大规模数据集下支持向量机训练样本的缩减策略*)

罗 瑜[1] 易文德[2] 王丹琛[3] 何大可[1]

(西南交通大学信息科学与技术学院 成都 610031)[1] (重庆文理学院数学与计算机科学系 永川 402160)[2]

(四川省信息安全测评中心 成都 610017)[3]

摘 要 大量数据下支持向量机的训练算法是SVM研究的一个重要方向和焦点。该文从分析SVM训练问题的实质和难点出发,提出一种在训练前先求出类别质心,去除非支持向量对应的样本,从而达到缩小样本集的方法。该方法在不损失分类正确率的情况下具有更快的收敛速度,并从空间几何上解释了支持向量机的原理。仿真实验证明了该方法的可行性和有效性。

关键词 支持向量机,分解算法,类别质心,准支持向量

Sample Reduction Strategy for Support Vector Machines with Large-Scale Data Set

LUO Yu[1] YI Wen-De[2] WANG Dan-Chen[3] HE Da-Ke[1]

(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031)[1]

(Dept. of Mathematics&Computer Science, Chongqing University of Arts and Sciences, Yongchuan 402160)[2]

(Sichuan Province Information Security Testing Evaluation Center, Chengdu 610017)[3]

Abstract Training algorithm for large-scale support vector machines(SVM) is an important and active subject in the field of SVM research. After the analysis of the nature and difficulties in training SVM, a new reduction strategy is proposed in this paper for training svm with large-scale sample set. In general, class centroid is solved before training and removing the samples corresponding to non support vectors. Through this method, the number of samples is reduced before training svm. This method is fast in convergence without accurate loss and propose the explanation of SVM theory from space geometry. The results of simulation experiments show the feasibility and effectiveness of this method.

Keywords Support vector machines, Decomposition algorithm, Reduction strategy, Centroid , Quasi-support vectors

1 引言

Vapnik[1,2]等人提出的统计学习理论是一种针对=小样本的学习理论,它避免了人工神经网络等方法的网络结构难以确定、过学习和欠学习以及局部极小等问题,被认为是目前针对小样本的分类、回归等问题的最佳理论。

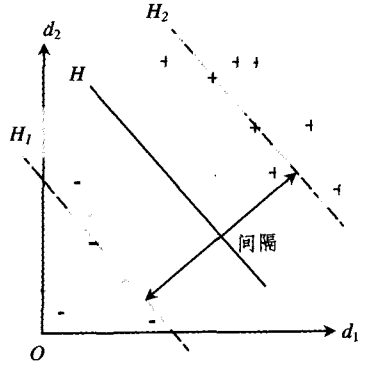

图1 分类超平面

支持向量机的基本思想是对于非线性可分样本,基于1909年Mercer核展开定理,将其输入样本通过非线性变换映射到另一个高维空间Z(Hilbert空间)中,在变换后的空间中构造一个最优的分类超平面H(图1),使其在保证分类精度(满足经验风险)的同时最大化超平面两侧的空白区域(即最大化置信范围),也即是H_1与H_2间的几何间隔(图1)。这使得分类的结果不但在训练集上得到优化,而且在整个样本集上的风险也有上界,这就是SVM的结构风险最小化思想。鉴于篇幅的原因,本文只针对两类问题进行探讨,但本文的结论同样适用于多类问题。

2 支持向量机理论基础

对于两类问题,给定样本集(x_i,y_i);$x_i\in R^n$,$y_i=\pm1$,$i=1,2,\cdots,l$和核函数$K(x_i,x_j)$。K对应特征空间Z(Hilbert空间)的内积,$K(x_i,x_j)=(\Phi(x_i)\cdot\Phi(x_j))$。设计基于SVM的分类器,就是在$Z$中寻找最优超平面$H$,即最大化两类的几何间隔。对于样本集线性可分和不可分,都可用下式来表达[1,3]:

*)基金项目:上海市特种光纤重点实验室科研项目,地铁CBTC无线接入安全认证算法研究。罗 瑜 博士生,研究方向为并行计算及模式识别。

2 曾黄麟.粗集理论及其发展[M].重庆:重庆大学出版社,1998
3 曾黄麟.智能计算[M].重庆:重庆大学出版社,2004
4 郑书富,管延勇,史开泉.分辨矩阵与它在非一致决策中的应用[J].山东大学学报,2005,35(2):86~89
5 Xiaohua H, Cercon N. Learning in relational data-bases: a rough set approach[J]. Computation Intelligence, 1995, 11(2): 323~337
6 叶东毅,陈昭炯.一个新的差别矩阵及其求核方法[J].电子学报,2002,31(7):1086~1088
7 王希雷,王磊.粗集中区分矩阵对不一致问题处理的研究[J].微机发展,2003,13(6):119~120
8 汪廷华,程从从.Rough集中基于分明矩阵的决策规则约简研究[J].计算机科学,2005,32(8.A):42~44
9 黄兵,周献中.不完备信息系统分配约简与规则提取的矩阵算法[J].计算机工程,2005,31(17):20~22
10 管延勇,薛佩军,王洪凯.不完备信息系统的可信决策规则提取与E-相对约简[J].系统工程理论与实践,2005,12:76~82

$$\min_{w,b,\xi} \tau(w)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t. } y_i((w\cdot\Phi(x_i))+b)\geqslant 1-\xi_i, i=1,\cdots,l \quad (1)$$

$$\xi_i\geqslant 0, i=1,\cdots,l$$

其中 w—权重向量，b—阈值，C—惩罚参数，ξ_i—松弛变量。上式又称为标准SVM(线性问题的 $\Phi(x_i)=x_i$)。由于特征空间的维数很大甚至是无穷的，并且 Φ 是未知的，一般方法并不能直接求解上式，而是通过求解此问题的 Wolfe 对偶问题

$$\min_{\alpha} f(\alpha)=\frac{1}{2}\alpha^T Q\alpha-e^T\alpha$$

$$\text{s.t. } 0\leqslant\alpha\leqslant Ce \quad (2)$$

$$y^T\alpha=0$$

来解决。其中 α_i 是 Largrange 乘子，Q 是 Hessian 矩阵，$Q_{ij}=y_iy_jK(x_i,x_j)$，Q 是半正定矩阵。这是凸二次规划问题，其最优解满足 KKT 条件[4]，求出解 α_i^* 后可直接计算 w 和 b[1~3]，并构造决策函数

$$f(x)=\sum_{i=1}^{l}y_i\alpha_i^* K(x,x_i)+b^*$$

$$y=\mathrm{sgn}(f(x)) \quad (3)$$

3 SVM 训练算法

支持向量机的训练过程，就是求解最优化问题的过程。支持向量机具有一些很好的特性：从图 1 可看出，由于支持向量仅是样本集中的很小一部分，因此其解具有稀疏性；另外，支持向量机是一个凸二次规划问题，这就保证解的存在和唯一性。虽然在理论上有许多求解二次规划的方法(比如内点法、既约梯度法等)，但是支持向量机中二次规划的变量维数等于训练样本的个数 l，从而使 Hessian 矩阵元素的个数是 l^2，更关键的是 Q 不是稀疏的，这就造成实际问题的求解规模过大，而使许多传统方法不适用。

近年来，学者相继开发出了很多 SVM 快速训练算法，例如 Vapnik 的块(Chunking)算法[2]，Osuna 的分解算法[6]，Joachims 的 SVM^{Light} 算法[7]，John Platt 的序贯最小优化(SMO)算法[8]等。这些算法的实质是将大规模的原问题分解为若干小规模的子问题，然后对子问题反复迭代求解从而构造出子问题的近似解，并使该解逐渐收敛到原问题的最优解。尽管快速训练算法的引入改善了支持向量机的训练速度，但是其计算复杂度和收敛速度仍然强烈依赖于样本数 l，在实时性要求比较高的场合(比如在实时模式识别环境下)，快速训练算法仍然难以满足实时要求。本文提出的方法就是在不影响训练结果和分类正确率的前提下，使得参与训练的样本集 $l'\ll l$，使得训练耗费时间成指数级的降低。

4 样本缩减策略

从式(3)可以看出，对决策分类面 $f(x)$ 有贡献的样本点是对应于 Largrange 乘子 $\alpha_i^*>0$ 的样本 x_i，分布在图 1 中 H_1 和 H_2 上，称为支持向量(Support Vectors，SVs)。对应于 $\alpha_i^*=0$ 的训练样本称为非支持向量。下面的定理指出非支持向量不仅与支持向量的决策无关，也不会影响支持向量机的训练过程和训练结果。

定理 1 支持向量机的训练过程和训练结果与非支持向量无关(证明见附录 1)。

分解算法将大规模的原问题分解为若干小规模的子问题，然后对子问题反复迭代求解。由于式(3)的解具有稀疏性，非支持向量的参与训练耗费了 SVM 训练的大部分时间(非支持向量数远远大于支持向量数)。因此，如果能够先验地去除部分非支持向量，将会大大提高支持向量机的训练速度，同时提高训练的收敛速度。

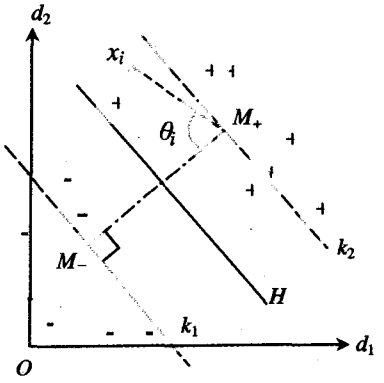

图 2 类别质心

引入力学中刚体的概念，在 N 维空间中，设正负样本分布分别分布在一个半径为 ζ^+ 和 ζ^- 的超球范围内，并且其分布为独立同分布的，则正负样本集分别表现为两个刚体，其类别质心分别记为 M_+ 和 M_-。对于线性问题，正样本集质心为 $M_+=\frac{1}{l_+}\sum_{i=1}^{l_+}x_i$，$l_+$ 是正样本个数，负样本集质心为 $M_-=\frac{1}{l_-}\sum_{i=1}^{l_-}x_i$，$l_++l_-=l$。对于非线性问题，正样本集质心为 $M_+=\frac{1}{l_+}\sum_{i=1}^{l_+}\Phi(x_i)$，负样本集质心为 $M_-=\frac{1}{l_-}\sum_{i=1}^{l_-}\Phi(x_i)$。

推论 1 假如两件类样本集是可分的，则最优分类超平面位于两类质心之间(图 2)，也即两类质心居于最优超平面两侧，并且，各类的支持向量分布于类别质心和最优超平面之间。

从类别聚类的观点来看，类别刚体的质心可以视为类别的聚类中心。假设两类质心都在最优分类超平面的一侧，则从样本分布来看，只有正负样本集区域重叠过多时才会发生，也即两类样本的特征太过相似，从而导致两类样本集是不可分的。这符合哲学上的观点：要解决问题，则问题必须是能解决的。同理，根据样本分布，支持向量位于最优超平面和质心之间。

观察图 2，k_1 和 k_2 是垂直于两类质心连线的方向，$|M_+M_-|$ 是连接两类质心的线段，定义正样本 x_i 夹角 θ_i^+ 是向量 $\overrightarrow{M_+M_-}$ 和 $\overrightarrow{M_+x_i}$ 的夹角(同理可定义负样本夹角)。支持向量机的原理，从几何上来看，就是在不知道两类质心的情况下获得垂直于 $|M_+M_-|$ 并且通过 $|M_+M_-|$ 中点的超平面。由于支持向量分布于类别质心和最优超平面之间，可得：

推论 2 两类的支持向量聚集于 k_1 和 k_2 之间，也即是 θ_i 不为钝角。

从而得到结论，当 θ_i 为钝角的样本为非支持向量，去除这些样本对支持向量机的训练不产生影响，并且能够提高最优化问题(3)的收敛速度。定义 $S^+=\{y_i=+1|i,i=1,2\cdots,l_+\}$ 和 $S^-=\{y_i=-1|i,i=1,2,\cdots,l_-\}$ 分别为正负样本序号的索引集，下面给出判别 θ_i 是否为钝角的方法。

对于正样本：

$$\cos\theta_i=\frac{\overrightarrow{M_+M_-}\cdot\overrightarrow{M_+x_i}}{|\overrightarrow{M_+M_-}||\overrightarrow{M_+x_i}|}$$

$$\overrightarrow{M_+M_-}=\frac{1}{l_+}\sum_{\substack{j=1\\ j\in S^+}}^{l_+}\Phi(x_i)-\frac{1}{l_-}\sum_{\substack{j=1\\ j\in S^-}}^{l_-}\Phi(x_i) \quad (4)$$

$$\overrightarrow{M_+x_i}=\frac{1}{l_+}\sum_{j=1j\in S^+}^{l_+}\Phi(x_j)-\Phi(x_i)$$

展开计算可得：

$$\overrightarrow{M_+M_-}\cdot\overrightarrow{M_+x_i}=\frac{1}{l_+^2}\sum_{j=1j\in S^+}^{l_+}\sum_{k=1j\in S^+}^{l_+}K(x_j,x_k)-\frac{1}{l_+l_-}\sum_{j=1j\in S^+}^{l_+}\sum_{k=1j\in S^-}^{l_-}K(x_j,x_k)-\frac{1}{l_+}\sum_{j=1j\in S^+}^{l_+}K(x_j,x_i)+\frac{1}{l_-}\sum_{j=1j\in S^-}^{l_-}K(x_j,x_i)$$

$$|\overrightarrow{M_+M_-}|=(\frac{1}{l_+^2}\sum_{j=1j\in S^+}^{l_+}\sum_{k=1k\in S^+}^{l_+}K(x_j,x_k)+\frac{1}{l_-^2}\sum_{j=1j\in S^-}^{l_-}\sum_{k=1k\in S^-}^{l_-}K(x_j,x_k)-\frac{1}{l_+l_-}\sum_{j=1j\in S^+}^{l_+}\sum_{k=1k\in S^-}^{l_-}K(x_j,x_i))^{\frac{1}{2}} \quad (5)$$

$$|\overrightarrow{M_+x_i}|=(\frac{1}{l_+^2}\sum_{j=1j\in S^+}^{l_+}\sum_{k=1k\in S^+}^{l_+}K(x_j,x_k)-\frac{2}{l_+}\sum_{j=1j\in S^+}^{l_+}K(x_j,x_i)+K(x_i,x_i))^{\frac{1}{2}}$$

$\cos\theta_i$ 的计算比较简单，其中计算复杂度为 $O(l^2)$ 的式子的值为常数，只涉及一次计算，整个 $\cos\theta_i$ 的计算复杂度为 $O(l^2+l)$。当 $\cos\theta_i<0$ 时 θ_i 为钝角，可以判定为可去除的非支持向量。在实际应用中可以适当放宽范围，限定 $\theta_i\geqslant\varepsilon(\varepsilon\geqslant\frac{\pi}{2})$。$\varepsilon$ 称为阈值，控制移去样本的范围，对于负样本的情况计算类似可得。我们把经过预选后参与支持向量机训练的样本集称作：

定义 1 把 $i\in\{\theta_i<\varepsilon|x_i\}$ 的样本集称为预选样本集，x_i 称为准支持向量(Quasi-SVs)。

5 实验仿真

选取 FERET 标准人脸图像库进行人脸分类实验。实验选取 500 人，每人采用 8 张图片建立人脸图像库，每幅图像提取宽、高为 27×13 的左眼与 9×13 的鼻子区域作为样本特征。这样实验训练样本数为 4000，样本维数为 468，另外选取这 500 人的各 2 张图片作为测试样本。实验采用 P4 2.4GHz，内存 1G 计算机，算法采用基于 SMO 的 Libsvm[9] 和基于本文方法改进的 F-Libsvm。由于本问题是多类问题，实验采用一对余多分类算法，采用径向基(RBF)核函数，SVM 参数 $C=100$，RBF 参数为 $\sigma=0.01$，实验结果如表 1 所示。

表 1 仿真实验结果

算法	支持向量数	预选样本数	训练时间(秒)	分类正确率(%)
Libsvm	124	—	78.6	53
F-Libsvm	124	623	38.3	53

由于人脸识别问题是多类问题，采用的多类分类方法是一对余算法，需要构造一系列两类分类机，其中的每一个分类机都把其中的一类同余下的各类分划开，因此表中记录支持向量数是各两类分类机的支持向量数的平均。本实验的目的是验证训练速度，所以只提取人脸的两个特征而不保证分类正确率。从实验结果来看，验证了本文所述方法的有效性。

结束语 本文提出了一种在训练前确定样本是否对训练结果产生贡献的方法，仿真实验结果表明：用比训练样本少得多的样本进行训练，可以使训练速度成指数级的提高，并且不损失训练结果的分类能力。本文提出的样本集质心观点对于训练结果的评价有指导的作用，本文确定样本集质心采用的是均值法。如何更好地表示质心关系到预选样本集的确定，是今后工作中的一个重点。同时，通过类别质心连线中点，法方向为类别质心连线的超平面与支持向量机中的最优超平面的关系，也是值得研究的一个方向。

【附录 1】

定理 1 证明：定义 $I_{sv}=\{i|\alpha_i>0\}$ 和 $I_{nsv}=\{i|\alpha_i=0\}$ 分别为支持向量和非支持向量对应样本序号的索引集，支持向量个数为 l'。引入只优化支持向量对应样本的问题

$$\min_{\alpha} g(\alpha)=\frac{1}{2}\alpha^TQ\alpha-e^T\alpha$$
$$\text{s.t. } 0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,l'\in I_{sv}$$
$$y^T\alpha=0 \quad (6)$$

要证明定理 1，只需要证明问题(2)和问题(6)同解。用反证法，假设问题(6)存在一个最优解 α' 使得 $g(\alpha')<g(\alpha^*)$。由于 α^* 是问题(2)的最优解，也即 α^* 是问题(6)的可行解，同样，α' 也是问题(2)的可行解。由 α^* 是问题(2)的最优解可得 $f(\alpha')>f(\alpha^*)$。又因为

$$f(\alpha')=\frac{1}{2}\sum_{i=1j=1}^{l}\sum^{l}\alpha'_i\alpha'_jy_iy_jk(x_i,x_j)-\sum_{i=1}^{l}\alpha'_i$$
$$=\frac{1}{2}\sum_{i=1j=1}^{l'}\sum^{l'}\alpha'_i\alpha'_jy_iy_jk(x_i,x_j)-\sum_{i=1}^{l'}\alpha'_i$$
$$=g(\alpha')$$
$$f(\alpha^*)=\frac{1}{2}\sum_{i=1j=1}^{l}\sum^{l}\alpha_i^*\alpha_j^*y_iy_jk(x_i,x_j)-\sum_{i=1}^{l}\alpha_i^*$$
$$=\frac{1}{2}\sum_{i=1j=1}^{l'}\sum^{l'}\alpha_i^*\alpha_j^*y_iy_jk(x_i,x_j)-\sum_{i=1}^{l'}\alpha_i^*$$
$$=g(\alpha^*)$$

由上式可得 $f(\alpha')=g(\alpha')<g(\alpha^*)=f(\alpha^*)$，这与 α^* 是问题(2)矛盾，定理 1 得证。注：α' 是 l' 维向量，带入 f 的时候拓展为 l 维向量，对于序号 $j\in I_{nsv}$ 的 $\alpha'_j=0$。

参考文献

1 Vapnik V N. 统计学习理论的本质[M]. 清华大学出版社，2000
2 Vapnik V N. 统计学习理论[M]. 电子工业出版社，2004
3 邓乃扬，田英杰. 数据挖掘中的新方法－支持向量机[M]. 科学出版社，2004
4 褚蕾蕾，陈绥阳，周梦. 计算智能的数学基础[M]. 科学出版社，2002
5 薛毅. 最优化原理与方法[M]. 北京工业大学出版社，2003
6 Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines[C]. In: Proc. IEEE Workshop on Neural Networks and Signal Processing, Piscataway, IEEE Press, 1997. 276～285
7 Joachims T. Making Large-Scale SVM Learning Practical[J]. In: Scholkopf B, Burges C, Smola A, eds. Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1999. 169～184
8 Platt J C. Fast training of support vector machines using sequential minimal optimization [C]. In: ScholKopf B, Burges C, Smola A, eds. Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1998. 185～208
9 Chang C C, Lin C J. LIBSVM: a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/～cjlin/libsvm, 2006-10-17