

文章编号:1001-9081(2007)02-0363-03

有效提高 SVM 参数搜索效率的样本集缩减策略

段崇雯,成礼智

(国防科技大学 理学院,湖南 长沙 410073)

(cynthia_1228@163.com)

摘要:核函数及相关参数的选择是支持向量机中的一个重要问题,它对模型的推广能力有很大的影响。当有大量样本参与训练的时候,寻找最优参数的网格搜索算法将消耗过长的时间。针对这一问题,提出一种舍弃非支持向量的样本点的策略,从而缩减了训练样本集。能够在基本保持原有测试准确度的前提下,将搜索时间减少一半。

关键词:支持向量;样本集缩减;网格搜索;最优参数选取

中图分类号:TP181 **文献标识码:**A

Sample set shrinking strategy efficiently improving parameters seeking of support vector machines

DUAN Chong-wen, CHENG Li-zhi

(Department of Science, National University of Defense Technology, Changsha Hunan 410073, China)

Abstract: The choice of kernel function and relative parameters plays an important role in Support Vector Machines (SVMs). It greatly influences the generalization performance of SVMs. It is time consuming to seek for optimal parameters when the training sample set is large. Concerning this problem, a sample set shrinking strategy was proposed. This method took some of the non-support-vector samples out of the training set; therefore efficiently reduced the set size. That is to say, with half the time consumed, a model can be constructed with testing accuracy just slightly changed.

Key words: support vector; sample set shrinking; grid searching; optimal parameters selection

0 引言

支持向量机(Support Vector Machine, SVM)是 20 世纪 90 年代由文献[1]提出的一种新的学习机器。与神经网络等人工智能领域现有的学习机相比,一方面, SVM 使用了结构风险最小化原则,因而具有较强的小样本学习能力和对新样本的推广能力;另一方面,由于采用了从数据空间到特征空间的非线性映射,不但增强了机器的非线性处理能力,避免了所谓的“维数灾难”,而且可以通过选择适当的映射函数避免算法陷入局部极小解。因此, SVM 在模式识别和函数估计等众多领域有着广泛的应用。

实验中发现, SVM 的性能在很大程度上依赖于相关模型,特别是模型参数的选取。一般用于寻找最优参数的网格搜索算法在遭遇较大的训练样本集合时,会产生搜索时间过长的问題。因此,国内外不少研究者设计了更为有效的搜索算法。文献[2]提出一种基于混合遗传算法的搜索方法;文献[3]改进了最优化参数选择方法;文献[4]根据 Margin/Radius 形式的误差上界和梯度下降法推导出解析的参数调整算式。

本文深入发掘支持向量在 SVM 模型中的决定性作用,以适当的支持向量的集合代替原训练样本集,并据此搜索最优参数,生成相应的 SVM,从而缩减了算法消耗的时间。同时,作为一种优化策略,它可以与许多基于训练样本的参数选择算法结合,因此具有广泛的适用性。

1 分类 SVM 及模型选择问题

1.1 分类 SVM 原理

已知训练样本集 $\{(x_i, y_i)\}_{i=1}^l$, 其中 $x_i \in R^m$ 对应第 i 个样本, $y_i \in \{-1, +1\}$ 表示该样本类别(这里仅考虑两类问题,多类问题可以转化为多个两类问题加以解决)。SVM 首先通过一个映射 φ (通常是非线性的), 将样本空间映射到所谓的特征空间, 然后在特征空间中寻找一个形如 $w_0^T \varphi(x) + b_0 = 0$ 的分类超平面, 对(特征空间中)线性可分的情形, 要求两类样本距分类面的最小间隔尽可能的大。而对于线性不可分情形, 增加松弛变量 ξ_i , 求解下面的约束优化问题:

$$\begin{aligned} \min_{w, b} \Phi(w) &= \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i \\ s. t. \end{aligned} \quad (1)$$

$$\begin{aligned} y_i(w^T \varphi(x_i) + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0, i = 1, 2, \dots, l \end{aligned}$$

C 是一个预先给定的常数, 称为误差惩罚系数。引入 Lagrange 乘子 α_i 转化成相应的对偶问题:

$$\begin{aligned} \max_{\alpha} Q(\alpha) &= \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ s. t. \end{aligned} \quad (2)$$
$$\begin{aligned} \sum_{i=1}^l \alpha_i y_i &= 0 \\ 0 \leq \alpha_i \leq C, i &= 1, 2, \dots, l \end{aligned}$$

收稿日期:2006-09-01;修订日期:2006-10-30 基金项目:国家自然科学基金资助项目(60573027)

作者简介:段崇雯(1981-),女,四川德阳人,硕士研究生,主要研究方向:计算数学、统计学习理论;成礼智(1962-),男,湖南常德人,教授,博士,主要研究方向:信息科学中新型算法与软件、小波变换与图像处理、应用数学。

其中, $K(x_i, x_j) \equiv \varphi^T(x_i)\varphi(x_j)$ 称为核函数。相应的分类超平面为:

$$\sum_{i=1}^l \alpha_i y_i K(x_i, x) + b_0 = 0 \quad (3)$$

从(2)、(3)两式可以看出, SVM 中仅仅涉及到了两个特征映射的内积 $K(x_i, x_j)$ 和 $K(x, x_i)$, 使我们可以完全不必考虑 $\varphi(x)$ 的具体形式, 从而避免了所谓的“维数灾难”。

1.2 模型选择

从上文可以看出, 确定一个 SVM 需要预先选择恰当的核函数并给定相应的参数以及误差惩罚系数 C , 此即 SVM 的模型选择问题。SVM 的模型选择问题包括选择恰当的核函数及其相应参数, 以及误差惩罚系数。

1.2.1 核函数的选择

目前得到研究的核函数主要有线性核函数、多项式核函数、Gaussian 核函数(也称径向基函数)以及 Sigmoid 核函数。文献[5]证明了, 当模型参数满足一定的关系时, 线性核函数的 SVM 可以视为 Gaussian 核函数 SVM 的一种特例; 文献[6]通过理论分析和数值实验指出, sigmoid 核函数 SVM 的表现不会比 Gaussian 的更好; 此外, 多项式核函数在阶数较高时可能导致计算上的困难。

另一方面, 文献[7]从函数回归问题中正则算子的角度, 分析指出 Gaussian 核函数相应的正则算子会对拟合函数任意阶的不光滑进行惩罚。因此在一般的光滑性假设下, 它会有出色的表现。

因此, 如果没有充分获得关于样本集和分类问题的先验信息, Gaussian 核函数是应用得最为广泛的 SVM 核函数。其具体形式为:

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

本文的实验数据也是基于 Gaussain 核 SVM 得到的。

1.2.2 核函数参数和误差惩罚系数 C 的选择

Gaussian 核函数中仅有一个参数 σ^2 , 它隐含地改变着映射函数 $\varphi(x)$, 从而控制着特征空间的性能。若将 Gaussian 核对应的特征映射视为从数据空间到一个再生核 Hilbert 空间的映射^[1]:

$$x \mapsto \varphi(x) = K(x, \cdot) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

可以看出, 若 $\sigma^2 \mapsto 0$, $\varphi(x)$ 退化成 Dirac 函数 δ_x , 因此映射样本的分布特征与原数据空间相同, 若 $\sigma^2 \mapsto \infty$, $\varphi(x) \equiv 1 \quad \forall x$, 从而将所有样本点归为一类。

而(1)式中的系数 C 控制着对样本错分的惩罚程度, 即体现了模型分类性能和复杂程度的某种折中。文献[5]中指出, $C \rightarrow 0$ 时, SVM 将所有样本都归入样本数目较多的那一类, 导致所谓的欠学习; 而若 $C \rightarrow \infty$, SVM 能够对两类样本进行完全正确的分类, 但这时的模型较为复杂, 训练很慢, 而且如果样本带噪, 则可能导致过学习, 不具备良好的推广性。

因此, 对参数 (C, σ^2) 值的选取要结合实际应用背景以及所给数据的先验知识。一般的做法是在一定取值范围内, 生成关于 (C, σ^2) 的二维网格, 在每一个网格点上采用交互检测^[9](Cross-Validation, CV)的方法求得相应误差, 而最终选取具有最小 CV 误差的一组参数。这种方法具有较好的适应性, 但缺点在于当训练样本集较大或维数较高时, 网格搜索所消耗的时间会相当长。对此, 可以采用样本集缩减和特征筛选^[10](Feature Selection, FS)加以解决。本文主要给出了一种

缩减训练样本集的思想。

2 训练样本集缩减策略

2.1 基本思想

对偶优化问题(2)的解中, 只有一部分 $\alpha_i \neq 0$, 对应的训练样本点称为支持向量。相比于其他样本点, 支持向量有如下性质:

第一点, 非支持向量的样本点的 Lagrange 因子:

$$\alpha_i = 0, i \notin SV \quad (4)$$

因此(3)式中的求和只需对 $i \in SV$ 的样本点进行, 也就是说, 支持向量能够将分类超平面完全确定下来;

第二点, 对非支持向量的样本点, 由(4)及对偶优化问题的 KKT(Karush-Kuhn-Tucker)条件求得相应的松弛因子

$$\xi_i = 0, i \notin SV$$

因此, 非支持向量的样本点总是会被正确地分类, 错误仅仅发生在支持向量样本点中。在原训练样本集上通过优化模型参数得到最小误差的过程, 可以简化到支持向量的集合上来进行;

第三点, 对于固定的样本集, 大部分样本的类别归属是可以由它们属性值(输入值)确定的, 也就是说, 即便是对应于不同参数的 SVM, 其支持向量集也有很大的相似性。所以, 某一对参数对应的支持向量集合, 可以被认为是其他参数对的支持向量集合的近似。

支持向量的这些特殊性质正是本文工作的基础。我们可以就此认为, 非支持向量的样本点在学习过程中的作用是几乎可以忽略的, 希望从训练样本集中提取一部分支持向量构成缩减的样本集用于训练, 而又能保持模型的本质属性, 进而提出下面的样本集缩减策略。

2.2 缩减训练集策略

对于缩减的训练样本集, 我们希望它所包含的支持向量的个数大约占到原样本集合的一半, 因为如果支持向量太多, 起不到缩减的目的, 而若太少, 又不能有效地反映分类问题的本质。当样本较多时, 若过分强调样本之间的差异(σ^2 较小), 会使模型过于复杂, 因此当训练集大小为 l 时, 取 $\gamma = \frac{1}{2\sigma^2} = \frac{1}{l}$ 是一个较为合理的选择。另一方面, 我们取 $C = 1$, 这意味着在(1)式中对目标函数的两个部分赋以相同的权重, 即认为分类准确度与推广能力同等重要。用这两个参数训练 SVM, 得到的支持向量将作为下一步网格搜索的训练集。

我们采用目前应用最为广泛的序列最小化优化算法^[11](Sequential Minimal Optimization, SMO)。该算法将优化过程分解成逐点进行, 因此训练时间与样本数目大约成正比。可以预见, 缩减训练集后的算法会比原来提高一倍的速度。

2.3 算法步骤

- 1) 取 $\gamma = \frac{1}{l}$ 和 $C = 1$ 进行初步训练, 将得到的支持向量作为缩减的样本集;
- 2) 在缩减的样本集上进行最优参数搜索, 得到最优参数;
- 3) 用 2) 中得到的最优参数重新训练学习机, 得到可以用于测试或预测的 SVM。

3 实验结果

我们采用 IDA 基准数据库^[12]中的 breast-cancer, banana,

splice,image 和 twonorm 五组数据,其训练和测试样本数及输入维数见表1。

表1 样本集

数据集名称	训练样本数	测试样本数	输入维数
breast - cancer	200	77	9
banana	400	4900	2
splice	1000	2175	60
image	1300	1010	18
twonorm	400	7000	20

从表1中看出,数据集 banana 和 twonorm 对模型的推广能力是很大的考验,而 splice 和 image 的训练时间将会成为最大的瓶颈。

按照上面描述的方法,首先取 $\gamma = \frac{1}{l}$ 和 $C = 1$ 进行初步训练,得到缩减的样本集(表2)。

表2 缩减后的样本集

数据集	缩减样本集大小	占原数据集比例(%)
breast - cancer	127	63.5
banana	167	41.75
splice	607	60.7
image	452	46.69
twonorm	120	30

进行网格搜索时,按 LIBSVM^[9]中的默认参数,取:

$$C \in \{2^{-5}, 2^{-3}, 2^{-1}, \dots, 2^{11}, 2^{13}, 2^{15}\}$$

$$\gamma = \{2^{-15}, 2^{-13}, 2^{-11}, \dots, 2^{-1}, 2^1, 2^3\}$$

比较用缩减前后样本集进行网格搜索所消耗的时间、得到的最优参数,以及用最优参数测试新样本得到的误差,结果如表3(表中每项上一行为缩减后结果,下一行为缩减前结果)。

表3 缩减样本集前后的实验结果

数据集	breast-cancer	banana	splice	image	twonorm
搜索时间(s)	56	283	532	530	19
	97	593	1192	1196	49
最优 C 值	128.0	2048.0	128.0	2048.0	2.0
	8.0	0.125	32.0	512.0	0.125
最优 γ 值	0.00048828125	0.5	0.001953125	0.03125	0.5
	0.03125	2.0	0.03125	0.03125	0.5
测试误差	68.8312%	87.0204%	89.7011%	97.5248%	96.6429%
	(53/77)	(4264/4900)	(1951/2175)	(985/1010)	(6765/7000)
	74.026%	87.5714%	90.1149%	98.0198%	97.7571%
	(57/77)	(4291/4900)	(1960/2175)	(990/1010)	(6843/7000)

从表3可以看出,用缩减后的样本集进行网格搜索得到的最优参数的办法,在大大缩短搜索时间的基础上(如对后四个样本集,搜索时间减少了一半以上),仍然可以得到很好的测试结果。

4 结语

SVM 较之神经网络等学习算法有着更深刻的理论基础,它具有较强的推广能力,有效地解决(避免)了“维数灾难”,并能保证最优解的全局性,这些优点使它在各个领域的应用日趋广泛。但实验中发现,目前最常用的求解 SVM 学习问题的序列最小化优化算法(SMO),虽然有效地解决了内存占用问题,却也陷入训练时间过长的困境。特别是当训练样本集很大或输入维数很高的时候,一次训练的时间便长得惊人,网格搜索最优参数的耗时更是令人生畏。因此,设计有效的样本集缩减和特征筛选算法,对 SVM 的推广有着至关重要的作用。本文提出的缩减样本集的方法,充分利用了支持向量对分类问题的决定性作用,方法简单易行,而且效果显著。此外可以看出,甚至可以在缩减的样本集中运用其他更为有效的参数优化算法(如文献[2]~文献[4]),以进一步提高速度。

参考文献:

- [1] VAPNIK V. Statistical Learning Theory[M]. New York: John Wiley & Sons, 1998.
- [2] 齐志泉,田英杰,徐志洁. 支持向量机中的核参数选择问题[J]. 控制工程, 2005, 12(4): 379-381.
- [3] 董春曦,饶鲜,杨绍全,等. 支持向量机参数选择方法研究[J]. 系统工程与电子技术, 2004, 26(8): 1117-1120.

- [4] KEERTHI SS. Efficient Tuning of SVM Hyperparameters Using Radius or Margin Bound and Iterative Algorithms[J]. IEEE Trans. on Neural Networks, 2002, 13(5): 1225-1229.
- [5] KEERTHI SS, LIN C J. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel[J]. Neural Computation, 2003, 15(7): 1667-1689.
- [6] LIN H T, LIN C J. A Study on Sigmoid Kernels for SVM and the Training of Non-PSD Kernels by SMO-type Methods[R]. Taipei: Department of Computer Science and Information Engineering, National Taiwan University, 2003.
- [7] SMOLA AJ, SCHOLKOPF B, MULLER KR. The Connection Between Regularization Operators and Support Vector Kernels[J]. Neural Networks, 1998, 11(4): 637-649.
- [8] CRISTIANINI N, SHAWE-TAYOR J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods[M]. Cambridge: Cambridge University Press, 2000.
- [9] CHANG CC, LIN CJ. LIBSVM: A Library for Support Vector Machines[EB/OL]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2006-04-01.
- [10] CHEN YW, LIN CJ. Combining SVMs with Various Feature Selection Strategies[A]. Feature extraction, foundations and applications[C]. Heidelberg: Springer, 2005.
- [11] PLATT JC. Fast Training Support Vector Machines Using Sequential Minimal Optimization[A]. Advances in Kernel Methods-Support Vector Learning[C]. Cambridge, MA: MIT Press, 1999. 185-208.
- [12] MÜLLER K-R. IDA Benchmark Repository[EB/OL]. <http://ida.first.fhg.de/projects/bench/benchmarks.htm>, 2002-07-19.