小样本数据的支持向量机回归模型参数及预测区间研究

陈 果, 周 伽
(南京航空航天大学, 江苏 南京 210016)

摘要: 支持向量机是由统计学习理论发展起来的机器学习算法, 它从结构风险最小化的角度保证了模型的最大泛化能力。文中运用支持向量机进行小样本数据回归分析研究。首先利用推广性的界理论指导支持向量机回归模型参数的选取, 以保证模型具有最大的推广能力; 其次, 运用基于正态分布和基于 t 分布的两种区间预测方法进行了预测值的区间估计; 最后, 利用模拟序列和真实的航空发动机油样光谱分析数据作为实验数据, 建立了支持向量机回归分析模型, 并与最小二乘法进行了比较。结果表明, 所提出的支持向量机模型参数选取和区间估计方法适用于小样本数据的回归分析, 具有较高的预测精度。

关键词: 计量学; 支持向量机; 小样本; 回归模型; 预测精度; 区间估计
中图分类号: TB9 文献标识码: A 文章编号: 1000-1158(2008)01-0092-05

Research on Parameters and Forecasting Interval of Support Vector Regression Model to Small Sample

CHEN Guo, ZHOU Jia
(Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu 210016, China)

Abstract: Support vector machine is a new machine learning method based on statistic learning theory (SLT), it can assure the most generalization on the foundation of structural risk minimization. The small sample data modeled with support vector regression (SVR) is described. Firstly, model parameters are chosen according to the bound theory of generalization performance in order to assure the most generalization of regression model; then, the two forecasting interval methods are applied, one is based on normal distribution, the other is based on t distribution; in the end, the SVR model is established by using simulated data and true aero-engine spectrometric oil analysis data, and it is compared with least square method. The result indicated that the parameter selection and interval estimation method of SVR regression model has high accuracy to regression analysis of small sample data.

Key words: Metrology; Support vector machine; Small sample; Regression analysess; Forecasting; Interval estimation

统计学习理论(Statistic Learning Theory: SLT)[1]是一种专门研究小样本情况下机器学习规律的基本理论和数学构架, 也是小样本统计估计和预测学习的最佳理论。它较好地解决了小样本、非线性、高维数和局部极小点等实际问题。由 Vapnik[1] 提出的基于结构风险最小的学习机器——支持向量机(Support Vector Machine: SVM), 从理论上保证了模型的最大泛化能力, 因此基于支持向量机的回归分析和函数拟合与最小二乘法、神经网络、灰色模型等模型相比, 往往具有更高的预测精度和预测效果[2~4]。

统计学习理论尽管从理论上得到了统计学习方法推广性的界的结论, 但是推广性的界是对于最坏情况的结论, 所给出的界在很多情况下是很松的, 尤其当 VC 维比较高时更是如此。而且推广的界由于函数 VC 维的计算非常困难而无法实施, 同时, 支持向量机回归模型的参数, 如核函数及其相关参数、惩罚因子 C 及损失函数 ε 等对模型的推广性具有很大影响。而实际的数据经常出现数据量少(20 个以

收稿日期: 2005-10-24; 修回日期: 2005-12-15
作者简介: 陈果(1972—), 男, 四川武胜人, 南京航空航天大学民航学院副教授, 博士, 主要从事航空发动机状态监测与故障智能诊断、专家系统、机器学习、模式识别等领域研究。cgzyx@263.net

下)、非等间隔、明显的非线性等特征, 显然, 支持向量机回归模型运用于实际问题求解, 尚存在模型参数选取和预测置信区间的求取等问题。因此, 运用支持向量机进行小样本数据的回归分析和函数拟合, 需要首先解决模型的泛化能力问题, 即选取模型参数以保证最大的推广性, 然后在此基础上, 建立小样本数据的置信区间计算模型, 获取预测点的置信区间。目前的相关研究[2~4]并未很好解决此问题。

本文利用支持向量机对小样本数据建立回归模型, 首先利用统计学习理论的推广性的界, 进行回归模型参数的选取, 然后, 在此基础上, 构造模型预测点置信区间的计算公式求取预测点的置信区间。从而完善支持向量机的回归分析理论。

1　支持向量机回归模型

支持向量回归[5~7]的基本思想是通过一个非线性映射将数据映射到高维特征空间, 并在这个空间进行线性回归。此模型是在分类模型的基础上引进一个修正距离的损失函数, 常用的损失函数有二次函数、Huber 函数、Laplace 函数和 ε 损失函数, 其中 ε 损失函数可以确保对偶变量的稀疏性, 同时确保全局最小解的存在和可靠泛化界的优化。因为这些较好的性质而得到广泛的应用。对于给定的训练样本 (x_i, y_i), $x_i \in R^d$, $y_i \in R$, $i=1, \cdots, n$, 回归的目标就是求下列回归函数: 其中$<w \cdot x>$为 w 和 x 的内积

$$f(x)=<w \cdot x>+b \tag{1}$$

求解以下优化问题:

$$\min \frac{1}{2}<w \cdot w>+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{2}$$

$$\text{s.t} \quad y_i-<w \cdot x_i>+b \leqslant \varepsilon+\xi_i;$$
$$<w \cdot x_i>-y_i+b \leqslant \varepsilon+\xi_i^*$$

其中, C 是预先给定的, 用于控制模型复杂度和逼近误差的折中, C 越大则对数据的拟合程度越高。ε 用于控制回归逼近误差管道的大小, 从而控制支持向量的个数和泛化能力, 其值越大, 则支持向量越少, 但精度会越低。将上述优化问题转化为其相应的对偶问题, 同时引进核方法则转化为求解如下约束问题的最大值, 解得 α_i, α_i^*

$$Q=(\alpha, \alpha^*)=\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)-\frac{1}{2}\sum_{i=1, j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i, x_j) \tag{3}$$

$$\text{s.t} \quad \sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0, \quad 0 \leqslant \alpha_i \leqslant C, i=1, 2, \cdots, n,$$

$$0 \leqslant \alpha_i^* \leqslant C, \quad i=1, 2, \cdots, n$$

出于稳定性考虑, b 的求解采用支持向量的平均值, 其中,

$$\hat{\alpha}=\varepsilon \cdot \text{sign}(\alpha_k-\alpha_k^*)$$

$$b=\text{average}_k\{\hat{\alpha}+y_k-\sum_i(\alpha_i-\alpha_i^*)\cdot K(x_i-x_k)\} \tag{4}$$

这样就得到目标的回归方程:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x, x_i)+b \tag{5}$$

2　支持向量机回归模型参数选取

推广性的界[7], 是统计学习理论中关于经验风险和实际风险之间关系的重要结论, 它分析了学习算法的性能。

由统计学习理论可知, 在经验风险最小化原则下学习机器的实际风险的组成为经验风险和置信范围, 其中置信范围不但受置信水平的 $1-\eta$ 的影响, 还受 VC 维 h 和训练样本数 l 的影响, 并且随着它的增加而单调减少。将其用公式表示如下:

$$R(w) \leqslant R_{\text{emp}}(w)+\Phi(l/h) \tag{6}$$

由上可知: 当 l/h 较小时, 置信范围 Φ 较大, 用经验风险近似实际风险就会有较大的误差, 用经验风险最小化取得的最优界可能就具有较差的推广性; 若 l/h 较大, 则置信范围就会很小, 此时经验风险最小化的最优解就接近实际的最优解。另一方面, 对一特定问题, 样本数 l 固定, 此时, VC 维越高, 则置信范围越大, 导致真实风险和经验风险之间的可能的差就越大, 因此我们在选择模型时, 不但要使经验风险最小化, 还有使 VC 维尽量的小, 从而缩小置信范围, 使期望风险最小。

由此可见, 对于支持向量机的回归模型, 为了使模型有效, 必须要具有一定的推广能力, 然而, 由于式(6)中的置信范围 $\Phi(l/h)$ 计算非常困难, 使得无法为在给定的模型参数下, 准确确定模型的泛化能力, 但是它却为模型参数的选取提供了重要思路, 根据模型推广性的界, 对于样本数目极小(通常小于 20)的数据回归分析, 支持向量机回归模型的参数选取应该遵循以下原则:

(1)为了避免 l/h 过小, 多项式核函数的多项式次数过高将使模型变得复杂, 虽然对已有数据能很好拟合, 但是不具有推广能力, 显然模型无效, 通常确定其取值范围为 1 ~ 2。

(2)损失函数的参数 ε 通过控制回归逼近误差

管道的大小，从而达到控制支持向量的个数和泛化能力的目的，其值越大，精度越低，则支持向量越少。为了在拟合精度和泛化能力之间平衡，ε 的取值范围一般为(0.000 1~0.01)。

(3)惩罚因子 C 用于控制模型复杂度和逼近误差的折中，C 越大，模型越复杂，则对数据的拟合程度越高。因此，为了控制模型复杂程度，通常 C 应取值偏小，但是为了使模型的经验误差不要过大，C 的取值又不能太小，通常确定其范围为(1~1 000)。

总之，要构造有效的支持向量机回归模型，不仅要使模型对已有数据达到很好的拟合，同时再使模型具有很好的推广能力。上述规则可以作为小样本数据下建立支持向量回归模型的基本原则。

3 支持向量机预测点置信区间计算

建立了有效的支持向量机回归模型后，对未知点进行预测，不能仅仅满足于预测点的取值，更重要的是要能够给出预测点在某一置信度下的置信区间。然而，不像经典的最小二乘法，支持向量机回归模型未能给出在对应预测模型参数下的置信区间计算公式。由于在建立了有效的回归模型后，对预测点的预测区间估计可以采用独立于模型之外的主元法。因此，本文基于主元法，引入基于正态分布和基于 t 分布的两种预测区间估计法。置信区间的主元法[8]的基本思想是：

(1)寻找一个样本的函数 $Z=Z(\xi_1, \cdots, \xi_n; \theta)$，此函数只含有待估参数 θ，而不含其它参数，并且 Z 的分布已知且不依赖参数 θ。

(2)对给定的置信系数 $1-\alpha$，定出常数 λ_1、λ_2，使得：

$$P(\lambda_1 \leqslant Z \leqslant \lambda_2)=1-\alpha$$

(3)从不等式 $\lambda_1 \leqslant Z(\xi_1, \cdots, \xi_n; \theta) \leqslant \lambda_2$ 中解得 $\theta_1 \leqslant \theta \leqslant \theta_2$，即为 θ 的置信系数为 $1-\alpha$ 的置信区间。

3.1 基于正态分布的预测区间求解

对于任意的真实值 y 和预测值 $\hat{y}$ 之间存在关系为 $y=\hat{y}+e$，e 为误差，不妨假设 $e \sim N(0, \sigma^2)$，于是 $y_0 \sim N(\hat{y}, \sigma^2)$，则有：$\dfrac{y_0-\hat{y}_0}{\sigma} \sim N(0, 1)$。

根据置信区间的主元法

$$Z=\frac{y_0-\hat{y}_0}{\sigma} \sim N(0, 1) \tag{7}$$

则可求得置信度为 95%即 $\alpha=0.5$ 的预测区间为：

$$[\hat{y}_0-1.96\sigma, \hat{y}_0+1.96\sigma]$$

3.2 基于 t 分布的预测区间求解

定理1：

设 $y \sim N(u, \delta^2)$，$y_1, \cdots, y_n$ 为 y 的样本，$\hat{y} \sim N(\hat{u}, \delta^2)$，$\hat{y}_1, \cdots, \hat{y}_n$ 为 $\hat{y}$ 的样本，且两个样本相互独立，样本方差：

$$S_1^2=\frac{1}{n_1}\sum_{i=1}^{n_1}(y-\bar{y})^2, \quad S_2^2=\frac{1}{n_2}\sum_{i=1}^{n_2}(\hat{y}-\bar{\hat{y}})^2$$

样本均值：

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \bar{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$$

则：

$$\sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} \cdot \frac{(\bar{y}-\bar{\hat{y}})-(u-\hat{u})}{\sqrt{n_1 S_1^2+n_2 S_2^2}} \sim t(n_1+n_2-2) \tag{8}$$

根据区间估计的主元法：

a. 令主元

$$Z=\sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} \cdot \frac{(\bar{y}-\bar{\hat{y}})-(u-\hat{u})}{\sqrt{n_1 S_1^2+n_2 S_2^2}}$$

b. 对给定的置信系数 $1-\alpha$，定出常数 λ，使得它满足：$P(|Z|<\lambda)=1-\alpha$，查 t 分布表可得：

$$\lambda=t_{\alpha/2}(n_1+n_2-2)$$

c. 解不等式 $|Z|<t_{\alpha/2}(n_1+n_2-2)$，则可以得到相应参数的区间估计。

由训练样本集 $T=\{(x_1, y_1), \cdots, (x_n, y_n)\}$ 所得到的模型的预测样本集为 $F=\{(x_1, \hat{y}_1), \cdots, (x_n, \hat{y}_n)\}$，令 $e_1=y_i-\hat{y}$ 则得到了误差集 $E=\{e_1, \cdots, e_n\}$，不妨假设误差总体 $e \sim N(0, \sigma^2)$，则有 E 为 e 的样本，同时令 $e_{n+1}=y_{n+1}-\hat{y}_{n+1}$ 为单点预测的误差，来自误差总体，且与 E 中元素相互独立，则由定理1可知：

$$\eta=\sqrt{\frac{n-1}{n+1}} \cdot \frac{e_{n+1}-\bar{e}}{S} \sim t(n-1) \tag{9}$$

其中：$\bar{e}=\dfrac{1}{n}\sum_{i=1}^{n} e_i$，$S^2=\dfrac{1}{n}\sum_{i=1}^{n}(e_i-\bar{e})^2$

取 $\alpha=0.05$，则求解 $P(|\eta| \leqslant \lambda)=1-\alpha$，得 $|\eta| \leqslant t_{\alpha/2}(n-1)$ 则有：

$$\bar{e}-t_{\alpha/2}(n-1)S\sqrt{\frac{n+1}{n-1}} \leqslant e_{n+1} \leqslant \bar{e}+t_{\alpha/2}(n-1)S\sqrt{\frac{n+1}{n-1}}$$

又因为 $e_{n+1}=y_{n+1}-\hat{y}_{n+1}$，则可得 y_{n+1} 的预测区间为：

$$\left[\hat{y}_{n+1}+\bar{e}-t_{\alpha/2}(n-1)S\sqrt{\frac{n+1}{n-1}}, \hat{y}_{n+1}+\bar{e}+t_{\alpha/2}(n-1)S\sqrt{\frac{n+1}{n-1}}\right] \tag{10}$$

3.3 两种预测区间估计方法的比较

下面用一算例对上述两种预测区间估计方法进

行比较。设函数原型为 $y=20x+1$, 由于真实数据存在着噪声, 因此利用函数 $y=20x+1+\xi$ 产生点数为 10、20、30、40 的序列, $x\in[0, 1]$, $\xi\sim N(0, 1)$。

利用支持向量机回归算法对其进行拟合, 运用上述基于正态分布和基于 t 分布的置信区间计算方法, 分别求得 $x=0.5$ 时在置信度为 95% 下的预测区间宽度, 同时利用一元线性最小二乘法求出 $x=0.5$ 时在置信度为 95% 下的预测区间宽度。其比较结果如表 1 所示。

从表 1 中可以看出, 样本点很少时, 基于正态分布的预测区间宽度比基于 t 分布的小, 随着样本点数的增加, 预测区间的宽度均在不断的变窄, 同时, 基于正态分布的置信区间宽度、基于 t 分布的置信区间宽度与最小二乘算法的预测区间宽度不断逼近。

事实上, 在实际应用中, 由于样本数目有限, 因此 σ 的估计值与真实值会有偏差, 特别是对小样本偏差会较大, 则此时基于正态分布所求解预测区间将产生较大误差, 然而对基于 t 分布的预测区间由于不依赖样本的 σ 估计, 从其求解结果应该更为准确, 随着样本数量的增加, 样本的 σ 估计将趋于准确, 因此基于正态分布和基于 t 分布的预测区间预测结果应该接近。由此可见, 本文提出的基于 t 分布的预测区间估计方法更为可靠。

表 **1**　预测区间宽度比较

样本点数	正态分布 95% 预测区间宽度	t 分布 95% 预测区间宽度	最小二乘法 t 分布 95% 预测区间宽度
10	2.368 7	2.812 4	3.009 8
20	1.885 1	2.053 4	2.103 3
30	1.916 1	2.028 5	2.028 5
40	1.632	1.703 2	1.701 5

4　应用实例

为了验证本文的方法有效性, 选取某航空发动机的光谱油样分析数据作为算例, 航空发动机油样光谱分析可以实现对发动机的磨损故障的早期预测, 而在实际使用中, 光谱数据具有不等间隔性、数据极少(20 个以内)、明显的非线性等特征。因此采用支持向量机进行光谱数据的回归分析具有较大优势。光谱分析所采用的仪器为美国 Bird 公司的原子发射光谱仪。

4.1　支持向量机与最小二乘法的回归效果比较

选取某型飞机发动机油样中的 Zn 和 Fe 元素含量数据作为样本数据, 该数据具有采样时间间隔长且非等间隔, 数据点少的特点, 根据支持向量机的模型选择原则, 应该采用尽量简单的模型以降低模型的复杂性提高模型的泛化能力, 从而使得模型有效。本文的支持向量机回归模型选取了多项式次数为一次的多项式核函数, 并与基于最小二乘法原理的一元线性回归模型[10] 进行比较, 从图 1、图 2、图 3 和图 4 中不难看出支持向量回归用于小样本数据回归分析的优越性。

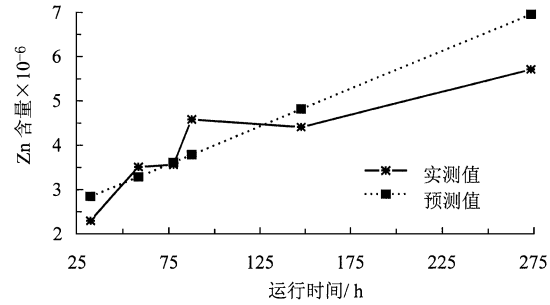

图 **1**　一元线性回归对发动机油样中 Zn 元素的预测

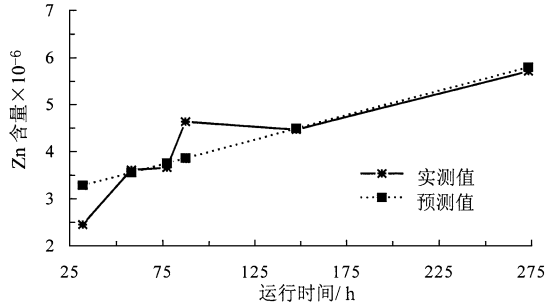

图 **2**　支持向量回归对发动机油样中 Zn 元素的预测

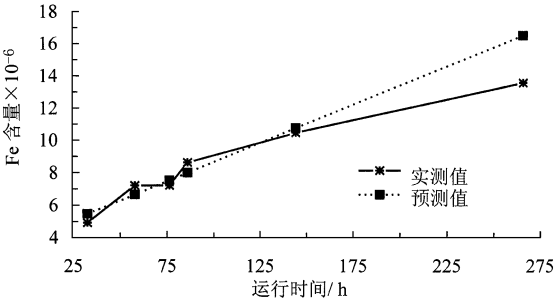

图 **3**　一元线性回归对发动机油样中 Fe 元素的预测

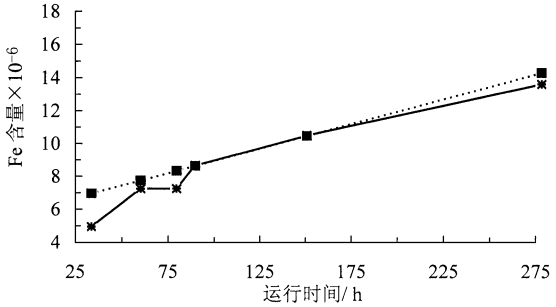

图 **4**　支持向量回归对发动机油样中 Fe 元素的预测

4.2 预测点置信区间的比较

将数据中发动机运行到 78 小时的点去掉，利用剩余的数据建立回归模型，并对 78 小时的数据点进行内插预测，用本文方法求得其 95%置信区间，并与最小二乘法进行比较。发动机运行时间在 78 小时时，测得的 Zn、Fe 含量的真实值分别为：Zn＝3.56×10^{-6}，Fe＝7.1×10^{-6}。预测区间如表 2 所示。

从表 2 中可以看出，最小二乘法的预测区间最宽，因此精度最低，而利用支持向量机回归算法后采用正态分布法计算得到的置信区间，虽然预测区间窄，但由于此方法在数据点极少时对方差估计存在较大误差，因此 SVM 的 t 分布预测区间计算方法更为可靠。

将数据中发动机运行到 273 小时的点去掉，利用剩余的数据建模，对 273 小时的点进行外推预测，并求得其 95%预测区间。发动机运行到 273 h 时，Zn、Fe 含量的真实值分别为：Zn＝5.67×10^{-6}，Fe＝13.38×10^{-6}。预测区间如表 3 所示。

从表 3 中，可以看出，预测点的值均在三个预测区间内，最小二乘法的预测区间最宽，说明了其模型的误差较大，预测精度较 SVM 低，同样，尽管 SVM 的正态分布假设下的置信区间最小，但是对于小样本，其可靠性低于 SVM 的基于 t 分布的置信区间计算结果。

表 **2** 元素含量在发动机运行 78 小时时的内插点预测区间 $\times10^{-6}$

元素	SVM 的正态预测区间	SVM 的 t 分布预测区间	最小二乘预测区间
Zn	[2.473 8, 4.861 7]	[1.822 6, 5.528]	[1.205, 5.987 7]
Fe	[5.755 2, 9.692 7]	[4.651 8, 10.761 8]	[4.211 7, 10.646 7]

表 **3** 元素含量在发动机运行 273 小时时的外推点预测区间 $\times10^{-6}$

元素	SVM 的正态预测区间	SVM 的 t 分布预测区间	最小二乘预测区间
Zn	[4.378 6, 6.755 3]	[3.718 4, 7.406 5]	[2.061 7, 11.734 7]
Fe	[13.194 2, 15.834 3]	[12.454 9, 16.551 6]	[11.342 7, 21.256 5]

5 结 论

(1)运用支持向量机进行小样本数据的回归分析，根据统计学习理论的推广性的界，提出支持向量机的小样本数据回归模型参数选取原则和取值范围。

(2)针对支持向量机的回归模型预测区间计算困难的问题，采用两种基于主元法的预测区间计算方法，即基于正态分布和基于 t 分布的预测区间计算方法，通过对模拟数据计算结果表明，二者在样本数据大时趋于一致，在样本数据较少时，基于正态分布的计算方法的可靠性较低；

(3)针对航空发动机油样光谱数据的非等间隔性、严重非线性及数据量少等特征，运用支持向量机建立了回归分析模型，采用本文的模型参数选取方法和置信区间估计方法对数据进行了回归分析，并与最小二乘法进行了比较，结果充分表明了支持向量机运用于航空发动机油样光谱数据建模的优势和有效性。

[参 考 文 献]

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer—Verlag, 1995, 52~123.

[2] Tay F E H, Lijuan Cao. Application of support vector machines in financial time series forecasting [J]. *Omega*, 2001, 29: 309~317.

[3] 阎辉，张学工，李衍达. 支持向量机与最小二乘法的关系研究[J]. 清华大学学报，2001，41(9)：77~80.

[4] 丁涛，周惠成，黄健辉. 混沌水文时间序列区间预测研究[J]. 水利学报，2004，12(12)：1~7.

[5] Gunn S. Support Vector Machines for Classification and Regression[R]. Technical Report of University of Southampton, 1998.

[6] Hongyu Sun, Chunming Zhang, Bin Ran. Interval prediction for traffic time series using local linear predictor [A]. 2004 IEEE Intelligent Transportation Systems Conference [C]. Washington, D. C., USA: 2004, 410~415.

[7] 边肇祺，张学工，等. 模式识别[M]. 北京：清华大学出版社，1998，284~304.

[8] 顾玉娣，杨纪龙. 概率论与数理统计[M]. 北京：航空工业出版社，2002，121~188.