双层模糊系统融合中心约束型最小包含球

徐 华
(江南大学物联网工程学院 无锡 214122)

摘 要 与传统的TSK模糊系统相比,改进的双层TSK模糊系统CTSK(Central TSK Fuzzy System)有如下优点:良好的可解释性、更好的鲁棒性、较强的逼近能力。但对于大样本或超大样本数据集,其时间复杂度和空间复杂度的开销都极大地限制了它的实用性。针对此不足,通过模糊系统融合中心约束型最小包含球(CCMEB)理论提出了CCMEB-CTSK(CCMEB-based CTSK)算法。该算法在继承CTSK优点的同时,又较好地实现了处理大样本和超大样本数据集的有效性和快速性。仿真实验研究分析了采用不同模糊规则数的CCMEB-CTSK的性能指标和运行时间的比较,以及训练样本不加噪声和加入噪声情况下CCMEB-CTSK泛化能力和鲁棒性能的测试。

关键词 模糊系统,中心约束型最小包含球,泛化,鲁棒性
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.037

Integration of Dual-layer Fuzzy System with Center-constrained Minimal Enclosing Ball

XU Hua
(School of IOT Engineering, Jiangnan University, Wuxi 214122, China)

Abstract This paper used the central TSK fuzzy system which is the improved double layer TSK fuzzy system. Compared with the traditional TSK fuzzy system, the CTSK fuzzy system(Centralized TSK Fuzzy System) adopts the upgraded dual-layer TSK fuzzy system and has the following advantages: better interpretability, stronger robustness and approximation capability. However, the time and space complexity overhead greatly limits large or super large data sets. In the light of this limitation, a new algorithm CCMEB-CTSK (CCMEB-based CTSK) was proposed here dealing with large data sets. The algorithm not only preserves the advantages of the CTSK fuzzy system, but also contributes to the high efficiency and rapidity in handling large and super large sample data sets. Through simulation experiments, an analysis of the difference in the performance index and running time of CCEMB-CTSK was made, including different fuzzy rule numbers as well as the difference in the generalization capability and performance robustness of CCEMB-CTSK in noiseless and noisy training samples.

Keywords Fuzzy system, CCMEB, Generalization capability, Robustness

1 引言

模糊系统[1,2]以模糊逻辑为基础,通过模仿人的模糊综合判断推理来处理常规方法难以解决的模糊信息处理的难题。

CTSK模糊系统是对TSK模糊系统的改进,其包含两层中心TSK模糊系统,是TSK模糊系统的一个特例[3]。但是CTSK模糊系统处理大样本数据集或超大样本数据集的时间复杂度、空间复杂度的开销是不容忽视的,处理的效率大大减弱。

中心约束型最小包含球(Center-Constrained Minimal Enclosing Ball,CCMEB)理论进一步扩大了最小包含球理论(Minimal Enclosing Ball,MEB)的适用范围,将改进的TSK模糊系统与中心约束型最小包含球CCMEB理论相融合形成新的算法,该算法的优点是在保持CTSK各优点的同时,对于大数据集和超大数据集具有更好的泛化能力和鲁棒性能。仿真实验的结果验证了该算法的有效性和快速性。

2 理论和方法

2.1 CTSK模糊系统规则

CTSK模糊系统[4]是对传统TSK模糊系统的改进,它与传统TSK模糊系统的区别主要在于:一方面,其采用的模糊规则为:

R^i: IF x_1 is $A_1^{(i)}$, x_2 is $A_2^{(i)}$, $\cdots$, x_n is $A_n^{(i)}$,
THEN $y_i = p_0^i + p^{(i)T}(x_i - u_n^{(i)})$

另一方面,其系统输出为:

$$f(x)=\sum_{i=1}^{m}(p^{(i)T}(x_i-u_n^{(i)})+p_0^i)\mu_{A(i)}(x) \quad (1)$$

根据文献[4],已推导并证明中心化TSK模糊系统CTSK的每一条模糊规则的结论部分的系数都可以看成是其每条规则加权输出函数在规则中心的一阶导数,并且每一条规则的结论部分等价于中心化TSK模糊系统的该条规则加权输出$f_i(x_1,\cdots,x_n)$在相应的规则中心$\boldsymbol{\mu}_i=(\mu_{i1}\cdots,\mu_{in})$的

到稿日期:2014-01-25 返修日期:2014-03-22 本文受国家留学基金委资助项目(201308320030),江苏省自然科学基金(BK20140165)资助。
徐 华(1978-),女,副教授,CCF会员,主要研究方向为水污染、人工神经网络、模糊系统等,E-mail:joanxh2003@163.com。

Taylor一阶展开，从这个角度来说，中心化 TSK 模糊系统 CTSK 比 TSK 模糊系统可以得更好的解释。

令

$$\overline{A_i}(x)=\mu_{A(i)}(x) \tag{2}$$

有

$$f(x)=\sum_{i=1}^{M}(p^{(i)T}(x_i-u_n^{(i)})+p_0^i)\overline{A_i}(x) \tag{3}$$

令

$$\varphi^i(x)=(\overline{A_i}(x)(x_i-u_n^{(i)})^T,\overline{A_i}(x))^T,$$
$$\varphi(x)=(\varphi^1(x)^T,\cdots,\varphi^M(x)^T)^T,$$
$$p^{(i)'}=(p^{(i)T},p_0^i)^T,P=(p^{(1)'T},\cdots,p^{(M)'T})^T$$

式(3)可写为：

$$f(x)=P^T\varphi(x) \tag{4}$$

其中，$\sum_i\mu_{A(i)}(x)\neq 0$。从式(3)和式(4)可以看出，CTSK 模糊系统也等价于一个模糊规则空间 $H\subset R^{M(n+1)}$ 上的线性回归系统。因此，可采用支持向量回归方法来构造模糊系统。

2.2 基于 L1 范数机器学习算法的 CTSK 模糊系统

L1-SVR[5](support vector regression)也是 L1 范数方法的一种。下面推导基于 L1 范数机器学习算法的 CTSK 模糊系统。

设训练集为 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，其中 $x_i\in X\subset R^n,y_i\in Y\subset R,i=1,\cdots,l$。根据结构风险最小化原则，将 CTSK 模糊系统的目标函数设定为：

$$\min\sum_{i=1}^{d}|y_i-P^T\varphi(x_i)|_\varepsilon+\tau P^TP/2 \tag{5}$$

其中，$\tau>0$，$|g|_\varepsilon$ 是 ε-不敏感损失函数。引入松弛变量 ξ 和 ζ，上式就转化为下面的二次规划问题 Quadratic programming，简称 QP 问题：

$$\begin{aligned}&\min_{P\in R^{M(n+1)},\zeta\in R^{2d}}\frac{1}{2}P^TP+\frac{1}{\tau}\sum_{i=1}^{d}(\xi+\zeta),\\&\text{s. t. } y_i-\varphi(x_i)^TP\leqslant\varepsilon+\xi\\&\quad\varphi(x_i)^TP-y_i\leqslant\varepsilon+\zeta,\\&\quad\xi\geqslant 0,\zeta\geqslant 0,i=1,\cdots,d\end{aligned} \tag{6}$$

Where:$\zeta=(\xi_1,\cdots,\xi_d,\zeta_1,\cdots,\zeta_d)^T$。这样，对式(4)中参数的求解问题就转化为对式(6)的二次规划 QP(Quadratic programming)的求解。

3 中心约束型最小包含球理论 CCMEB

CCMEB[6]理论进一步扩大了最小包含球理论的适用范围。CCMEB 求解可表示为如下约束优化问题：

$$\begin{aligned}&\arg\min_{c,r}r^2\\&\text{s. t. }(\varphi(x_i)-c)'(\varphi(x_i)-c)+\delta_i^2\leqslant r^2,i=1,\cdots,N\end{aligned} \tag{7}$$

其对偶形式：

$$\begin{aligned}&\arg\max_\alpha\alpha'(diag(K)+\Delta)-\alpha'K\alpha\\&\text{s. t. }\alpha'1=1,a\geqslant 0\end{aligned} \tag{8}$$

其中，

$$\Delta=[\delta_1^2,\cdots,\delta_N^2]'\geqslant 0 \tag{9}$$

类似于 MEB，由式(8)的最优解 α，可得 CCMEB 的中心点 c 和半径 r：

$$r=\sqrt{\alpha'(diag(K)+\Delta)-\alpha'K\alpha},c=\sum_{i=1}^{N}\alpha_i\varphi(x_i) \tag{10}$$

由于 $\alpha'1=1$，故在式(8)的目标函数中增加一项 $-\eta\alpha'1(\eta\in R)$将不会影响最优解的值，于是得式(11)：

$$\begin{aligned}&\arg\max_\alpha\alpha'(diag(K)+\Delta-\eta 1)-\alpha'K\alpha\\&\text{s. t. }\alpha'1=1,a\geqslant 0\end{aligned} \tag{11}$$

得到的式(11)对我们意义重大，这时任何形如式(12)包含一次项的 QP 问题都可以通过变形转换成式(11)：

$$\begin{aligned}&\arg\max_\alpha\alpha'f-\alpha'K\alpha\\&\text{s. t. }\alpha'1=1,a\geqslant 0\end{aligned} \tag{12}$$

此时，$\Delta=-diag(K)+\eta 1+f$，只要选取足够大的 η，确保 $\Delta\geqslant 0$ 即可。

文献[7]指出任何形如式(11)且满足 $\Delta\geqslant 0$ 的 QP 问题均可视作一个 CCMEB。

注：CCMEB 的限定条件是 QP 问题只需满足式(8)或式(11)。

4 CTSK 和 CCMEB-CTSK 算法

4.1 基于 L2 范数机器学习算法 CTSK

下面首先推导基于 L2 范数机器学习算法[8,9]的 CTSK 模糊系统。将基于 L1 范数的 CTSK 模糊系统的二次规划问题(Quadratic programming，QP)即式(6)转化为基于 L2 范数的 CTSK 模糊系统的二次规划问题 QP：

$$\begin{aligned}&\min_{P\in R^{M(n+1)},\zeta\in R^{2d}}\frac{1}{\tau d}\sum_{i=1}^{d}(\xi^2+\zeta^2)+\frac{1}{2}P^TP+\frac{2\varepsilon}{\tau}\\&\text{s. t. } y_i-\varphi(x_i)^TP\leqslant\varepsilon+\xi\\&\quad\varphi(x_i)^TP-y_i\leqslant\varepsilon+\zeta\\&\quad i=1,\cdots,d\end{aligned} \tag{13}$$

基于 L2 范数的 CTSK 模糊系统比基于 L1 范数的 CTSK 模糊系统有如下两点改进之处：1)$\xi\geqslant 0,\zeta\geqslant 0$ 会自动满足，而不用像式(6)中的人为设定；2)ε 参数不用人为设定，能够通过最优化自动获得。

由拉格朗日乘数法的标准应用程序产生式(14)：

$$\begin{aligned}&\max[a'\ a^{*\prime}]\begin{bmatrix}2\tau y\\-2\tau y\end{bmatrix}-[a'\ a^{*\prime}]\widetilde{K}\begin{bmatrix}a\\a^*\end{bmatrix}\\&\text{subject to}[a'\ a^{*\prime}]1=1,a,a^*\geqslant 0\end{aligned} \tag{14}$$

其中，$1=[1,\cdots,1]'$ 为 N 维列向量。$y=[y_1,\cdots,y_d]'$，$a=[a_1,\cdots,a_d]'$，$a^*=[a_1^*,\cdots,a_d^*]'$ 是双变量(拉格朗日乘子)，并且

$$\widetilde{K}=[\widetilde{K}(z_i,z_j)]=\begin{bmatrix}K+11'+\tau dI & -(K+11')\\-(K+11') & K+11'+\tau dI\end{bmatrix}$$

是一个 2d×2d 核矩阵。

4.2 CCMEB-CTSK

CTSK 的时间复杂度是 $O(N^3)$，这对于大样本数据集来说计算开销是相当可观的。但通过观察式(14)，我们发现 CTSK 可视作一个 CCMEB 问题，正是这个发现可以降低算法的复杂度。

根据下面定理[10,11]：

定理 1 给定一半正定矩阵 $\boldsymbol{M}$ 和一正数 $\alpha>0$，则 $\boldsymbol{M}+\alpha\boldsymbol{I}$ 必正定。

我们可以对式(14)进行优化：

令：

$$\Delta=-diag(\widetilde{K})+\begin{bmatrix}2\tau y\\-2\tau y\end{bmatrix}+\eta 1$$

式(14)可转化为：

$$\max[a'\ a^{*\prime}](diag(\widetilde{K})+\Delta-\eta 1)-[a'\ a^{*\prime}]\widetilde{K}\begin{bmatrix}a\\a^*\end{bmatrix}$$
$$\text{subject to}[a'\ a^{*\prime}]1=1, a, a^*\geqslant 0 \tag{15}$$

令：

$$\alpha=\begin{bmatrix}a\\a^*\end{bmatrix}$$

式(15)可转化为：

$$\max \alpha'(diag(\widetilde{K})+\Delta-\eta 1)-\alpha'\widetilde{K}\alpha$$
$$\text{subject to } \alpha'1=1, \alpha\geqslant 0 \tag{16}$$

只要选择足够大的 η，使 $\Delta\geqslant 0$ 即可。此时式(16)即是一个标准的 CCMEB 问题，于是就得到了本文的 CCMEB-based CTSK(简称 CCMEB-CTSK)算法。该算法为实现基于大数据集的快速模糊建模提供了可能。

CCMEB-CTSK 算法具体实现如下：

1. 选择和设置合适的参数，设置模糊规则数 m 和正常数 η，设置近似参数 ε。

2. 使用模糊聚类的回归数据集输入空间从总量为 N 的数据集 S 中随机选取 $\gamma(\gamma\leqslant N)$个点构成 CTSK 的输入，代入式(16)解得最优解 α_0，由 α_0 经式(10)得 $B(c_0,(1+\varepsilon)r_0)$。下面进入反复迭代过程，设 t 为迭代次数，初值 $t=0$。

3. 计算 S-CoreSet$_t$ 内所有点 x_ℓ 在核空间中到球心 c_t 的距离 d_ℓ，选择 $d_\ell>(1+\varepsilon)r_t$ 且最相异于 CoreSet$_t$ 的点 x^*，构建 $CoreSet_{t+1}=CoreSet_t\cup\{x^*\}$；若不存在这样的点，算法终止。

4. CoreSet$_{t+1}$经由式(16)得 α_{t+1}，将其代入式(10)得 $B(c_{t+1},(1+\varepsilon)r_{t+1})$；为 CCMEB 设置大的正常数 η。

5. $t=t+1$；根据当前 CoreSet$_t$ 内异类点的分布情况，若有需要，适度调整精度 ε 的值；最后回到第 3 步反复迭代直至算法终止。

6. 算法终止时的解 α_t 即为 CCMEB-CTSK 的解。用 $\alpha=\begin{bmatrix}a\\a^*\end{bmatrix}$和 $p^{(i)'}=(p^{(i)T},p_0^i)^T$，$P=(p^{(1)'T},\cdots,p^{(M)'T})^T$得到 CCMEB-CTSK 相应的参数，产生理想的 CCMEB-CTSK 模糊系统。

5 实验结果和分析

5.1 采用性能指标评价

为了检验 CCMEB-CTSK 的有效性，在以下的几个实验中把它与 CTSK、传统的 TSK 模糊系统进行了有效的比较，结果显示，在处理大数据集时，CCMEB-CTSK 比 CTSK、传统的 TSK 具有更好的逼近性能和抗噪声的鲁棒性能。

为了更好地评价模糊系统的性能，采用了以下的性能指标：

$$J=\sqrt{\frac{\sum_{l=1}^{N}(\hat{y}_l-\hat{y}_l^d)^2}{\sum_{l=1}^{N}(\hat{y}_l^d-\bar{y})^2}} \tag{17}$$

其中，$\bar{y}=\frac{1}{N}\sum_{l=1}^{N}\hat{y}_l^d$，$N$ 为样本总数，$\hat{y}_l^d$ 为第 l 个样本的期望输出，$\hat{y}_l$ 表示第 l 个样本的实际输出。

5.2 实验 1：不同模糊规则数的 CCMEB-CTSK 的比较

不同模糊规则数的 CCMEB-CTSK 的性能指标 J 与 CPU 运行时间的比较。在这个实验中，近似参数 ε 设置为 10^{-6}，正常数 η 设置为 100。在我们所做的几千次的实验中，这两个参数的值发挥了非常好的效果。而且从实验中我们发现，如果设置 η 为更小的值，实验效果不好；设置 η 为更大的值，效果不是很明显。设置 ε 为较大的值，实验效果不好；设置 ε 为足够小的值，实验错误率明显下降。所有其它参数的设定均由 20% 样本的验证集来决定。所有的算法均由 MATLAB 代码来实现。

实验 1：在这个实验中，我们研究基准正弦函数运行采样数据的实验情况和结果分析：

$$y=\frac{\sin(x)}{x} \tag{18}$$

运用式(18)运行的数据集的大小为 1e2 到 1e6。式(18)所采用的数据对是均匀统一的。

实验结果如图 1 和图 2 所示。随着训练数据集的增大和模糊规则数的增多，性能指标 J 逐渐变小，CPU 的训练时间相应地增大。但是，太多的模糊规则反而会是模糊模型的一大软肋。从图 1 和图 2 可以看出，当训练数据集的大小从 1e5 变化到 1e6 时，增加的 CPU 的训练时间不超过 50 秒。这充分表明了本文提出的 CCMEB-CTSK 算法能非常有效地应用于大型数据集。

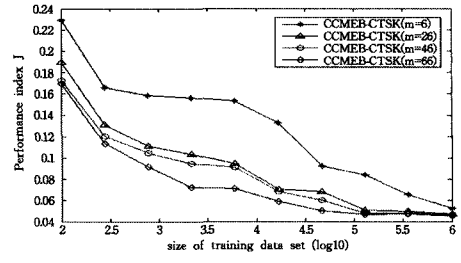

图 1 性能指标 J 的比较

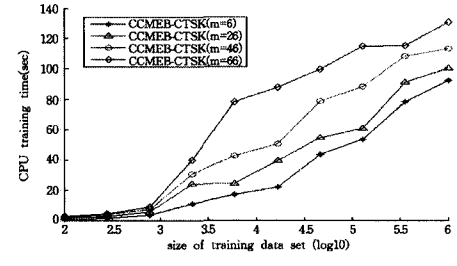

图 2 CPU 训练时间的比较

5.3 实验 2：训练样本没有噪声

实验 2：在这个实例中，我们选用了预测将来值的 Mackey-Glass 时间序列函数作为测试，Mackey-Glass 时间序列为微分延迟函数，定义如下：

$$x(t)=\frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)}-0.1x(t) \tag{19}$$

这是一个不可预测且不收敛，初始化十分敏感的时间序列。在此实验中，假设 $\tau=17$ 且 $x(0)=1.0(x(t)=0$ 当 $t<0$ 时)。

我们从该时间序列选取了 20000 个输入—输出数据对 $\{x^d,y^d\}$，其形式如下：

$$\{x^d=[x(t-18),x(t-12),x(t-6),x(t)],y^d=x(t+6)\} \quad (20)$$

其中，t 的取值为 118 到 20117。综合考虑性能和速度，我们选用适合的 $m=26$，并取前 10000 个数据对作为训练样本，后 10000 个数据作为测试样本。

表 1 给出了 CCMEB-CTSK($m=26$)、CTSK 与 TSK 的对训练样本以及测试样本的性能指标比较，从中可以看出，对于大数据集，CCMEB-CTSK 比 CTSK 和 TSK 具有更好的泛化能力。

表 1 CCMEB-CTSK、CTSK 和 TSK 对 Mackey-Glass 时间序列的预测性能比较

算法	训练次数	模糊规则数	测试样本的性能指标 J
CCMEB-CTSK	10000	40	0.044877
CTSK	10000	40	0.158562
TSK	10000	40	0.258533

图 3－图 5 分别给出了 CCMEB-CTSK、CTSK 与 TSK 的逼近效果。从图中能直观看出，CCMEB-CTSK 的逼近效果优于 CTSK、TSK。

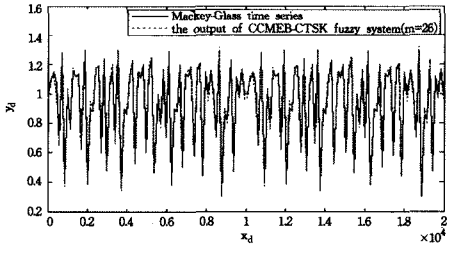

图 3 CCMEB-CTSK 模糊系统对于 Mackey-Glass 时间序列的预测结果

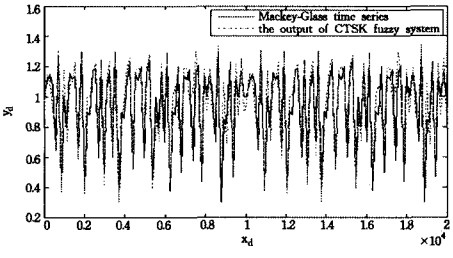

图 4 CTSK 模糊系统对于 Mackey-Glass 时间序列的预测结果

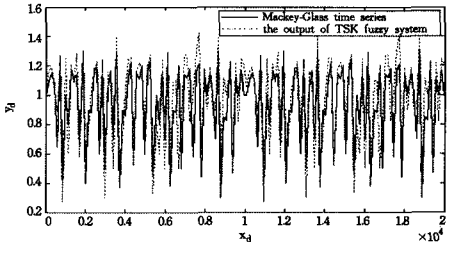

图 5 TSK 模糊系统对于 Mackey-Glass 时间序列的预测结果

5.4 实验 3：训练样本加入 5%的高斯白噪声

为了检验 CCMEB-CTSK 的鲁棒性能，我们在样本中加入了均值为 0、方差为 1 的高斯白噪声，实验结果显示，CCMEB-CTSK 比 CTSK 和 TSK 具有更好的鲁棒性能。

实验 3：表 2 给出了加入 5%的高斯白噪声后，CCMEB-CTSK、CTSK 以及 TSK 模糊系统对于 Mackey-Glass 时间序列函数的预测效果。从结果不难看出，当加入噪声后，预测结果比没有噪声情况下稍差，但与 TSK 和 CTSK 相比，CCMEB-CTSK 预测效果仍更好，也就是说 CCMEB-CTSK 有更强的鲁棒性。

表 2 有噪声情况下 CCMEB-CTSK、CTSK 与 TSK 对 Mackey-Glass 时间序列的预测性能

算法	训练次数	规则数	测试样本的性能指标 J
CCMEB-CTSK	10000	40	0.295471
CTSK	10000	40	0.586951
TSK	10000	40	0.815413

结束语 根据大数据摩尔定律，数据每年一直都以很快的速率不断增长，这意味着人类将有越来越多的数据需要处理。“大数据”作为时下最热门的 IT 行业的词汇，海量、高增长率和多样化的信息资产，需要有更强、更准确、更高效的处理工具。本文提出的方法在面对大样本或超大样本数据集时，仍可拥有较强的泛化能力和鲁棒性能，这为解决大数据问题提出了参考。

参 考 文 献

[1] 侯越. 基于改进 T-S 模糊神经网络的交通流量预测[J]. 计算机科学，2014，8(1)：121-126

[2] 冯定芸，于福生，王晓. 模糊规则组的谐调度[J]. 计算机科学，2013，40(5)：45-47

[3] 徐华，薛恒新. 中心化模糊系统 CTSK 的分析及应用[J]. 计算机工程，2008，34(23)：7-16

[4] Chung K F L，Duan J C. On multistage fuzzy neural network modeling[J]. IEEE Trans. Fuzzy systems，2000(8)：125-142

[5] 蔡前凤，郝志峰，刘伟. 基于模糊划分和支持向量机的 TSK 模糊系统[J]. 模式识别与人工智能，2009，22(3)：411-416

[6] 钱鹏江，王士同，邓赵红，等. 基于最小包含球的大数据集快速谱聚类算法[J]. 电子学报，2010，38(9)：2035-2041

[7] Tsang I W-H，Kwok J T，Zurada J A. Generalized Core Vectore Machines[J]. IEEE Transactions on Neural Networks，2006，17(5)：1126-1139

[8] Zhang Ying-song，Kingsbury N. FAST L0-BASED SPARSE SIGNAL RECOVERY[C]//2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2010). 2010：403-408

[9] 刘建伟，李双成，罗雄麟. p 范数正则化支持向量机分类算法[J]. 自动化学报，2012，38(1)：76-87

[10] Lee C-H，Zaïane O R，Park H H，et al. Clustering High Dimensional Data：A Graph-based Relaxed Optimization Approach[J]. Information Sciences，2008，178(23)：4501-4511

[11] 刘向东，骆斌，陈兆乾. 支持向量机最优模型选择的研究[J]. 计算机研究与发展，2005，42(4)：576-581