摘　要

支持向量机方法和数据挖掘领域是现在国内外学术界的研究热点。数据挖掘在许多商业应用中都取得了十分理想的效果，但是在流程工业生产过程中，应用数据挖掘成功的例子还不多见。本论文在经典的数据挖掘算法中结合了支持向量机方法，并针对每个算法给出了用于实际工业控制项目的例子，讨论了基于支持向量机的数据挖掘方法在工业应用中的利弊。论文的主要内容包括以下三个方面：

首先，论文描述了基于核主元分析结合支持向量机的工业建模预测。复合肥生产工艺过程比较复杂，采用传统方法对复合肥养分含量建模难以达到理想效果。在直接使用支持向量回归建模时，数据预处理、核函数参数选择是两个难点。论文提出一种 KPCA-SVR 方法，结合了两种核方法的优点，又提出了一种核参数选择的改进算法，并融合到 KPCA-SVR 方法中，通过对实际工业数据的仿真研究，结果表明该方法取得了很好的效果。

其次，论文重点描述了基于支持向量机的关联规则提取。文中提出了一种基于支持向量机的关联规则提取方法，通过支持向量聚类、数据域描述等方法来归类样本数据，利用得到的支持向量来提取规则。该方法充分发挥了支持向量机处理小样本非线性能力强、泛化性能好的优势，并克服了其分类函数可理解性差的缺点，同时把经典 SMO 算法的思想引入来提高关联规则提取的执行效率。在标准数据集和实际数据的仿真中取得了较好的效果，为关联规则提取提供了一个新的思路。

最后，在文中详细介绍了一个具体的工业数据挖掘的应用实例——株洲冶炼集团硫酸厂的数据挖掘软件。该软件将关联规则、分类、聚类、回归等数据挖掘算法应用到铅烧结烟气 WSA 制酸过程中。另外，该软件还包括三维图像显示、在线指导等其他功能。通过使用数据挖掘软件，可以得到更多的铅烧结烟气 WSA 制酸过程中的信息，使变量的关系更加清晰，也方便了工作人员的操作。

关键字：工业数据挖掘，支持向量机，关联规则，聚类，主元分析，数据挖掘软件

Abstract

Support vector machines and data mining are two popular research topics studied by domestic and overseas scholars nowadays. Data mining has achieved lots of accomplishment in business world and other fields. However, process data mining combined with SVM has few successful examples in industry. In this paper, some data mining algorithms combined with SVM are proposed and applied in industry. There are some main works in this paper as rendered below.

First of all, a method which combines with SVM and KPCA is proposed and applied in the compound fertilizer production which is so complex that the traditional methods could not get a good result on modeling of fertilizer's component. Using support vector regression directly, data preprocess and kernel parameters selections are hard problems. In this paper, a KPCA-SVR method is proposed. The kernel principal component analysis is used to process data in order to get nonlinear principal component and get noise data off. Then an improved parameters selection method is introduced to predict the final component of the compound fertilizer. Simulation results based on practical industrial data show the effectiveness of the proposed modeling approach.

Secondly, a new way of association rules extraction based on SVM is proposed in this paper. The SVC and data description is used to analyze the sample data, and the obtained support vectors are used to get the association rules in this method. It takes advantage of the abilities of SVM which can deal with limited samples and nonlinear data and have a good generalization performance. At the same time, it overcomes the unintelligible problems of SVM's classifiable function. And the program efficiency is improved by introducing the classic SMO algorithm. Simulations based on industrial data have been done and the results showed great effectiveness of this proposed modeling approach which could provide a novel idea to get the association rules.

Thirdly, industrial process data mining software which is used in a WSA process and developed by myself is introduced. It includes several main functions, such as

association rules, classification, clustering and regression. Other functions like three dimension display and online direction are also involved in the software. The data mining software makes it easy to control the process.

Key words: Process data mining, Support vector machines, Association rules, clustering, Principle component analysis, Process data mining software

第一章　绪论

1.1　工业数据挖掘

1.1.1　数据挖掘概念

数据挖掘(Data Mining)是从大量的、不完全的、有噪声的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[1]。随着信息技术的高速发展，人们积累的数据量急剧增长，动辄以 TB 计，如何从海量的数据中提取有用的知识成为当务之急。数据挖掘就是为顺应这种需要应运而生发展起来的数据处理技术，是知识发现(Knowledge Discovery in Database)的关键步骤。

数据挖掘的任务主要是关联分析、聚类分析、分类、预测、时序模式和偏差分析等。

(1)关联分析(Association Analysis)

关联规则挖掘是由 Agrawal 等人首先提出的。两个或两个以上变量的取值之间存在某种规律性，就称为关联。数据关联是数据库中存在的一类重要的、可被发现的知识。关联分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网。一般用支持度和可信度两个阈值来度量关联规则的相关性，还不断引入兴趣度、相关性等参数，使得所挖掘的规则更符合需求。

(2)聚类分析(Clustering Analysis)

聚类是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。

(3)分类(Classification)

分类就是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，并用这种描述来构造模型，一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得分类规则。分类可用于规则描述和预测。

(4)预测(Predication)

预测是利用历史数据找出变化规律建立模型，并由此模型对未来数据的种类

及特征进行预测。预测关心的是精度和不确定性，通常用预测方差来度量。

(5)时序模式(Time-series Pattern)

时序模式是指通过时间序列搜索出的重复发生概率较高的模式。与回归一样，它也是用已知的数据预测未来的值，但这些数据的区别是变量所处的时间不同。

(6)偏差分析(Deviation Analysis)

在偏差中包括很多有用的知识，数据库中的数据存在很多异常情况，发现数据库中数据存在的异常情况是非常重要的。偏差检验的基本方法就是寻找观察结果与参照之间的差别。

数据挖掘对象

根据信息存储格式，用于挖掘的对象有关系数据库、面向对象数据库、数据仓库、文本数据源、多媒体数据库、空间数据库、时态数据库、异质数据库以及internet 等。

数据挖掘流程

(1)定义问题：清晰地定义出业务问题，确定数据挖掘的目的。

(2)数据准备：数据准备包括：选择数据——在大型数据库和数据仓库中提取数据挖掘的目标数据集；数据预处理——进行数据再加工，包括检查数据的完整性及数据的一致性、去噪声、填补丢失的域、删除无效数据等。

(3)数据挖掘：根据数据功能的类型和数据的特点选择相应的算法，在净化和转换过的数据集上进行数据挖掘。

(4)结果分析：对数据挖掘的结果进行解释和评价，转换成为能够最终被用户理解的知识。

(5)知识的运用：将分析所得到的知识集成到业务信息系统的组织结构中去。

数据挖掘技术是一个年轻且充满希望的研究领域，商业利益的强大驱动力将会促使它不停地的发展。每年都有新的数据挖掘方法和模型问世，人们对它的研究正日益广泛和深入。尽管如此，数据挖掘技术仍然面临着许多问题和挑战，如：数据挖掘方法的效率亟待提高，尤其是超大规模数据集中数据挖掘的效率；开发适应多数据类型、容噪的挖掘方法，以解决异质数据集的数据挖掘问题；动态数据和知识的数据挖掘；网络与分布式环境下的数据挖掘等；另外，近年来多媒体

数据库发展很快，面向多媒体数据库的挖掘技术和软件今后将成为研究开发的热点。下面给出一些常用的数据挖掘方法的分类示意图[2]。

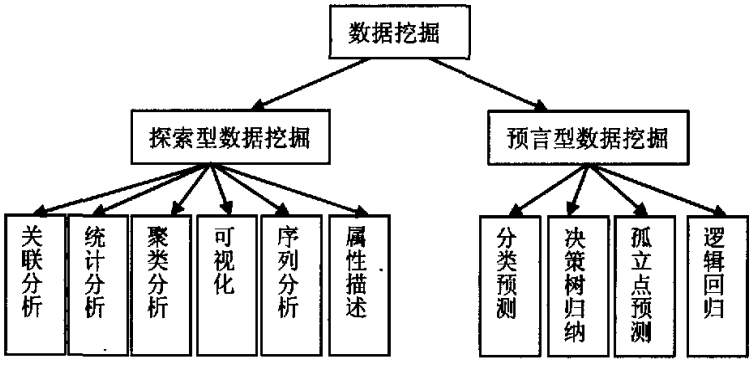


图 1-1 数据挖掘分类结构图

1.1.2 工业数据挖掘

数据挖掘在商业、生物、制药等许多领域已经取得了重大成果，其中一些思路和方法为其向工业控制系统的应用扩展提供有益的借鉴。国内流程工业控制行业已经有了一些数据挖掘的应用例子。

SAS Enterprise Miner 是一种通用的数据挖掘工具(在第四章中将详细介绍)，在我国企业中比较典型的应用有：上海宝钢配矿系统应用和铁路部门在春运客运研究中的应用；由中控软件公司承担的浙江省重大科技招标项目“流程工业数据挖掘技术及软件在三唑磷合成过程中应用研究”项目利用具有自主知识产权的 DCS 系统、实时数据库和数据挖掘软件为企业提供综合自动化整体解决方案，取得了实质性的突破。该项目采用浙大中控的 JX-300 DCS、ESP-iSYS 实时数据库和 ESP-Miner 数据挖掘软件已在浙江新农化工有限公司成功投运，提高了三唑磷收率和含量，在数据挖掘技术应用于农药化工领域处于国内领先水平。

但是，在过程工业中，数据挖掘的应用却有很大的限制和障碍，现有的一些应用也没有取得在商业、金融、保险、生物学等领域中那样好的效果。这是由于过程工业的以下四个特点数据特点所决定的，

(1) 数据量大，数据关联性强；

工业数据挖掘所需要处理的数据量巨大，实际得到的数据样本维数众多。以冶炼工业为例，WSA 数据由数据采集系统每隔几秒就在实时数据库中产生一组数据，每组数据中包含上万个记录，如何从这些海量数据中得到有用的知识，对于常用的数据挖掘算法的性能是一个挑战，也正是现有的数据挖掘算法的一个难点。而且与商业上事务处理数据库中的数据不同，过程工业数据的属性之间往往存在复杂的非线性关系以及相互耦合的现象。

(2) 数据时间性强，随着时间波动且滞后严重；

在过程工业的连续生产中，由于工况的改变、操作控制的调整和原料的改变，数据时刻在变化，这增加了从中提取知识的难度，而且过程工业数据的质量指标和操作参数之间存在较大的时间滞后。在很多情况下，由于装置是连续生产的，因此很难得到操作参数与质量指标之间精确对应的数据记录；

(3) 数据的不完整性严重；

与商业数据不同，过程工业数据中往往存在较大噪声或者孤立点。由于原料改变、工艺改变、人为因素、生产装置故障以及测量仪表问题，过程工业数据中不可避免地存在孤立点，进一步增加了数据挖掘的难度，使得可用于分析的高质量数据少。在过程工业的数据挖掘中，通常需要对一些质量指标如产量、能耗、收率、纯度、杂质含量、环境影响等进行评价，然而这些指标中，有些往往是无法直接测量的如产率、收率等，有些虽然可以通过实验室分析化验得到，如纯度、杂质含量等，但这些指标采样周期长，通常只有一天两到三次，造成数据质量不高。

(4) 工业模型复杂，软件技术支持不成熟。

现有的数据挖掘软件大多以商业过程中事务处理为背景，而这些事务本身相对比较简单，目的也比较明确。例如在超市购物分析中，我们想知道同时购买啤酒和尿布的人具有什么样的特征，只需要将这些数据检索出来分析就可以了。但过程工业数据挖掘则复杂得多，例如在当前操作参数下，产品的某些质量指标比较差，想通过数据挖掘来找出原因。对于这样的问题，首先需要对工艺有一个比较全面的了解，需要了解影响质量指标的参数有哪些，而且当影响质量指标的操作参数过多时，需要对参数进行筛选；其次需要考虑这些参数之间的耦合关系；最后还要考虑时间的滞后。由于工业过程的特点，所需要的参数往往不容易

找到，在这种情况下我们可以采用一些其他的方法如建模，来得到需要的参数，但这样会使得最后挖掘出来的知识置信度较低。因此，这些因素决定了过程工业数据挖掘的难度。

为了解决这些困难，我们思考了如下一些解决的方法：

采用支持向量机方法对小样本问题进行建模，可以发挥它处理小样本能力强的优势，建立更加准确的模型，该方法也是论文的研究重点之一，将在第二章和第三章中详细阐述；通过与软测量结合的方法来得到一些不易测量的数据点；在工业数据预处理的过程中，更加注重与专家的交流；采用 SOM 数据辨识方法来分析数据；用特征选择和特征提取方法来降低维数避免耦合；通过经典的滤波方法、小波变换以及聚类等方法来去除孤立点等等。

总之，工业数据无论数量还是维度都比较复杂，而且很多与目的变量相关的重要数据不易取得，因此在数据预处理阶段需要做的工作远远多于商用数据的预处理，这也是工业数据挖掘的难点之一。此外，数据结果的呈现也有别于商用数据挖掘软件，应该根据工业生产的自身特点开发出有工业数据挖掘特点的软件，这将在第四章中详细介绍。

数据挖掘用于商业领域只有短短十几年的时间，但已经发挥出了无与伦比的巨大作用。我们可以预见：在不远的将来，它必将给流程工业领域带来一次全新的变革。

1.2 支持向量机

2.1.1 支持向量机概念

支持向量机最初是由Vapnik从两类分类问题中求最优分类面的问题中提出来的[3]，问题描述如下：给定 l 个训练样本 $(\mathbf{x}_1, y_1),...(\mathbf{x}_i, y_i),...,(\mathbf{x}_l, y_l)$，$\mathbf{x}_i \in \mathbf{R}^n, i = 1,...,l$ 和类标号 $y_i \in \{-1,1\}$ (代表分类问题中的两类样本)，在线性可分的情况下，最优分类面问题即最大化两类样本之间的距离，也就是所谓的最大化分类间隔，图1-2 中给出了在两维空间中求取最优分类面的示意图。

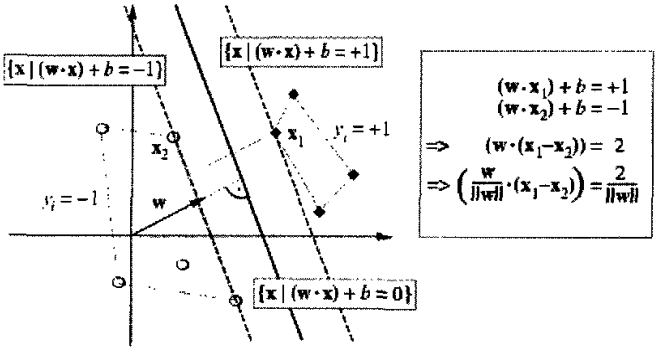


图 1-2 最优分类面问题图

从图上可以看出，两类样本之间分类间隔 $\rho(\mathbf{w},b)$ 为，

$$\rho(\mathbf{w},b)=2/\|\mathbf{w}\| \tag{1-1}$$

因此在线性可分的情况下，求取最优分类面的问题可以归结为如下的一个二次规划的问题，

$$\begin{aligned}\min\quad & \frac{1}{2}\|\mathbf{w}\|^2\\ \text{s.t.}\quad & y_i(\mathbf{w}\bullet\mathbf{x}_i)+b-1>0, i=1,2,...,l.\end{aligned} \tag{1-2}$$

其中 $\mathbf{w}$，b 分别为权向量和阈值。在线性不可分的情况下，引入松弛因子 C，对错分的样本进行惩罚，问题 (1-2) 可以重新写为，

$$\begin{aligned}\min\quad & \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{l}\xi_i\\ \text{s.t.}\quad & y_i(\mathbf{w}\bullet\mathbf{x}_i)+b-1>\xi_i, i=1,2,...,l.\end{aligned} \tag{1-3}$$

其中 ξ_i 被称为松弛变量。该优化问题的Lagrange对偶为，

$$\begin{aligned}\max\quad & W(a_i)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\bullet\mathbf{x}_j\\ \text{s.t.}\quad & \sum_{i=1}^{l}y_i\alpha_i=0\\ & 0\le\alpha_i\le C, i=1,...l\end{aligned} \tag{1-4}$$

其中 $\alpha_1,...,\alpha_l$ 为每个训练样本所对应的Lagrange系数，最优解中不为零的

Lagrange系数所对应的样本称为支持向量(Support Vectors, 简称 SVs)。若α^*为最优解, 则$\mathbf{w}^* = \sum_{i=1}^{l} \alpha^* y_i \mathbf{x}_i$, 分类决策函数可以写为,

$$f(x) = \mathrm{sgn}(\mathbf{w}^* \bullet \mathbf{x} + b) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha^* y_i \mathbf{x}_i, \mathbf{x} + b^*) \tag{1-5}$$

以上的结果是在线性可分的情况下得出的结论, 如何将它运用到非线性的问题中呢, 在解决非线性问题中, 通过引入非线性变换$\psi: \mathbf{R}^n \to \mathbf{H}$将输入空间$\mathbf{R}^n$映射到一个高维特征空间$\mathbf{H}$, 将非线性问题转换为线性问题, 在这个高维特征空间中采用同样的方法求取最优分类面。

引入映射后的二次优化对偶问题的目标函数可以写为,

$$W(a_i) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \psi(\mathbf{x}_i) \bullet \psi(\mathbf{x}_j) \tag{1-6}$$

当特征空间维数巨大时, 直接计算内积的复杂度是巨大的。那么如何解决向高维空间映射带来的维数复杂性问题? 从上面的推导过程中我们可以看出, 二次优化问题中的目标函数和最后得到的最优分类面的决策函数的表达式中, 仅包含特征空间中两个样本的点积而与空间映射的具体形式无关, 因此, 如果能够求得高维空间中两个样本的点积, 那么就可以解决向高维空间映射带来的复杂性问题, 这就是SVM将样本向高维空间映射的本质所在。为了避免在高维空间中直接计算内积, 将高维空间中的内积转化为在输入空间中的某个函数进行, 这个函数就是核函数。 事实上利用核函数解决空间复杂性问题的方法被称为核方法(Kernel Method), 在其他领域也得到广泛的应用。

定义不同的映射就可以得到不同类别的核函数, SVM 中核函数的选择是一个重要的研究方向。在 SVM 中常用的核函数有: 线性核 $K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x} \bullet \mathbf{x}_i$、多项式核$K(\mathbf{x}, \mathbf{x}_i) = ((\mathbf{x} \bullet \mathbf{x}_i) + 1)^d$、高斯径向基核 $K(\mathbf{x}, \mathbf{x}_i) = e^{-\|\mathbf{x} - \mathbf{x}_i\|^2 / (2\sigma^2)}$、二层神经网络核$K(\mathbf{x}, \mathbf{x}_i) = \tanh(k\mathbf{x} \bullet \mathbf{x}_i - \delta)$等。

只有满足Mercer定理[4]的映射函数才能成作为核函数, 引入核函数后的最优分类面问题可重写为,

$$\begin{aligned} \max \quad & Q(\alpha_i)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}a_i a_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^{l} y_i\alpha_i=0 \\ & 0\le\alpha_i\le C, i=1,...l \end{aligned} \tag{1-7}$$

相应的分类决策函数如下，

$$f(x)=\mathrm{sgn}(\mathbf{w}^*\bullet\mathbf{x}+b)=\mathrm{sgn}(\sum_{i=1}^{l}\alpha^* y_i K(\mathbf{x}_i,\mathbf{x})+b^*) \tag{1-8}$$

式(1-8)就被称为支持向量机。

2.2.2 SVM 的研究现状

作为近年的研究热点，现今存在许多关于 SVM 方法的研究方向，主要包括算法本身的改进和 SVM 的实际应用。其中算法改进主要集中在核的选择和参数的选择上面，另外提高训练速度也是一个难点问题；对于应用，大多是利用支持向量机的优良性能解决一些模式识别问题。下面介绍当前 SVM 的几个主要研究方向及进展，包括 SVM 训练算法、SVM 模型选择、SVM 派生方法、SVM 应用研究等[5]。

2.2.2.1 SVM 训练算法改进

支持向量机最大的问题之一就是训练速度慢，这也是众多学者研究的重点。虽然在 1999 年之前有很多学者提出了一些改进方法，但是效果并不是十分理想。

Platt 提出 SMO(Sequential Minimal Optimization)算法来解决大规模训练样本的问题[6]，将工作样本集的规模减为两个样本。可以说在 SVM 训练算法上 SMO 具有里程碑的意义与地位。之所以需要两个样本，是因为等式线性约束的存在使得同时至少有两个 Lagrange 乘子发生变化。由于只有两个变量，迭代过程中每一步的子问题的最优解可以直接用解析的方法求出来，从而避免了多样本的情况下数值不稳定及耗时问题。Platt 还设计了一个两层嵌套循环分别选择进入工作样本集的样本，这种启发式策略大大加快了算法的收敛速度。

Platt 借鉴了 SVM^{light} 中的一些思想，对 SMO 算法进行了改进：一方面利用 shrinking 缩小工作集的搜索范围，提高搜索速度；另一方面利用核函数缓存的策略，减少重复计算核函数矩阵中的元素的次数。

Keerithi 等人提出一种采用通过两个阈值来确定工作集的方法[7]，采用该方法可以大大减少计算核函数的次数，提高 SMO 算法的效率。随后 Keerthi 等人提出了一种基于违背样本对(Violating Pair)和有限违背样本对的泛化 SMO 算法 GSMO[8]，在 GSMO 算法每次迭代中选择任意的有限违背对作为当前的工作样本集。Keerthi 证明了该算法的收敛性。此外，还有一些学者提出了分解算法[9]。

2.2.2.2 SVM 的模型选择

SVM 的另一个难点也是瓶颈问题即模型选择(Model Selection)[10]问题。在 SVM 中有两个关键的参数，一个是惩罚因子C，另一个是核函数参数p。所谓模型选择就是选择合适的模型参数，通常通过最小化算法的泛化误差这一指标来选择最佳的参数，当然也有一些其他的方法。

在模型选择中留一法(Leave-One-Out)以及交叉检验(Cross-validation)是最基本的方法[11]，Joachims 提出了泛化误差的 Xi-Alpha 边界[12]，它使用 SVM 中的 Lagrange 系数α_i和松弛变量ξ_i来计算泛化误差的上界。

Wahba 提出了一种泛化近似交叉检验 GACV(Generalized Approximate Cross Validation)的方法[13]来确定泛化误差的界，基于统计学习理论中结构风险最小化原则，Burges 提出了一种基于 VC 维的近似的泛化误差的边界 $h \le D^2 \| w \|^2 + 1$，D 为包含所有样本的最小球直径。

Chapelle 等人提出了所谓的硬间隔(Hard Margin)边界[14]，指出 SVM 算法的泛化误差小于$0.25D^2 \| w \|^2 / l$，其中为w为权向量。Chappelle 将它扩展到软间隔(Soft Margin)中，提出了一种修改的径向一间隔边界(Modified Radius-Margin Bound)。Vapnik 提出了一种叫做 Span-Rule 的近似 LOO 方法。

2.2.2.3 SVM 的拓展方法

支持向量机最初的提出是以分类问题为背景，但是后来许多学者将该理论拓展到了回归、聚类等领域，提出了一些新的思路。

Schölkopf 等提出了一种新的支持向量机ν-SVC 用于分类问题[15]，与C-SVC 支持向量机有两个需要选择的参数不同，在ν-SVC 中只有一个需要选择的参数$\nu \in (0,1]$。Schölkopf进一步证明：(1) 参数ν是分类错误率的上界；(2) 参

数ν是支持向量比率的下界。由于ν-SVC具有的这一特性，对于SVM中的参数选择问题，只需要预先指定一个最大的分类错误率，就可以得到最优的参数。进一步将这一思想扩展到回归预测中，提出了ν-SVR算法，与ν-SVC一样采用一个参数$\nu \in (0,1]$来控制支持向量的个数。

Cortes 和Vapnik提出SVM 最初用于分类问题[16]，后来Vapnik在定义了ε 不敏感损失函数的基础上，提出了ε 支持向量回归算法，将SVM方法从分类问题拓展到回归预测中。

针对高维空间中的属性预测(Distribution Estimation)问题，Schölkopf 提出了一种 One-Class SVM 的方法[17]。在 One-Class SVM 中，样本被映射到一个高维空间中，寻找样本点和高维空间原点的最大间隔。

D.M.J. Tax 提出了支持向量数据域描述方法[18]，为了建立样本的数据域描述模型，将输入空间的样本映射到一个高维的特征空间，寻找样本的最小包含超球。这一算法可以用于孤立点检测。

此外，Ben-Hui将SVM的思想引入到聚类问题中[19]，提出了基于支持向量数据域描述的支持向量聚类算法。该算法可以发现任意边界形状的类，而且不需要预先给定类的数目，只需给定一个错误率的上界，即可自动决定样本集中类的个数。还有一些学者提出了模糊支持向量机概念[20]。

2.2.2.4 SVM 的应用研究

由于 SVM 具有良好的泛化性能，并能够有效地解决非线性和维数灾难等一系列难题，使得 SVM 在许多领域如人脸识别、基因序列分析、孤立点检测中得到很好的应用。在工业方面，主要应用 SVM 进行模式识别。

陈念贻给出了关于 SVM 在化学化工中应用的综述[21]，分别介绍了 SVM 在多变量数据校正、数据建模、商品检验、相图和新化合物的计算机预报、新材料制备的实验设计、环境污染的建模和预报以及分子设计、药物设计等领域的一些应用。陆文聪等开发了一个处理化学化工数据的 SVM 算法软件 ChemSVM[22]。陈念贻介绍了 SVM 和其他一些核函数算法在化学计量学中的应用[23]。丁亚平等将 SVM 用于食品分析和湿法冶金中的多变量数据校正问题中[24]。许建华等提出了 SVM 在油气判别中的应用[25]。

1.3 论文主要内容

本论文从内容上分为三个部分，首先采用支持向量机结合了核主元分析方法建模，提出了一种 KPCA-SVR 算法，该方法充分利用了核方法的优势，并将其应用到复合肥生产的过程中，取得了很好的效果。其次，介绍了一种探索型的数据挖掘方法，提出了一种基于支持向量机的关联规则算法，并将其应用到了冶炼工业过程中，通过对实际采集的数据进行仿真得到了一些有价值的结论。最后，详细介绍了作者参与开发的一个数据挖掘软件，详细描述了软件的整个制作过程和大部分功能，为数据挖掘在工业中的应用给出了一个实例。

数据挖掘属于交叉学科[26]，它涉及到了多种方法与领域的结合，是由实际应用产生的学科。工业数据挖掘是其中的一个分枝领域，是数据挖掘方法与过程控制领域的结合。而支持向量机作为统计学习理论的一种新方法[27]，也是一门交叉学科，其理论基础坚实、实用性强、应用效果好。论文涉及到的这两个概念在整个知识领域的关系如下图所示，

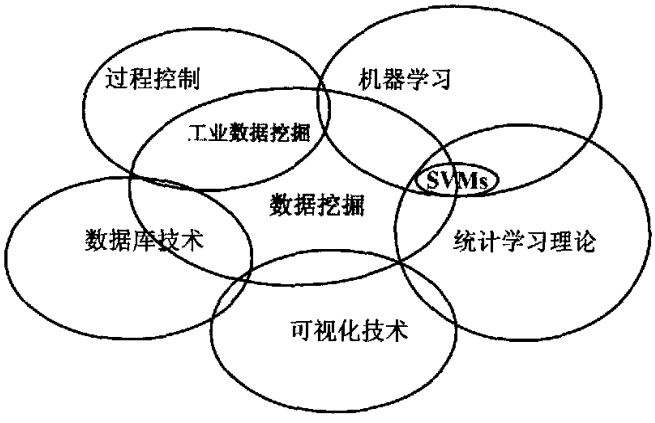


图 1-4 论文概念在知识领域的关系图

在每章的布局上，在第二、三章中，都从研究背景、算法介绍和应用实例或仿真实验三方面加以介绍和阐述，层层递进。在第四章中，从数据挖掘任务到程序的设计，很详细的介绍了设计软件的全过程，可以让读者对工业数据挖掘软件的设计了解的更加透彻。

在结构上，论文第二章讲解了一种支持向量机理论的应用和模型选择难点分析与改进，是对支持向量机核方法的研究，第三章是对支持向量机理论在数据挖掘方面的推广，第四章是数据挖掘的实例，这三个部分是由理论创新到应用推广的过程。下面给出整个论文的框架图，从中可以清晰的看出各个章节的内容。

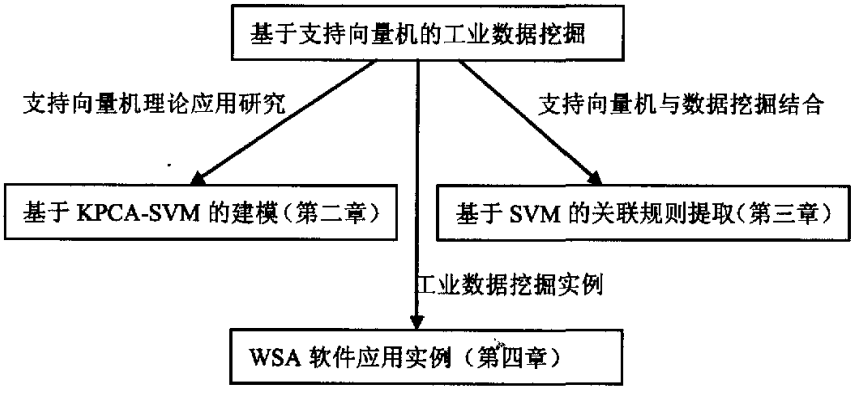


图 1-5　论文框架图

1.4　小结

在绪论中，重点介绍了数据挖掘、工业数据挖掘、支持向量机的基本概念、研究现状和进展情况。同时文章也阐述了它们在工业中的发展情况、实际运用情况、存在的优势和不足之处，并在章末给出了全文所要介绍的研究重点和文章的整体框架。

第二章 基于支持向量机与核主元分析的建模研究

2.1 引言

在数据挖掘算法中，数据预处理是第一步，也是最重要的一步，因为数据挖掘的各种算法都是建立在有效合理的数据基础上的，只有高质量的数据才能产生高质量的决策。数据预处理主要分为数据清理、数据集成、数据变换和数据规约。

数据规约包括数据聚集、维规约、数据压缩和数字规约。其中维规约是从众多属性中去掉一些不相关的属性。为了去除数据仓库中相关性较低的属性，可以使用多种技术对属性之间的相关性进行分析处理，主要方法有双变量统计、主成份分析、模糊技术、信息增益法、粗糙集方法、相关系数法等。在模式识别或决策支持系统等许多实际应用中，识别相关属性是数据挖掘的关键。而实际上，对于特定的数据挖掘任务，用户往往只对属性集的某些子集感兴趣，对属性子集的操作也减少了信息的处理量，所以要对属性进行特征选择。这其中包括了很多方法，主元分析就是其中之一。

在本章中，将介绍一种基于核方法的建模方法，即 KPCA-SVR 算法。它采用核主元分析来降维处理数据，然后利用改进的支持向量机回归方法进行建模。该方法有两大主要优势：第一是能够利用两种方法都是基于核函数的特点，在中间步骤中对核函数矩阵等相关变量进行存储，大大提高了算法的效率；第二是能够自动选择参数，减少了人为经验对程序结果的影响，能够实现在线更新模型。并在复合肥生产数据中进行了仿真实验，取得了很好的效果。

2.2 核主元分析

2.2.1 PCA 方法

主元分析法(Principal Component Analysis，简称 PCA)[28]是一种能够处理相关性的统计分析技术。对于高维原始变量 **X**。通过可逆变换 **T**，使 **TX** 不仅表征了 **X** 的重要信息。而且 **TX** 的某些分量具有较低方差，这样对 **TX** 的舍取在均方差意义下最优。

主成份分析对属性进行约简后，属性的特征化为隐性的。我们先来看传统的

主成分分析方法。对于输入为$n*m$维的矩阵**X**，我们可以把它表示为，

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{T}+\mathbf{E} \tag{2-1}$$

其中**T**为主元变量，**P**为主元向量

$$\mathbf{T}=\mathbf{X}\mathbf{P} \tag{2-2}$$

t_k是**X**的线性函数。设$\mathbf{X}=(x_1,x_2,\cdots,x_n)$，具体计算方法如下，

步骤1　计算属性均值$\bar{x}$，并从所有的数据中减去以校正这些点$\bar{x}$，求得归一化矩阵；

步骤2　计算方差-协方差矩阵$(x_i-\bar{x})(x_i-\bar{x})$；

步骤3　确定非零向量$\mathbf{P}_i$和正实数k，相对于实对称矩阵**C**有$Cp=Lp$，L为特征值；

步骤4　求出相应的所有特征向量；

步骤5　排序特征值，使得每个特征向量$\mathbf{P}_i$都保留剩余方差中；

步骤6　计算每个主成份$\mathbf{P}_i$的方差贡献率。

主元分析是通过搜索能代表原数据的正交向量，创立一个替换的、较小的变量集来组合属性的精华，因而原数据可以投影到这个较小的集合中，导致数据集的压缩[29]。从统计模式识别的观点来看，主元分析实际上是降维处理过程。它忽略了具有较小方差的线性组成部分，保留具有较大方差项，从而减小了有效数据表示的维数。但它是一种线性算法，只能提取数据中的线性相关特性。考虑到工业数据都是比较复杂的情况，我们下面继续讨论非线性PCA方法.

2.2.2　非线性 PCA 方法

对于大量掺有过程误差、噪声、冗余信息及各变量间存在强相关性的非线性数据，从中提取其重要特征、消去误差及冗余信息、大大降低数据维数是数据处理的一项重要任务。如何有效地解决大量非线性数据的特征提取、降低数据维数，从而进行生产输出质量指标的预测将是论文讨论的重点。

线性 PCA 是一种线性算法，只考虑二阶统计特性，只能提取数据中的线性关系，从而将高维原始数据的线性特征映射到低维特征空间。主成分分析是将多个变量综合成少数变量的一种多元统计方法，为解决多指标的综合评价提供了一

个很好的手段。但现实的指标之间的关系往往是非线性的，线性 PCA 评价方法中可能出现各指标的贡献率过于分散的情况，找不到具有全面综合能力的指标。主成分的所谓综合，实际上只是对变量间共性的一种提取，因此要综合相关性不大的变量，采用线性主成分分析法是不妥的。这时就需要考虑采用非线性 PCA 方法。非线性 PCA 的方法有很多，有基于神经网络的非线性方法等等。论文主要采用一种基于核函数的 KPCA [30]。核主成分分析通过一个非线性变换，首先将原变量空间映射到高维特征空间，在这个高维特征空间中进行线性主成分分析。通过核技巧，KPCA 评价方法只需在原空间进行点积计算，而不必知道非线性变换的确切形式。

设 $x_k \in R^p$（$k=1,2,\cdots,l$）为样本数据向量。假设 $\sum_{k=1}^{l} x_k = 0$，线性 PCA 就是对矩阵 $\mathbf{C}=\frac{1}{l}\sum_{k=1}^{l} x_k \bullet x_k^T$ 求特征值和特征向量。推广到非线性情况，首先把原空间的数据通过非线性变换 ϕ。投影到特征空间 F。非线性 PCA 就可以看成是在 F 中对矩阵 $\bar{\mathbf{C}}=\frac{1}{l}\sum_{k=1}^{l}\phi(x)\phi(x)^T$ 进行线性主元分析。显然所有特征值 λ 和特征向量 V 都在 $\phi(x)$ 张成的子空间内。有 $\lambda(\phi(x)\bullet\mathbf{V})=\phi(x)\bullet\bar{\mathbf{C}}\mathbf{V}$，且存在系数 $\alpha_1,\alpha_2,\cdots,\alpha_l$。使得 $\mathbf{V}=\sum_{i=1}^{l}\alpha_i\phi(x_i)$。定义矩阵 K，使得 K(l*l)。$\mathbf{K}_{ij}=\phi(x_i)\phi(x_j)$。可以得到 $l\lambda\alpha=\mathbf{K}\alpha$。

对于主成分提取，只需要计算一个测试点 $\phi(x)$ 在特征向量上的投影。

$$(\mathbf{V}_k\cdot\phi(x))=\sum_{i=1}^{l}\alpha_i^k\cdot(\phi(x_i)\phi(x)) \tag{2-3}$$

基于上述问题，给出了一种称之为 KPCA 的非线性主元评价模型，KPCA 方法将观察变量空间 X 通过一个非线性变换由原空间映射到高维持征空间 F，在 F 中进行线性主成分分析。通过核技巧，KPCA 评价方法只需在原空间进行点积计算，而不必知道确切形式 M。

KPCA 与 PCA 具有本质上的区别：PCA 是基于指标的，而 KPCA 是基于样本的。KPCA 的优势是可以最大限度地抽取指标的信息；但是 KPCA 抽取指标的

实际意义不是很明确，计算也比 PCA 复杂。

核函数方法的基本原理是通过非线性函数$\phi(\cdot)$把输入空间映射到高维空间，在特征空间中进行数据处理，其关键在于通过引入核函数，把非线性变换后的特征空间内积运算转换为原始空间的核函数计算,从而大大简化了计算量[31]。

2.3 KPCA-SVR 方法

尽管SVM计算复杂度与训练样本变量的维数无关，能够有效地对付高维问题，但是输入变量之间的线性相关性影响模型的精度和泛化能力，而PCA等特征提取方法能够有效处理变量之间的共线性，降低输入变量维数。因此，本论文提出了基于KPCA-SVR的软测量建模方法，即以输入数据的主元作为SVM模型的输入，这样既结合了KPCA的特征提取能力，又利用了SVM的良好的非线性函数逼近能力。最后针对工业过程数据进行了应用研究，研究结果表明KPCA-SVR软测量模型的性能要优于SVM软测量模型[32]。

下面简要介绍一下支持向量机回归方法—SVR[33]（support vector regression）

支持向量回归根据训练样本点，拟合出一条曲线，使大部分样本点都在支持向量回归曲线所处的不敏感带中，从而得到建立模型的参数，给出回归函数。其数学描述为，

$$\min_{w\in R^n,\xi^{(*)}\in R^{2l},b\in R} \tau(w,\xi^{(*)}) = \frac{1}{2}\|w\|^2 + C\cdot\frac{1}{l}\sum(\xi_i+\xi_i^*) \tag{2-4}$$

$$s.t.\ \ ((w\cdot x_i)+b)-y_i \le \varepsilon+\xi_i, i=1,2,\cdots,l,$$

$$y_i-((w\cdot x_i)+b) \le \varepsilon+\xi_i^*, i=1,2,\cdots,l,$$

$$\xi_i \ge 0, \xi_i^* \ge 0, i=1,2,\cdots,l$$

其中l表示样本数量，ε表示不敏感带宽度，ξ为松弛变量，C为惩罚因子。求解其对偶问题

$$\min\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i) \tag{2-5}$$

$$st.\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \quad 0\le\alpha_i,\alpha_i^*\le\frac{C}{l}, i=1,2,\cdots,l$$

得到最优解 $\overline{\alpha}=(\overline{\alpha}_1,\overline{\alpha}_1^*,\cdots,\overline{\alpha}_i,\overline{\alpha}_i^*)^T$，即可构造出决策函数

$$f(x)=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)K(x_i,x)+\overline{b} \tag{2-6}$$

其中 $K(x_i,x_j)$ 为核函数，$\overline{b}$ 为平均阈值。在支持向量回归的方法中，数据的相似程度是通过内积进行估价的，选择不同的核函数，相当于选择了不同的内积，不同的空间映射，不同的标准对相似性进行估价。论文选择泛化能力较好的高斯核函数，$K(\mathbf{x},\mathbf{x}_i)=e^{-\|\mathbf{x}-\mathbf{x}_i\|2/(2\sigma^2)}$。

工业采集到的数据如果直接用来进行支持向量回归很不理想，而 KPCA 不仅能够去除噪声，进行一定程度的数据校正，而且比线性 PCA 能够提取出更多的样本信息[34]。有研究表明，在达到相同分类性能的前提下，KPCA 所需的主元个数要少于线性 PCA。由于得到的样本数比较少，采用支持向量回归的方法，来发挥其处理小样本能力强、泛化能力好的优势[35]。

KPCA-SVR 方法综合了上面的两种核方法，利用了它们的很多共同点，例如，核函数参数矩阵可以通过同一个函数来计算，都需要求解拉格朗日乘子，对训练数据的处理思路也相同。编写共用的核函数处理、拉格朗日乘子函数，以及类似的训练数据处理函数，把两种方法有机地结合起来，简化了程序和整个计算步骤。而且通过 KPCA 部分处理后的数据的维数大大降低了，有利于提高 SVR 部分的运算速度和准确度，使算法的性能大大提高了。

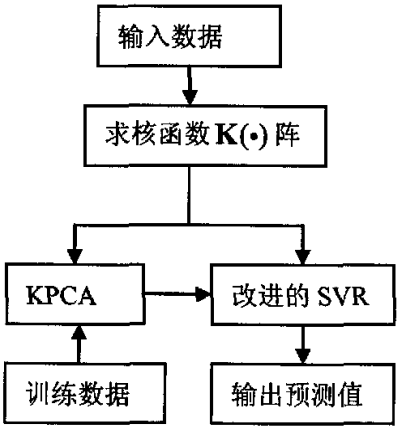


图 2-1 KPCA-SVR 流程图

改进的梯度参数选择方法

在支持向量回归中，推广能力是指学习机对未知数据进行回归时的回归误差[36]。推广能力估计是指通过训练集给出回归函数能力的预测，而好的推广能力估计方法是实现 SVR 参数选择的基础。找到一个好的推广能力估计才能计算参数，从而选择到最优值。对于支持向量回归的误差界存在很多估计方法，最基本的方法为留一法：$loo=\sum_{i=1}^{l}\left|f^{t}(x_t)-y_t\right|$，其中 l 表示样本数，$f^{t}(x_t)$ 表示去掉 t 个样本后训练其他样本得到的对第 t 个样本的函数估计值。由于留一法计算量非常大，因此比较常用的方法是半径间隔法[37]需要估计的核参数为 θ，在高斯核函数中，θ 即为 C（惩罚因子）和 q（核参数）。在下面的求解中，可以把 C 看成核参数的一部分来求解。误差估计的上界简单推导如下，

在支持向量回归中，可以得到支持向量回归的一个误差界[38]

$$RM(\alpha,\theta,\varepsilon)=4R^2e^T(\alpha+\alpha^*)+l\varepsilon \tag{2-7}$$

ε 的含义明确，其值容易选择，下面只讨论 C，q 的选择。

采用最速下降法，上面界对参数 $\theta(C,q)$ 求导得，

$$\frac{\partial}{\partial\theta}RM(\alpha,\theta,\varepsilon)=4\frac{\partial R^2}{\partial\theta}e^T(\alpha+\alpha^*)+4R^2\frac{\partial e^T(\alpha+\alpha^*)}{\partial\theta} \tag{2-8}$$

可以得到一种求解 θ 的递推方法，每一步更新的公式为，

$$\Delta\theta=-\eta\frac{\partial}{\partial\theta}RM(\alpha,\theta,\varepsilon) \tag{2-9}$$

由于所求问题为非凸集，采用梯度法计算的时候，初值的选择十分重要，可能导致得到的解仅为局部最优解而非全局最优解。这个问题能够通过遗传算法来解决[39]，但是计算量大，比较耗时。

在留一法的原理中，当去掉的点为非支持向量的时候，产生的误差小于 ε[40]；当去掉的点为支持向量的时候，产生的误差相对较大。因此，我们可以采用支持向量比率（即 sv/l，其中 sv 表示支持向量数目，l 表示样本数目）来判断回归误差。当支持向量率大的时候，误差较大，反之亦然。支持向量率计算方便，可以先采用支持向量率估计出参数的初值范围，再利用梯度法求解出较精确的值。整个程序流程图如下，

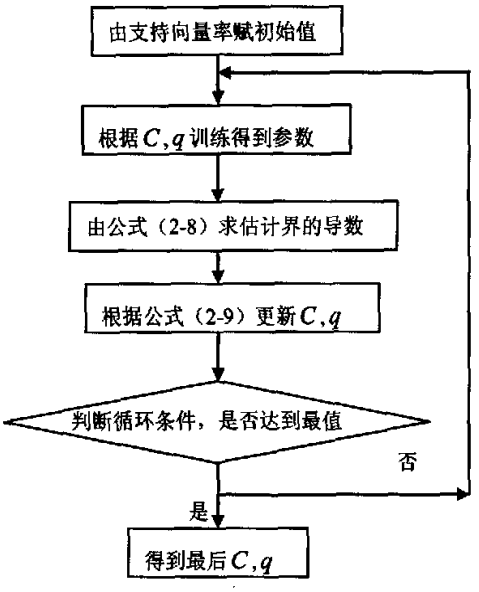


图 2-2 改进的支持向量回归参数选择流程图

这种改进算法的特点是，计算量较小，结果也比较准确，具体的实验结果将在下节具体给出。

2.4 KPCA-SVR 在复合肥生产中的应用

复合肥生产的特点是流程长、易堵、腐蚀性强，生产中装置间相互影响比较大，关联性强，有许多干扰源。各部分装置常具有大滞后、强耦合及非线性等特性。复合肥养分含量等质量指标难以直接测量，影响复合肥的生产调节及成品质量的控制。复合肥装置由三部分组成：磷酸工段、氢钾工段、复合肥工段。磷酸工段的原料是磷矿石，生产合格的稀磷酸；氢钾工段的原料是氯化钾和 98%以上的浓硫酸，生成的硫酸氢钾与磷酸工段送来的稀磷酸制成混酸送往复合肥工段；复合肥工段将氢钾工段送来的合格混酸生成复合肥。从复合肥混酸槽开始，分成 A、B 两条生产线，经常存在分别停车状态，存在间歇性的情况，使得工况更加复杂。要通过生产数据预测复合肥中的氮、磷、钾养分，采用机理建模等一些传统方法相当困难[41]。

本章提出的一种 KPCA-SVR 方法，根据 KPCA 和 SVR 都是基于核方法的特点，所提出的 KPCA-SVR 方法，有机的结合了两种方法的共同点。使得

KPCA-SVR 方法既具有 KPCA 提取非线性主元成分，可以最大程度地去除噪声的性能，又具有 SVR 处理小样本能力强、泛化性能好的特点，而且算法的执行效率高。同时针对 SVR 部分提出了一种核参数选择的改进算法，结合到 KPCA-SVR 方法中，在复合肥的养分预测仿真研究中取得了很好的效果。

本章采用的数据是某股份有限公司硫酸厂复合肥 1 个月的生产数据，氮、磷、钾养分含量为化验分析值。预测复合肥氮、磷、钾养分所用到的相关数据构成了输入空间，属性为：氯化钾含量，硫酸流量以及 A、B 线（参见附录）的相关数据等，根据输入空间建立复合肥养分的预测模型。

2.4.1 仿真 1： SVR 、PCA-SVR 和 KPCA-SVR 方法对比

从现场共采得 110 个数据，具有 10 个属性，先对数据进行 KPCA 预处理，得到了 110×110 的 $K(\cdot)$阵，对其求特征值和特征向量，得到 110×2 的降维矩阵，两个主元贡献率分别为 58.20%和 39.47%。选择高斯径向基核作为 KPCA 的核函数，其核参数 $q=1.312\times10^{-6}$。关于 KPCA 的核函数参数选择，目前还没有很好的方法，论文是利用了试探性的方法给出了 KPCA 的参数选择，得到了一些选择的规律：当 q 的值较小的时候，得到的 $\mathbf{K}(\cdot)$ 阵趋近于元素全为 1 的矩阵，反之 $\mathbf{K}(\cdot)$ 阵中会出现很多 0，显然都提取失去了意义；可以通过在训练样本上计算 $\|x_i-x_j\|$ 的平均值得到 q 的粗略范围，以保证 $\mathbf{K}(\cdot)$ 阵的大部分数值都比较正常，即不会出现上面两种情况，有利于提取主元。

从采得的 110 个数据中任取 70 个数据作为训练样本，其余 40 个作为预测样本，分别用 KPCA-SVR、SVR 和 PCA-SVM 三种方法进行处理，核函数均为高斯径向基核函数。

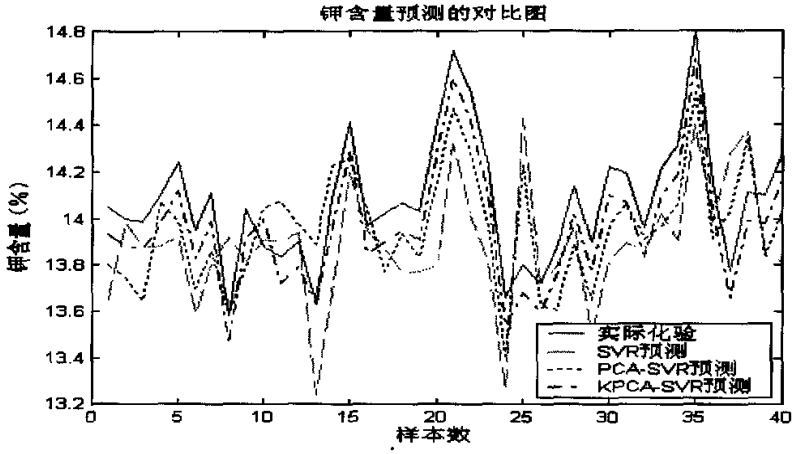


图 2-3(a) 钾含量对比

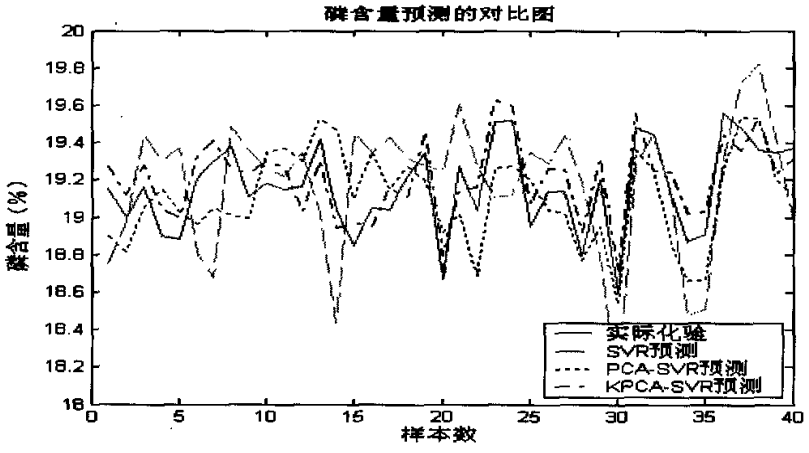


图 2-3(b) 磷含量对比

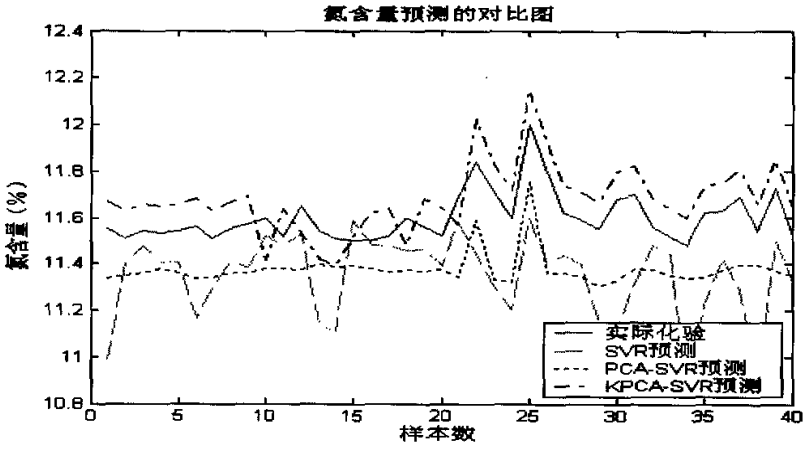


图 2-3 (c) 氮含量对比

图 2-3 SVR, PCA-SVR 和 KPCA-SVR 三种方法预测曲线对比

表 2-1 SVR,PCA-SVR 和 KPCA-SVR 三种方法误差率对比（磷）

	相对误差率	均方差	超调量	C	q
SVR	0.0562	0.7868	0.5114	580	0.012
PCA-SVR	0.0332	0.5124	0.2325	353	0.343
KPCA-SVR	0.0168	0.2534	0.1132	623	0.032

从表 2-1 中可以看出，采用了 KPCA 预处理之后的方法，效果要明显好于前面的方法。限于篇幅，此处只列了磷的错误率表。由图 2-3 可以清晰的看出，钾和氮也具有类似的比较好的预测效果，说明了 KPCA-SVR 方法的有效性。从 KPCA-SVR 预测曲线与实际曲线的逼近程度，说明了论文自动选择的支持向量参数的有效性。

表 2-2 论文方法与穷举法对比（KPCA-SVR，磷）

	C	q	运算时间	均方差
穷举法	691	0.036	3531s	0.2643
本文方法	623	0.032	142s	0.2534

从表 2-2 中可以看出，采用论文方法选择的参数，从数值和性能上，都与穷举法比较接近，而时间却大大缩短了。

2.4.2 仿真 2： 采用带参数选择的 SVR 进行模型更新

通过仿真 1 证明了 KPCA-SVR 方法的有效性。由于能够自动选择参数，KPCA-SVR 方法还可以实现在线模型更新，建立更新模型，即每次更新加入样本，除去相同数目的最早的旧样本，更新核参数。在仿真 2 中，由于数据点比较少，更新数据时不去掉旧的数据点。（每周更新一次，初始 50 个样本点）得到的结果如下，

表 2-3 更新过程中的 C，q 值

	初始值	第一次更新	第二次更新	第三次更新
C	641	731	564	643
q	0.03	0.035	0.032	0.036

更新数据时，参数的值发生了一定的变化，但由于采用的只是一个月的生产数据，所以更新所得到的参数值变化不大，不过仍然可以得到较好的性能。

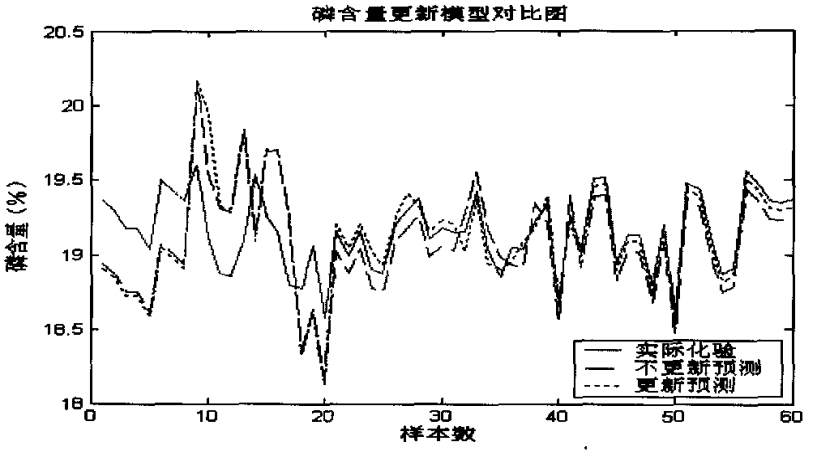


图 2-4 更新模型的磷的预测图

从图 2-4 中可见，由于建模的初始样本点较少（仅有 20 个），所以偏差较大。随着样本点数和核参数数值的不断更新，得到的结果越来越准确，优于不更新数据建模所得结果。模型的在线更新很重要，尤其是对于数据波动较大的参数而言。

2.5 小结

本章提出的 KPCA-SVR 方法成功地解决了工况复杂的复合肥养分预测问题，并通过与 SVR 和 PCA-SVR 的比较，体现了其性能优势。KPCA-SVR 方法不但利用了两种方法的共同点，大大提高了效率，也可以自动选择参数，降低了使用支持向量机的难度。利用该方法来更新预测模型，可以得到更加广泛的应用。进一步的研究工作是从理论上提出更加有效、简便的方法来解决支持向量机参数选择问题。

第三章　基于支持向量机的关联规则提取

3.1　引言

如何从大量的数据中提取出有效的规则一直是数据挖掘的热点之一。在现有的方法中，比较常用的方法有经典的通过频繁集获取规则、基于粗糙集理论的规则提取、采用神经网络提取规则和利用决策树等分类算法提取规则等等，这些方法都存在训练速度慢的缺点。

支持向量机在数据挖掘的分类、聚类问题中已经取得了很好的应用。它具有理论基础坚实，处理小样本能力强，可以处理高维、非线性数据，泛化性能好的优点。但是由于其分类函数存在可理解性差的黑箱问题，因此不能直接用来提取规则。

本章提出了一种新的方法很好地解决了这个问题，该方法利用数据点在高维空间的数据域描述和得到的支持向量来提取规则，能够较快的提取关联规则。通过标准数据集的检验证明了该方法的有效性，同时通过对实际工业数据的仿真，进一步验证了该方法的实用性，为其在工业领域的应用提供了一个思路。

3.2　关联规则

关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。它在数据挖掘中是一个重要的课题，最近几年已被业界所广泛研究。

关联规则挖掘的一个典型例子是购物篮分析。关联规则研究有助于发现交易数据库中不同商品（项）之间的联系，找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。分析结果可以应用于商品货架布局、货存安排以及根据购买模式对用户进行分类。

Agrawal 等于 1993 年首先提出了挖掘顾客交易数据库中项集间的关联规则问题，以后诸多的研究人员对关联规则的挖掘问题进行了大量的研究。他们的工作包括对原有的算法进行优化，如引入随机采样、并行的思想等，以提高算法挖掘规则的效率，对关联规则的应用进行推广。

最近也有独立于 Agrawal 的频集方法的工作[42]，以避免频集方法的一些缺陷，探索挖掘关联规则的新方法。也有一些工作[43]注重于对挖掘到的模式的价值

进行评估，他们提出的模型建议了一些值得考虑的研究方向。

基本概念

设$I=\{i_1,i_2,\cdots i_m\}$是项集，其中$I_k(k=1,2,\cdots,m\}$可以是购物篮中的物品，也可以是保险公司的顾客。设任务相关的数据 D 是事务集，其中每个事务 T 是项集，使得$T\subseteq I$。设 A 是一个项集，且$A\subseteq T$。

关联规则是如下形式的逻辑蕴涵：$A\subseteq B$，$B\subseteq I$，且$A\cap B\neq\varnothing$。关联规则具有如下两个重要的属性：

支持度: P(A∪B)，即 A 和 B 这两个项集在事务集 D 中同时出现的概率。

置信度: P(B | A)，即在出现项集 A 的事务集 D 中，项集 B 也同时出现的概率。

同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。给定一个事务集 D，挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则，也就是产生强规则的问题。

关联规则种类

1) 基于规则中处理的变量的类别，关联规则可以分为布尔型和数值型。

布尔型关联规则处理的值都是离散的、种类化的，它显示了这些变量之间的关系。

数值型关联规则可以和多维关联或多层关联规则结合起来，对数值型字段进行处理，将其进行动态的分割，或者直接对原始的数据进行处理，当然数值型关联规则中也可以包含种类变量。

2) 基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

在单层关联规则中，所有的变量都没有考虑到现实的数据是具有多个不同的层次的。

在多层关联规则中，对数据的多层性已经进行了充分的考虑。

3) 基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。

在单维关联规则中，我们只涉及到数据的一个维，如用户购买的物品；

在多维关联规则中，要处理的数据将会涉及多个维。

经典的频集算法

Agrawal 等于 1993 年提出了一个挖掘顾客交易数据库中项集间的关联规则的重要方法，其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。

所有支持度大于最小支持度的项集称为频繁项集，简称频集，频集算法的基本思想为：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。经典的 Apriori 算法，简单介绍如下：

频繁项集：给定一个最小支持度 min_sup，如果一个项集的出现频率不小于最小支持度，则称该项集为频繁项集。

1. 关联规则：一个形如 A→B 的蕴涵式，其中 A 和 B 均为项集，且 A∩B=Φ。A∪B 的出现频率与事务数据库中记录总数之比称为改规则的支持度，而规则的置信度是事务数据库中包含 A 的事务同时也包含 B 的比例，即为条件概率 P（B | A）；
2. 强规则：同时满足最小支持度和最小置信度的规则称为强规则；

TID	项集
101	1, 2, 5
102	2, 4
103	2, 3
104	1, 2, 4
105	1, 3
106	2, 3
107	1, 3
108	1, 2, 3, 5
109	1, 2, 3

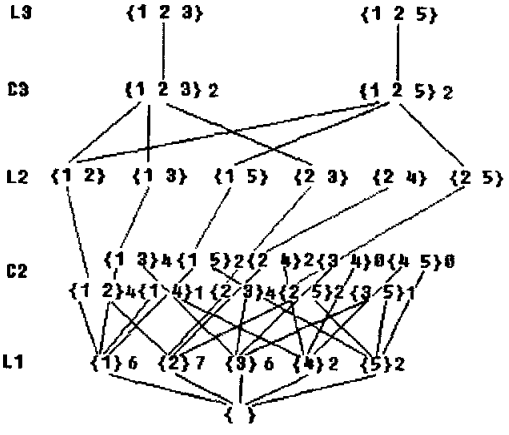


图 3-1 Apriori 算法示例图

3.3 基于支持向量机的关联规则提取算法

3.3.1 支持向量聚类

在数据域描述[44]的基础上，可以得到基于支持向量的聚类算法[45]。给定l个球形分布训练样本$\{x_1,...,x_i,...,x_l\}\subset\chi$，其中$\chi$为输入空间，$\chi\in\mathbb{R}^n, i=1,...,l$，寻找这些样本的一个最小包含球，球半径为$R$，球心为$a$。数学描述为：$\|x_i-a\|^2\le R^2 \quad \forall i$。在现实的数据集的分布中，不可避免存在一些孤立点，为了使得这种描述不至于对孤立点过分敏感，允许一些孤立点的成在，引入一种被称为松弛变量的参数ξ_i，数学描述变为 $\|x_i-a\|^2\le R^2+\xi_i,\xi_i\ge 0 \quad \forall i$。通常情况下，在数据空间$\chi$，即使忽略所有的孤立点，数据也不一定是球形分布的，在这种情况下，为了得到输入空间数据的一个描述，采取的方法是通过映射$\Phi:\chi\to F$将输入空间的数据映射到一个高维（甚至是无限维）的特征空间F，在这个特征空间中寻找数据的最小包含球，这一问题可以归结为一个求解二次规划的问题，

$$\min \quad W(\xi_i,R,a)=R^2+C\sum_{i=1}^{l}\xi_i \tag{3-1}$$

$$\text{s.t.} \quad \|\Phi(x_i)-a\|^2\le R^2+\xi_i \tag{3-2}$$

$$\xi_i\ge 0, i=1,2,...,l. \tag{3-3}$$

其中，C是松弛因子，用于在R和$\sum_{i=1}^{l}\xi_i$之间进行平衡。

该优化问题的Lagrange形为，

$$\min \quad L=R^2+\sum_{i}^{l}(R^2+\xi_i-\|\Phi(x_i)-a\|^2)\beta_i-\sum_{i=1}^{l}\xi_i\mu_i+C\sum_{i=1}^{l}\xi_i \tag{3-4}$$

$$\text{s.t.} \quad \|\Phi(x_i)-a\|^2\le R^2+\xi_i \tag{3-5}$$

$$\xi_i\ge 0,\beta_i\ge 0,\mu_i\ge 0, i=1,2,...,l \tag{3-6}$$

其中β_i，μ_i是Lagrange系数。为了最小化（3-4），我们需要找到L的鞍点，令L对于主变量ξ_i,R,a和对偶变量β_i,μ_i的导数等于零，得到如下的等式，

$$\sum_{i=1}^{l}\beta_i=1, a=\sum_{i=1}^{l}\beta_i\Phi(x_i),\ \beta_i=C-\mu_j \tag{3-7}$$

将上面的等式带入（3-4）得到其Wolfe对偶为，

$$\max \quad Q(\beta)=\sum_{i}^{l}\Phi(x_i)^2\beta_i-\sum_{i,j}^{l}\beta_i\beta_j\Phi(x_i)\cdot\Phi(x_j) \tag{3-8}$$

$$\text{s.t.} \quad 0\le\beta_i\le C, i=1,2,...,l$$

为了避免直接在高维特征空间中计算内积，通常将高维空间中的内积转化为输入空间中的核函数进行计算，核函数满足 $K(x_i,x_j)=\psi(x_i)\cdot\psi(x_j)$，常用的核函数有：线性核 $K(x,x_i)=x\cdot x_i$，多项式核 $K(x,x_i)=((x\cdot x_i)+1)^d$，高斯径向基核 $K(x_i,x_j)=e^{-p\|x_i-x_i\|^2}$，引入核函数后公式（3-8）可以重写为，

$$Q(\beta)=\sum_i^l K(x_i,x_i)\beta_i-\sum_{i,j}^l \beta_i\beta_j K(x_i,x_j) \tag{3-9}$$

最优解中不为零的Lagrange系数所对应的样本称为支持向量（Support Vectors 简称 SVs），在支持向量中，满足 $0<\beta_i<C$ 条件的样本称为非边界支持向量，$\beta_i=C$ 的支持向量称为边界支持向量。

在特征空间中的一个点 $\Phi(x_i)$ 到球心 a 之间的距离定义为 $D(x_i)=\sqrt{\|\Phi(x_i)-a\|^2}$，公式（3-7）代入上式得到，

$$R(x_i)=\sqrt{\sum_{i,j}^l \beta_i\beta_j K(x_i,x_j)+K(x_i,x_i)-\sum_j^l K(x_j,x_i)\beta_j, i,j=1,...l} \tag{3-10}$$

特征空间最小包含球球半径为 $R=R(x_i)$，其中 x_i 为任一支持向量。当 R 和 a 确定后，我们就得到了给定训练数据集的一种数据描述，对于特征空间中的一个未知点 $\Phi(z_i)$，根据它到 a 的距离 $D(z_i)$，我们可以判断它是否是属于这个数据集，还是一个孤立点。如果 $D(z_i)\le R$，则 z_i 属于这个数据集，否则被认为是一个孤立点。从而达到了我们的要求。

在输入空间中属于同一类的数据点，连接它们的路径上的任何点都属于同一类，把这些路径上的点映射到高维空间，都应该包含在最小超球中。反之，若两点之间的点映射到高维空间中，落在了超球外，那么这两个点不属于一类。基于此，可以构成一个输入样本空间的连接矩阵（行，列均代表原始数据点，若属于同类，则标0，否则标1）。

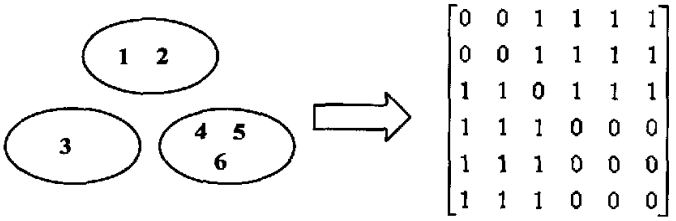


图3-2　支持向量聚类示例图

根据连接矩阵，可以判断出聚类的数量和每一类中对应的点。聚类的数量，和相似程度的定义可以通过调整支持向量参数来控制，聚类的边界通过支持向量来描述。

3.3.2 支持向量关联规则提取方法

现有的规则提取方法需要大量训练样本来得到规则。支持向量机的优势在于处理非线性、高维数据能力强，且推广能力好[46]。把这些特点应用于关联规则提取中，能够很好的提取出非线性、高维复杂数据中的规则[47]。采用支持向量聚类来对样本进行分析，得到的规则的数量和精确度可以通过支持向量参数来调节，比较灵活。

对于分类得到的每一类数据点，都可以提取出一条规则。采用数据域描述可以更好地细化规则，更好地聚集数据，去除孤立点。同时找到支持向量，为提取出有效的规则打下基础。由于所用方法都是基于支持向量机的，程序中可以编写相同的功能模块来提高效率。

支持向量机利用核函数把数据从原始空间映射到高维空间，计算都是在原始输入空间进行，这就为直接利用支持向量来提取关联规则提供了方便[48]。利用高维空间得到的支持向量，通过其原始空间各属性值的极值来确定样本的属性范围。提取过程中，规则的支持度是数据域描述中的样本占总样本的数量的比值，置信度是数据域样本和由满足提取出规则的样本的比值[49]。可以通过调整支持向量的参数来调整置信度和支持度。

支持向量提取规则算法步骤：

1. 训练样本归类（聚类）

通过聚类把训练样本点分开，而不是用分类算法，增加了最后得到规则的灵活性，也更能反映实际情况。采用支持向量聚类，充分利用支持向量机的优势，使得聚类的结果，推广能力强。这个过程中，参数的选择很重要，可以改变聚类的数量以及准确度。由于现在还没有很好的方法来确定怎样调整参数聚类能得到较高的准确度，论文试验中采用试探性的方法，并利用了参数具有的一定的意义[50]。

2. 每类样本数据域描述（聚集数据，得到支持向量）

对聚类后得到的每类样本，通过数据域描述来进行聚集，得到边界支持向

量 $\mathbf{x}_i$。去掉孤立点之后，使得到的规则更加精确，由于算法与第一步聚类相似，程序实现中，采用相同模块提高效率。

3. 提取关联规则（利用支持向量属性极值得到属性范围）

利用每一类得到的支持向量来确定每一条规则各属性值的范围。$x_{ij} \in (x_{j\_\min}, x_{j\_\max})$。其中 $x_{j\_\min}, x_{j\_\max}$ 表示得到的支持向量在一个属性上的极值。求出每个分类的各个属性值域，得到了规则在这个属性上的范围。

4. 细分规则（对 3 步得到的规则进行细分）

对于一些空间分布非凸的数据集来说，由 3 步得到的规则可能存在支持度很大、置信度很小的情况（支持度可以通过第一步来计算，置信度则由第二步求得），这时需要对规则进行细分。预先设置当置信度小于 50%的时候，则对于该类样本进行 1、2、3 步循环处理，直到得到的规则的置信度满足条件。

5. 调整得到的规则（根据准确率，以及支持度与置信度）

得到细分的规则后，通过测试样本来计算出准确率。如果得到的指标（准确率等）不合要求，则通过调整支持向量参数来调整最初的规则的数目和精确程度，整个算法的流程图如下。

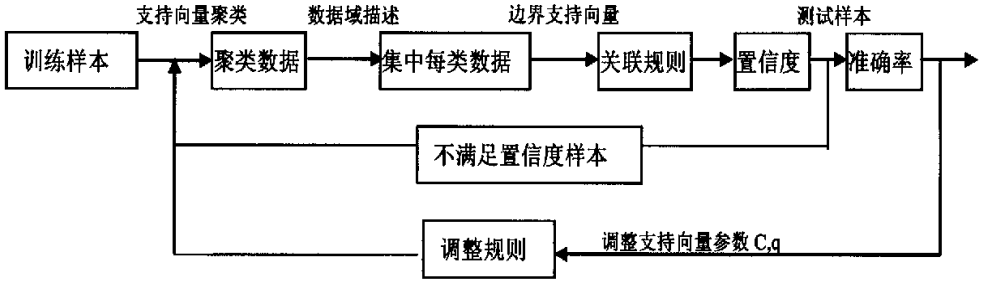


图 3-3 支持向量提取规则流程图

3.4 仿真研究

3.4.1 仿真 1：标准数据集仿真

试验采用的标准的 benchmark 数据集[51]中的 IRIS 和 WINE 数据特征如下，

表 3-1 试验标准数据集

数据集	属性数	类数	数据量
IRIS	4	3	150
WINE	13	3	178

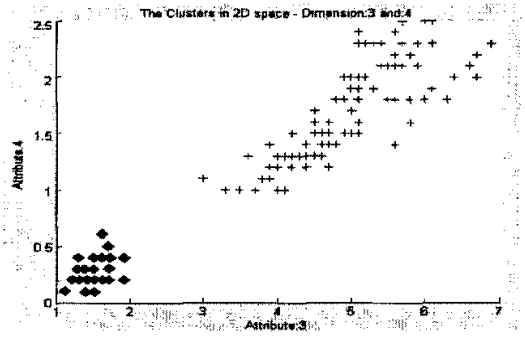
(Iris 是一种鸢尾属植物) 在数据记录中,每组数据包含 Iris 花的四种属性：萼片长度，萼片宽度，花瓣长度和花瓣宽度，三种不同的花各有 50 组数据。这样总共有 150 组数据或模式。

而 WINE 数据集是对意大利 3 个不同地区生产的一种葡萄酒在酿造过程中，通过分析其重要的 13 个化学属性而得到的数据集。

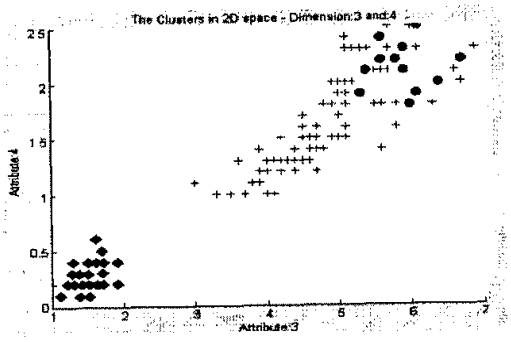
在算法中，经过聚类后，数据被分成多类，聚类的数量直接决定了得到规则的数量，这可以通过调整支持向量的参数来确定。下面以 IRIS 数据集为例来描述支持向量关联规则提取方法,以及参数调整对结果地影响。

惩罚因子 C 取 0.8，q 的选择采用试探性方法，取值越大聚类越细，得到的规则越多，图 3-4 给出了根据 q 的变化聚类数目相应变化的图形。

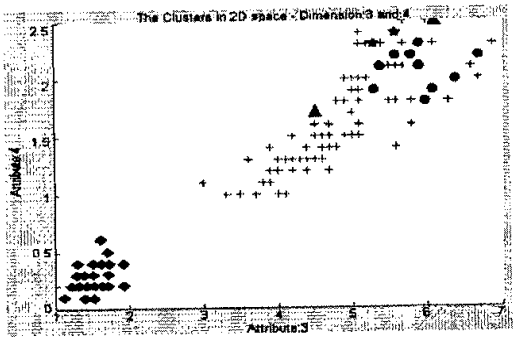
以聚成三类为例，其中每一类通过数据域描述,找出支边界持向量把数据点进行聚集，去除孤立点，如图 3-5（其中实心圆表示支持向量，方形表示孤立点）。



二类（q=1）



三类(q=3)



五类(q=5.5)

图 3-4 IRIS 数据集支持向量聚类图(属性 3，4 维上的平面图)

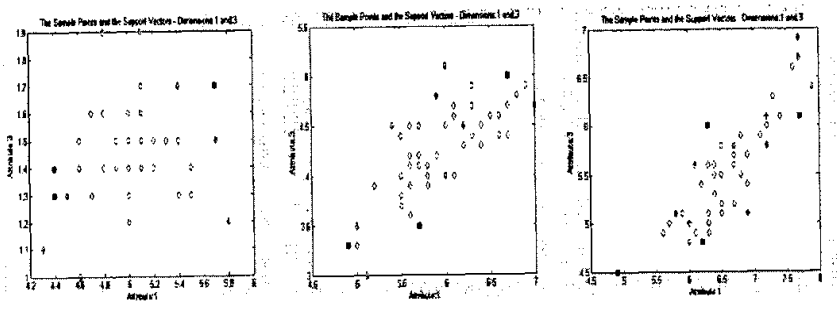


图 3-5 IRIS 数据集分成 3 类后，每类的支持向量数据域描述图（属性 1，3 维的平面图）

根据数据域描述中得到支持向量，通过其各属性的极值来得到关联规则如下(其中 C 表示 Class),

表 3-2 IRIS 数据集聚类成三类得到的规则

IRIS	规则	支持度	置信度
C1	4.5<sepal-length<5.5, 3.1<sepal-width<3.8 1.0<petal-length<1.7, 0.1<petal-width<0.5	34.33%	96.50%
C2	4.9<sepal-length<6.2, 2.0<sepal-width<3.4 3.3<petal-length<4.9, 1.0<petal-width<1.8	33.66%	88.50%
C3	4.4<sepal-length<7.9, 2.2<sepal-width<3.5 4.1<petal-length<6.6, 1.8<petal-width<2.5	35.24%	84.00%

对表 3-2 的结果进行测试，预测结果准确率达到了 87.6%。说明了方法提取出的关联规则的正确性。

表 3-3 IRIS 数据集聚类成五类得到的规则

IRIS	规则	支持度	置信度
C1	4.5<sepal-length<5.5, 3.1<sepal-width<3.3 0.8<petal-length<1.6, 0.1<petal-width<0.3	34.33%	96.50%
C2	4.9<sepal-length<7.6, 2.0<sepal-width<3.4 3.3<petal-length<6.0, 1.2<petal-width<2.3	48.61%	83.20%
C3	6.3<sepal-length<7.7, 2.6<sepal-width<3.0 5.0<petal-length<6.6, 1.8<petal-width<2.3	15.42%	82.32%
C4	6.7<sepal-length<7.3, 2.8<sepal-width<3.0 5.5<petal-length<6.2, 1.9<petal-width<2.1	3.33%	86.51%
C5	7.4<sepal-length<7.7, 2.6<sepal-width<2.8 6.4<petal-length<6.6, 1.8<petal-width<2.2	3.33%	100.00%

表 3-3 说明利用支持向量聚类，通过调整支持向量参数可以得到的更加细致的规则。测试得到准确率为 88.5%。表 3-4 中给出了 WINE 数据集分成 3 类的情况，测试得到的准确率为 86.4%。可以去掉一些归类后具有交叉值的属性来简化规则，如上面的 sepal-length 和 s-width。

表 3-4 WINE 数据集聚类成三类得到的规则

WINE	规则	支持度	置信度
C1	c1[12.78,14.51], c2(0.83,2.4) c3(2.04,3.05), c4(17.8,20.6) c5[60.0,88.5], c6(1.04,3.22) c7[1.57,2.31], c8[0.13,0.54] c9(1.23,1.43), c10[3.46,7.55) c11(0.54,1.41), c12[2.27,2.34] c13[987.5,1450.0]	42.15%	87.23%
C2	c1[12.15,12.78), c2(1.43,2.1) c3(1.45,2.03), c4[17.9,25.6] c5(67.2,88.5), c6(0.98,3.82) c7[1.57,2.31]c8(0.34,0.66) c9[1.27,3.58], c10(1.28,7.0) c11[0.97,1.29), c12(1.56,3.75) c13(289.0,1004.0)	40.89%	85.22%
	c1(13.4,14.83), c2(0.74,5.2) c3[2.03,3.11], c4(24.1,30.0)	26.97%	86.17%

	c5(80.0,162.0),c6(1.02,3.66)
C3	c7(0.34,0.97)，c8(0.23,0.66)
	c9(1.35,3.20),c10(5.30,12.0)
	c11(0.58,0.78),c12(1.27,2.15)
	c13(567.0,1568.0)

支持度和置信度也通过支持向量参数来调整，聚类数目越小，支持度越大；数据域描述越严格，置信度越小。通过对未知样本按规则归类，可以得到方法的准确率。两种数据集的多类提取规则的准确率都达到了80%以上。采用了SMO算法之后，程序执行的效率大大提高了[52]，和不采用SMO方法的程序相比，计算时间缩短了50%以上。

3.4.2 仿真2：某冶炼企业制酸浓度关联规则挖掘

输入属性为某冶炼公司的烧结烟气WSA制酸过程生产数据，整个流程分为三段转化、熔盐换热、采用普通空气作为冷凝介质冷凝成酸。试验数据以冷凝酸浓度为目标，共包括9个相关属性，450个稳态化后的训练样本。先对一些连续属性离散化，采用论文方法，得到结果如下表（表中T表示温度，Q表示流量）

表3-5 制酸浓度关联规则表

浓度	规则	支持度	置信度
高	烟气Q(29.5,40.0),净烟气T(30.1,39.75) SO2含量(1.44,2.25),换热器T(411.0,413.9) 入T(405.9,407.3)，出T(470.9,481.8) 烟入T1(350.2,400.2),烟入T2(402.0,430.0) 尾气T(81.6,81.8)	11.51%	85.38%
中	烟气Q(39.5,43.0),净烟气T(41.0,42.5) SO2含量(1.44,2.25),换热器T(378.0,402.8) 入T(395.0,403.2)，出T(461.3,468.9) 烟入T1(352.0,410.0),烟入T2(433.0,501.0) 尾气T(82.3,83.8)	75.35%	87.40%
低	烟气Q(72.5,79.7),净烟气T(43.3,42.99) SO2含量(1.44,2.25),换热器T(352.0,366.2) 入T(390.2,392.3)，出T(454.0,459.0) 烟入T1(354.0,431.0),烟入T2(401.0,503.0) 尾气T(84.3,84.5)	17.18%	83.60%

实际工业过程中，由于工况复杂，部分数据难以采集，得到的训练样本数量

较少，所以可以发挥支持向量机处理小样本的优势[53]。经过筛选，去掉有交叉值的属性,并分析实际工况得到实用的规则：净化后烟气温度(37.61,39.75)，换热器出口温度(411.9,430.3)，转化塔入口烟气温度(405.9,407.7)，尾气温度(81.6,81.8)，得到的制酸浓度高。

分析规则发现换热器温度高，得到酸浓度高，这是以前操作没有发现的规律。经过理论分析，换热器出口温度在(411.9,430.3)范围内时，进入反应的烟气温度处在催化剂的最佳温度上，从而提高了反应效率，得到了很好的结果。这对于生产具有很强的指导意义，也证明了方法的实用性。采用 SMO 算法后，训练的速度大大提高了，使得算法更具实时性。而且本方法还可以应用于解决多目标优化问题[54]。

3.5 小结

本章提出的支持向量提取规则的方法，充分发挥了支持向量机的优势，与其他关联规则提取方法相比，具有提取规则速度快的优势，为关联规则提取提供了一个新思路。通过仿真证明了其有效性和可行性。此外，支持向量机方法在小样本上的优势使得本方法在故障诊断问题中也可以得到较好的应用。如何能够得到最合适的规则数目以提高准确率是进一步的工作。

第四章 株冶数据挖掘软件

4.1 引言

现今，软件市场中存在着多种数据挖掘工具，其中具有代表性的如 SAS、SPSS 等国外的大型统计软件，但是国产的数据挖掘软件则是难得一见，专门针对工业数据挖掘的软件更是凤毛麟角。国内的大多数企业一般都采用国外大型的统计软件，但是这需要付出相当高的软件购买和使用费用，实际应用效果也不甚理想。

本章中详细介绍了为湖南株洲冶炼集团硫酸厂开发的工业数据挖掘软件。内容从项目任务分析到软件的功能模块设计都做了比较详细的阐述，并重点介绍了作者的主要工作。大量的图示不仅能使读者直观的看到软件的设计思路，而且能够更好的理解工业数据挖掘软件的开发过程，建立更加深刻的印象。文字部分主要针对作者的主要工作、项目的开发过程和功能进行了大量的说明，并在一些章节中介绍了部分主要功能的设计思路和算法流程。

4.2 数据挖掘软件发展

数据挖掘从提出以来，就相应产生了大量的数据挖掘软件。至今为止，归纳起来可以分为四代产品[55]。下面就它们的发展过程做简要介绍。

4.2.1 数据挖掘软件发展

4.2.1.1 第一代数据挖掘软件

特点：

1. 支持一个或少数几个数据挖掘算法；
2. 挖掘向量数据（vector-valued data）；
3. 数据一般一次性调进内存进行处理。

缺点：在数据足够大，并且频繁变化的情况下，需要利用数据库或者数据仓库技术进行管理，第一代系统显然不能满足需求。

典型代表：CBA

4.2.1.2 第二代数据挖掘软件

特点：

1. 与数据库管理系统（DBMS）集成；
2. 支持数据库和数据仓库，和它们具有高性能的接口，具有高的可扩展性；
3. 能够挖掘大数据集、以及更复杂的数据集；
4. 通过支持数据挖掘模式（data mining schema）和数据挖掘查询语言增加系统的灵活性；
5. 典型的系统如 DBMiner，能通过 DMQL 挖掘语言进行挖掘操作。

缺点：只注重模型的生成，如何和预言模型系统集成导致了第三代数据挖掘系统的开发。

典型代表：DBMiner

4.2.1.3 第三代数据挖掘软件

特点：

1. 和预言模型系统之间能够无缝的集成，使得由数据挖掘软件产生的模型的变化能够及时反映到预言模型系统中；
2. 由数据挖掘软件产生的预言模型能够自动地被操作型系统吸收，从而与操作型系统中的预言模型相联合提供决策支持的功能；
3. 能够挖掘网络环境下（Internet/Extranet）的分布式和高度异质的数据，并且能够有效地和操作型系统集成；

缺点：不能支持移动环境。

典型代表：SPSS Clementine

4.2.1.4 第四代数据挖掘软件

特点：

1. 目前移动计算越发显得重要，将数据挖掘和移动计算相结合是当前的一个研究领域；
2. 第四代软件能够挖掘嵌入式系统、移动系统、和普遍存在（ubiquitous）计算设备产生的各种类型的数据。

目前还没有很成形的代表。

4.2.2 数据挖掘软件市场占有率

现在国际市场上常用的数据挖掘软件大约有 150 多种，最知名的当数 SAS

和 SPSS，下图是一个抽样调查，在一定程度上反映了国外在从事数据挖掘工作中对数据挖掘软件的使用情况。

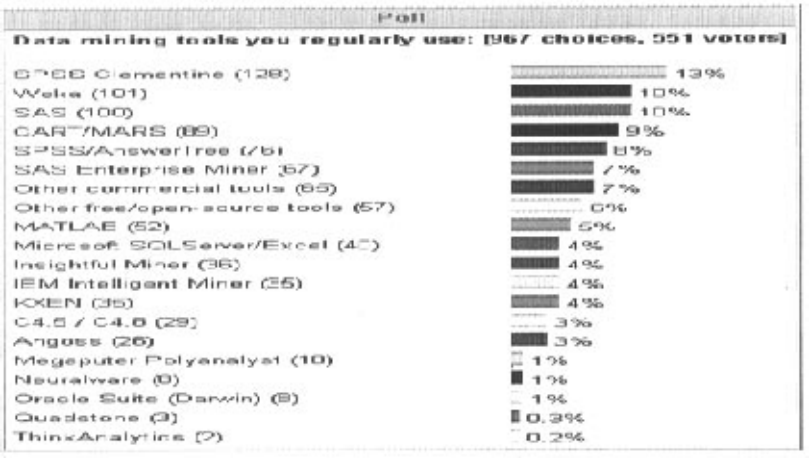


图 4-1 抽样调查的数据挖掘软件使用情况

比较常见的数据挖掘算法有：聚类分析、决策树、神经网络、规则归纳。一般比较好的数据挖掘工具都会支持这几种算法。国内企业常用的数据挖掘工具有以下几种[56]：

(1)SAS 系列产品

SAS/EM(Enterprise Miner)是一个图形化界面、菜单驱动的、拖拉式操作的、对用户非常友好且功能强大的数据挖掘集成环境。其中集成了：数据获取工具，数据抽样工具，数据筛选工具，数据变量转换工具，数据挖掘数据库，数据挖掘过程，多种形式的回归工具，为建立决策树的数据剖分工具，决策树浏览工具，人工神经元网络，数据挖掘的评价工具等。

SAS 在数据挖掘方面地位最高，但价格过于昂贵。而且每年都要交付使用费，不符合国内用户的习惯。在国内，上海宝钢配矿系统和铁路部门在春运客运研究中都应用 SAS 软件。

(2)IBM DB2 Intelligent Miner

Intelligent Miner 采用了多种统计方法和挖掘算法，主要有：单变量曲线、双变量统计、线性回归、因子分析、主变量分析、分类、分群、关联、相似序列、序列模式、预测等。

Intelligent Miner 通过其独有的世界领先技术，例如自动生成典型数据集、发现关联、发现序列规律、概念性分类和可视化呈现，可以自动实现数据选择、数据转换、数据挖掘和结果呈现这一整套数据挖掘操作。若有必要，对结果数据集还可以重复这一过程，直至得到满意结果为止。根据 IDC 的统计，Intelligent Miner 目前是数据挖掘领域最先进的产品之一。它采取客户/服务器（C/S）架构，并且它的 API 提供了 C++类和方法。 IBM 在国内的成功案例有大连农行和几个省市的移动通信公司的例子。

(3)SPSS 系列产品

Clementine 是 SPSS 的核心挖掘产品，它提供了一个可视化的快速建立模型的环境，被誉为第一数据挖掘工具。使用它，企业可以将数据分析和建模技术与特定的商业问题结合起来，找出其他传统数据挖掘工具可能找不出的答案。组成部分包括数据获取、探查、整理、建模和报告都使用一些有效、易用的按钮表示，用户只需用鼠标将这些组件连接起来建立一个“数据流”，可视化的界面使得数据挖掘更加直观和具有交互性，从而可以将用户的商业知识在每一步中更好的利用。Clementine 所使用的分析技术包括神经元网络、关联规则和规则归纳技术。Clementine 支持顾客剖析、时序分析、市场售货篮分析和欺诈行为侦测。

SPSS 的另一种重要的挖掘产品 AnswerTree 可以帮助用户确认细分市场及其模式，建立顾客档案资料，挖掘隐藏市场趋势。应答树应用的分析运算法则：两类 CHAID、分类和回归树、QUEST。 DecisionTime 2.0 及 WhatIf 2.0 帮助用户建立准确的预测，并利用此预测制定计划。

Clementine 功能相对稍弱，但胜在简单易用，而且价格比 SAS 更加经济。所以，在国外使用 Clementine 的人比使 SAS 的人多。Clementine 以前的版本不支持中文，所以在国内销量很低。不过新的 Clementine 7 是用 Java 语言编写的并支持中文，国内有几个省市的移动通信公司正在使用。

在了解了数据挖掘软件的使用现状之后，下面介绍一下开发的株冶数据挖掘软件的情况和作者在其中的主要工作。

4.3 株冶数据挖掘软件整体框架设计

整个软件的框架根据用户指定的任务来设计，开发过程也是遵循这个任务框架体系来完成的。图 4-2 中只给出了整个软件的框架图，纵向上看，完全是一个

数据挖掘的流程，从数据采集开始一直到结果呈现；横向上看，则是根据软件的四大部分功能设计，完成了用户所需要的任务。当然在开发过程中，还增添了不少细节性的内容，框架图中并没有给出。

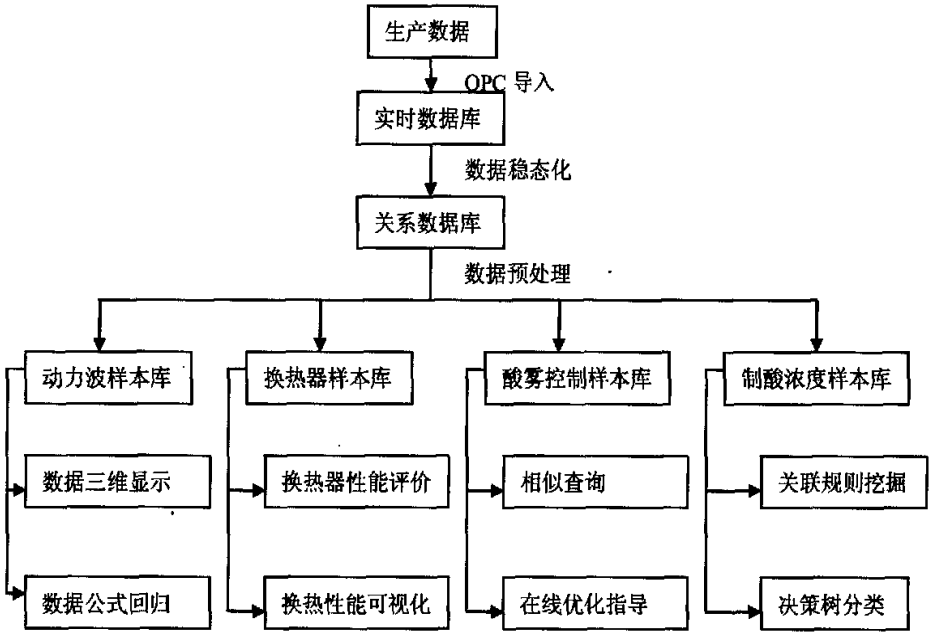


图 4-2 软件框架图

本软件从结构上看，采用了典型的数据挖掘过程来设计，即数据预处理，数据挖掘算法，知识呈现三个层次，如上面的框架图所示，即从采集生产数据的实时数据库开始，将数据导入样本数据库，进行样本数据库中的数据预处理，存储处理过的可以进行挖掘的完备数据。得到完备数据后，在这之上进行各种数据挖掘和替他算法设计，以及用户界面设计，通过各种数据挖掘手段得到信息，最后向用户呈现出数据挖掘结果，即知识表达。

本软件从技术上看，利用了微软公司最新的.net 技术进行开发，语言选用主推语言 C#。以 XML Web 服务为前提而设计的.NET 与 Web 服务具有极高的亲和性，这使得开发出的程序具有更高的通用性，尤其是为以后该数据挖掘程序的维护和二次开发提供了便利，而且其简单易用的特点也加快了开发的速度。实时数据库采用浙大中控软件公司自行开发的 ESP-iSYS-A 实时数据控制平台，通过 OPC 接口采集生产数据。对于样本库的选择，程序最初拟采免费的 ACCESS 数

据库，但是经过能力测试与估算，发现其性能和容量有限，难以满足工业数据量和本软件的要求，故最终改用功能更强的数据库 SQL-SERVER。

本软件从所涉及到的理论算法上看，主要包括了关联规则提取、数据分类、样本聚类、回归预测等常用的数据挖掘算法。此外还包括了数据三维图像显示、阀位值在线优化指导、组态位号、历史数据相似查询等其他算法。其中涉及到了很多理论与实际相结合的部分，如数据库的数据预处理方法等等。总之，可以称为一个典型的工业数据挖掘软件。

本软件从内容上看，根据 WSA 过程，以及用户提出的需要，分为四个主要部分：动力波数据三维图像显示、换热器性能评估、酸雾控制优化与在线指导、制酸浓度关联规则挖掘与决策树分类等主要功能。软件设计的四个主要菜单也是针对的这四个部分。此外为了完善一些相关的其他功能，软件中又加入了历史记录查询和样本点聚类等算法。在下一节中，将主要介绍作者参与设计并实现的三部分内容来介绍本软件的各功能块。

4.4 软件设计与实现

4.4.1 数据库的设计

在数据挖掘软件中，数据库的设计十分重要，因为在数据库部分，程序要做很大一部分工作。可以毫不夸张地说，如果软件是一座大厦的话，数据库部分则是整个软件的基石。只有打好了下层基础，才能建设更好的上层建筑。这是作者在软件中的主要工作，其主要内容涉及两部分工作，

1 数据库的选择

现有的数据库种类很多，下面对一些常用的数据库做一下简单的介绍，

ORACLE 在技术上一直处于领先地位，可以用于不同的操作系统，是名副其实的通用数据库，是大型商务系统的首选数据库之一；SYBASE 数据库在 90 年代初期使用率很高，现在市场占有率在缩小，不过仍然有一些老客户在使用。它可以用于不同的操作系统；DB2 原来用于 IBM 的大中型机器上，现在也可以安装在 PC 机上。由于 IBM 的原因，其市场销售份额很高。软件界面类似 MSSQL，操作相对比较复杂；SQL-SERVER 是目前一个界面友好、上手较快的数据库软件，性能也随着版本一步步提高，随着微软的扩张（也因为其易用性）一步一步

的扩大了自己的领地。但是目前只能用在 windows 操作系统上；MYSQL 是免费的数据库软件，它是最快的 WEB 数据库，但在功能上相对前面的数据库较弱一些；FOXPRO 是传统的数据库软件，操作简洁，开发速度快，也支持 SQL 查询；ACCESS 是微软的桌面型数据库，支持的数据量比较小。

数据库的选择要根据项目的需要来确定，量体裁衣。出于费用和使用方便的考虑，本项目最初考虑采用 ACCESS 数据库，它的数据最大容量为 2GB。在实际选择的时候，我们首先对株冶的 WSA 过程的数据量进行估算。实时数据库设置为 10 秒存储一次数据（每次存储共有 80 个位号）。这样算来每分钟大约 2KB 的数据量，一年的数据量大约 1.05GB。在数据预处理后数据量大约为原来总量的 1/3。随着存储数据量的增加，数据库在查询插入等性能上有一定程度的下降。如果用户导入的是 2 年以上的数据，则不能保证 ACCESS 数据库还能正常使用。另外通过测试，在查询存储速度方面， ACCESS 数据库的表现也并不令人满意（导入 1 个月的数据进入样本库需要 30 分钟左右）。因此，项目最终采用功能更强的 SQL-SERVER，其理论上存储量达到 10T 以上，而且通过测试，其查询存储速度也非常快（导入 1 个月的数据只需要 10 秒种）。

总之，数据库的选择很重要，正是由于选择了合适的数据库，本项目才能继续正常的开发。下面介绍一下数据库的结构和在软件中如何操作。

2 结构和操作的设计

选择了数据库之后，需要考虑如何设计数据库的结构，即总共需要的表和表与表之间的关联。本项目的思路是：每次用户从实时数据库导入一定时间段的数据到样本数据库的总表 zhuye_total 中，数据预处理也在这步完成（此步一般导入比较大的时间段，通常为一个季度或者一年，比较耗时但较少使用）。然后每次每部分进行数据挖掘操作的时候，都要设置预挖掘数据的时间段（包含在前面导入的时间当中），即在样本数据库中从总表根据用户设定的时间段把数据从 zhuye_total 导入各个部分的子表中（wave, conversion, control, thickness），算法所用数据实际上是在子表中进行操作的。由于数据库中的表比较简单，并没有很强的关联性，项目中没有涉及到数据库之间的关联情况。

在工业数据挖掘软件中，数据库涉及到了实时数据库和样本数据库两个。所以如何在它们之间过渡数据也十分重要。在实时数据库导入样本数据库的过程

中，通过预处理欲找出数据中的稳定工况点，作为进一步分析的对象。具体采用固定长度窗口法对各变量属性值进行分段，对段内数据点进行线性回归，根据回归误差以及斜率值来判断该时间段内的点是否处于稳定状态。在数据线性回归后，还要进行匹配，只有在一个时间段内，所有的属性都具有稳定数值才是有效的数据。其中，可供调整（用户定义）的参数有：窗口长度，误差率，扰动率。在项目中，根据试验的实际效果取窗口长度采用 5 分钟，误差率采用 0.08，扰动率采用 0.2。通过初步的预处理，就可以在样本库中得到稳定的数据点的集合。

在软件中，也允许用户在一定的权限内查看和更改数据库，以方便用户及时发现问题或者人工去掉孤立点。程序中对操作数据库设计了 2 个方式，首先在主菜单中，设置有数据库菜单如下，

图 4-3 数据库菜单

在这个菜单项下面，可以对本项目的数据库进行操作，三部分的功能为，

数据预处理参数：设定主要设备的延时和一些工况参数。

样本库操作：设定样本库的参数，把从实时库导入的数据进行预处理。由于换热器部分不是从实时库导入数据，在这部分提供了 EXCEL 表导入接口。

实时数据导入：从实时库中导入实时数据到样本库的总表当中，这也是程序的最初始部分。

其次，在动力波，换热器，酸雾控制，制酸浓度每一部分具体使用数据挖掘算法的时候，都设计了数据库操作的子菜单。主要实现两大功能：样本库数据导入，把数据从 zhuye_total 表中根据用户所给的时间导入到子表中，即导入需要挖掘的时间段的数据；实现数据的手工处理，用户可以根据需要把样本库中不合格的数据删除，另外也可以根据需要修改一些数据的数值。

完成了对数据库的设计之后，可以在此基础上设计数据挖掘算法。

4.4.2 数据挖掘算法设计

本软件中涉及到了四个典型的数据挖掘算法。下面分别介绍了各种算法的实际设计情况，这些也是作者工作的重点之一。

1 制酸浓度关联规则挖掘

根据用户输入的参数（支持度，置信度），查询出制酸浓度与整个过程中主要变量之间的关联来，并通过程序处理以最简单的方式呈现出来。程序中采用了经典的 apriori 算法（详细内容如第三章所述）来进行关联规则挖掘。实现过程中，先做好 apriori 子程序，然后嵌入到主程序中。

算法的伪码如下，

```
procedure AprioriAlg()
begin
    L1 := {frequent 1-itemsets};
    for ( k := 2; Lk-1 0; k++ ) do {
       Ck= apriori-gen(Lk-1) ; // new candidates
    for all transactions t in the dataset do {
      for all candidates c Ck contained in t do
            c:count++
      }
      Lk = { c Ck | c:count >= min-support}
    }
    Answer := k Lk
end
```

结果首先以表格的形式显示出来，如 4-4 左图所示，为了方便用户理解，在显示的过程中，作者加入了很多解释性的语言，使得结果更加便于理解。同时为了让用户能够更加清楚明了的看出和理解得到的关联规则，又设计了关联规则图来表示属性之间的关联，即把各个属性都列在横轴上，纵轴显示了属性的数值范围，用直线把有关联的属性数值之间连接起来，最后一个纵轴为目标属性，其效果如 4-4 右图所示。

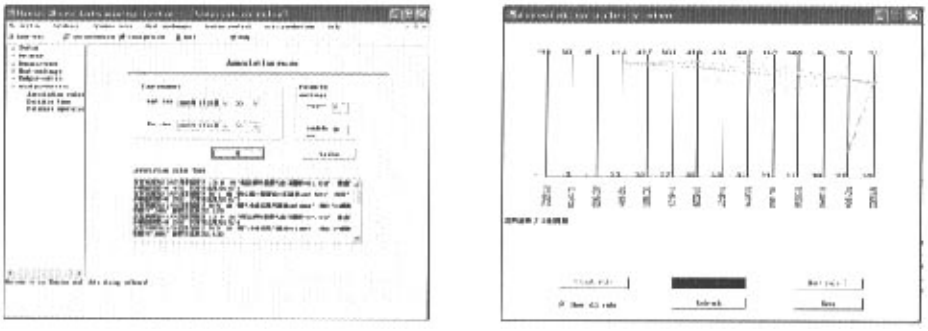


图 4-4 制酸浓度关联规则图

2 制酸浓度决策树分析

根据输入数据参数，画出数据的决策树，能够比较直观的显示出数据的关系。

分类是通过分析历史数据，推导出数据的一个抽象的推广描述，并利用该描述来对未来数据进行预测并进而进行决策。分类在数据挖掘中是非常重要的一项任务，分类的目的是利用一个分类函数或分类模型把数据库中的数据项映射到给定类别中的某一个，通过对训练样本数据的分析处理，发现指定的某一样本是否属于某一个特定数据子集的规则。

分类算法涉及许多领域，包括统计学方法、机器学习方法、神经网络方法等等。其中，机器学习方法中的决策树分类方法是分类中最常见的一种方法，它具有速度快、准确性高、模型简单易理解、容易转换成分类规则、容易转换成数据库查询语言等许多优点，是一种较为实用的分类方法。目前，国内外对于决策树分类方法的研究较为成熟，已经提出了 ID3、C4.5、CLS、CART、SLIQ、SPRINT 等多种算法。本例子采用一种 C4.5 决策树分类方法，得到的结果如下图。算法的流程图如下，

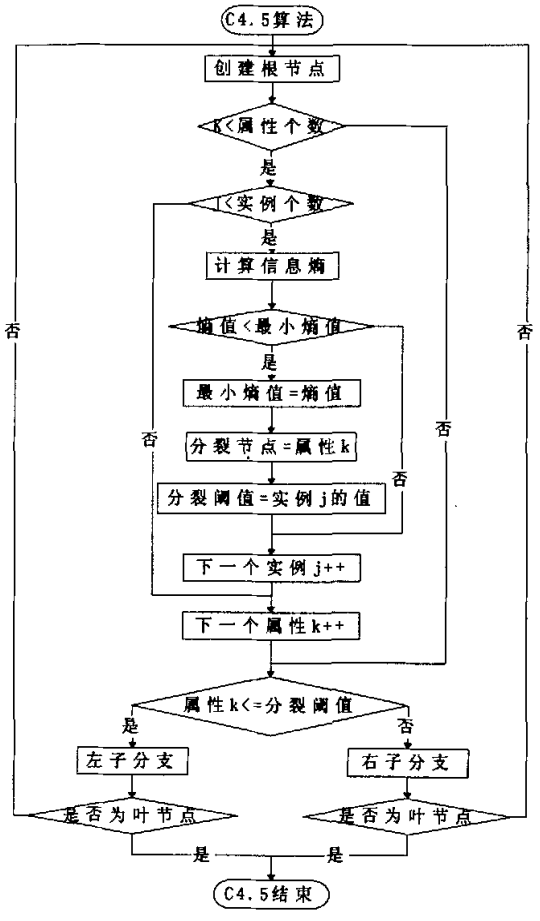


图 4-5 C4.5 算法流程图

软件界面如下所示，根据输入参数得到最终的决策树，

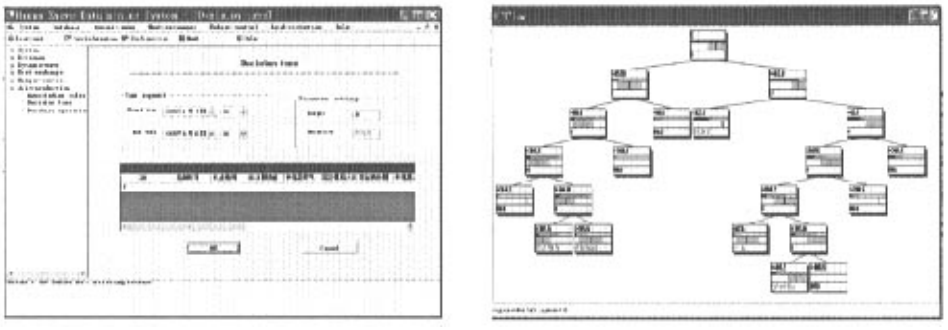


图 4-6 制酸浓度决策树图

3 酸雾控制典型样本聚类

酸雾控制部分的数据量较大，该子菜单为这部分提供了数据聚类功能，可以根据聚类得到的结果，选择出一定数量的“典型样本”来进行数据分析，看出数据库中样本点的特征。聚类算法采用的是 k-mean 算法来进行，在编程实现的过程中， k-mean 算法属于最著名，最常用的划分方法的一种。它最吸引人的地方就在于它在处理大数据集时的高效性。算法的主要思想是：给定一个包含 n 个对象的数据库，以及要生成的簇的数目 k。k-mean 算法将数据对象组织为 k 个划分($k \leqslant n$)，其中每个划分代表一个簇。

算法以距离作为划分准则，使得在同一簇中的对象是“相似的”，而不同簇中的对象是“相异的”。k-mean 算法的处理流程为：

1.随机的选择 k 个对象，每个对象初始地代表了一个簇的平均值（即中心）；

2.对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇；

3.在新产生的 k 个簇的基础上，更新各个簇的平均值，即计算每个簇中对象的平均值；

4.重复 2，3，直到均值不再发生变化。或者直到准则函数收敛，通常采用平方误差准则。

基于此，在选择典型样本之前，用户需要指定出需要得到聚类的样本数量。程序的界面如下图所示：

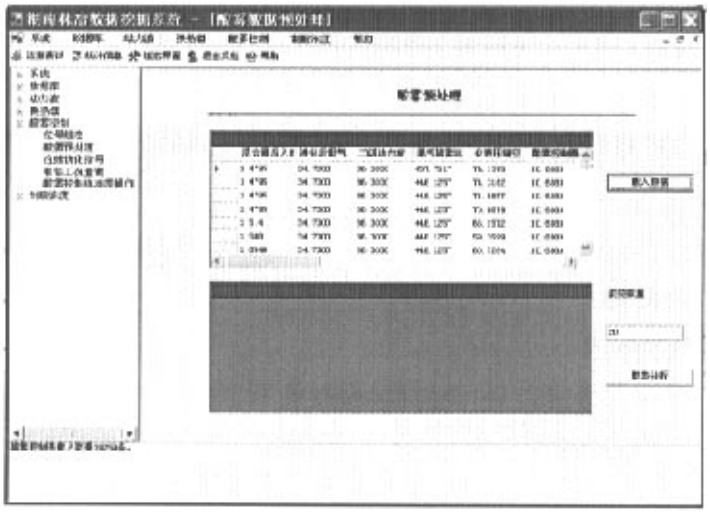


图 4-7 酸雾处理聚类界面

4 动力波洗涤器的压降-流量回归

根据动力波部分三个重要变量：烟气入口压力、压力损失、喷液压力，提供数据回归预测功能。这部分不涉及系统机理分析内容，主要是为用户提供一个直观参照。按参考说明动力波洗涤器的有效操作中存在一个（入口气速、气液比）的驻波区，在驻波区边界上气相压力损失应随着流体动力学特性的变化而产生较明显影响。该部分可以实时预测动力波喷液压力数据的数值，为生产提供参考。具体的算法则是采用线性最小二乘回归的方法，来建立动力波模型，对用户选择区域的样本给出二次多项式回归的解析表达。

最小二乘法是一种数学优化技术，它通过最小化误差的平方和找到一组数据的最佳函数匹配。它是一种采用最简的方法求得一些绝对不可知的真值，从而令误差平方之和为最小的方法。

由于本功能也属于方便用户的附加功能，意在为用户提供一个参考信息，因此投入的工作不是很大。由于采用的线性模型比较简单，效果并不理想，还需要根据动力波部分的具体情况，建立更好的模型，比如可以采用其他的更有效的建模方法。采用最小二乘法得到的回归公式界面如 4-8 所示：

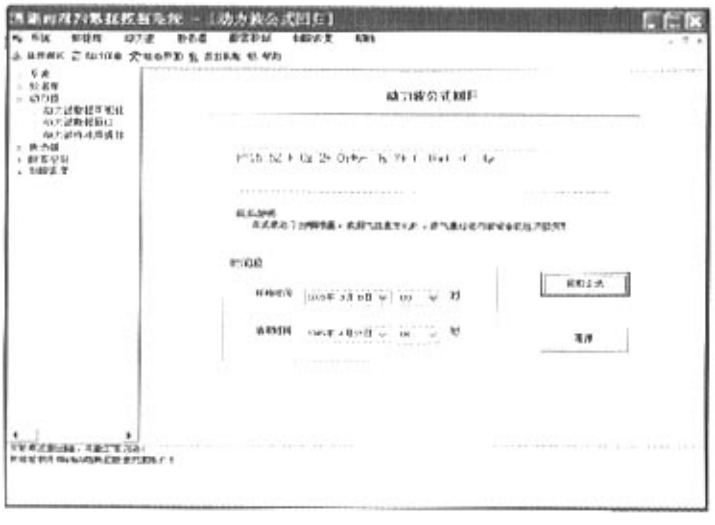


图 4-8 动力波回归界面

4.4.3 软件的界面设计

软件的大部分 GUI 都是作者参与编写的。下面介绍一些主要的界面。

按照项目任务的要求，共设计七个主要菜单项，分别为系统、数据库、动力波、换热器、酸雾控制、制酸浓度和帮助。下面的章节中将重点讲述动力波，换热器，酸雾控制这 3 个主要的功能菜单项，制酸浓度的界面前面部分已经给出。过程描述见附录。

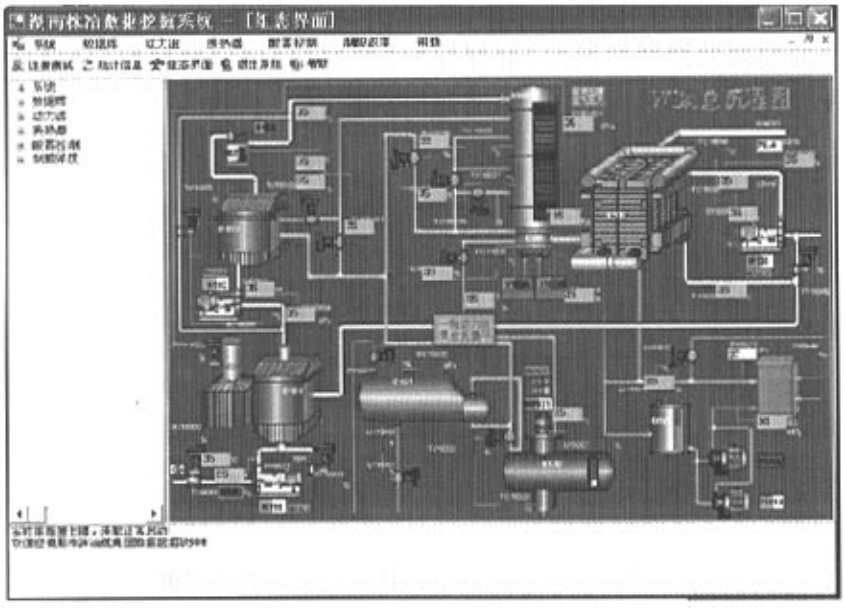


图 4-9 程序主界面

软件的主界面如图 4-9 所示，分为三部分，左侧树状部分能够展开显示整个菜单项功能，方便用户操作；右侧为主要功能显示区域，不运行数据挖掘算法的时候，显示 WSA 的整个工况，方便用户了解整个生产情况；运用算法的时候，显示各个功能设置界面以及大部分数据挖掘结果界面。最下面一栏显示目前程序所处状态，类似于 VC 的状态栏一样，可以显示一些基本的程序运行情况、数据和操作记录情况。

4.4.3.1 动力波三维显示与回归

动力波菜单项如图 4-10 所示，分为三个子菜单，动力波数据可视化，动力波数据回归，动力波样本库操作，下面主要介绍可视化部分界面，

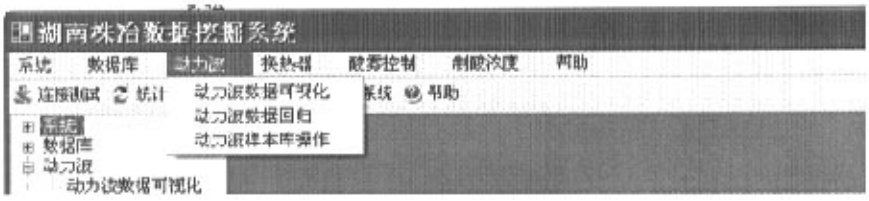


图 4-10 动力波菜单项

动力波数据可视化

在数据可视化部分，完成的任务主要是三维显示数据的功能。软件设计提供给用户动力波部分六个重要参数选择的功能，可以从中任意选出其中三个参数来，查看这三组数据之间的空间分布关系。

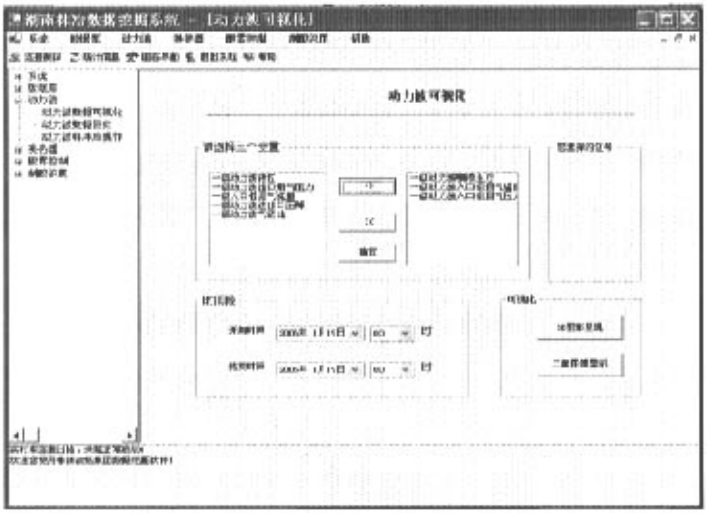


图 4-11 动力波三维显示设置面

这部分涉及到的数据挖掘算法不多，主要是图像显示功能，开发过程采用 DirectX 技术用 VC 语言编写好子程序，在主程序中调用这个子程序，这也是整个软件的思路，即大多数算法都是先编写好子程序，在主程序中调用。在运行的时候，程序设置界面如 4-11 所示。

动力波部分的三维显示是用户根据空间的数据点来观察属性的一个功能，并不需要很多的理论算法研究，而是需要大量的实际操作经验作为基础才能从中发掘出有用的信息，这也正是数据挖掘的一个最大的特点。各种算法只是辅助手段，要得到有价值的结论，最根本的还是需要依赖于专家经验。在程序中，动力波显

示图如下所示，

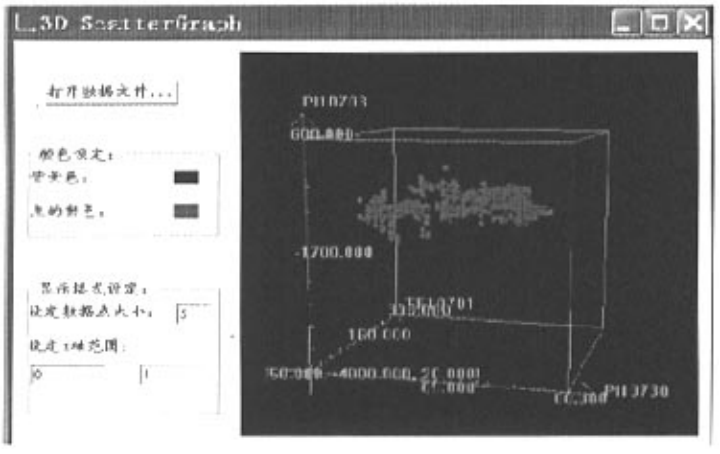


图 4-12 动力波三维图像显示

此外，为了能够更好地方便用户观看数据之间的关系，软件中又添加设计了类似等高线图的二维显示图形，能够把空间的三维图形通过二维平面显示出来。通过它可以更加方便的看出数据点的分布情况，为用户观察动力波数据提供了多种选择。

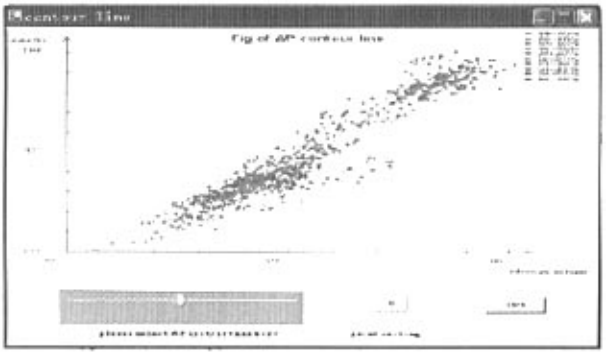


图 4-13 动力波二维图像显示

4.4.3.2 换热器性能评估

换热器菜单项如图 4-14 所示，分为三个子菜单：换热器能力评估、换热器参数可视化、换热器样本库操作，

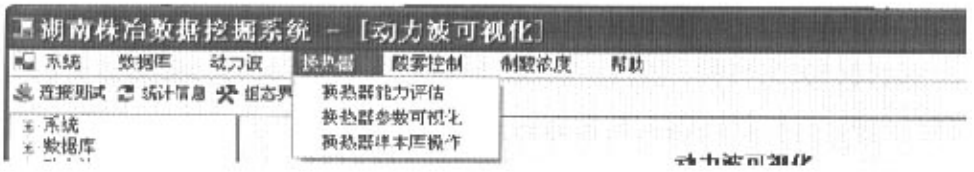


图 4-14 换热器菜单项

换热器能力评估

换热器性能的评估最好的方法就是知道当前的换热系数，与正常的换热系数进行比较。该子菜单提供了计算换热系数的功能，根据输入工况的情况，通过公式计算出相应的换热系数之后，有利于评价换热器当前情况。计算换热系数的难点在于得到计算公式所需的各个参数，这需要和操作人员进行深入交流，查阅一些相关的设计文献，但是在算法和程序上面并没有多大难度可言，程序的界面如下，

图 4-15 换热系数计算

换热器参数可视化

换热器参数可视化的主要功能也是评价当前的换热器性能，只是想为用户提供一个更加直观，更有参考性的途径。算法的基本原理是，查询出和输入工况相似的各个历史工况，查出这些工况的换热系数，并画出曲线，其中当前换热系数用红点标出。根据曲线的趋势和当前所处的位置来评价和判断当前换热器的状况，如果现在的点处在平稳的曲线上，则换热器性能良好，如果处在下降的曲线上，则换热器性能较差。在查询过程中，由于相似工况很容易在集中的时间段出

现，故应该设置一定的时间间隔。

在程序实现过程中，这部分也不存在难点，只是涉及到了历史数据库查询和画图功能，实现起来也相对比较简单。

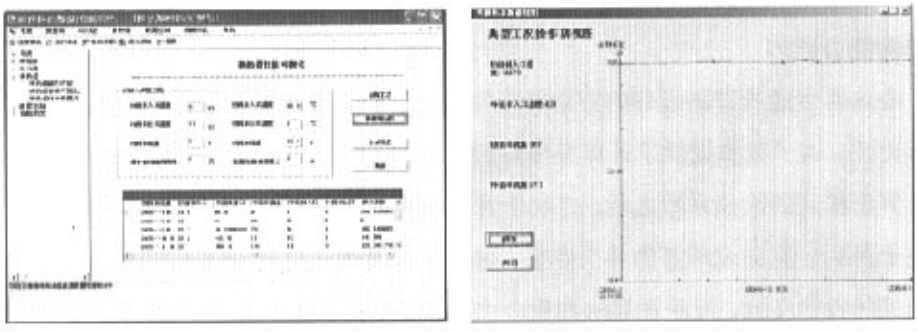


图 4-16 换热器性能评估

4.4.3.3 酸雾控制在线优化指导

酸雾控制菜单项如图 4-17 所示，分为五个子菜单：位号组态、数据预处理、在线优化指导、相似工况查询和酸雾控制样本库操作。下面主要介绍一下在线优化指导的界面和思路，

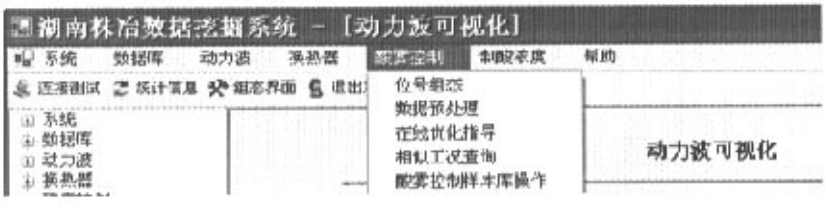


图 4-17 酸雾控制菜单

在线优化指导

根据当前实时数据，提出酸雾控制两个操作阀的在线的操作建议值。在线优化过程的思路是：根据酸雾控制历史数据，通过相似查询的方法来找到到历史数据库中的效果的最佳值，即我们所要给出的建议值。

这部分程序实现主要也是数据库查询，但是难点在于细节处理，由于相似查询得到的样本点容易集中，所以在程序中也需要设定时间间隔。相似度也是用户根据自己的需要来设定的一个数值，需要有一定的经验并不断调整；另外需要根

据目前的操作设定阀位的域值；查询方式也可以多种选择，欧式距离，绝对距离等等。在线优化的界面如下所示，

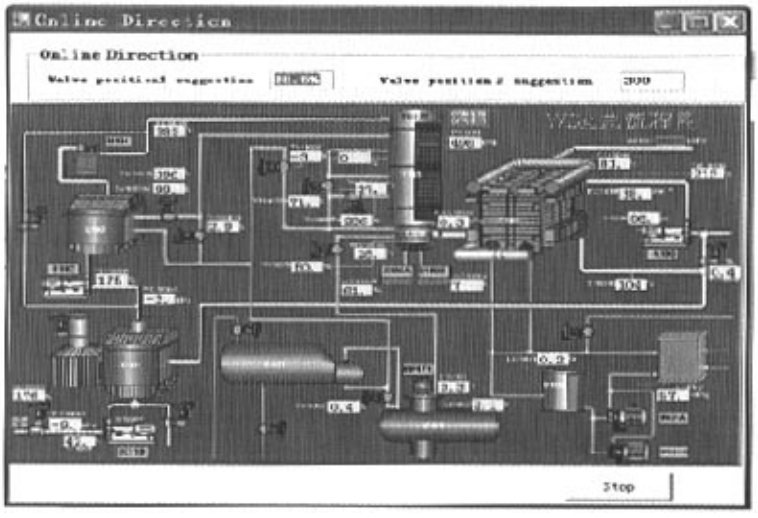


图 4-18 在线优化指导

4.5 小结

本章详细讲述了整个数据挖掘软件的设计、功能和体系结构，并给出了大量的图示加以讲解说明。通过对该软件的编写和设计，加深了作者对工业数据挖掘的整个思路和体系结构的理解和认识，也更加清楚了工业数据挖掘的重点和难点，即数据预处理和专家经验的重要性。当然，本软件还有一些功能没有完善，有许多功能还需要进一步细化处理，而且软件还没有正式投入生产，很多算法只是在很小的数据集上做了性能测试，还没有得到很多有价值的信息，所以还需要时间和运行情况来调整软件的部分算法，并与业内专家多沟通来得到有价值的结论，从而达到数据挖掘的最终目的。

第五章　结论与展望

5.1　论文内容总结

论文从支持向量机和工业数据挖掘的基本概念谈起，具体论述了如何在工业数据挖掘中应用支持向量机方法，文中谈到了两个新的算法并给出了实际应用，最后给出了一个工业数据挖掘软件应用实例。

在第二章中，从数据预处理谈起，论述了如何将支持向量机和主元分析方法结合，在程序中如何实现，讨论了两种核方法结合的优势，并给出了复合肥生产的例子来具体说明如何应用。在第三章中，将支持向量机方法拓展到了数据挖掘领域，提出了一种基于支持向量机的关联规则挖掘算法，并从经典数据集和实际数据库中抽取数据验证，仿真结果显示了该方法的有效性。在第四章中，论文着重阐述了一个数据挖掘软件应用的实例－湖南株洲冶炼集团硫酸厂数据挖掘软件。从软件的设计到各项功能的开发，以及一些数据结果的可视化、软件开发过程中遇到的问题等都做了非常详尽的描述，使读者能够更加具体的了解工业数据挖掘的流程，更加直观的了解数据挖掘软件。

下面再次列出论文的框架图，以方便读者回顾整个论文的内容体系结构，

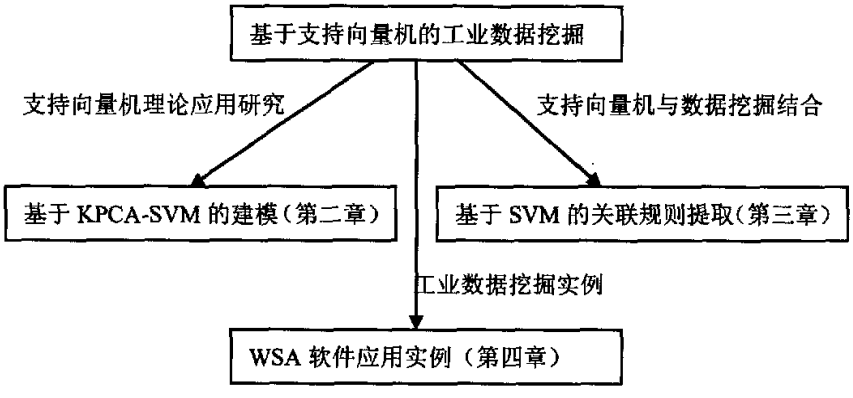


图 5-1 论文结构框架图

5.2　展望

回顾两年来的理论与实际研究工作，可以看到支持向量机和数据挖掘技术

在经历了十多年时间的发展之后，虽然取得了一定的成果，但在无论在理论方面还是在实际应用中都不是很成熟。

就 SVM 来讲，模型选择问题，核函数的选择以及其理论上的意义还需要进一步深入研究探讨，这需要具备坚实的数学基础和严谨的科学态度，这部分理论也是目前支持向量机发展的瓶颈；在应用方面，虽然 SVM 在各个行业都进行了应用，但是并不是所有的效果都十分理想，主要的成果集中在模式识别一块，其他大部分领域都还只是尝试性的研究。

就工业数据挖掘而言，现在国内外应用成功的例子还不多见。另外正如文中讨论的那样，由于工业数据本身的特点，在一定程度上面限制了大多数数据挖掘方法的发挥，这也是论文中探讨支持向量机方法在数据挖掘中应用可行性的目的所在。目前在实际生产中，人们还是更加信任传统的理论方法。在软件开发方面，工业数据挖掘软件还处在刚刚起步的阶段，既没有成型的大型软件也缺少专家经验，还需要不断的积累开发经验。

总之，这两个研究热点虽然被人们所看好，但是要想在实际应用中取得更加显著的成果，还需要做大量的工作，在理论上还有很长的路要走，论文也只是在其中的一些方面做了一些尝试性的研究，希望能给读者一些参考。

作者在攻读硕士学位期间发表论文

已经发表论文：

1 “基于KPCA-SVR方法的复合肥养分含量建模研究”，马朝阳，苏宏业，傅永峰，褚健，中国科学技术大学学报自动化专辑，2005.11. page: 314-321

2 “SVM-Association Rules Extraction Applied in Metallurgical Industries”, Chaoyang Ma, Hongye Su, Jian Chu, International Conference on Intelligent Systems and Knowledge Engineering, 2006, Shanghai.

已经录用论文：

1 “基于支持向量机的关联规则提取”，马朝阳，任佳，苏宏业，褚健，第六界全球智能控制与自动化大会，2006.6 ，大连.

2 “Application of Data Mining Methods in Eluxyl Process”, Jia Ren, Chaoyang Ma, Hongye Su, Jian Chu, The 6th World Congress on Intelligent Control and Automation , 2006.6, Dalian.

致　谢

在浙大、在杭州的六年人生最美好的时光中，母校给了我太多太多。回首2000 年刚入校的时候，我只是一个懵懂的憧憬大学美好生活的少年。而今，我感到自己的蜕变，少了些激情，多了些沉稳，但我依然渴望挑战，渴望新生活。在母校的时光即将告一段落了，至此论文撰写之际，我想借此机会向在这六年中给我帮助的人一一表示感谢。

首先要感谢的是我的导师褚健教授和苏宏业教授，他们严谨的治学态度，渊博的学识，对学科发展动态的准确把握，对问题实质的洞察入微和对行业发展的高屋建瓴以及学术上的高标准严要求，都使我受益匪浅。更为重要的是，两位老师为我提供了非常便利的实验条件和宽松的研究环境，使我能够全身心的投入到学习中去。

在研究生的两年中，先控所和控制系，以及学院里的很多老师、同学在生活、学习上都曾给予我帮助和关心，也使得我有机会认识了很多好朋友。特别是，同实验小组的贾涛，张英，任佳，杨黎刚等师兄师姐，在学术研究和论文撰写方面给了我很大帮助，在这里表示深深的感谢。

最后，我想感谢我的父母和家人，他们在精神上的大力支持，是我学习和前进的动力，也使得我顺利完成本论文的工作。我把论文献给他们，可能他们不是最能理解它内容的人，但却是最希望看到它的人。

参考文献

1. 范明, 孟小峰译, 数据挖掘概念与技术, 北京: 机械工业出版社, page: 14-20, 2001.
2. 邵风晶等, 数据挖掘原理与算法, 北京: 中国水利水电出版社, 2003.
3. Vapnik Vladimir N., The Nature of Statistical Learning Theory, Springer-Verlag, New York, Inc, 2000.
4. Burges C.J.C., A Tutorial on Support Vector Machines For Pattern Recognition, Data Mining and Knowledge Discovery, 2(2) page:1-47, 1998.
5. 张英, 基于支持向量机的过程工业数据挖掘技术研究, 浙江大学博士学位论文, page: 17-20, 2005.
6. Platt J.C., Using Analytic QP and Sparseness to Speed Training of Support Vector Machines, Advances in Neural Information Processing Systems 11, 557-563, 1999.
7. Keerthi S., Shevade S., Bhattcharyya C, et al., Improvements To Platt's SMO Algorithm For SVM Classifier Design, Neural Computation, 13(3), page: 637-649, 2001.
8. Keerthi S., Gilbert E., Convergence Of A Generalized SMO Algorithm For SVM Classifier Design. Machine Learning, 46(1/3), page: 351-360, 2002.
9. LIN Chihjen., On The Convergence Of The Decomposition Method For Support Vector Machines, IEEE Transactions on Neural Networks, 12(6), page: 1288-1298, 2001.
10. Vapnik V. and Chapelle O., Bounds on error expectation for support vector machine, Alexander J. Smola, Peter L. Bartlett, Bernhard Scholkopf , Dale Schuurmans, Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 1999, page: 261-281.
11. Tsuda K., Rätsch G., Mika S., and Müller K. R., Learning To Predict The Leave-One-Out Error Of Kernel Based Classifiers, In Proceedings of ICANN'01, page: 331-338, 2001.
12. Joachims T., The Maximum-Margin Approach To Learning Text Classifiers:

12. Joachims T., The Maximum-Margin Approach To Learning Text Classifiers:

Method, Theory And Algorithms, Ph.D. Thesis, Department of Computer Science, University of Dortmund, 2000.

13. Wahba G., Lin Y., and Zhang H., GACV for support vector machines, Advances in Large Margin Classifiers MIT Press, Cambridge, MA, 1999.
14. Chapelle O., Vapnik V., O. Bousquet, S. Mukherjee, Choosing Kernel Parameters For Support Vector Machines, Machine Learning 46, page: 131-160, 2001.
15. Schölkopf B., Bartlett P., Smola A., Williamson R., Support Vector Regression with Automatic Accuracy Control, In L. Niklasson, M. Bodén, and T. Ziemke, editors, Proceedings of ICANN'98, Perspectives in Neural Computing, page: 111-116, Berlin, Springer Verlag, 1998.
16. Cortes C., Vapnik V., Support-Vector Network, Machine Learning, 20, page: 273–297, 1995.
17. Schölkopf B., Smola A., Williamson R. C., Bartlett P.L., New Support Vector Algorithms, Neural Computation, 12: 1207-1245, 2000.
18. Tax D.M.J., Duin R.P.W., Data Domain Description By Support Vectors, In M. Verleysen, editor, Proceedings ESANN, page: 251-256, Brussels, 1999.
19. Ben-Hui A., Horn D., Sidgelmann H.T., et.al. Support Vector Clustering, Journel of Machine Learning Research, 2: 125-137, 2001.
20. Lin ChunFu, Wang ShengDe, Fuzzy Support Vector Machines, IEEE Trans. on Neural Networks, 13(3), page: 466-471, 2002.
21. 陈念贻, 丁亚平, 叶晨洲等, 支持向量回归一吸光光度法同时测定溶液中的 Ph, Cd, Zn, 计算机与应用化学, 19(6), page: 717-718, 2002b.
22. 陆文聪, 陈念贻, 叶晨洲, 李国正, 支持向量算法和软件介绍, 计算机与应用化学, 19(6), page: 697-702, 2002a.
23. 陈念贻, 陆文聪, 支持向量机算法在化学化工中的应用, 计算机与应用化学, 19(6), page: 674-676, 2002a.
24. 丁亚平, 陈念贻, 吴庆生等, 导数光谱一支持向量机回归法同时测定 NO_3^-、NO_2^-, 计算机与应用化学, 19(6), page: 752-754, 2002.

25. 许建华, 张学工, 李衍达, 基于最小二乘支持向量机的油气判别技术, 模式识别与人工智能, 15(4), page: 507-510, 2002.
26. Hand D. J., Statistics and Data Mining: Intersecting Disciplines, http://diagnosis.xjtu.edu.cn/ acdemic/kdd/others_KDD_Abstract.html.
27. 邓乃扬, 田英杰, 数据挖掘中的新方法－支持向量机, 北京, 科学出版社, 2004.
28. 张杰, 多变量统计过程控制, 化学工业出版社, 2000.
29. 王宏漫, 欧宗瑛, 采用 PCA/ICA 特征和 SVM 分类的人脸识别, 计算机辅助设计与图形学学报, 4-15, page: 416-422, 2003.
30. B.Scholkopf, A.Smola, K.R.Miller, Nonlinear component analysis as a kernel eigenvalue problem, Neural computation, 10(5), page: 1299-1319, 1998.
31. 孔薇, 杨杰, 基于神经网络的非线性 PCA 方法, 计算机仿真, 20-7, page: 94-97, 2003.
32. 李尔国, 俞金寿, 一种基于输入训练神经网络的非线性 PCA 故障诊断方法, 控制与决策, 18-2, page: 183-185, 2003.
33. 朱国强, 刘士荣, 俞金寿, 基于支持向量机的数据建模在软测量建模中的应用, 华东理工大学学报, 9-28, page: 6-9, 2002.
34. 赵广社, 张希仁, 基于主成分分析的支持向量机分类方法研究, 计算机工程与应用, page: 37-40, 2004.
35. 陶卿, 曹进德, 孙德敏, 基于支持向量机分类的回归方法, 软件学报, 15, page: 1024-1026, 2002.
36. Thomas Philip Runarsson, Asynchronous Parallel Evolutionary Model Selection for Support Vector Machines, Neural Information Processing-Letters and Reviews, 3(3), page: 59-67, 2004.
37. Chapelle.O., Vapnik V., Bousquet O., and Mukherjee S., Choosing kernel parameters for support vector machines, Machine Learning, 46, page: 131-160, 2001.
38. Chung K.M., Kao W.C., Sun C.L., Wang L.L., and Lin C.J., Radius margin bounds for support vector machines with the RBF kernel. Neural

Computation, 15, page: 2643-2681, 2003.

39. M-W.Chang, C.J.Lin, Leave-one-out Bounds for Support Vector Regression Model Selection, Neural Computation, 17, page: 1188-1222, 2005.
40. De Jone K. A., Evolutionary Computation for Discovery, Communications of the ACM, 42(11), page: 51-53, November 1999.
41. 刘瑞兰，软测量技术若干问题的研究及工业应用，浙江大学博士学位论文, page: 54-56, 2004.
42. Han.J., Pei.J., Yin Y., Mining frequent patterns without candidate generation, In proc2000 ACM-SIGMOD Int.conf.management of data, dalas, TX, May 2000.
43. Kleinberg J., Papadimitriou C., Segmentation problem, Proceedings of the 30th annual Symposium on theory of computing, ACM, 1998.
44. Daewon Lee, An Improved Cluster Labeling Method for Support Vector Clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence Volume 27, Issue 3, page: 461-464, March 2005.
45. Asa Ben-Hur Biowulf, David Horn, Hava T.Siegelmann, Vladimir Vapnik, Support Vector Clustering, Journal of Machine Learning Research, 2, page: 125-137, 2001.
46. Glenn Fung, Sathyakama Sandilya, R.Bharat Rao, Rule extraction from linear support vector machines, Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, page: 32-40, 2005.
47. 李占斌，孟庆春，一种可提取等式规则的SVM规则提取算法，计算机应用, 24, page: 280-282, 2004.
48. 王鹏，朱小燕，基于 RBF 核的 SVM 的模型选择及其应用，计算机工程与应用, 24, page: 72-73, 2003.
49. 赵妍，文东丽等，从样本数据中提取模糊规则的算法研究，石油化工高等学校学报, 9-17, page: 83-86, 2004.
50. Klemettinen M., Mannila H., Ronkainen P., Toivonen H., Verkamo I., Finding interesting rules from large sets of discovered association rules, In

Proceedings of the Third International Conference on Information and Knowledge Management(CIKM'94), page: 401-407, Gaithersburg, Maryland, ACM Press, 2001.

51. Haydemar Nunez, Rule extraction from Support Vector Machines. Proceedings of ESANN', page: 107-112, 2002.
52. Keerthi S S. et al, Improvement to Platt's SMO Algorithm for SVM classier Design, TR CD-99-14, Dept of Mechs. and Prod.Engin. National CnL of Singapore, 1999.
53. 王华忠，俞金寿，核函数方法机器在过程控制中的应用，石油化工自动化，1，page: 25-30, 2005.
54. Bhaskar V., Gupta S. K., Ray A. K., Applications of multi-objective optimization in chemical engineering, Rev. Chem. Eng., 16(1), 1-54, 2000a.
55. Goebel M., Gruenwald L., A Survey Of Data Mining And Knowledge Discovery Software Tools, ACM SIGKDD Explorations, 1(1) page: 20-33, 1999.
56. 恽爽，胡南军，董俊，陈道蓄，数据挖掘软件现状研究，计算机工程与应用，8, page: 189-193, 2003.

附录 论文所有工艺背景资料

1 复合肥生产装置工艺简介

巨化硫酸厂复合肥装置投资达6700万元，年产20万吨粒状硫基氮磷钾高效复合肥，于2001年投产。采用国内复合肥生产新技术——氯化钾低温转化和喷浆造粒干燥流程生产S-NPK复合肥，即把生产磷酸、磷氨与硫酸钾有机结合起来，不仅取消了磷酸浓缩或料浆浓缩装置，而且降低了高温法生产硫酸钾过程的难度，具有流程短，“三低一少”（防腐材质要求低、能耗低、成本低和投资少）的特点。

复合肥装置由三部分组成：磷酸工段、氢钾工段、复合肥工段。磷酸工段的原料是磷矿石，生产合格的稀磷酸；氢钾工段的原料是KCL和98%以上的浓硫酸，生成的硫酸氢钾与磷酸工段送来的稀磷酸制成混酸送往复合肥工段；复合肥工段将氢钾工段送来的合格混酸生成复合肥。

复合肥生产流程简介

磷酸工段流程简介

生产原理：在合适的工艺条件下，硫酸和磷矿进行反应，生成磷酸及磷石膏，产生的含氟尾气用水进行洗涤，洗涤达标后排放。

主反应方程：$Ca_5F(PO_4)_3+5H_2SO_4+溶液=3H_3PO_4+5CaSO_4\cdot 2H_2O+HF+溶液$

生产流程：由原料岗位、反应岗位和过滤岗位构成。

原料岗位工艺流程：由火车运进的磷矿石经卸料后在磷矿仓库堆存，磷矿经桥式起重机，送入块矿贮斗后，经板式给料机送入粗碎鄂式破碎机，粗碎分筛后，由胶带输送机送入细碎鄂破碎机再次破碎，破碎后的矿粉由胶带输送机送入细矿贮斗，经电子皮带称计量后，和一定配比量的水一起送入球磨机，磨成合格的矿浆后由砂浆泵输送给制酸岗位使用。

反应岗位工艺流程：矿浆贮槽内的磷矿浆由两台矿浆泵分别送至1#、2#反应槽，第一反应槽与罐区送来的硫酸及过滤岗位返回的回磷酸反应，溢流到第二反应槽，与第二反应槽内磷矿浆和硫酸反应后的料浆混合，溢流到消化槽，使料浆得以充分反应，进一步提高磷石膏结晶的过滤性能，消化槽内的料浆由料浆泵送至过滤岗位进行过滤。

二个反应槽和消化槽产生的尾气与过滤岗位产生的尾气一起进入文丘里洗涤器、尾气洗涤器，洗涤合格后，由尾气风机排到烟囱。

过滤岗位工艺流程：由反应岗位送来的约含 $P_2O_5$22%的磷酸料浆，通过抽真空的转台式过滤机过滤后，滤盘下由成品磷酸接管引出的合格成品稀磷酸送到氢钾岗位，由初滤液和一洗液接管引出的洗液作为回磷酸返回反应系统，滤饼经过3级洗涤后，干法排渣至排渣场。过滤机上的尾气送至反应岗位的文丘里洗涤器。

氢钾工段流程简介

生产原理：在合适的工艺条件下，氯化钾和硫酸进行反应，生成硫酸氢钾和氯化氢。

主反应方程：$KCL+H_2SO_4=KHSO_4+HCL\uparrow$ +热量

生产流程：由转化岗位、吸收岗位构成。

转化岗位工艺流程：来自原料仓库的KCI经螺旋计量称计量后，并经二级预热器预热后，和来自硫酸贮槽的浓硫酸，按一定配比连续不断的加入到 1#KCL转化釜中进行反应，进行转化反应的同时，物料连续地溢流至2#KCI转化釜中继续反应。反应生成物硫酸氢钾浆料连续地溢流至氢钾混酸槽中，和一定配比的含五氧化二磷19%~23%的磷酸制备成混酸，合格的混酸被送往复合肥工段。反应生成的HCI气体用水经多级吸收后制得31%的盐酸，另一路，经冷冻干燥后送往氯磺酸装置。

硫酸氢钾在不同的温度下可以由两种不同的状态存在，即固态和液态，当温度约大于100℃时，硫酸氢钾呈溶液状，但较粘稠；当温度低于100℃时，硫酸氢钾结成硬而有韧性的固体，当温度大于130℃时，硫酸氢钾呈流动性较好的溶液状态。氯化钾的转化程度随反应温度的升高而增大。当反应温度小于110℃时转化率低于85%；当反应温度在110℃~130℃时，转化率再85%~90%；反应温度大于130℃时，转化率可达到95%以上。

在氢钾工段有 1#、2#两个磷酸贮槽，磷酸工段生产出来的稀磷酸轮流进入两个贮槽。当使用其中一个磷酸贮槽的磷酸时，该磷酸贮槽的磷酸注入阀门关闭，底部的磷酸使用阀门打开，磷酸工段过来的稀磷酸流入另一个磷酸贮槽。当正在使用的磷酸贮槽中的磷酸快用完时，对正在注入磷酸的磷酸贮槽取样，分析磷酸的浓度和密度，并以该浓度和密度作为这一整槽磷酸的浓度和密度。当上一槽磷酸用完后，将磷酸工段过来的磷酸切换到空槽开始注入。而已经分析好磷酸的浓度和密度的磷酸贮槽停止注入，并打开底部的阀门开始使用。

吸收岗位工艺流程：转化槽产生的HCL气体经E0301一级石墨酸洗冷却，产

生冷凝盐酸，待冷凝盐酸合格后，打入粗盐酸贮槽。净化的气体由离心风机鼓入两只并联的石墨吸收塔进行循环冷凝吸收，吸收后盐酸进入盐酸中间槽，在盐酸中间槽中的盐酸合格后，打入盐酸贮槽，在石墨吸收塔中未被完全吸收的氯化氢依次进入一吸塔、二吸塔、三吸塔，进行三次吸收后排空。吸收液由三吸塔、顺序向二吸塔、一吸塔、盐酸中间槽溢流，新鲜水由三吸塔补充。由离心风机出来的另一路 HCL 气体去冷冻干燥后，由压缩机送往氯磺酸装置。盐酸贮槽及盐酸灌装产生的尾气经风机升压后送到二吸塔进口。

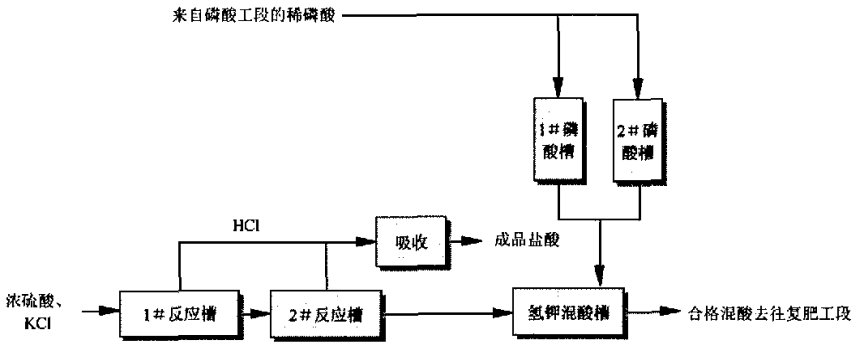


图 1. 氢钾工段流程图

复合肥工段流程简介

生产原理：由硫酸氢钾、浓硫酸以及磷酸组成的混酸与氨气进行中和反应，反应后的料浆经喷浆造粒后制成复合肥。

主反应方程：$NH_3+H_3PO_4=(NH_4)H_2PO_4+$热量

$NH_3+(NH_4)H_2PO_4=(NH_4)_2HPO_4+$热量

$NH_3+KHSO_4=KNH_4SO_4+$热量

$2NH_3+H_2SO_4=(NH_4)_2SO_4+$热量

生产流程：由中和岗位、造粒岗位、干线岗位、涂膜岗位组成。

中和岗位工艺流程：外界的常温常压空气经空气压缩机压缩后，进入压缩空气缓冲罐，供各工段使用，产生的压缩空气通过无热再生干燥器后，供各工段仪表使用；接受合成氨厂送来液氨先贮存，后经过液氨蒸发器加热蒸发为气氨，进

入气氨缓冲罐，随时供中和岗位使用。

混酸与氨气进入管式反应器进行中和反应，反应后的料浆进入闪蒸槽除去水分，满槽后溢流至喷浆槽，由喷浆泵送至造粒岗位。

中和尾气通过中和尾气风机，再经过中和尾气洗涤塔洗涤后排空。

造粒岗位工艺流程：空气经炉底风机进入热风炉，经过煤燃烧加热为热空气后由热风机送至造粒岗位。

来自中和岗位的料浆与空压站送来的压缩空气通过造粒机机头喷枪以雾化状态喷入造粒机，与以料幕的形式下落的返料接触，在热风炉送来的热风和造粒机的转动下形成复合肥颗粒，最后从造粒机机尾送出。

从造粒机出来的尾气先通过文丘里洗涤器经过洗涤后，进入洗液循环槽，通过尾气风机再到尾气洗涤塔经过再次洗涤后，从烟囱排放。

干线岗位工艺流程：造粒机出来的颗粒由斗提机送入滚筒筛，经过筛分后小颗粒直接到刮板输送机送至造粒机；大颗粒进入破碎机破碎后，进入刮板输送机；成品进入流化床冷却器由风机送入的空气冷却后由皮带机送至涂膜岗位。其中流化床出去的空气通过旋风除尘器除尘后，经过冷却引风机后，送至热风炉岗位。

涂膜岗位工艺流程：经过筛分、冷却、粒度合格的成品颗粒进入涂膜机，油剂通过蒸汽加温后通过计量泵打入涂膜机（有时在油剂加入的同时，涂膜粉也加入涂膜机），对成品颗粒进行涂膜后，送入斗提机，至成品料斗，最后进行称量包装。

从复肥混酸槽开始，分成 A、B 两条生产线，其中一些非常重要的变量，比如中和度、混酸流量，喷浆量等都是 A 线、B 线分别测量的，但是，最后氮、磷、钾的含量却是两条线的产品混合后检测的，也就是说两条线的输入变量值是不同的，但是对应的输出变量的值却是相同的。

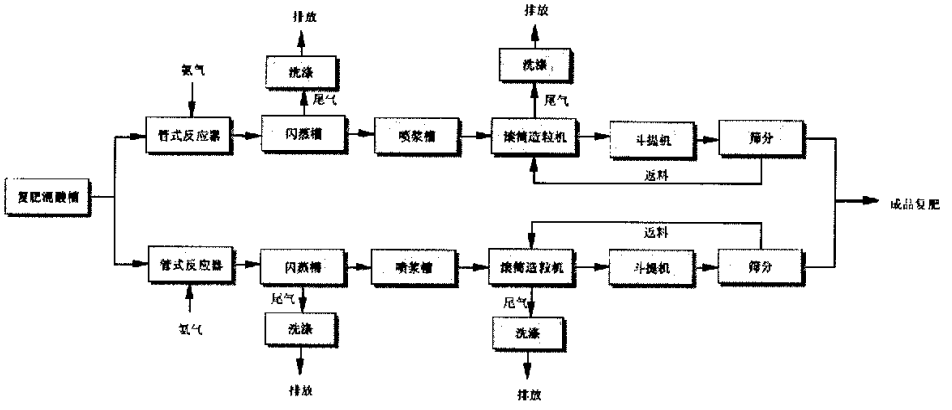


图 2. 复肥工段流程图

2 WSA 过程

烧结烟气 WSA 制酸过程工艺简介

株冶铅粗炼分厂采用丹麦托普索公司开发的 WSA 制酸技术和装备及美国孟山都的动力波气体净化技术和装备，应用于含有二氧化硫的铅锌烟气混合制酸。

株冶的整个 WSA 流程为典型的 WSA 湿法制酸流程，三段转化、熔盐换热、采用普通空气作为冷凝介质冷凝成酸。来自铅烧结与锌沸腾炉的混合烟气首先通过二级动力波洗涤及湿式电除雾器完成洗涤净化脱除灰尘及酸雾等杂质，通过原料气预热器 E101 与来自 WSA 冷凝器 E100 的热空气换热后升温，通过风机升压在熔盐换热器 E103 中与熔盐换热，使气体温度达到催化剂的起燃温度后进入转化器。转化器设三段，在 VK-WSA 催化剂的催化下，烟气中的二氧化硫转化成三氧化硫，经过两段反应后的烟气通过各自的中间层冷却器换热降温后去三段，离开三段的烟气通过工艺气冷却器进一步降温，温度降至 300℃后进入 WSA 冷凝器。在冷凝过程中，所有的三氧化硫与水合成硫酸并沿着冷凝器的玻璃管冷凝成酸。冷凝的浓硫酸通过换热器降温，由酸泵送到酸槽贮存。

该系统在工序检测、控制及总线网络中，生产管理过程和监控层的硬件和软件都采用美国费希尔.罗斯蒙特公司的 Delta V 系统，FF 现场总线仪表变送器、Profibus 现场总线设备。

熔盐控制系统由温度控制、循环及联锁控制组成。温度控制是由挂在 DeltaV 系统 FF 总线上的 5 台 3244 总线型温度变送器和 7 台总线型气动调节阀组成.控

制系统的热平衡。循环及联锁控制，是由挂在 DeltaV 系统 Profibus—DP 总线上的 PLC 和变频器组成. 分别控制盐泵电动机和紧急排放电磁阀。

烧结烟气 WSA 制酸过程的现状

WSA 工艺流程的一个重要特点是充分利用了转化器中的化学转化热能，去加热上游工段新进入的原料气。工艺要求的几个检测点温度控制范围比较窄：反应器出口温度要求在 290±5℃，冷凝器出口温度 100±5℃。转化器的温度是采用控制熔盐阀门开度来调节熔盐流通量，进而控制熔盐从转化器中带出的热量来调整转化器温度。如当转化器的出口温度变化得偏高时，控制系统调节作用使得阀门开大，增加熔盐的流通量，加大热载体带走的热能量值，促使转化器的出口温度回到正常值.计算机中的多套调节控制回路针对相应的控制对象，制定了其特定的控制策略，其中还包括了几个解耦控制，总体的控制效果基本满足了工艺要求。

本项目主要关注的设备和对象情况如下：

A. 动力波洗涤器

动力波洗涤净化技术由一套动力波设备组成，它包括初级逆喷洗涤器、气体冷却器、末级逆喷洗涤器和湿式电除雾器。动力波洗涤属于典型的湿法捕集范畴，其特点是利用气液两相的强烈接触过程除去气体中的固体尘粒，同时对烟气具有大幅降温能力。其基本过程是：将洗涤液喷入气流，气流和液体相撞，迫使液体呈幅射状自里向外射向筒壁，在气一液界面区形成强烈的湍动区(泡沫区)，使洗涤液和气体的动量达到平衡，气液紧密接触而产生稳定的驻波，驻波浮在气流中，随气、液相对动量的大小而升降。在泡沫区，由于气体与表面极大的且更新迅速的液体表面接触，而产生颗粒捕集、气体吸收和气体急冷等作用。

目前烟气洗涤净化的风机负荷已经达到最大，由于要维持主烟气管路的负压操作要求，洗涤净化的压力损失是限制烟气处理量的主要因素。因此技术人员希望提供烟气洗涤压力损失的操作辅助信息。

B. 熔盐换热器

由托普索公司引进的 WSA 装置中系统的换热是由一个熔盐系统换热系统来完成的。用熔盐而不用蒸汽可以无须依赖于蒸汽压力就可精确地控制温度。经过熔

盐换热器E103后的烟气视需要可通过一个旁路备用的工艺燃烧器加热以达到进入 WSA 反应器所需的起燃温度。株冶的 WSA 反应器有三个催化剂层，每一层催化剂下方出口都设一个熔盐换热器以挪走多余的反应热。经过三层催化反应大约 99%的 SO_2 可被转化成 SO_3。在气体进入 WSA 冷凝器前．反应器内的熔盐换热器应将它冷却至 300℃以下。传递热的熔盐用泵循环，在各个热交换器之间传递热量。系统的热平衡由熔盐冷却器来调整，过多的热在熔盐冷却器中以蒸汽（1.5Mpa）的形式释放出系统。

工艺气冷却器E103中熔盐流道是由若干根翅片管组焊成多组蛇形管并排列入换热器壳体内，管内介质为熔盐，为 KNO_3、KNO_2和 $NaNO_3$的混合熔融物，其具体配比以及物性参数不详。由 SO_2反应器反应后形成以 SO_3为主成分的混合气体进行冷却。

由于熔盐换热器的现场安装特性，换热蛇管的焊口在安装现场采用全氩弧焊，施工质量与穿管及焊接的顺序、保护气流量、喷嘴至工件距离、侧向风、电极尺寸都有关系。据现场工艺人员判断，目前株冶的WSA熔盐换热系统E103在焊接质量上可能存在问题，导致熔盐换热器在使用期间出现逐渐扩大的换热管堵塞情况，直接影响到整个 WSA 制酸系统的热平衡。因之而增加的工艺燃烧器使用已经带来一些经济损失和安全隐患。因此厂方技术人员需要得到在操作区间内各种工况下换热器换热能力的评估以指导系统热平衡操作。

C. WSA 反应器

WSA(Wet gas Sulfuric Acid)工艺是一个催化过程，该过程将 SO_2 转化为 H_2SO_4，这些过程是：

$$(1)\quad SO_2 + 1/2O_2 \rightarrow SO_3 + 22Kcal/mole$$

$$(2)\quad SO_3 + H_2O \rightarrow H_2SO4(气) + 24Kcal/mole$$

$$(3)\quad H_2SO_4(气) + 0.17H_2O(气) \rightarrow H_2SO_4(液体) + 15Kcal/mole$$

反应(1)是在含钒的氧化催化剂作用下发生的，托普索系列氧化催化剂的起燃温度约为 400℃，这也是 WSA 反应器规定的烟气入口温度。因此，须将净化后的洁净烟气加热到起燃温度后，送入催化剂层进行反应。在 WSA 转化反应器中，工艺烟气要通过三层催化剂及三个熔盐热交换器构成的转化器，使烟气中的 SO_2

转化成SO_3。由于在转化反应过程放热，为了保证催化剂的正常使用，烟气每通过一层催化剂即每经过一次转化后必须经过一个熔盐热交换器降低其温度才能进入下一层催化剂。在前两部分之间，烟气由热交换器 E106 和 E107 中的循环熔盐冷却。在最后的冷却器 E108 中，烟气被冷却到 290℃，SO_3按(2)式与水蒸气反应气体被冷却，进入湿气制酸冷凝器，H_2SO_4按(3)被冷凝。

由于这些化学反应都是放热反应，因此通过热交换，反应热又可用来加热工艺用气。反应热的多少取决于供气中SO_2的含量。在正常情况下，其浓度是以满足整个工艺自热的要求，氧化热、水合热及硫酸的部分冷凝热在系统内部全部被利用，系统多余的热量能通过余热锅炉即熔盐冷却器回收，可副产高压蒸汽，工艺具有很高的热回收率。但某些情况下系统热平衡不能保证时，还必须用外部热源加热，如株冶直接使用天然气炉。

WSA 工艺采取的是湿式催化转化，对催化剂要求非常严格，现场采用 VK-WSA 专用催化剂。经过反应的烟气中硫回收率可达 99%以上，尾气中SO_2小于 200ppm，酸雾小于 10ppm，可达标排放。

WSA 制酸工艺具有突出的特点：不论原料烟气SO_2浓度的高低，其均可产出浓度大于 96%的工业用浓硫酸，不仅治理了低浓度二氧化硫排放的问题，同时具有一定的经济效益。目前基本可以稳定生产 97%工业用浓硫酸，但尚达不到直接生产 98%工业用浓硫酸的水平，从商业化标准角度而需要采用调和手段进行处理。

D、酸雾控制器

从反应器出来的烟气在温度降至 300℃后进入 WSA 冷凝器。在冷凝过程中，气态硫酸与达到凝点的水合成硫酸并沿着冷凝器的玻璃管降至酸液储槽。为增强冷凝效果，在冷凝器前采用工艺压缩气喷入高温气化硅油以作为烟气内冷凝酸的结晶核。在尾气出口处设有在线酸雾检测仪以检验酸雾冷凝效果。

目前工艺压缩气出口阀位采用人工操作，没有闭环控制回路。

3. 目标和效益

株冶铅烧结烟气 WSA 制酸过程数据挖掘项目的目标，是要通过历史数据分析解决设备的负荷能力评价和操作参数优化问题。

通过对株冶铅烧结烟气 WSA 制酸过程及相关工艺技术的分析了解，在与现场工艺人员的交流基础上，初步确定烧结烟气 WSA 制酸过程实施数据挖掘的目标为：

■ 为动力波洗涤器的过程变量提供可视化描述，在其关联复杂程度有限情况下给出最高二次多项式的解析表达及参数估计。

■ 提供 WSA 制酸系统的熔盐换热器 E103 基于历史数据的换热能力评估参数，并提供可视化呈现。

■ 提供酸雾控制器的优化操作指导，从而根据实时操作工况优化冷凝效果。在稳定工况下通过优化操作指导，使用后的平均尾气酸雾检测指标接近或超过近期操作的稳态最好水平与平均水平的均值。

■ 提供冷凝酸浓度与原料、相关过程参数的关联分析并提供用户界面呈现。

实现该过程所使用 DCS 系统的实时数据自动采集和存储，以达到项目内容所必需的数据水平。