第17卷 第7期
2007年7月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.17 No.7
Jul. 2007

OAO－SVMs的训练时间性能分析及算法改进

张 耿,张桂新
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要:支持向量机(SVM)算法是统计学习理论中最年轻的分支。结构风险最小化原则使其具有良好的学习推广性。但在实际应用中,训练速度慢一直是支持向量机理论几个亟待解决的问题之一,这一点在SVM向多类问题领域推广时表现的尤为明显。文中将从样本分布与类别数量两方面入手,对传统的SVM多分类OAO算法进行训练时间性能上的分析,并引入分层的思想,提出传统OAO－SVMs算法的改进模型H－OAO－SVMs。通过与其他常见多分类SVMs训练时间的比较表明:改进后的H－OAO－SVMs模型具有更优的训练时间性能。

关键词:支持向量机;多分类算法;训练速度
中图分类号:TP181 **文献标识码**:A **文章编号**:1673－629X(2007)07－0024－04

Analysis of OAO－SVMs' Training Time and Its Algorithm Improvement

ZHANG Geng, ZHANG Gui-xin
(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: SVM (support vector machine) algorithm is the newest branch of statistic learning theory. Because the structural risk minimization principle makes SVM exhibit good generalization. But in practice, training slowly is one of the questions which are not solved satisfactorily in the field of SVMs. Moreover, the defection is enlarged when expanding SVMs to multi－category classification. Analyses the training time performance of the traditional OAO－SVMs based on swatch distributing and swatch number, presents the hierarchical OAO SVMs model having the better training speed, and compares it to the others' multi－class algorithm based on the SVMs.

Key words: support vector machine; multi－class algorithm; training speed

0 引 言

支持向量机[1](Support Vector Machine, SVM)是基于VC(Vapnik－Chervonenkis Dimension)维理论和结构风险最小原理的机器学习算法。它很好地克服了维数灾难、局部最优以及过拟合等传统算法所不可规避的问题,并保证了出众的推广能力。但是,SVM理论本身还有几类问题亟待解决,训练算法速度慢即为其中之一。这一点在SVM向多类问题领域推广时表现的尤为明显。

文中将通过对OAO算法训练时间性能的分析,提出训练时间性能更优的H－OAO－SVMs模型,并与其他常见多分类SVMs做训练时间性能上的比较。

收稿日期:2006－10－09

作者简介:张 耿(1977－),男,湖北人,硕士研究生,研究方向为支持向量机;张桂新,副教授,硕士,主要从事电气传动及控制、模式识别的研究。

1 常见的多分类SVMs算法

SVM在实质上是两类问题的分类器,而现实中多分类问题却更为常见。将SVM推广到多分类领域是一个正在研究中的问题。当前主流的算法主要是构造一系列的两类分类器,然后按照一定的规则将它们组合起来完成多类问题的分类。其具体形式主要有以下几种(文中主要讨论训练阶段)。

1.1 One－Against－All(OAA)

OAA算法对已知M类样本,分别以每个类别为正类构建M个子分类器[2]。

1.2 One－Against－One SVMs(OAO－SVMs)/Directed Acyclic Graph(DAGSVMs)

OAO－SVMs与DAGSVMs[3]算法在训练阶段,在已知M类样本的训练集T上,对所有的类别组合构造$M(M-1)/2$个SVM。

1.3 Error Correcting Codes(ECC－SVMs)

ECC－SVMs是在M类样本的基础上,构建一系列(总数为L)的两类问题,并为每个两类问题建立一

个决策函数,得到 L 个 SVM[4]。

1.4 Hierarchical SVMs(H－SVMs)

H－SVMs 针对训练样本的 M 类数据,按照一定的规则(如聚类),采用分级的形式,逐步将某一类与其他类分开。最终生成有 $M-1$ 个中间结点的二叉树拓扑结构,每个中间结点对应一个 SVM[5]。

2 OAO－SVMs 及其训练时间分析

2.1 OAO－SVMs 算法

One－Against－One SVMs(OAO－SVMs)[2]是最常见的一种多分类 SVM 算法。在训练阶段,针对 M 类样本的训练集 T 中的所有的类别组合 $(i,j) \in \{(i,j) \mid i \leqslant j, i,j = 1,\cdots,M\}$,以($i$(正类),$j$(负类)) 为样本集分别构造 $M(M-1)/2$ 个 SVM:

$$g^{i-j}(x) = \sum_{i=1}^{l_{i-j}} y_i \alpha_i^* K(x,x_i) + b^*$$

令 $M=5$,OAO－SVMs 的拓扑结构如图 1 所示。

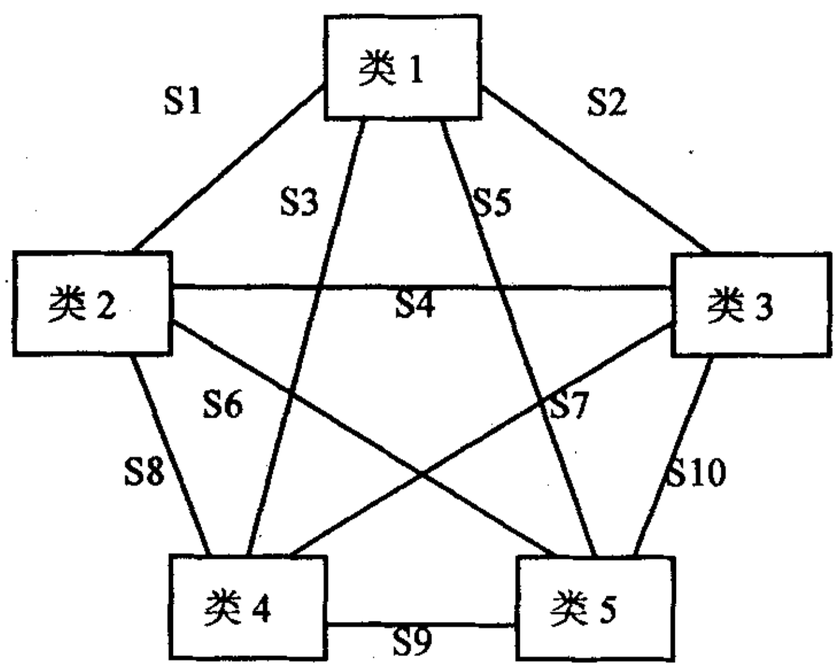


图 1　5 类样本的 OAO－SVMs 拓扑结构

图中边集 S_i 表示以其邻接点(类别)为样本集的 SVM。

2.2 OAO－SVMs 训练时间分析

根据文献[6],样本训练时间 T 与样本集大小 S 之间的关系为:

$$\log_{10} T \approx a(\log_{10} S)^2 \quad (a \text{ 为常数})$$

单个 SVM 的训练时间主要取决于训练样本数量的多少。当多个 SVM 采用组合的方式进行多类问题的分类时,因为 SVM 组合形式不同,造成各种算法训练时间上的差异。由图 1 可见,OAO－SVMs 算法中所有的类别在拓扑结构中处于对称状态。因此该算法的训练时间不受拓扑结构的影响,下文将只从类别的样本分布与类别数等方面来讨论 OAO－SVMs 训练时间的特性($a=0.06$)。

2.2.1 样本分布对 OAO－SVMs 训练时间的影响

为了找出样本分布与训练时间之间的联系,令总样本数 $S=10000$,类别数 $M=8$。按正态分布($\mu=1250$,$\sigma^2 \in \{100\ 200\ 300\ 400\ 500\}$)随机产生 5 组样本分布(经排序处理)如图 2 所示。

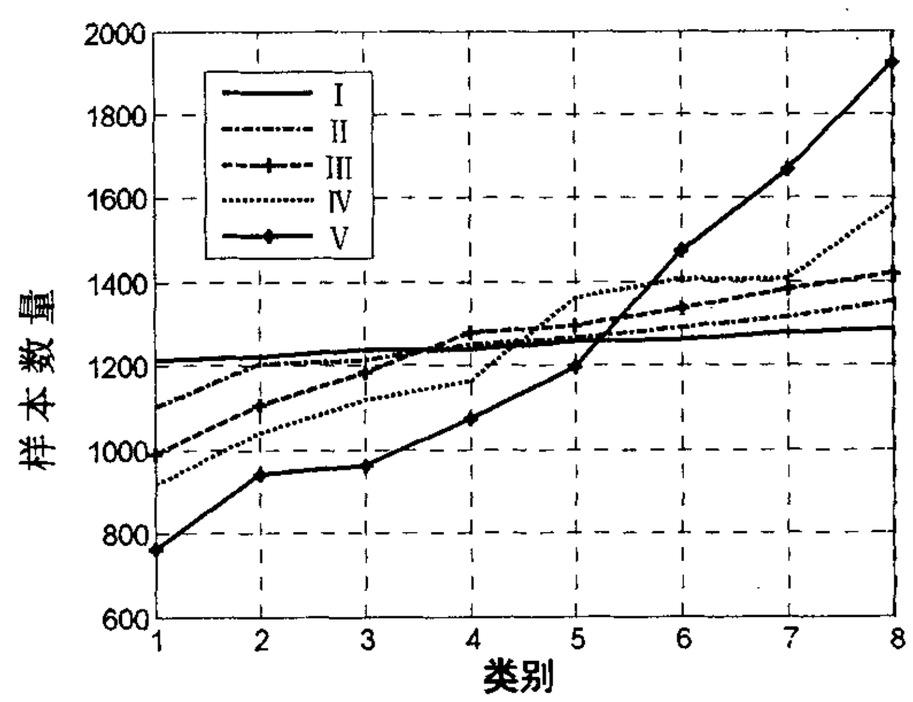
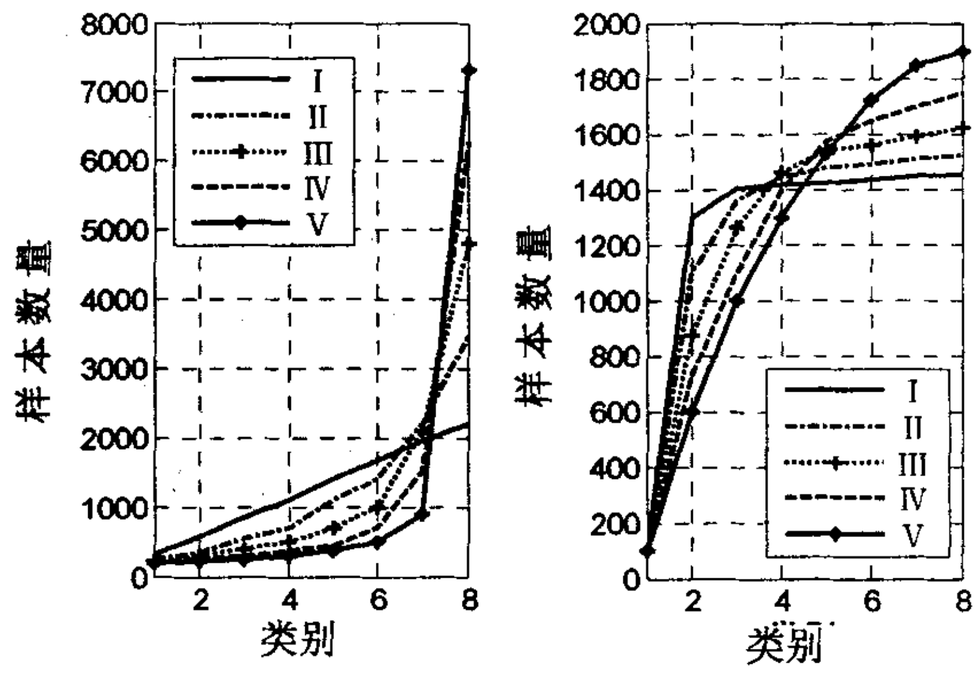


图 2　5 组随机正态样本分布

样本分布Ⅰ～Ⅴ对应的训练时间依次为 753.4,753.9,755.3,757.8,767.6。可见在 OAO－SVMs 算法中,样本分布越均匀,训练时间越短。为了验证这一结论,给出 10 组人工生成的样本分布数据如图 3 所示。



(a)分布数据　　(b)分布数据

图 3　10 组人工样本分布数据

相应的训练时间如表 1 所示。

表 1　人工样本分布对应的 OAO－SVMs 训练时间

样本＼分布	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
(a)	793.2	860.0	961.3	1112.6	1250.8
(b)	772.6	774.8	778.6	783.9	790.8

表 1 数据进一步验证了关于 OAO－SVMs 算法训练时间的特性:样本的分布越均匀,训练时间越短;反之样本分布越不均匀,训练时间越长。

2.2.2 类别数量对 OAO－SVMs 训练时间的影响

OAO－SVMs 算法针对 M 类样本,构建 $M(M-1)/2$ 个 SVM。令样本分布为均匀分布,总样本数 $S=10000$,得类别数量与训练时间的关系如表 2 所示。

表 2　类别数对 OAO－SVMs 训练时间的影响

类别数量	4	8	16	32
训练时间	490.5	753.4	1088.4	1553.4

表 2 数据说明在 OAO－SVMs 算法中,样本类别数与训练时间成正比。

3 OAO-SVMs 算法的模型改进

由于 OAO-SVMs 算法中,训练时间与样本类别数成正比,而样本类别数又与子 SVM 的数量成正比,因此,训练时间与子 SVM 的数量成正比。为了改善 OAO-SVMs 的训练时间性能,有必要降低该算法中子 SVM 的数量。针对这种情况,通过引入分层的思想,提出一种在树型结构的不同层次上进行两两分类的 H-OAO-SVMs 算法。其核心思想包括子分类器的构建与层次结构的确立两个方面。

3.1 子分类器的构建

H-OAO-SVMs 算法在构建子分类器时,规则与 OAO-SVMs 一致,均是在总样本集上针对所有类别,两两间训练子分类器。但是 H-OAO-SVMs 算法的总样本集概念并不同于 OAO-SVMs 中的全体样本。在这种新的基于树型拓扑结构的算法中,任意一个非终端结点,其子结点中包含的所有样本为当前总样本集。在此基础上,每一个子类(子类的集合)只与其父结点下的其他兄弟结点按 OAO 算法构建子 SVM。具体形式如图 4 所示。

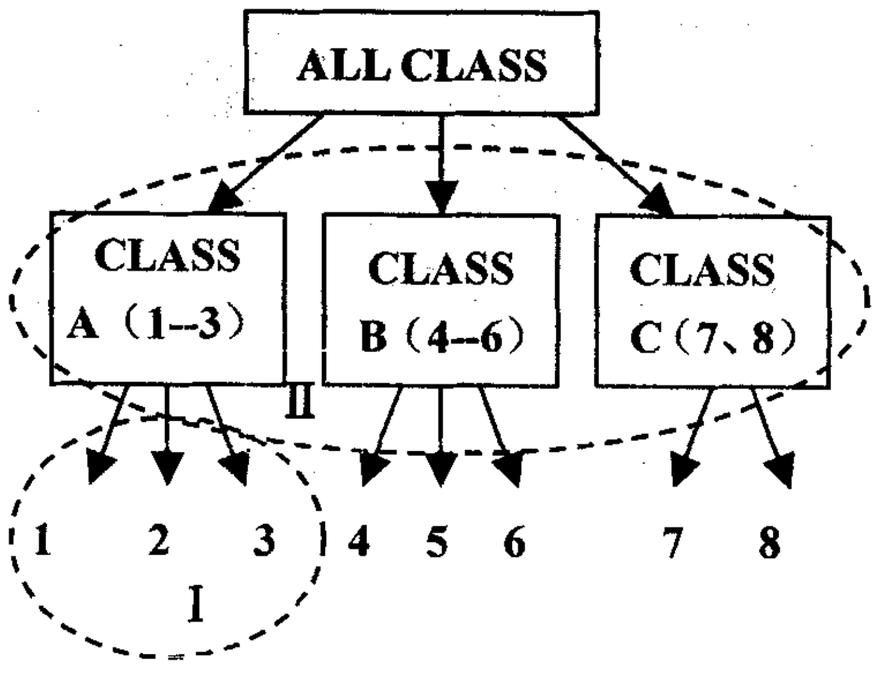


图 4 H-OAO-SVMs 算法拓扑结构

图 4 所示模型为原始类别数 $M=8$ 时的一种 H-OAO-SVMs 拓扑结构。在图中可见,所有原始类别均处于拓扑结构的终端。通过一定的聚类规则,若干类别向上聚合成新的类别,直至所有类别最终在根结点聚合成一类。在该模型中,类别 1~8 不再作为总样本进行 28 个子 SVM 的训练,而是改由非终端结点包含的类别来进行子分类器的训练。如图 4 虚线框Ⅰ中,以类 1,2,3 为总样本集按 OAO 算法构建子 SVM。虚线框Ⅱ中,以类 A,B,C 为总样本集按 OAO 算法构建子 SVM。其他亦然。通过这种层次型的 OAO 算法,子 SVM 的数量减少为 10 个,训练时间也会相应下降。

3.2 层次结构的确立

M 类样本在向上聚合形成树型拓扑结构的过程中,有一个问题值得关注:

哪些类别应该被聚合成一个类。在对 OAO-SVMs 算法的训练时间性能分析中已知:样本分布的均匀程度与训练时间成正比。为了在 H-OAO-SVMs 算法中取得训练时间的进一步减少,在构建树型拓扑结构时,应遵循以下两条规则:

(1)样本容量相似的类别尽量被聚合到同一个类中;

(2)所有层次上非终端结点下的子类样本分布应尽量均匀。

第一条规则在微观上保证了在某一非终端结点下按 OAO 算法训练 SVM 时,样本的分布趋于均匀。而第二条规则在宏观上保证了任意层次上的任意一组 OAO 模型中,子类样本分布都趋于均匀。通过这样的聚类规则,可以使得 H-OAO-SVMs 的整体训练时间减少。

3.3 H-OAO-SVMs 算法的训练时间性能分析

H-OAO-SVMs 算法最大的特点是根据样本的分布自适应地调整树型拓扑结构,并在兄弟结点间采用 OAO 算法来进行子分类器的训练。其训练时间的性能主要与样本分布与类别数相关。

为了考察样本分布对 H-OAO-SVMs 训练时间的影响,并在此基础上与传统的 OAO-SVMs 算法进行比较,仍然采用样本分布数据(见图 2、图 3)。得相应的训练时间如表 3 所示。

表 3 样本分布对 H-OAO-SVMs 训练时间的影响

样本 \ 分布	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
图 2	581.8	581.7	581.6	582.4	585.8
图 3(a)	577.7	549.6	503.9	420.2	350.6
图 3(b)	581.2	580.6	581.0	582.8	584.1

同理,令样本分布为均匀分布,总样本数 $S=10000$。得类别数量与训练时间的关系如表 4 所示。

表 4 类别数对 H-OAO-SVMs 训练时间的影响

类别数量	4	8	16	32
训练时间	417.9	581.8	708.2	741.1

通过对表 3、4 中数据的分析,并与表 1、2 进行比较,可对 H-OAO-SVMs 的时间性能及其相对传统 OAO-SVMs 算法的性能改进做如下总结:

(1)H-OAO-SVMs 算法中,样本类别数与训练时间成正比。

(2)当样本分布极端不均匀时,H-OAO-SVMs 算法的训练时间相比一般情况有明显减少。其余情况下,样本分布的均匀程度对训练时间的影响并不明显。

(3)在条件相同的情况下,H-OAO-SVMs 算法的训练时间相比较于 OAO-SVMs 算法有明显的优

势。

3.4　H－OAO－SVMs 训练时间性能与其它常见多分类 SVM 算法的比较

3.4.1　DAG－SVMs 算法、OAA 算法与 ECC－SVMs 算法

(1) DAG－SVMs 在训练阶段与 OAA 算法相同。训练时间性能劣于 H－OAO－SVMs 算法。

(2) OAA 与 ECC－SVMs 算法在训练阶段,针对每个子分类器,全体样本 S 都参加训练,总训练时间仅仅与 OAA 算法中的类别数量 M 与 ECC－SVMs 算法中的码字长度 L 相关。则在 $S = 10000, M = 8$ 时,OAA 算法训练时间为 2035.2,而 ECC－SVMs 算法训练时间为 $254.4 * L, (L \in (3,127))$。两者同样在训练时间性能上劣于 H－OAO－SVMs 算法。

3.4.2　H－SVMs 算法

M 类分类问题,H－SVMs 将构建 $M-1$ 个 SVM。因为分级策略不同,以这些 SVM 作为中间结点的二叉树的拓扑结构也不同。根据二叉树的性质,包含 $M-1$ 个中间结点二叉树,其层次的范围为:

$$(\lfloor \log_2 M \rfloor + 1, M)$$

H－SVMs 根据聚类规则的不同,生成的树型拓扑结构的层数在此范围内波动。下文仅讨论拓扑结构最理想的情况,即层数最小时,样本分布(见图 2、3)对 H－SVMs 算法训练时间的影响如表 5 所示。

表 5　样本分布对 H－SVMs 训练时间的影响

分布 / 图	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
图 2	525.6	525.9	526.8	529.0	535.4
图 3(a)	555.2	592.2	624.1	653.0	668.7
图 3(b)	531.0	534.17	538.7	544.6	551.4

通过比较样本分布对 H－SVMs 与 H－OAO－SVMs 训练时间的影响,可知在样本分布极端不均匀的情况下,H－OAO－SVMs 的训练时间明显优于 H－SVMs 算法,而在其他情况下,H－OAO－SVMs 算法的训练时间非常接近于 H－SVMs 的理想训练时间。

4　结束语

SVM 是一种极具理论与实用价值的模式识别方法,训练速度慢是它几个亟待解决的问题之一。对于多类识别问题,文中分析了 OAO－SVMs 算法的训练时间性能,并在此基础上提出了改进 H－OAO－SVMs 算法。通过对相关数据的分析比较,证明这种新模型在训练时间上具有比传统的多类 SVM 算法更为优秀的性质。

参考文献:

[1] Cristianini N, Shawe－Taylor J. An Introduction to Support Vector Machines and Other Kernel－based Learning Methods [M]. Cambridge :Cambridge University Press, 2004.

[2] 邓乃扬,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2004.

[3] Platt J C, Cristianini N, Shawe－Taylor J. Large Margin DAGs for multiclass classification[C]// In Advances in Neural Information Processing Systems. [s. l.]: MIT Press, 2000: 547－553.

[4] Dietterich T G, Bakiri G. Solving Multiclass Learning Problems via Error－Correcting Output Codes[J]. Journal of Artificial Intelligence Research,1995(2):263－286.

[5] Azimi－Sadjadi M R, Zekavat S A. Cloud Classification Using Support Vector Machines[C]// In Proc of the 2000 IEEE Geoscience and Remote Sensing Symposium(IGRASS 2000). Honolulu, Hawaii:[s. n.],2000:669－671.

[6] Plat J C. Fast Training of Support Vector Machines using Sequential Minimal Optimization[C] // In: Advances in Kernel Methods——Support Vector Learning. [s. l.]: MIT Press, 1999:185－208.

(上接第 23 页)

性,使得异构系统、异构数据可以通过一个统一的简单的渠道进行交换,并使数据交换双方能够理解所交换的数据。文中所给出的两种具体解决方案,分别适用于不同的环境,可以根据网络条件、数据实时性要求、系统改造成本等因素选择不同的方案。对于数据交换过程中的数据安全性没有做详细分析,有待于进一步研究。

参考文献:

[1] 李　强,王延章.基于元数据的电子政务数据交换的研究[J].计算机工程与应用,2003(28):205－207.

[2] 印　鉴,陈忆群,张　钢.基于 CWM 的数据挖掘服务中心设计[J].计算机工程与应用,2004(32):177－180.

[3] Object Management Group. Common Warehouse Metamodel (CWM) Specification Version1[M]. USA:OMG Press,2003.

[4] Poole J, Chang Dan, Tolbert D, et al. Common Warehouse Metamodel Developer's Guide[M]. USA:OMG Press,2003.

[5] Zhao Xiaofei,Huang Zhiqiu,Shen Guohua,et al. Metadata Integration of Engineering Data Warehouse System Based on Metamodel[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2006,38(3):341－346.

[6] 郑洪源,周　良.基于 CWM 的标准 ETL 的设计与实现[J].吉林大学学报:自然科学版,2006,24(1):50－55.