文章编号：1001-0920(2010)02-0282-05

一种基于主动学习的 SVM 增量训练算法

徐海龙，王晓丹，廖 勇，权 文
（空军工程大学 导弹学院，陕西 三原 713800）

摘 要：针对 SVM 训练学习过程中难以获得大量带有类标注样本的问题，提出一种基于距离比值不确定性抽样的主动 SVM 增量训练算法(DRB-ASVM)，并将其应用于 SVM 增量训练. 实验结果表明，在保证不影响分类精度的情况下，应用主动学习策略的 SVM 选择的标记样本数量大大低于随机选择的标记样本数量，从而降低了标记的工作量或代价，并且提高了训练速度.

关键词：支持向量机；增量训练；主动学习；被动学习；监督学习
中图分类号：TP181 **文献标识码**：A

Incremental training algorithm of SVM based on active learning

XU Hai-long, *WANG Xiao-dan*, *LIAO Yong*, *QUAN Wen*
(Institute of Missile, Air Force Engineering University, Sanyuan 713800, China. Correspondent: XU Hai-long, E-mail: xhl_81329@163.com)

Abstract: To the problem that large-scale labeled samples are difficult to be acquired in the course of SVM training, the active learning strategy is used in the SVM training and an incremental training algorithm of active SVM based on the uncertainty based sampling of distance ratio is proposed in the paper. The experimental results show that the active SVM learning strategy can considerably reduce the labeled samples and costs compared to the passive learning method. And at the same time, it can ensure the accurate classification performance kept as the passive SVM and also expedite the SVM training.

Key words: Support vector machines; Incremental training; Active learning; Passive learning; Supervised learning

1 引 言

支持向量机(SVM)基于坚实严谨的统计学习理论，比传统学习方法具有更好的学习性能和泛化能力，因而得到广泛的应用. 在 SVM 训练学习中，现有的 SVM 训练算法都是基于大量带有类标记样本的监督学习，学习算法以外界给定的已标注样本集作为训练集进行训练，从中归纳出模型. 在一些现实应用中，对样本集进行标注代价昂贵、枯燥乏味或十分困难，而获取未被标注的样本则相对容易. 例如针对基因序列的测试，标注一段基因序列需要进行代价昂贵的实验，而获取基因片段的代价则小得多[2]. 大量未标注样本含有丰富的有助于学习器的信息，如能有效地利用未标注样本，无疑将在一定程度上提高学习算法的性能[3].

传统的监督学习(即被动学习)方法构建正确率满足要求的分类器十分困难. 因此，机器学习领域中的主动学习方法便应运而生，用于有效地解决这类问题. 主动学习方法一般分为 f 和 q 两个独立部分，f 为学习器(学习引擎)，q 为选择函数(选择引擎). 在主动SVM的学习中，学习引擎采用支持向量机已标记的样本集 T 进行训练学习，而选择引擎则根据学习进程，以某种选择策略从候选样本集 U 中主动选择对分类器性能最有利的样本.

主动学习与传统被动学习的最大区别在于选择函数 q. 如何选择新的训练样本，直接关系到整个算法的性能. 不同的样本选择策略对应于不同的主动学习算法. Cohn 等[4] 提出了误差减少的抽样方法；后经 Roy 等[5] 改进，选择使当前分类器对测试集分类误差最小的样本作为候选样本. Lewis 等[6] 提出了不确定性抽样(Uncertainty based sampling, UBS)选择策略，以当前分类器最不能确定其分类的样本作为候选样本. 此外，还有版本空间缩减方

收稿日期：2009-03-18；**修回日期**：2009-05-12.
基金项目：国家自然科学基金项目(60975026)；陕西省自然科学基金项目(2007F19).
作者简介：徐海龙(1981—)，男，陕西韩城人，博士生，从事智能信息处理、支持向量机的研究；王晓丹(1966—)，女，陕西汉中人，教授，博士生导师，从事智能信息处理、雷达目标识别等研究.

法，如 Seung 等[7] 和 Freund 等[8] 提出了委员会投票选择(Query by committee，QBC)，QBag，QBoost 和 Active decorate 算法等[11-13].

文献[4] 提出的基于误差减少的主动学习方法，虽然其学习准确率高，但选择了过多的冗余样本，且在每次选择样本之前，必须搜索整个样本空间才能确定选择哪些样本，因此其学习时间长，计算复杂度高. 文献[7,8] 提出的 QBC 主动学习方法不需要检测整个样本空间，计算复杂度较低，学习速度较快，但基于 QBC 的方法和 UBS 方法都易于选择奇异点样本，这些样本具有一定的不确定性，加入后会加大分类器的分类误差.

本文受文献[6,9,10] 的启发，在分析 SVM 支持向量分布特点的基础上，提出一种基于距离比值的不确定性抽样主动SVM增量训练算法. 该算法能抑制由于加入不确定性样本而产生的误差传播问题以及选择过多冗余样本问题，以尽可能少的标记样本获得较高的 SVM 分类准确率.

2　SVM 增量学习中支持向量的变化

SVM 是从线性可分情况的最优分类面发展而来的. 如图 1 所示两类分类问题的最优分类面 H_{12}：$g(x)=0$，其中线性判别函数 $g(x)=\omega x+b$，其最优分类决策函数为

$$f(x)=\mathrm{sgn}\{(w^* x)+b^*\}=\mathrm{sgn}\Big\{\sum_{i=1}^{n}\alpha_i^* y_i(x_i x)+b^*\Big\}. \tag{1}$$

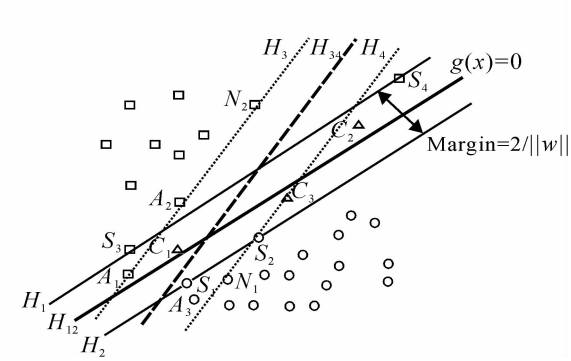

图 1　SVM 支持向量分布

根据 SVM 的机理可知，式(1) 最后仅对少量不为零的 α_i 对应的向量 x_i(即在 H_1 和 H_2 上的支持向量) 进行求解，而与其他向量无关. 根据 Karush-Kuhn-Tucher 条件(简称 KKT 条件)[1]，新增样本加入后，部分违背 KKT 条件的新增样本或原样本将转化为满足 KKT 条件的样本甚至支持向量；部分满足 KKT 条件的新增样本将转化为支持向量；部分满足 KKT 条件的原样本集的非支持向量将转化为支持向量；部分原支持向量将转化为非支持向量(参见图 1).

在图 1 中，$S_1\sim S_4$ 为最优分类面 H_{12} 原支持向量集，$A_1\sim A_3$ 为新增样本，$C_1\sim C_3$ 为 H_1 和 H_2 分类边界间隔的样本. 训练后的最优分类面 H_{34} 的支持向量集由 N_1,S_2,C_3,A_1,A_2,N_2 组成. 满足 KKT 条件的新增样本 A_2，在训练后转化为支持向量；满足 KKT 条件的原样本 N_1 和 N_2 在训练后转化为支持向量；违背 KKT 条件的新增样本 A_1 以及处于原分类间隔的样本 C_2 和 C_3，在训练后转化为满足 KKT 条件的样本，并且 A_1 和 C_3 成为支持向量；原支持向量集中的 S_1,S_3,S_4 在训练后转化为满足 KKT 条件的样本.

由 SVM 的支持向量几何分布可知，图 1 中处于分类间隔的样本 A_1,C_1,C_2,C_3 等，分布于分类边界附近的样本 A_2,N_2,N_1,A_3 等，以及支持向量集在 SVM 的训练学习过程中最不确定的样本，所包含的信息量最为丰富. 若在 SVM 的训练学习过程中，主动地选择这些最不确定的样本，则不仅能加快学习速度，减少存储空间，而且可降低标注样本的代价. 正是受此启发并根据文献[6] 的 UBS 主动学习策略，本文提出了基于距离比值的不确定性主动 SVM (Distance ratio based-active SVM，DRB- ASVM) 增量学习方法.

3　DRB-ASVM 算法思想

从几何理论上分析，样本空间通常呈聚集分布，或经合适的非线性映射后在特征空间呈聚集分布. 支持向量集或分类边界附近的样本只是样本集的一部分，根据第 2 节的分析，这部分样本在主动 SVM 训练学习中是对学习进程最有利的样本. 那么这部分样本在聚集区域如何分布呢？

如图 1 所示，由 SVM 的机理可知，对于所有支持向量，有 $\|g(x)\|=1(g(x)=\omega x+b)$；对于分类面上的样本点，有 $g(x)=0$，对于其他任意样本点，由 SVM 理论有

$$x=x_p\pm d\frac{\omega}{\|\omega\|}. \tag{2}$$

其中：x_p 为向量 x 相对于分类面 $\omega x+b=0$ 的垂足，d 为样本点 x 到分类面的距离，正负号表示样本点在分类面的两侧不同. 式(2) 代入 $g(x)=\omega x+b$，有

$$g(x)=\omega^{\mathrm{T}}x+b=\omega^{\mathrm{T}}\Big(x_p\pm d\frac{\omega}{\|\omega\|}\Big)+b=\omega^{\mathrm{T}}x_p+b\pm\omega^{\mathrm{T}}d\frac{\omega}{\|\omega\|}=\pm d\|\omega\|. \tag{3}$$

即有

$$d=g(x)/\|\omega\|. \tag{4}$$

由图 1 及式(4) 可以看出，样本与最优分类面之间距离的远近，可作为样本属于不同类别的概率的

标准,即可通过 SVM 决策函数中的 $g(x)$ 值度量样本的后验概率. 通过对大量分类模式样本的研究,认为支持向量集及分类边界附近的一部分样本是最不确定性样本,这部分样本是模式的边界向量集合的子集. 本文采用中心距离比值法作为主动学习选择策略,以选择SVM主动学习进程中对学习机最有利的不确定样本.

为了叙述方便,首先给出涉及到的一些定义:

定义 1 某一类样本的平均特征称为该类样本的中心 m. 已知样本向量组 $\{x_1,x_2,\cdots,x_n\}$,则其中心为 $m=\frac{1}{n}\sum_{i=1}^{n}x_i$. 对于非线性可分的模式,采用非线性映射 Φ 将输入空间映射到某一特征空间 H,其中心为

$$m_\Phi=\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i). \tag{5}$$

定义 2 两个样本之间的特征差异称为样本距离. 已知两个 N 维样本向量 x_1 和 x_2,其样本距离为

$$d(x_1,x_2)=\|x_1-x_2\|_2=\sqrt{\sum_{i=1}^{N}(x_1^i-x_2^i)^2}.$$

对于非线性可分模式,已知两个向量 z_1 和 z_2,经非线性映射 Φ 作用后,映射到特征空间 H,则这两个向量在特征空间的 Euclidean 距离为

$$d^H(z_1,z_2)=\sqrt{K(z_1,z_1)-2K(z_1,z_2)+K(z_2,z_2)},$$

其中 $K(\cdot,\cdot)$ 为核函数.

定义 3 中心距离是指各样本到其中心的距离. 假设一类模式的 N 维训练样本向量为 $\{x_1,x_2,\cdots,x_n\}$,其中心为 m,则中心距离为

$$d(x_i,m)=\|x_i-m\|_2=\sqrt{\sum_{j=1}^{N}(x_i^j-m^j)^2}. \tag{6}$$

平均中心距离是指各样本中心距离的平均值. 假设一类模式的 N 维训练样本向量为 $\{x_1,x_2,\cdots,x_n\}$,其中心为 m,则其平均中心距离为

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}d(x_i,m).$$

定义 4 中心距离比值是指某一样本的中心距离与该类模式平均中心距离的比值,即

$$\mathrm{Ratio}_{x_i}=d(x_i,m)/\bar{d}. \tag{7}$$

对于非线性可分模式,假设训练样本集为 $\{x_1,x_2,\cdots,x_n\}$,经非线性映射 Φ 作用后映射到某一特征空间 H. 由于不知映射 $\Phi(x_i)$ 的具体表达形式,无法根据式(5)求得中心,则映射后在特征空间的中心距离采用下式计算:

$$d^H(x,m_\Phi)=\sqrt{K(x,x)-\frac{2}{n}\sum_{i=1}^{n}K(x,x_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,x_j)}. \tag{8}$$

其平均中心距离和中心距离比值如下:

平均中心距离

$$\bar{d}^H=\frac{1}{n}\sum_{i=1}^{n}d^H(x_i,m_\Phi), \tag{9}$$

中心距离比值

$$\mathrm{Ratio}_{\Phi(x_i)}=d^H/\bar{d}^H. \tag{10}$$

定义 5 已知模式的训练样本为 $\{x_1,x_2,\cdots,x_n\}$,经非线性映射 Φ 作用后映射到某一特征空间 H,得到它们在空间 H 的中心距离比值. 设定阈值 r_x,则模式的边界向量集合为

$$\{x_i\mid \mathrm{Ratio}_{\Phi(x_i)}>r_x,\ i=1,2,\cdots,n\}. \tag{11}$$

定义 6 将 $y_ig(x_i)\leqslant 1$ 称为违背广义 KKT 条件(即违背 KKT 条件)的样本和支持向量集的并集. 对应的满足广义 KKT 条件为 $y_ig(x_i)>1$,对应的样本为位于分类间隔之外且被分类器正确分类的样本.

根据SVM支持向量分布理论和主动学习策略,距离分类面越近的样本点所具有的信息越丰富. 在本文提出的算法中,首先使用式(4)进行预选,以减少标注样本数量;然后使用中心距离比值,选择模式的边界向量作为主动学习中的选择策略,据此在主动SVM学习过程中尽量选择边界向量集中的样本. 该算法称为基于距离比值的主动 SVM(DRB-ASVM)增量学习算法.

4 DRB-ASVM 算法描述

4.1 符号意义

X_o^k:第 k 次训练所用原始样本集;

Ω_o^k:第 k 次训练用原始样本集得到的 SVM 分类器;

X_o^{sv} 分类器 Ω_o^k 的支持向量集;

X_o^{sk}:第 k 次训练原始样本集中满足 Ω_o^k 的广义 KKT 条件的样本集;

X_o^{vk}:第 k 次训练原始样本集中违背 Ω_o^k 的广义 KKT 条件的样本集;

X_I^k:第 k 次取出的新增样本,当 $k=0$ 时表示初始样本集;

X_I^{sk}:新增样本中满足 Ω_o^k 的广义 KKT 条件的样本集;

X_I^{vk}:新增样本中违背 Ω_o^k 的广义 KKT 条件的样本集;

X_r^{sk}:根据式(7)和阈值 r_x 保留仅大于 r_x 的样本后,剩余的满足 Ω_o^k 的广义 KKT 条件的样本集;

X_u^{sk}:X_o^{sk} 和 X_I^{sk} 的并集.

4.2　运算过程

输入：已标注样本集 L（至少正负样本各一个），未标注或候选样本集 L_U，每次采样的样本数 m，边界阈值 r_x，距离阈值 d，终止条件 S；

输出：分类器 Ω，预标注样本.

Step1：判断第 k 次训练结束后，SVM 分类器 Ω_o^k 是否达到终止条件 S. 是则结束训练，否则转 Step2.

Step2：判断 L_U 中是否有未标注样本，是则首先对未标注样本用 Ω_o^k 预先标注；然后根据式(4) 选择 m 个最不确定的样本，作为下一次增量训练样本 X_I^k；最后将 X_I^k 中未标记的样本交由领域专家进行正确标注.

Step3：检验 X_I^k 中的样本是否违背 Ω_o^k 的广义 KKT 条件. 根据检验结果，X_I^k 被分为 X_I^{sk} 和 X_I^{vk}.

Step4：将 X_o^{sk} 和 X_I^{sk} 合并得 X_u^{sk}. 根据标示符，将集合分为正例样本集 A^+ 和负例样本集 A^-. 根据式(10) 及阈值 r_x 进行处理，保留仅大于 r_x 的样本，得到剩余正例样本集 A_r^+ 和剩余负例样本集 A_r^-，合并二者得到 X_I^{sk}.

Step5：将 X_I^{sk}，X_I^{vk} 和 X_o^{vk} 合并得到 X_o^{k+1}，对其训练得到新的分类器 Ω_o^{k+1}，并生成 X_o^{sk+1} 和 X_o^{vk+1}，$k=k+1$，转 Step1.

5　实验结果及分析

为验证算法的有效性，通过实验比较 DRB-ASVM，QBC，随机采样和错误缩减采样在增量学习过程中的算法性能. 实验数据取自机器学习领域权威且标准的 UCI 数据集.

为便于验证算法及对未标注样本的处理，对多类数据集仅取其中两类并对一些样本去掉标记. 验证 SVM 使用 Steve Gunn SVM Toolbox，$C=100$，核函数为多项式函数，参数值为 2，算法中的阈值 d 根据初始训练的分类器决定. 对每个数据集重复5次采样，取其平均值.

表 1　各算法在增量学习过程中的采样次数比较

数据集	算法				正确率/%
	随机采样	错误缩减采样	QBC	DRB-ASVM	
balance-scale	65.4	28.9	56.8	**27.5**	83.6
breast-cancer-wisconsin	160.3	78.5	89.7	**67.5**	89.8
Mushroom	2340.2	560.3	**78.7**	81.7	93.02
ionosphere	50.5	47.8	**25.9**	45.6	89.8
house-votes-84	175.7	88.5	112.7	**69.5**	94.5
hepatitis	90.6	56.3	48.2	**38.9**	85.8
credit-screening	450.6	**180.5**	250.3	190.5	90.1
glass	160.4	70.5	56.5	**48.6**	89.5
soybean-large	259.4	302.4	278.3	**235.3**	93.5

4 种算法在 9 个数据集上的采样次数和正确率如表 1 所示，其中各算法对数据集采样次数较少者以黑体标出. 可以看出，本文提出的 DRB-ASVM 增量学习方法在整体上好于基于 QBC 和错误缩减采样的主动学习策略. 各种主动学习策略的采样次数均少于随机采样法，减小了标注代价，加快了训练学习进程.

4 种算法在 balance-scale 数据集和 breast-cancer-wisconsin 数据集的实验结果如图 2 和图 3 所示. 可以看出，本文提出的 DRB-ASVM 增量学习算法对这两个数据集的分类正确率高于其他算法的分类正确率. 另外，错误缩减采样的部分正确率低于其他算法的正确率. 这是由于采用损失函数来估计未来期望误差率，不够精确而导致估计失真度较大.

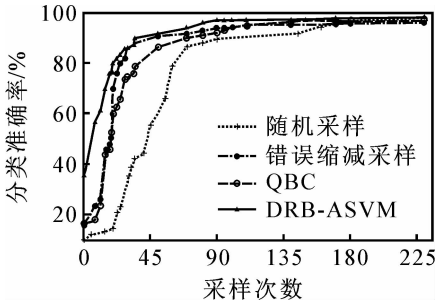

图 2　balance-scale 数据集实验

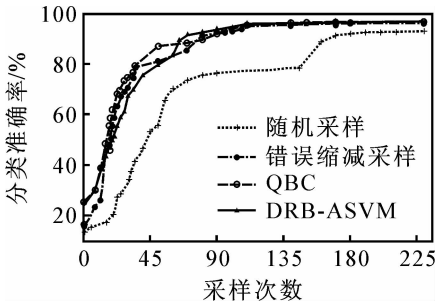

图 3　breast-cancer-wisconsin 数据集实验

算法中使用的阈值 r 和 d 对实验结果也有重要影响：一方面影响最不确定样本的选择范围，另一方面影响算法的训练精度. 在算法实验中，本文根据初始分类器计算分类间隔，阈值取最大分类间隔的 2 倍，随着 SVM 增量训练过程，根据最大分类间隔的变化进行阈值修正，以加快训练进程.

6　结　　论

本文在分析 SVM 支持向量几何分布特性以及主动学习原理的基础上，提出了基于距离比值的不确定主动 SVM 学习方法. 将其应用于 SVM 增量训练学习过程，通过实验验证了算法的有效性和可行性. 该方法可有效减少主动学习算法的采样次数，降低标注代价.

参考文献(References)

[1] Vapnik V. The nature of statistical learning theory[M]. New York: Springer Press, 1995.

[2] 龙军, 殷建平, 祝恩, 等. 选取最大可能预测错误样例的主动学习算法[J]. 计算机研究与发展, 2008, 45(3): 472-478.
(Long J, Yin J P, Zhu E, et al. An active learning algorithm by selecting the most possibly wrong-prediction instances[J]. J of Computer Research and Development, 2008, 45(3): 472-478.)

[3] 龙军, 殷建平, 祝恩, 等. 主动学习中一种基于委员会的错误分类采样算法[J]. 计算机工程与科学, 2008, 30(4): 69-72.
(Long J, Yin J P, Zhu E, et al. A committee-based misclassification sampling algorithm in active learning[J]. Computer Engineering & Science, 2008, 30(4): 69-72.)

[4] Cohn D A, Ghahramani Z, Jordan M I. Active learning with statistical models[J]. J of Artificial Intelligence Research, 1996, 4: 129-145.

[5] Roy N, McCallum A K. Toward optimal active learning through sampling estimation of error reduction[C]. Proc of 18th Int Conf on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 441-448.

[6] Lewis D D, Gale W. A sequential algorithm for training text classifiers[C]. Proc of 17th Annual Int ACM-SIGIR Conf on Research and Development in Information Retrieval. Dublin: Springer-Verlag, 1994: 3-12.

[7] Seung H S, Opper M, Sompolinsky H. Query by committee[C]. Proc of 15th Annual ACM Workshop on Computational Learning Theory. Pittsburgh: Morgan Kaufmann, 1992: 287-294.

[8] Freund Y, Seung H S, Samir E, et al. Selective sampling using the query by committee algorithm[J]. Machine Learning, 1997, 28(2/3): 133-168.

[9] 张翔, 肖小玲, 徐光祐. 基于最大熵估计的支持向量机概率建模[J]. 控制与决策, 2006, 21(7): 767-770.
(Zhang X, Xiao X L, Xu G Y. Probabilistic outputs for support vector machines based on the maximum entropy estimation[J]. Control and Decision, 2006, 21(7): 767-770.)

[10] 徐海龙, 王晓丹, 史朝晖, 等. 一种基于距离比值的支持向量机增量训练算法[J]. 空军工程大学学报, 2008, 9(4): 29-33.
(Xu H L, Wang X D, Shi Z H, et al. An incremental training algorithm of SVM based on the distance ratio [J]. J of Air Force Engineering University, 2008, 9(4): 29-33.)

[11] 朱红斌, 蔡郁. 基于主动学习支持向量机的文本分类[J]. 计算机工程与应用, 2009, 45(2): 134-136.
(Zhu H B, Cai Y. Text categorization based on active learning support vector machines [J]. Computer Engineering and Application, 2009, 45(2): 134-136.)

[12] 胡正平. 基于最佳样本标记的主动支持向量机学习策略[J]. 信号处理, 2008, 24(1): 105-107.
(Hu Z P. An active learning strategy of SVM via optimal selection of labeled data[J]. Signal Processing, 2008, 24(1): 105-107.)

[13] 陈耀东, 王挺, 陈火旺. 半监督学习和主动学习相结合的浅层语义分析[J]. 中文信息学报, 2008, 22(2): 70-75.
(Chen Y D, Wang T, Chen H W. Combining semi-supervised learning and active learning for shallow semantic parsing [J]. J of Chinese Information Processing, 2008, 22(2): 70-75.)

(上接第 281 页)

[6] Yager R R. On the Dempster-Shafer framework and new combination rules[J]. Information Science, 1989, 41(2): 93-137.

[7] 孙全, 叶秀清, 顾伟康. 一种新的基于证据理论的合成公式[J]. 电子学报, 2000, 28(8): 117-119.
(Sun Q, Ye X Q, Gu W K. A new combination rules of evidence theory[J]. Acta Electronica Sinica, 2000, 28(8): 117-119.)

[8] 张多林, 潘泉, 张洪才, 等. 一种基于信息源可信度的证据组合新方法[J]. 系统工程与电子技术, 2008, 30(7): 1210-1213.
(Zhang D L, Pan Q, Zhang H C, et al. Combination rule of evidence theory based on credibility of sensor[J]. Systems Engineering and Electronics, 2008, 30(7): 1210-1213.)

[9] 韩崇昭, 朱洪艳, 段战胜, 等. 多源信息融合[M]. 北京: 清华大学出版社, 2006.
(Han C Z, Zhu H Y, Duan Z S, et al. Multi-source information fusion[M]. Beijing: Tsinghua University Press, 2006.)

[10] Pratap Misra, Per Enge. Global positioning system — Singal, measurements and performance[M]. 2ed. Lincoln: Ganga-Jamuna Press, 2006.