i$ , 其中

$\omega_i = \left( \sum_{r=1}^n (a_r(x_q) - a_r(x_i))^2 \right)^{-\frac{1}{2}}$  为距离平方的倒数计算预测总精度  $R_0^2 = \left( \sum_{q=1}^k (f(x_q) - \bar{f}(x))^2 - \sum_{q=1}^k (f_1(x_q) - f(x_q))^2 \right) / \sum_{q=1}^k (f(x_q) - \bar{f}(x))^2$ , 其中  $f(x_q)$  为原输出值,  $\bar{f}(x) = \frac{1}{k} \sum_{q=1}^k f(x_q)$  为原输出均值.

对于每个因素依次循环 FOR ( $i = 1, \dots, n$ ) 计算去掉因素  $i$  的精度差值  $\Delta_i = R_i^2 - R_0^2$ ,  $R_i^2$  为去掉该因素  $i$  后求得的预测精度,  $R_0^2$  为原精度. 如果  $\Delta_i > \varepsilon$ ,  $\varepsilon$  为已设定的正值, 则去除该因素  $i$ , 反之保留.

经过以上过程, 确定影响拉伸强度的 KIV (关键输入变量) 为钢卷重量、粗轧温度、C 元素值等 16 个变量, 如表 2 所示.

表 2 SS400 牌号钢卷的拉伸强度训练样本集  
Tab.2 Training sample set of tensile strength of JIS SS400

	D001	D002	D003	...	D2097	D2098	D2099
钢卷重量/kg	24460	11610	12350	...	19760	19390	24710
粗轧温度/℃	1056	1076	1098	...	1100	1078	1109
粗轧宽度/mm	1512	1497	1504	...	1054	1052	1503
抽出温度/℃	1218	1196	1203	...	1203	1209	1214
粗轧厚度/μm	453156	364393	354434	...	383960	384079	364223
精轧温度/℃	869	866	895	...	880	883	882
精轧宽度/mm	1508	1489	1504	...	1050	1047	1502
精轧厚度/μm	11502	2749	2499	...	2149	2150	2750
卷取温度/℃	627	658	656	...	657	657	653
凸度	50	51	61	...	41	33	28
C 元素值/%	0.15	0.168	0.147	...	0.158	0.158	0.151
Si 元素值/%	0.045	0.05	0.071	...	0.06	0.06	0.074
Mn 元素值/%	0.56	0.66	0.66	...	0.65	0.65	0.66
S 元素值/%	0.014	0.005	0.023	...	0.011	0.011	0.015
P 元素值/%	0.016	0.016	0.025	...	0.014	0.014	0.018
Al 元素值/%	0.024	0.02	0.02	...	0.033	0.033	0.035
拉伸强度/MPa	435	345	360	...	545	545	565

3.2 参数预测

把表 2 中 16 个 KIV 作为输入端, 拉伸强度作为输出端, 采用多元回归、BP 神经网络、支持向量机 (SVM), 以及经过降维预处理的支持向量机<sup>[6]</sup> (PSVM) 方法建立 KIV 与 KOV 之间的关系模型, 用于拉伸强度的预测, 采用以上 2099 条数据进行测试, 得到结果如表 3.

表 3 不同算法的测试结果  
Tab.3 Test results of different methods

	多元线性回归	BP 神经网络	SVM	PSVM
检验系数 $R^2$	0.564	0.771	0.813	0.910

经验证, 经过降维预处理的支持向量机效果最好, 具体算法描述如下:

Step1: PCA 降维处理:

- a: 计算样本点输入空间样本协方差矩阵  $\mathbf{S} = (s_{ij})_{n \times n}$  及样本相关矩阵  $\mathbf{R} = (r_{ij})_{n \times n}$ ;
- b: 为了消除量纲的影响, 对样本进行标准化, 求出标准化样本主成分;
- c: 选取前  $t$  ( $t < n$ ) 个样本主成分, 使其累计贡献率达到一定的要求, 一般取 80%~90%.

Step2: SVM 回归方法 [6]:

采用 Vapnik 提出的  $\varepsilon$ -SVR. 给定  $\{(\mathbf{x}_1, \mathbf{z}_1), \dots, (\mathbf{x}_l, \mathbf{z}_l)\}$ ,  $\mathbf{x}_i \in \mathbf{R}^n$  作为输入,  $\mathbf{z}_i \in \mathbf{R}^l$  作为目标输出.  $\varepsilon$ -SVR 的标准形式是:  $\min_{\omega, b, \xi, \xi^*} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^l \xi_i + C \sum_{i=1}^l \xi_i^*$ , 约束为  $\omega^T \phi(\mathbf{x}_i) + b - \mathbf{z}_i \leq \varepsilon + \xi_i$ ,  $\mathbf{z}_i - \omega^T \phi(\mathbf{x}_i) - b \leq \varepsilon + \xi_i^*$ ;  $\xi_i, \xi_i^* \geq 0, i = 1, \dots, l$ . 其中,  $\omega \in \mathbf{R}^n$ ,  $b \in \mathbf{R}$ ,  $\xi, \xi^*$  为松弛变量,  $C$  为指定的常数, 主要在提高泛化能力和减小误差之间起调控作用,  $\varepsilon$  为已设定的正数, 主要是用来控制算法希望达到的精度. 转换为对偶问题就是:

$$\min_{\alpha, \alpha^*} \left( \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l \mathbf{z}_i (\alpha_i - \alpha_i^*) \right)$$

$$\sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \quad 0 \leq \alpha_i, \alpha_i^* \leq C, \quad i = 1, \dots, l$$

这里  $Q_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi^T(\mathbf{x}_i) \phi(\mathbf{x}_j)$ .  $\alpha, \alpha^*$  为这个对偶问题的参数.

则判别函数为:  $\sum_{i=1}^l (-\alpha_i + \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b$

这里选取 Libsvm [7] 工具箱, 采用  $\varepsilon$ -SVR, 其中核函数采用径向基内积函数, 取  $\varepsilon = 0.01$ ,  $C = 10$ , 其它采用默认设置. 取 1499 条记录作为训练集, 其余 600 条记录作为测试集, 得到的  $R^2$  为 0.910.

3.3 误差分析

将拉伸强度的实际值、预测值进行比对得到训练误差, 如表 4.

平均绝对误差为  $MAE = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n} = 5.95$ , 平均

相对误差绝对值  $MAPE = \frac{\sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|}{n} = 0.012$ , 均等

系数  $EC = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i + \hat{y}_i)^2} = 0.992 > 0.9$ .

以上误差分析结果表明预测模型结果在企业控制要求的可接受范围内. 同时, 对误差进行正态分布检验, 得到检验值  $P$  为  $0.372 > 0.1$ , 可以认为服从正态分布. 图 2 是误差分布直方图.

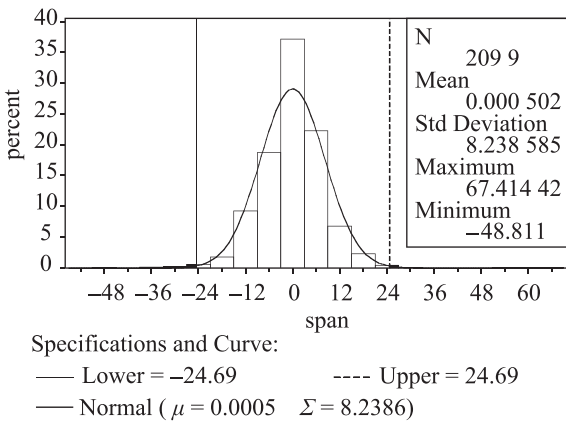


图 2 预测误差的直方分布图

Fig.2 Distribution histogram of prediction errors

由于 SVM 方法是对原值的无偏估计, 故误差的均值近似为 0, 得到标准方差  $\delta$  为 8.239.  $6\delta$  的置信区间概率为 99.74%,  $4\delta$  置信区间概率为 95.44%. 它的含义是当采用上述生成模型进行拉伸强度预测时, 真实值落在以预测值为中心的  $\pm 24.72$  MPa 区间内的概率为 99.74%, 落在  $\pm 16.48$  MPa 区间内的概率为 95.44%, 以上结果可满足企业需求, 达到了分析目的.

3.4 维度分析

维度分析主要针对离散型变量, 可以将连续型变量经过聚类分析划分为不同区间, 转为离散型变量. 以粗轧温度为例, 经过聚类得 1003、1040、1082、1124 四点为中心的区间. 按此方法将以上 16 个连续变量转化为离散型的区间变量, 作为拉伸强度的分析维度, 如粗轧温度维、精轧宽度维等, 生成数据立方, 利用 OLAP [8] 技术, 可以从不同的维度组合计算拉伸强度的最值、均值、方差、极差等常用统

表 4 训练样本的误差值

Tab.4 Errors of training sample set

	D001	D002	D003	D004	...	D2097	D2098	D2099
预测值 $\hat{y}$	434.12	346.14	355.61	528.09		556.44	547.26	564.82
实际值 $y$	435	345	360	530	...	545	545	565
误差 $e$	-0.88	1.14	-4.39	-1.91		11.44	2.28	-0.18

计值,并通过企业标准比对,利用公式

$$C_{pk} = \text{Min}[(USL - AVG)/3\delta, (AVG - LSL)/3\delta]$$

计算过程能力指数  $C_{pk}$ ,判断生产工艺过程的稳定程度,其中  $USL$  为标准上线,  $LSL$  为标准下线,  $AVG$  为分析参数的均值。

还可通过拉伸强度的实绩值与企业标准值进行比对,生成合格与不合格的二值决策变量。维度分析常用的算法有决策树、朴素贝叶斯分类、遗传算法、粗糙集<sup>[9]</sup>等。这里采用粗糙集算法,它有着严密的数学基础,能在保留关键信息的前提下对数据进行化简并求得知识的最小表达。通过建立分类变量与决策变量的决策信息表,进行属性约简,提取出分类变量和决策变量之间的一系列规则<sup>[10]</sup>。将拉伸强度是否符合标准作为决策变量,其余离散化后的维度作为属性变量,利用粗糙集得到如下规则:

当精轧厚度在 [6174.39,8653.87] 区间内、粗轧温度在 [1065.80,1121] 区间内、粗轧厚度在 [405204.74,434383.19] 区间内、精轧温度在 [860.34,895.12] 区间内、Si 元素值在 [0.041,0.069] 区间内、S 元素值在 [0.006,0.011] 区间内时,钢卷拉伸强度不符合标准;...

当精轧厚度在 [2298.4,3582.06] 区间内、粗轧温度在 [1065.80,1121] 区间内、粗轧厚度 [364288.1,377003.5] 区间内、精轧温度在 [860.34,895.12] 区间内、Si 元素值在 [0.041,0.069] 区间内、S 元素值在 [0.011,0.021] 区间内时,钢卷拉伸强度符合标准;...

应用上述方法,针对合格钢卷正确预测的准确率在 91.3%,生成的部分规则已经应用于生产,取得了不错的效果。

#### 4 总结 (Conclusion)

数据仓库已在国内外各大行业广泛应用。相比传统的 SPC 系统,基于数据仓库平台构建的过程数据分析系统,充分利用了平台上充足的数据和强大的数据处理和分析功能,可以全面整合相关数据,

深入挖掘数据之间的关系,同时方便数据分析结果的保存和共享。由于过程数据分析系统构建基于数据仓库的数据和软硬件设备,实施过程周期短、投资少,其成果可迅速被企业利用。在国内各大企业特别是钢铁企业的应用中取得了显著效果。目前,在数据仓库平台上构建专用决策和分析系统已成为企业信息化的发展方向。

#### 参考文献 (References)

- [1] 张公绪,孙静.现代质量控制与诊断工程[M].北京:经济科学出版社,1999.
- [2] Kimball R, Ross M. The data warehouse toolkit: The complete guide to dimensional modeling[M]. 2nd ed. New York, NY, USA: John Wiley & Sons, 2002.
- [3] 嵇晓,鲍玉斌.工业数据仓库设计方法及其在质量分析中的应用[J].控制与决策,2001,16(02): 229-232.
- [4] 汪永生,邵惠鹤. CIPS 中的数据仓库技术[J].化工自动化仪表,2000,27(1): 36-40.
- [5] Vapnik V N. The nature of statistical learning theory[M]. 3rd ed. New York, NY, USA: Springer, 2005.
- [6] 赵荣泳,张浩,张辉,等.一种新的机器学习方法——PSVM 应用于数控磨床智能诊断的研究[J].制造业自动化,2005,27(1): 39-41.
- [7] Chang C C, Lin C J. LIBSVM – A library for support vector machines[EB/OL]. [2002-12-04]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [8] Giovinazzo W A. Object-oriented datawarehouse design: Building a star schema [M]// Upper Saddle River, NJ, USA: Prentice-Hall, 2000.
- [9] Witten I H, Frank E. Data mining: Practical machine learning tools and techniques[M]. 2nd ed. San Francisco, CA, USA: Morgan Kaufmann, 2006.
- [10] 李翠玲,张浩,赵荣泳.粗糙集理论及其在机电行业中的应用潜力分析[J].机电一体化,2005,11(6): 24-26.

#### 作者简介:

赵相东(1979-),男,博士生.研究领域为钢铁企业信息  
系统构建及相关数据分析.

陆剑峰(1973-),男,副教授.研究领域为智能生产系统.

张浩(1962-),男,研究员,博士生导师.研究领域为智  
能生产系统.

# 基于数据仓库的钢铁企业过程数据分析系统

作者: 赵相东, 陆剑峰, 张浩, [ZHAO Xiangdong](#), [LU Jianfeng](#), [ZHANG Hao](#)  
作者单位: 赵相东, 陆剑峰, [ZHAO Xiangdong](#), [LU Jianfeng](#) (同济大学CIMS研究中心, 上海, 200092), [张浩, ZHANG Hao](#) (同济大学CIMS研究中心, 上海, 200092; 上海电力学院电力与自动化工程学院, 上海, 200090)  
刊名: [信息与控制](#) [ISTIC PKU](#)  
英文刊名: [INFORMATION AND CONTROL](#)  
年, 卷(期): 2010, 39(2)  
被引用次数: 0次

## 参考文献(10条)

1. 张公绪. 孙静. 现代质量控制与诊断工程 1999
2. [Kimball R. Ross M The data warehouse toolkit: The complete guide to dimensional modeling](#) 2002
3. 嵇晓. 鲍玉斌. [工业数据仓库设计方法及其在质量分析中的应用](#) 2001(2)
4. [汪永生. 邵惠鹤 CIPS中的数据仓库技术](#) 2000(1)
5. [Vapnik V N The nature of statistical learning theory](#) 2005
6. [赵荣泳. 张浩. 张辉 一种新的机器学习方法-PSVM 应用于数控磨床智能诊断的研究](#) 2005(1)
7. [Chang C C. Lin C J LIBSVM-A library for support vector machines](#) 2002
8. [Giovinazzo W A Object-oriented datawarehouse design: Building a star schema](#) 2000
9. [Witten I H. Frank E Data mining: Practical machine learning tools and techniques](#) 2006
10. [李翠玲. 张浩. 赵荣泳 粗糙集理论及其在机电行业中的应用潜力分析](#) 2005(6)

## 相似文献(10条)

1. 学位论文 [刘伟 基于过程数据的企业集成关键技术研究与应用](#) 2008

流程企业综合自动化是实时生产管理集成优化的核心, 而综合自动化的基础是企业生产过程数据的有效集成。企业生产过程数据主要包括生产运行与管理涉及到的实时和历史数据、事件、消息等。与传统企业集成不同的是过程数据具有不同的时间周期、不同的概念外延、生产工艺知识约束以及实时性要求。此外, 企业集成环境的复杂性, 如传感器数据高噪音、异步采样等也增加了过程数据集成的难度。<br>

论文以流程企业过程数据集成背景, 研究模型驱动的过程数据集成技术, 重点研究过程数据集成模型、模型驱动的过程数据集成集成框架、QoS自适应的实时发布/订阅机制以及基于反馈的多传感器数据融合算法, 力图研发一个支持不同尺度的过程数据集成的工具。<br>

论文首先分析流程企业生产过程数据集成特点和传统企业数据集成建模方法的不足, 提出了一种基于模型驱动的企业过程数据集成方法, 采用领域本体的方法从时态对象、集成过程、语义集成三个角度建立过程数据集成模型, 并对这三个模型进行了形式化描述, 此外, 通过定义映射规则, 实现应用本体间关系的映射。<br>

在分析分布式事件通知服务体系结构的基础上, 提出了基于发布/订阅的过程数据集成框架, 通过采用事件—条件—动作(ECA规则)来支持模型驱动的企业过程数据集成方法。为了简化过程数据集成模型的建模, 提出了一种可视化的ECA规则描述规范, 并设计开发了相应的编译工具。<br>

针对传统的分布式事件处理不能满足流程企业中过程数据集成的实时性要求和QoS保障的问题, 论文在传统分布式事件处理之上扩展设计了QoS保障策略和带截止期的ECA规则(RECA), 提出了一种自适应发布/订阅的机制, 该机制通过动态调整系统参数, 可以同时提供多层次的服务质量。实验数据验证了该机制能够提高多服务请求并发情况下, 不同QoS等级的响应处理和可预测性。<br>

进一步, 针对企业过程数据集成中面临的传感器采集高噪音、异步采样等问题, 论文给出了多传感器数据融合的一种数学描述, 在比较分析典型的多传感器数据融合的算法的基础上, 提出了一种基于反馈控制原理的多传感器数据融合算法, 并给出了算法实现及实验数据验证。<br>

最后, 基于上述研究成果, 论文设计并开发实现了一个适应大规模分布式流程企业生产过程数据集成的自适应实时发布/订阅服务系统, 并作为流程企业生产执行系统(SMES)的核心构件, 在多家石化企业得到了成功应用。

2. 学位论文 [张波 企业级过程数据库](#) 2006

过程数据库(Process DataBase)是对项目数据进行定义、收集和分析的一种工具, 基于历史数据来进行估算, 制定评价计划, 确定项目可行性, 找出项目改进的潜在区域等。对企业来讲, 数据是企业决策的核心依据, 而对数据详尽的分析, 则是企业生存和发展的生命力所在。研发过程数据是公司的重要知识资产。建立研发度量数据库系统、收集和分析过程数据有利于建立研发过程能力, 以确定研发能力水准; 支持进行估算, 以制订和评价项目计划; 及早识别问题, 以调整决策。

本文结合一个大型企业改进研发过程的实际案例, 对国内外大型企业过程数据库情况进行了分析和研究, 提出了企业级过程数据库的解决方案, 详细描述了如何在企业中建立一个过程数据库来改进企业的研发过程能力。

论文首先进行了一定的理论研究, 阐述了过程数据库和能力成熟度模型的关系, 度量的概念和对企业和过程数据库的意义。接着论文通过数据库解决方案引出了过程数据库的解决方案, 并对过程数据库进行了概要设计, 包括系统设计的思路, 原则, 解决方案和具体的实施步骤和过程模型。然后论文详细描述了如果通过度量来收集数据和建立中心数据库和选择产品, 最后在中心数据库的基础上开发了基于Web的用于管理和客户使用的过程数据库管理及客户系统。

3. 学位论文 [关强 基于Web的管理与过程数据集成及演示系统的实现](#) 2002

该文依据从开发实践中总结出来的流程企业综合自动化系统体系架构, 从三方面, 也可以说三个层次, 讨论了在实现管理与过程的数据集成时必须面对的问题, 并提出了一定的解决方法并加以实践。1、过程数据获取中的系统异构。OPC技术是解决这一问题的最佳方案, 该文对OPC机制进行了深入的分析并实现了基本的OPC数采演示系统, 而且将OPC技术运用到工程实践中, 开发了以OPC技术为基础的先进过程控制教学实验系统。2、管理数据与过程数据基于Web的集中访问。该文通过对OLEDB技术的研究, 实现了以OLEDB技术为基础的数据访问组件, 对数据访问过程进行了面向对象的封装, 实现了多数据源的集成, 同时数据访问组件与ASP技术互为补充, 构成了三层结构的中间层。3、管理数据与过程数据的综合。通过前面两部分的工作, 可以实现管理与过程数据初步的集成, 而进一步实现管理与过程数据的高度综合, 对企业具有更深远的意义。近年来迅猛发展的数据仓库系统为实现这一目的带来了新的思想和方法。该文通过实现一个简化的数据仓库系统, 体现了数据仓库系统解决方案的基本思想—面向主题与集成。

4. 期刊论文 [何海江. 魏善沛. 肖杰. HE Hai-jiang. WEI Shan-pei. XIAO Jie 基于组件的工业过程数据仓库 -计算机工程与设计](#) 2007, 28(8)

工业企业自动化系统所产生的过程数据, 分散在异构系统上, 不能统一存储、调用和管理, 造成数据资源浪费。基于组件和分布式应用技术, 设计和实现了工厂级的工业过程数据仓库, 包括: 集成各类自动化软件的过程数据, 数据压缩存储, 可视化分析工具开发。通过该系统, 可集成和开发工业过程历史数据资源, 从中挖掘对生产有用的知识。

5. 学位论文 [杨涛 基于数据仓库的业务过程智能核心组件的构建](#) 2007

本文对基于数据仓库的业务过程智能核心组件的构建进行了研究。文章对该系统中业务过程和绩效指标的数据仓库表示及存储结构进行了分析和设计, 提出了较为完整的业务过程管理系统的体系结构, 对该体系结构进行了层次分解并逐层介绍了各个组成部分的功能作用。将数据仓库技术应用于业务过程管理系统, 可以帮助企业对已有过程数据进行分析、监测、预测、控制和优化。通过某公司的案例分析, 对提出的业务过程挖掘过程进行了验证, 对该公司的技术支持部门绩效水平做了合理的评估, 分析了问题发生的原因, 并给出了具体的建议。

6. 学位论文 [张立权 基于模糊推理系统的工业过程数据挖掘](#) 2006

工业过程数据收集和存储技术的迅猛发展导致快速增长的大量数据, 这些数据存储在数据库、数据仓库或其它种类的数据存储介质中。从海量的数据中挖掘隐藏的、有用的信息, 能够为工业过程的在线监测、故障诊断、模型辨识、控制策略设计和预测等提供强有力的决策支持。

工业过程数据挖掘应用的主要任务是选择和建立有效的、适合工业过程数据特征的挖掘方法。基于模糊推理系统的数据挖掘方法能够使用同一模型结构(模糊IF-THEN规则)分别



执行描述式和预测式数据挖掘任务,提取的规则模式易于操作人员的理解和管理者的决策支持。对于复杂的工业过程,它能够以一种自然的方式评价输入变量的重要性,以选择最相关的变量描述系统的动态行为。定义隶属度函数的灵活性有助于在不同的粒度空间上寻找系统的操作模型,挖掘工业过程变量之间内在的关系和规律,有效地解决工业过程的实际问题。论文主要的研究工作如下:

(1) 针对传统的、基于梯度的模糊推理系统学习方法中存在的收敛速度和振荡之间的冲突问题,以及动量项学习方法中动量项因子的选择难题,提出一种改进的、基于梯度的、用于模糊推理系统参数优化的实时学习算法(G-RTL)。通过引入与均方误差相关的动态误差传递因子,使得在相同学习率系数的前提下,与传统的、基于梯度的学习方法和动量项学习方法相比,在处理大规模样本集时具有较高的收敛速度和精度,并且学习过程是稳定的,非常适用于工业过程的在线学习。通过经典的倒车控制问题和与经典。BP网络逼近性能比较的仿真结果表明本方法是有效的。

(2) 针对工业过程系统高维数据、非线性特点,提出一种基于归一化方差信息的自适应模糊规则挖掘方法(NV-AMFR)。基于数据挖掘技术和Mamdani模糊模型,从一个简单的初始结构出发,使用G-RTL学习算法优化模糊推理系统的参数向量,利用优化后获得的模糊规则的置信度度量及隶属度函数归一化方差信息,确定输入空间中模糊规则的密度需要加强的区域,以及用于划分论域上模糊子集数目的需要增加的输入变量,从而可以在不同的粒度空间上有效地挖掘过程变量之间内在的联系和规律,而且还能够以一种易理解的方式评价输入变量对系统输出的影响程度,以选择最相关的变量描述系统的动态行为,给出了一个新的、更合适的模型结构。非线性函数逼近的数值例子仿真验证了本方法的有效性。

(3) 针对工业过程数据库中普遍存在的不一致性、不完整性和历史性,提出一种推理空缺模糊规则的最邻近扩散方法(ND-EMR)。基于模糊推理系统和样本数据分布的先验知识,使用改进的G-RTL实时学习算法,通过确定最优输出模糊子集的质心和模糊规则的置信度度量,推理样本数据未覆盖区域上的空缺模糊规则,并构造一个完备的模糊规则集,从而有效地解决了样本数据未覆盖的区域上系统的不可预测问题。结合混沌时间序列预测问题以及WM方法比较的仿真结果,表明本方法不仅有效而且可以适于不能预测的情况。

(4) 基于模糊T-S预测模型,结合多种数据挖掘技术,提出一个基于数据挖掘的复杂工业过程智能控制新方法。使用G-RTL实时学习算法,快速而准确地辨识模糊T-S预测模型;基于所辨识的模糊T-S预测模型,对于批过程的受限非线性最优控制,运用平行分布补偿算法和最小值原理,把一个复杂的非线性系统最优控制设计问题转化为局部线性子系统的最优控制问题,从而给出一种有效和简单的最优模糊控制方法。将所提出的方法结合一个半连续反应器的建模和最优控制进行了仿真研究,结果表明本方法具有较高的建模精度,并且能够获得更高的主产品产量。

7. 期刊论文 [何小东. 何海江. He. Xiaodong. He. Haijiang 基于.NET的工业过程可视化数据仓库的研究 -微计算机信息](#)2006, 22 (18)  
研究在.NET框架下如何设计和构造厂级工业过程数据仓库,包括:集成各类自动化软件的过程数据,数据压缩,提供可视化分析工具。

8. 学位论文 [郑刚 数据质量分析系统中元数据管理的研究与实现](#) 2007

数据仓库作为支持决策制定过程的重要手段,近几年来得到了迅速发展,并已经成功应用到制造业、零售业、金融服务、电信、运输等多个行业。随着数据仓库的深入应用,数据质量问题成为关系到数据仓库建设成败和数据能否有效应用的重要关键问题。由于数据仓库主要由各种数据组成,这些数据的质量直接影响着数据仓库的质量,而数据仓库质量严重影响着数据仓库使用者的信心,因此,由数据质量引起的问题越来越受到政府机构、企业、个人的关注。目前,国外已经把数据质量技术作为一个独立领域来研究,并且成立了一些相关机构和组织,定期开展活动和会议来发展数据质量技术。

基于数据仓库的数据质量分析系统就是为了研究和改善数据质量问题而开发的,元数据管理在其中起着重要作用。元数据是关于数据的数据,通过它可以准确和完整地表述数据与业务之间的关系,使用户了解数据仓库中数据的来龙去脉,这为了解数据的质量状况和数据之间的关系、有效分析数据仓库中数据质量奠定了基础。而在现实的许多数据仓库项目中,元数据管理没有得到应有的重视,很多企业甚至都没有建立相应的数据管理体制。这种情况导致在数据仓库建设中,元数据分散在系统中的各个组成部分,缺乏统一和集中管理的基础,无法形成独立的层次。

针对以上问题,本文对数据质量分析系统中的元数据管理进行了研究,主要研究内容与创新点如下:

1. 引入了CWM模型并进行了扩展。根据项目特点,在研究了几种通用的元数据管理策略后,选取了公共仓库元模型(Common Warehouse Metamodel, CWM),并根据系统的需求,通过定义一系列的新类和类之间的关联,采用面向对象技术中的继承机制对CWM模型进行了扩展。
2. 对各种元数据进行集中管理,并建立了统一的元数据存储库。对系统中的ETL(Extract, Transform, Load)过程、数据质量分析过程、数据资源等元数据进行了集中统一管理,并基于CWM标准,运用“对象关系映射”技术,将类、对象以及属性映射存储到数据库中,建立了统一的元数据存储库,方便了元数据的管理。
3. 采用了灵活的实现技术,方便系统扩展。在系统实现中,采用Web Service和XML等技术实现不同平台之间元数据的共享与交换,不但简化了系统的维护,而且方便以后对系统进行扩展。

9. 学位论文 [苑野 基于数据仓库的顾客服务模型的设计与实现](#) 1999

当今社会正迈入知识经济时代,随着信息技术的高速发展,信息资源对于企业的经济价值和社会价值所起的作用越来越明显,而数据仓库技术的出现将使企业获得商业智能变得更为容易,从而使企业在市场竞争中具有竞争优势。结合知识经济时代顾客服务的新特征,即强调顾客占有率和行销方式从一对多方式向一对一方式的转变,该文提出以顾客服务为目标决策分析的基本内容空和采用的基本方法与技术,并提出基于数据仓库技术的顾客服务模型,阐述了面向顾客服务数据仓库的具体实现过程,同时介绍了当前较为流行的两种数据仓库建模方法,对这两种建模方法的理论基础、设计过程和适用范围做了较为详细的说明与比较。该文以一实际的数据仓库案例为基础,通过对当前较为流行的数据仓库产品的广泛调查与研究,确定较为合适的数据仓库产品来实现面向顾客服务的企业数据仓库,结合案例的特点提出了具体的数据模型建模方法,交实际解决了建立数据仓库中数据抽取过程数据更新问题。

10. 期刊论文 [刘仑. LIU Lun 基于项目过程管理的数据仓库分析与建立 -技术与创新管理](#)2006, 27 (3)

随着科学技术的飞速发展,科技评估已成为现代科技管理的必要手段和决策的重要依据,在现代科技项目管理中发挥越来越重要的作用。通过分析反映项目执行过程实时运行数据和历史数据的数据模型和数据存储,将数据仓库理论和数据挖掘技术应用于科技项目评估,使隐含在项目计划执行过程数据中的一些特性可以被及时发现,从而对项目计划顺利执行和达到预期目标提供强有力的支持。

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_xxykz201002020.aspx](http://d.g.wanfangdata.com.cn/Periodical_xxykz201002020.aspx)

授权使用: 复旦大学图书馆(fddx1wxsjc), 授权号: 5278c7a2-cd43-4831-a9ff-9e6a00e4ac6b

下载时间: 2011年1月13日