文章编号: 1006-2475(2012) 02-0005-03

支持向量机的缺陷及改进算法

郭光绪

(南京航空航天大学计算机科学与技术学院,江苏 南京 210016)

摘要: 传统支持向量机通常关注于数据分布的边缘样本,支持向量通常在这些边缘样本中产生。本文提出一个新的支持向量算法,该算法的支持向量从全局的数据分布中产生,其稀疏性能在大部分数据集上远远优于经典支持向量机算法。该算法在多类问题上的时间复杂度仅等价于原支持向量机算法的二值问题,解决了设计多类算法时变量数目庞大或者二值子分类器数目过多的问题。

关键词: 支持向量机; 稀疏性; 多类问题; 推广性能

中图分类号: TP301.6　　**文献标识码**: A　　**doi**: 10.3969/j.issn.1006-2475.2012.02.002

Deficiencies of Support Vector Machines and Its Improved Algorithm

GUO Guang-xu

(College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, China)

Abstract: Traditionary support vector machines (SVMs) usually focus on edge patterns of data distribution, and support vectors (SVs) usually generates from these patterns. This paper proposes an alternative algorithm, which generates SVs from all training patterns. The sparsity of the algorithm is validated on most data sets far better than typical SVMs. The complexity of the algorithm in multi-class problems is merely equivalent to two class SVMs, which greatly solves the problems of too many variables or too many binary classifiers in multi-class SVMs.

Key words: support vector machines; sparsity; multi-class problems; generalization

0 引　言

支持向量(SVs) 的数目是评价支持向量机(SVMs) [1-2] 算法的重要指标。少的支持向量数目能够提高分类器的分类速度,因此一些稀疏算法如 l-norm SVM[3]、Lp-SVM[4]、自适应 Lp-SVM[5] 等被相继提出。然而这些算法所获得的支持向量仍然倾向于数据分布的边缘样本,为了刻画出数据的分布,所需的支持向量数目仍然很多。

对于多类问题,SVM 通常有两类解决方法: (1) 通过构造多个二值子分类器,如“one-against-all”、“one-against-one”[6]、DAGSVM[7] 和 ECOC SVM[8-9]。如果有 k 类样本, “one-against-all”需要训练 k 个二值子分类器,而“one-against-one”则需训练 k(k-1) /2 个二值子分类器。(2) 在一个目标函数中同时考虑所有子分类器的优化参数[10-13]。如果有 n 个训练样本,这种方法通常需要求解一个 k × n 个变量的二次规划问题,当训练样本较多时,实现较为困难。

本文提出一个新颖的 SVM 算法,该算法从全局数据分布中选择支持向量,大大减少了支持向量的数目。并且在多类问题的应用上,仅需要一个分类器,且算法的时间复杂度仅等价于 SVM 的两类问题。

1 改进算法

给定训练集合 $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$,其中有 m 个训练样本。首先考虑两类问题,即 $y_i = \pm 1$。本文使用文献[13]中的无偏置(unbiased) 分类器框架:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x) \quad (1)$$

其中 $K(\cdot,\cdot)$ 为核函数,也可以看作是 x_i 与 x 间的相似性度量; α_i 为第 i 个训练样本 x_i 的权重,令向量 $\alpha = [\alpha_1 \ldots \alpha_m]^T$, $y = [y_1 \ldots y_m]^T$。优化如下的目标

收稿日期: 2011-10-26

作者简介: 郭光绪(1987-) ,男,江苏如皋人,南京航空航天大学计算机科学与技术学院硕士研究生,研究方向: 模式识别,人工智能。

函数:

$$\min \sum_{i=1}^{m}(y_i f(x_i)-1)^2 \quad (2)$$

$$\text{s.t. } \mathbf{1}^T\alpha \le D, \alpha \ge 0 \quad (3)$$

其中参数 D 用来调节 α 的稀疏性[14],通过交叉验证来确定。

定义矩阵 $K=[K(x_i,x_j)]$ 为 m × m 的核矩阵,定义向量 $f=[f(x_1)\ldots f(x_m)]^T$,则 f = K* diag(y) * α,且式(2)可以写成矩阵的形式:

$$\min \alpha^T \text{diag}(y) KK^T \text{diag}(y) \alpha - 2\alpha^T \text{diag}(y) Ky \quad (4)$$

约束条件仍为式(3)。最终算法需要求解一个 m 阶的线性约束二次规划问题。算法的时间复杂度等同于 SVM。

2 算法的多类问题应用

不同于两类问题,在 k 类问题中 $y_i \in \{1,\cdots,k\}$。定义

$$f_c(x)=\sum_{y_i=c}\alpha_i K(x_i,x) \quad (5)$$

$$f_{-c}(x)=\sum_{y_i\neq c}\alpha_i K(x_i,x) \quad (6)$$

则 $f_c(x)$ 可看做样本 x 属于 c 类的置信度度量,$f_{-c}(x)$ 为样本不属于 c 类的置信度度量。因此,如果一个样本 x 属于 c 类,希望 $f_c(x)$ 越大越好,而 $f_{-c}(x)$ 越小越好。定义 m 阶对角矩阵矩阵 I_c 和 I_{-c}:

$$I_c(i,i)=\begin{cases}1, & \text{if } y_i=c\\ 0, & \text{otherwise}\end{cases} \quad I_{-c}(i,i)=\begin{cases}0, & \text{if } y_i=c\\ 1, & \text{otherwise}\end{cases} \quad (7)$$

则

$$f_c(x_j)=\alpha^T I_c K_{:j} \quad (8)$$

$$f_{-c}(x_j)=\alpha^T I_{-c} K_{:j} \quad (9)$$

首先使用类似“one-against-all”的多类分类处理策略,将一类训练样本看作正类样本,其余训练样本都看作负类。优化下面的目标函数:

$$\min \sum_{c=1}^{k}\sum_{y_i=c}[f_c(x_j)-f_{-c}(x_j)-1]^2 \quad (10)$$

$$\text{s.t. } \mathbf{1}^T\alpha \le D, \alpha \ge 0 \quad (11)$$

将式(8)和式(9)代入式(10),得到:

$$\min \sum_{c=1}^{k}\alpha^T(I_c-I_{-c})KI_cK^T(I_c-I_{-c})\alpha - 2\sum_{c=1}^{k}\alpha^T(I_c-I_{-c})KI_c\mathbf{1} \quad (12)$$

输出的分类器形式为:

$$\hat{y}=\arg\max_c f_c(x) \quad (13)$$

因为目标函数中仅有 m 个变量,变量的数目与类别数目无关,优化仅需求解一个 m 阶的二次规划问题,它的时间复杂度等价于二值 SVM 分类器求解。而通常的多类 SVM 算法至少需要求解 k × m 个变量(k 个子分类器)。

同样也可以采用类似“one-against-one”的多类分类思想,优化如下目标函数:

$$\min \sum_{j=1}^{m}\sum_{c\neq y_i}[f_{y_j}(x_j)-f_c(x_j)-1]^2 \quad (14)$$

$$\text{s.t. } \mathbf{1}^T\alpha \le D, \alpha \ge 0 \quad (15)$$

式(14)写成矩阵的形式为:

$$\min \sum_{j=1}^{m}\sum_{c\neq y_i}\alpha^T(I_{y_j}-I_c)K_{:j}K_{:j}^T(I_{y_j}-I_c)-\sum_{j=1}^{m}\alpha^T(2I_{y_j}-I)K_{:j} \quad (16)$$

不同于“one-against-one”的 SVM,该问题同样仅仅需要求解 m 个变量。

3 实 验

在下面的实验中 K(·,·)采用 RBF 核函数,即:

$$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2),\gamma>0 \quad (17)$$

验证参数 γ 的范围是 $\{2^{-10},2^{-9},\ldots,2^{10}\}$,参数 D 的范围选择 $\{2^1,2^2,\ldots,2^{10}\}$,γ 和 D 通过 five-fold 交叉验证确定。

首先在人工数据集上比较本文的算法和 SVM 支持向量选择的不同。

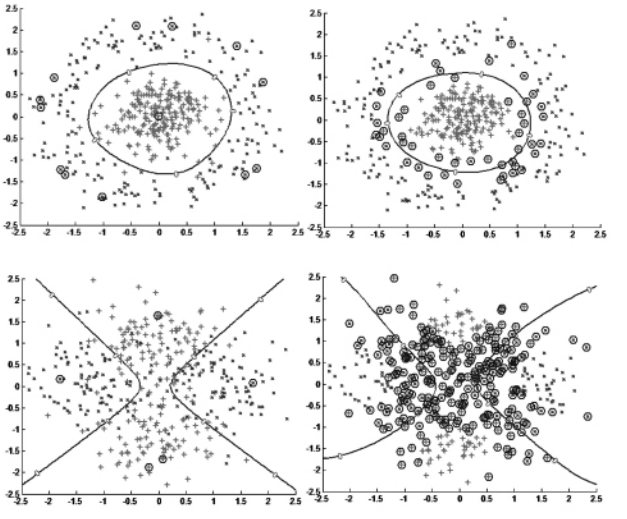

图 1 人工数据集的实验结果(左边为本文算法,右边为 SVM。圈表示支持向量)

从图 1 可以看出,本文算法选择的支持向量数目明显少于 SVM 的支持向量。在图 1 上面的数据集上,本文算法的支持向量数目为 13 个,SVM 为 47 个;在图 1 下面的数据集上,本文算法的支持向量数为 5 个,而 SVM 为 254 个。并且可以看出支持向量的分布也明显不同,SVM 的支持向量主要出现在数据分布的边缘或者说是数据分布的重叠区域,本文算法主要出现在数据分布的中心。

第二个实验,笔者选择 UCI 数据库中的 8 个数据集(后 4 个为多类),来观察算法在真实数据机上的分类情况。对于多类数据,本文算法选择优化式

(10)、(11),SVM 采用"one-against-all"策略。实验重复 10 次,实验结果如表 1 所示。

表 1　UCI 数据集上的实验结果

数据集	本文算法		SVM	
	测试误差	SV 数目	测试误差	SV 数目
Automobile	13.8 ±4.8	43	16.7 ±4.0	50
Bupa	30.9 ±2.6	21	30.1 ±2.9	118
Heart	16.7 ±2.6	25	17.4 ±2.1	65
Pima	23.1 ±1.6	24	24.2 ±1.3	208
Iris	4.2 ±1.7	23	4.7 ±1.7	77
Tae	49.0 ±5.3	48	49.9 ±5.7	168
Thyroid	5.9 ±2.0	39	5.2 ±1.9	45
Yeast	41.6 ±1.1	149	40.8 ±1.7	1601

从表 1 可以看出,本文算法就测试误差而言与 SVM 相当。而本文算法的支持向量数目在这 8 个数据集中都少于 SVM,并且本文算法在其中 6 个数据集中的支持向量数目都小于 SVM 的一半,甚至在一些数据集(如 Pima, Yeast)上,本文算法的支持向量数大约只有 SVM 的十分之一。

4　结束语

本文提出一个新颖的支持向量机算法,算法的支持向量从全局数据分布中产生。本文算法能够简单地推广到多类问题中,且相应的多类问题的时间复杂度仅仅等价于 SVM 的二值问题,所要求解的变量数目等于训练样本数目,与类别数无关。实验结果表明,本文算法的推广性能和 SVM 相当,而所获得的支持向量数目却远远小于 SVM。

参考文献:

[1] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[2] Vapnik V. The Natural of Statistical Learning Theory [M]. New York: Springer, 1995.

[3] Zhu J, Rosset S, Hastie T, et al. 1-norm support vector machines [C]//Advances in Neural Information Processing Systems. 2004: 49-57.

[4] Bradley P S, Mangasarian O L. Massive data discrimination via linear support vector machines [J]. Optimization Methods Software, 2000, 13(1): 1-10.

[5] Liu Y, Zhang H H, Park C, et al. Support vector machines with adaptive Lq penalty [J]. Computational Statistics and Data Anualysis, 2007, 51(12): 6380-6394.

[6] Chin K K. Support Vector Machines Applied to Speech Pattern Classification [D]. Univ. Cambridge, Cambridge, UK, 1998.

[7] Platt J C, Cristianini N, Shawe-Taylor J. Large margin DAG's for multiclass classification [C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 547-553.

[8] Pujol O, Escalera S, et al. An incremental node embedding technique for error correcting output code [J]. Pattern Recognition, 2008, 41(2): 713-725.

[9] Pujol O, Radeva P, et al. Discriminant ECOC: A heuristic method for application dependent design of error correcting output codes [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(6): 1007-1012.

[10] Vapnik V. Statistical Learning Theory [M]. New York: Wiley, 1998.

[11] Weston J, Watkins C. Multi-class Support Vector Machines [R]. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, 1998.

[12] Crammer K, Singer Y. On the learnability and design of output codes for multiclass problems [J]. Machine Learing, 2002, 47(2-3): 201-233.

[13] Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines [J]. Journal of Machine Learning Research, 2002, 2(2): 265-292.

[14] Tibshirani R. Regression shrinkage and selection via the lasso [J]. J. Royal Statistical Soc. B, 1996, 58(1): 267-288.

(上接第 4 页)

Advances in Natural Processing Research on Computing Science, 2006, 18: 151-162.

[9] 骆正清,陈增武,王泽兵,等. 汉语自动分词研究综述 [J]. 浙江大学学报: 自然科学版, 1997, 31(3): 306-312.

[10] Salton G, Wong A, Yang C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[11] 徐凤亚,罗振生. 文本自动分类中特征权重算法的改进研究 [J]. 计算机工程与应用, 2005, 41(1): 181-184, 220.

[12] James Auen. Natural Language Understanding [M]. The Benjamin/Cummings Publishing Company, 1991.

[13] 胡鑫. 中文文本分类的特征选取研究 [J]. 甘肃科技, 2006, 22(5): 119-120.

[14] 宋枫溪,高林. 文本分类器性能评估指标 [J]. 计算机工程, 30(13): 107-109, 127.