文章编号:1001 - 9081(2008)09 - 2382 - 04

# 基于支持向量机的质量控制软测量建模

姜贤林,郭秀清
(同济大学 电子与信息工程学院,上海 201804)
(jxl82@163.com)

**摘 要**:在具体研究支持向量机理论的基础上,提出了一种基于支持向量机的软测量控制方法。针对工业过程变量无法在线测量和大滞后的问题,建立了相应的支持向量机回归模型,将此方法用于合成反应器的质量控制中,实现了输出值的在线预估,并分析了参数调整和核函数的选择对建模的影响,得到了泛化良好的模型仿真结果。

**关键词**:支持向量机;软测量;质量控制;建模
**中图分类号**: TP274 **文献标志码**:A

# Soft-sensor modeling of quality control based on support vector machine

JIANG Xian-lin, GUO Xiu-qing
(*College of Electronics & Information Engineering, Tongji University, Shanghai* 201804, *China*)

**Abstract:** On the basis of studying Support Vector Machine (SVM) theory, a soft-sensor controlling method based on Support Vector Machine wass presented. In order to solve the problem of getting the important parameter that is hard to be measured online and has long time-delay, a soft-sensor controlling method based on support vector machine was presented. In the control process, modeling techniques have been studied intensively, and then RBF kernel function was chosen to establish an exact support vector machine model. On the background of quality control in a company, the online estimate of output value was realized. Under the circumstance of changing and choosing different parameters and through a lot of research and simulation, a relatively better generalization result model was established.

**Key words:** Support Vector Machine (SVM); soft-sensor; quality control; modeling

## 0 引言

自 20 世纪 60 年代开始,贝尔实验室以 Vapnik[1] 为代表的研究人员开始研究有限样本情况下的机器学习问题,形成了“统计学习理论”(Statistical Learning Theory,SLT),并在该理论基础上发展了一种新的模式识别方法——支持向量机(Support Vector Machine,SVM)。支持向量机表现出了很多优于传统统计方法的性能,引起了学者的关注,并在理论研究和应用研究方面涌现出了大量的成果,广泛用于诸如数字识别、人脸检测、指纹鉴别等领域中。

将支持向量机应用于工业过程软测量技术,具有非常重要的意义。在工业过程控制中,一些与产品质量密切相关或对保证安全生产有重要作用的过程变量,由于某些技术或经济的原因难以通过传感器在线直接测量,或能测量却具有较大的测量滞后,因此需要根据某种最优规则,选择一组与该参数有密切联系又容易测量的工艺参数,通过构造相应的数学关系,用计算机软件实现对该参数的在线实时估计,这就是软测量技术的实现思想。而支持向量机的软测量建模方法对工艺机理复杂系统的非线性建模起了非常重要的作用。文献[2]将支持向量用于炭黑工艺建模,并与主成分回归、反向传播人工神经网络以及径向基神经网络建模方法相比较,结果表明 SVM 更适合于较强非线性系统的建模。文献[3]讨论了基于最小二乘支持向量机的软测量数据建模原理和方法,并将其应用在汽车排放的氮氧化合物的软测量中,通过与基于神经网络的软测量方法进行比较,结果显示出 SVM 明显的优势。文献[4]提出了一种改进的最小二乘支持向量机回归方法,大大简化模型复杂程度,同时将这一方法应用于生物发酵过程,建立青霉素发酵过程中产物浓度的软测量模型,实现青霉素浓度的在线预估。支持向量机已逐渐成为新的研究热点,并将推动现代工业生产工艺的不断发展。

## 1 机器学习及统计学习理论[5]

### 1.1 机器学习问题

机器学习的目的是根据给定的训练样本求对系统输入输出之间依赖关系的估计,使它能够对未知输出做出尽可能准确的预测。

设有 $n$ 个独立同分布的训练样本$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,从一组函数集$f(x,\omega)$中选择出能够使$\bar{y}$最好地逼近训练器响应 $y$ 的函数$f(x,\omega_0)$,使预测的期望风险最小:

$$R(\omega) = \int L(y,f(x,\omega))\,\mathrm{d}F(x,y) \tag{1}$$

其中:$f(x,\omega)$ 为预测函数集,$F(x,y)$ 为联合分布函数,$\omega \in \Psi$ 为函数的广义参数,$L(y,f(x,\omega))$ 是给定输入 $x$ 下训练器响应 $y$ 与学习机器给出的响应$f(x,\omega)$ 之间的损失函数。

学习的目标在于使期望风险最小化,但是在实际的机器学习问题中,已知信息只有独立分布样本$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,式(1)中定义的期望风险无法直接计算得出,传统的学习方法中采用了所谓的经验风险最小化(Empirical Risk Minimization,ERM)原则。用对参数 $\omega$ 求经验风险最小值来代替求期望风险 $R(\omega)$ 的最小值,从而求得学习机器的参数 ,其中经验风险定义为:

$$R_{\mathrm{emp}}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L(y_i,f(x_i,\omega)) \tag{2}$$

**收稿日期**:2008 - 03 - 31;**修回日期**:2008 - 06 - 13。

**作者简介**:姜贤林(1982 - ),女,山东青岛人,硕士研究生,主要研究方向:过程控制、计算机控制; 郭秀清(1965 - ),女,内蒙古丰镇人,副教授,博士,主要研究方向:过程控制、计算机控制。

事实上,用 ERM 准则代替期望风险最小化并没有经过充分的理论论证,一味追求训练误差小并不总能达到好的预测效果。某些情况下,训练误差过小反而会导致推广能力下降,产生过学习问题。究其原因,一是因为样本不充分,二是学习机器设计不合理。在有限样本情况下,经验风险最小并不意味着期望风险最小,学习机器的复杂性不但与研究的系统有关,也要和有限的学习样本相适应,这是相互关联的两个方面。

### 1.2　统计学习理论

为了研究函数集在经验风险最小化原则下的学习一致性问题和推广性,统计学习理论中采用 VC 维作为衡量函数集学习性能的指标之一。VC 维越大,学习机器越复杂(容量越大)。学习机器的实际风险由两部分组成:一部分是经验风险(训练误差);另一部分是置信范围,用 $\Phi$ 表示,它与学习机器的 VC 维及训练样本数有关,并且随着 $h/n$ 的增加而减小,可以简单地表示为:

$$R(\omega) \leqslant R_{\text{emp}}(\omega) + \Phi(h/n) \tag{3}$$

其中:$h$ 为函数集的 VC 维,$n$ 为采样数。

要同时最小化经验风险和置信范围,传统的 ERM 准则在样本数有限时就暴露出不足。为解决这个问题,统计学习理论提出了一种新的策略,即结构风险最小化(Structural Risk Minimization, SRM)准则。SRM 准则同时考虑经验风险和函数的 VC 维,在经验风险和置信范围之间寻求折中,从而最小化实际风险。实施方案为:设计函数集的某种结构使每个子集中都能取得最小的经验风险,然后只需选择适当的子集使置信范围最小,则这个子集中使经验风险最小的函数就是最优函数。支持向量机方法实际上就是这种思想的具体体现。

## 2　支持向量机

### 2.1　SVM 的基本思想

SVM 方法最开始是从线性可分情况下的最优分类面提出的。设 $d$ 维空间内线性可分样本集为 $(x_i,y_i)$,$i = 1,2,\cdots,n$,$x \in R^d$,$y \in \{+1, -1\}$。分类面方程为:

$$(w \cdot x) + b = 0 \tag{4}$$

要把所有样本分开,必须满足下列条件:

$$y_i[(w \cdot x) + b - 1] \geqslant 0 \tag{5}$$

满足上式就能保证经验风险最小,使等号成立的训练样本即为支持向量(Support Vectors)。此时分类间隔等于 $2/\|w\|$,那么使分类间隔最大就等价于使 $\|w\|^2$ 最小,从而得到最小的置信范围。这相当于一个约束优化问题,即在式(5)的约束条件下,求式(6)的最小值:

$$\Phi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}(w \cdot w) \tag{6}$$

利用 Lagrange 函数求解,并将原问题化为其对偶问题后进行计算,最终得到的最优分类函数为:

$$f(x) = \text{sgn}\{(w^* \cdot x) + b\} = \text{sgn}\left\{\sum_{i=1}^{m} \alpha_i^* y_i (x_i \cdot x) + b^*\right\} \tag{7}$$

在线性不可分情况,通过引入一个松弛项 $\xi_i \geqslant 0$ 和惩罚参数 $C > 0$,将原来的优化问题转化为下面的等价问题:

$$\text{最小化 } \Phi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{8}$$

$$\text{满足 } y_i(w \cdot x_i + b) - 1 + \xi_i \geqslant 0; \quad i = 1,2,\cdots,n \tag{9}$$

便得到线性不可分情况下的最优分类面,即广义最优分类面。它在最小经验风险和推广能力之间寻求折中,$C$ 取值较大,可得到较小的经验风险(分类错误率低);$C$ 值较小,则能得到较好的推广能力(分类间隔较大)。

对非线性问题,可以通过事先选择好的某一个非线性映射 $\Phi: R^d \to H$,将输入向量映射到高维特征空间,在特征空间 $H$ 中构造一个最优分类超平面,此即 SVM 的实现思想。

在高维空间 H 中构造最优超平面时,只需进行内积运算 $\Phi(x_i) \cdot \Phi(x_j)$。找到一个函数 $K$,使得 $K(x_i,x_j) = \Phi(x_i) \cdot \Phi(x_j)$,高维空间的内积运算就可以用原空间中的变量直接计算了。这样甚至无需知道非线性映射 $\Phi$ 的形式,即使变换空间 $H$ 的维数增加,在求解最优分类面时计算复杂度的增加也并不多。根据泛函的有关理论,只要一个函数 $K(x_i,x_j)$ 满足 Mercer 条件,它就对应某一变换空间中的内积,这样用函数 $K(x_i,x_j)$ 代替点积 $(x_i \cdot x_j)$,将问题从原空间变换到一个高维空间,其相应的判别函数变为:

$$f(x) = \text{sgn}\{(\omega \cdot \boldsymbol{x}) + b\} = \text{sgn}\left\{\sum_{i=1}^{m} \alpha_i^* y_i K(x_i,x) + b^*\right\} \tag{10}$$

内积函数 $K(x_i,x_j)$ 称为核函数,构造上述判别函数的学习机器就称为支持向量机。其最终判别函数只包含训练输入与支持向量的内积求和,所以构造学习机器的复杂度仅取决于支持向量的个数,而不是特征空间的维数。其基本思想可用图 1 表示。

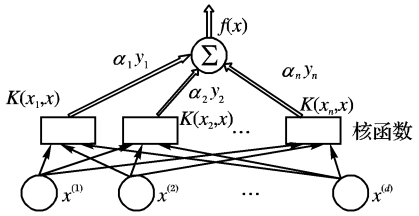


图 1　SVM 示意图

目前常用的核函数主要有四类。

1) 线性核函数:$K(x,x_i) = x \cdot x_i$。

2) 多项式核函数:$K(x,x_i) = [(x \cdot x_i) + 1]^d$; $d = 1,2,\cdots,n$。

3) 径向基核函数(RBF):$K(x,x_i) = e^{-\frac{\|x-x_i\|^2}{2\sigma^2}}$。

其中,$\sigma$ 为核函数的宽度。

4) S 形核函数(Sigmoid):$K(x,x_i) = \tanh(v(x \cdot x_i) + C)$。

### 2.2　支持向量机回归[6]

支持向量方法是通过分类问题提出的,但它同样可以应用到回归问题中,并且仍保留分类问题中最大化间隔算法的主要特征:非线性函数可以通过核特征空间中的线性学习器得到,同时系统的容量由与特征空间维数不相关的参数控制。

考虑一个数据集 $(x_i,x_j)$,$i = 1,2,\cdots,n$,$x_i \in R^d$,$y_i \in R$ 的拟合问题,设要寻求的回归函数为:

$$f(x) = (w \cdot x) + b \tag{11}$$

支持向量机回归引入了 $\varepsilon$ 不敏感损失函数,以忽略真实值某个上下范围内的误差,定义如下:

$$L_\varepsilon(y) = \begin{cases} 0, & |y - f(x)| \leqslant \varepsilon \\ |y - f(x)| - \varepsilon, & \text{其他} \end{cases} \tag{12}$$

求回归函数的问题就是求式(13)的最小值:

$$\Phi(w,\xi,\xi^*) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\xi_i^*\right) \tag{13}$$

约束条件为:

$$\begin{cases} f(x_i) - y_i \leqslant \xi_i + \varepsilon \\ y_i - f(x_i) \leqslant \xi_i^* + \varepsilon \end{cases} \tag{14}$$

其中:$\xi_i,\xi_i^* \geqslant 0$; $i = 1,2,\cdots,n$

优化式(13),其第一项使函数更为平坦,从而提高泛化能力;第二项则可以减少误差。常数 $C$ 控制对超出误差 $\varepsilon$ 的样本的惩罚程度。

采用类似的方法,回归问题可以转化为其对偶形式,即:

$$\text{最大化 } -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(x_i \cdot x_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \quad (15)$$

$$\text{约束为 } \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0;\ 0 \leqslant \alpha_i, \alpha_i^* \leqslant C, i = 1,2,\cdots,n \quad (16)$$

对于非线性回归,首先使用一个非线性映射 $\Phi$ 把数据映射到一个高维特征空间,然后在高维特征空间中进行线性回归。即用一个核函数代替内积 $(x_i \cdot x_j)$,则非线性回归要求下列函数的最大值:

$$Q(\alpha,\alpha^*) = -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i,x_j) - \varepsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) \quad (17)$$

约束条件仍然为式(16)。

由此,可得回归函数表达式为:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x,x_i) + b \quad (18)$$

### 2.3 SVM 的特点

支持向量机是在统计学习理论下导出的,与神经网络等传统方法相比有很大的优越性,它体现了结构风险最小化原则的设计思想,具有较好的推广能力;其次,将原始空间的非线性问题转化为高维特征空间的线性问题,在高维空间中构造线性判别函数,简化了问题的求解,而运算复杂度并没有增加;另外,支持向量机算法最终转化为一个凸优化问题,有唯一的全局最优解,避免了神经网络等容易陷入局部极值的缺点;最后,支持向量机可以有效的解决小样本学习问题。

## 3 SVM 在反应质量控制建模中的应用[7]

### 3.1 数据采集及软测量建模

本论文的实验建模基于某石油化工厂的质量控制生产过程。在合成反应器中,选取反应器的反应质量为主导变量,通过工艺机理分析,选取了与控制反应器反应质量相关的变量,如物料的分压和浓度等作为辅助变量,由于工艺上的原因,在反应过程中变量的在线测量非常困难,无法有效地进行在线控制,因此应用支持向量机算法建立合成反应器的质量控制的回归数学模型。

首先,采集合成反应相关的过程变量值,并对所得的生产数据进行分析检验,剔除坏的数据,进行归一化,选出具有均匀性、代表性的 100 组数据,选取前 80 组数据作为训练样本,用于 SVM 建模,剩下的 20 组数据作为测试样本,进行模型的泛化检验。

### 3.2 参数调整对建模的影响

选用 RBF 核函数建立输出值的支持向量机回归模型,在不同的模型参数下进行仿真实验和测试。

首先,选取参数 $C = 1000, \varepsilon = 0.5, \sigma = 2$,训练结果和测试结果如图 2 和图 3 所示。

改变参数,使参数 $C = \infty, \varepsilon = 0.5, \sigma = 5$,训练结果和测试结果如图 4 和图 5 所示。

对两次不同参数下的仿真结果进行误差分析,得出的误差曲线如图 6 所示,训练和测试的 100 组数据误差均统计包括在内。

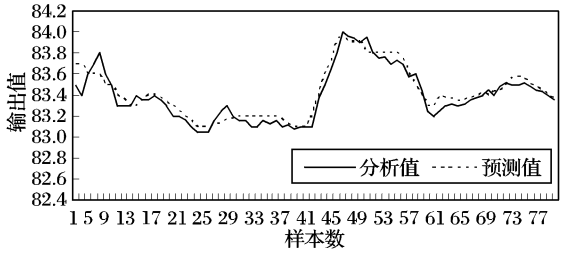

图 2 SVM 模型训练结果

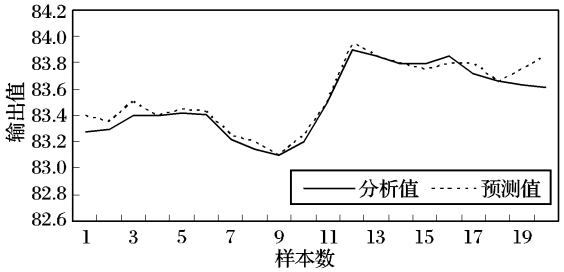

图 3 SVM 模型测试结果

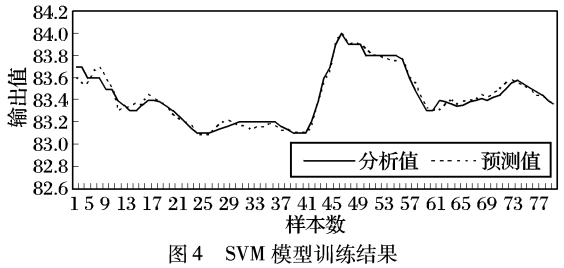

图 4 SVM 模型训练结果

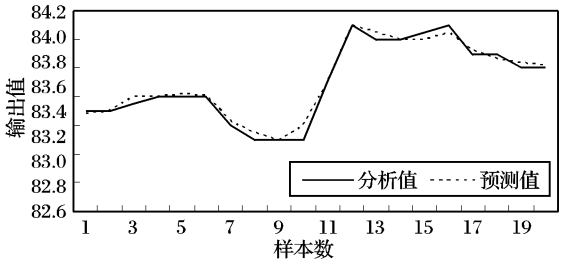

图 5 SVM 模型测试结果

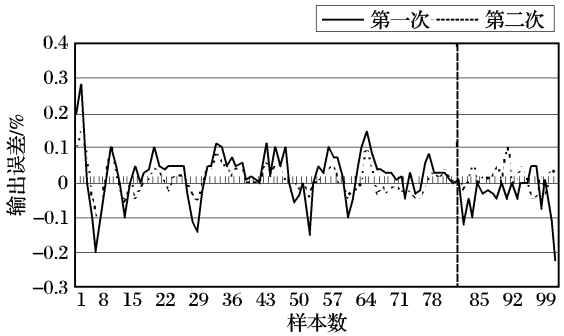

图 6 误差曲线

比较以上两组结果可以看出,无论训练还是测试过程,模型的预测输出值都能较为准确地拟合实际离线分析测量的输出值,误差在允许的范围内。并且,随着惩罚参数 $C$ 增大,核函数宽度参数 $\sigma$ 增大,模型测试结果误差变小。在质量控制的过程中,可以通过调整模型参数,得到不同的模型性能。

### 3.3 核函数的选择对建模的影响[8]

以上是以 RBF 核函数为例,现选择其他核函数进行软测量建模并进行比较。实验结果如表 1 所示。

通过表 1 可以明显看出,选择不同的核函数,对模型性能的影响很大,因为不同的核函数决定了支持向量个数的不同,从而导致训练时间和拟合精度的差异。在本论文的仿真过程中,通过对不同核函数参数下仿真结果的比较,最终决定选用 RBF(径向基)核函数建模,虽然用时较长,但是最终的预测误差达到最小,表现出了良好的拟合效果。

**表 1　不同核函数对建模质量的影响**

| 核函数 | 分类器参数 | 训练时间/s | 预测误差/% | 最大误差/% |
|---|---|---|---|---|
| Poly(多项式) | $q = 3$ | 36.5 | 1.278 4 | 2.018 3 |
| RBF(径向基) | $\sigma = 5$ | 164.1 | 0.863 7 | 1.041 5 |
| Sigmoid(S 形) | $v = 0.000\,01$ | 123.6 | 1.435 5 | 2.534 0 |

目前,尚未有一个通用的方法或者明确的理论用来说明如何选择核函数达到最优,因为各种核函数的适用对象不同,必须在实际的工业生产过程中,针对具体的控制对象和控制变量,分析其控制指标,并通过以往的经验和仿真,来选择不同的核函数进行建模,比较性能的优劣,有关这方面的研究还需要进一步地深入。

## 4　结语

本文对基于 SVM 理论的软测量建模方法进行了具体的理论分析,并深入研究了支持向量机的建模及仿真过程。在实际的工业生产过程中,对建立的合成反应器反应质量的网络模型进行仿真测试,比较训练、测试结果与实际分析得到的输出值之间的误差,拟合曲线效果良好。在采用不同的核函数的条件下,通过分析比较输出的结果,肯定了在此质量控制过程中选择 RBF 核函数建模性能的优越性。仿真结果表明,通过不断调整模型参数,可以减少输出值的误差,得到泛化较好的模型结果。

支持向量机建模问题往往是针对生产过程为庞大、时变、非线性的复杂系统,不同的应用领域有不同的建模方法和参数的选择问题,需要根据实际生产情况进行分析。另外,本文的 SVM 建模过程大部分还停留在仿真阶段,怎样将有效的算法应用到实际在线的生产中,还需要做进一步地探讨和深入地研究。

**参考文献:**


[1] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer Verlag, 1995.

[2] 李梦龙, 刘军红, 黎金明, 等. 基于支持向量机的炭黑工艺建模[J]. 应用基础与工程科学学报, 2005, 13(1): 51 - 57.

[3] 谭超. 基于支持向量机的软测量技术及其应用[J]. 传感器技术, 2005, 24(8): 77 - 79.

[4] 常玉清, 王福利, 王小刚, 等. 基于支持向量机的软测量方法及其在生化过程中的应用[J]. 仪器仪表学报, 2006, 27(3): 241 - 244.

[5] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32 - 42.

[6] 王定成, 方廷健, 高理富, 等. 支持向量机回归在线建模及应用[J]. 控制与决策, 2003, 18 (1): 88 - 95.

[7] 陈鑫. OXO 反应质量软测量[D]. 上海: 华东理工大学, 2005.

[8] 马勇, 黄德先, 金以慧. 基于支持向量机软测量建模方法[J]. 信息与控制, 2004, 33(4): 417 - 421.

[9] CRISTIANINI N, SHAWE-TAYLOR I. An introduction to support vector machines and other kernel-based learning methods[M]. London: Cambridge University Press, 2000.

[10] ANARWAL M, JADE A M. Support vector machines: A useful tool for process engineering applications [J]. Chemical Engineering Process, 2003, 99(1): 57 - 62.


(上接第 2377 页)

说明 live 方式比 non_live 方式总迁移时间稍长但可以忍受,而且被迁移虚拟机的服务中断时间要少很多,可以提高虚拟机在迁移时的可用性。

3)不同负载对迁移时间的影响。从图 1 ~ 图 3 可以看出,在内存大小相同的情况下,负载的不同导致总迁移时间和停机时间也不相同。由表 1 ~ 表 4 可知,I/O 密集和内存密集型的总迁移时间分别为无负载时的 1.07 倍、1.2 倍,而 CPU 密集型负载是无负载时的 2.2 倍;I/O 密集和内存密集型的停机时间分别为无负载时的 2.15 倍、1.84 倍,而 CPU 密集型负载是无负载时的 7.75 倍。说明 CPU 密集型负载对 XEN 的动态迁移影响比较大。

## 4　结语

本文在总结相关工作的基础上,研究了虚拟机动态迁移方法,并设计了 4 种不同负载下的迁移实验。实验数据表明 CPU 密集型负载对迁移的影响最大,I/O 和内存密集型影响较小。动态迁移方式的停机时间总是远远小于传统静态迁移方式的停机时间,内存越大,静态迁移停机时间越长,但动态迁移停机时间变化不大,说明动态迁移可以显著改善服务性能。本文主要在局域网进行迁移,今后会向广域网迁移发展,为了保持虚拟机的网络连接,需要采用移动 IP 或者动态 DNS 等网络重定向技术[5];如果在集群中要迁移整个磁盘内容,可以在目的主机上建立一个 COW 的磁盘镜像,支持块设备迁移[6]。

**参考文献:**


[1] BARHAM P, DRAGOVIC B, FRASER K, *et al.* Xen and the art of virtualization[C]// Proceedings of the nineteenth ACM symposium on Operating Systems Principles (SOSP19). New York: ACM Press, 2003: 164 - 177.

[2] CLARK C, FRASER K, HAND S, *et al.* Live migration of virtual machines[C]// Proceedings of the 2nd ACM/USENIX Symposium on Networked Systems Design and Implementation (NSDI). Boston: ACM Press, 2005: 273 - 286.

[3] OSMAN S, SUBHRAVETI D, SU G, *et al.* The design and implementation of zap: A system for migrating computing environments [C]//Proceedings of the 5th USENIX Symposium on Operating Systems Design and Implementation (OSDI-02). New York: ACM Press, 2002: 361 - 376.

[4] HANSEN J G, HENRIKSEN A K. Nomadic operating systems: Master's thesis[D]. Denmark: Department of Computer Science, University of Copenhagen, 2002.

[5] BRADFORD R, KOTSOVINOS E, FELDMANN A, *et al.* Live wide-area migration of virtual machines including local persistent state [C]// Proceedings of the 3rd International Conference on Virtual Execution Environments. New York: ACM Press, 2007: 169 - 179.

[6] SAPUNTZAKIS C P, CHANDRA R, PFAFF B, *et al.* Optimizing the migration of virtual computers[C]// Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI-02). New York: ACM Press, 2002: 377 - 390.