

一种改进的快速支持向量机分类算法研究

秦玉平^{*1,2}, 王秀坤¹

(1.大连理工大学 电子与信息工程学院, 辽宁 大连 116024
2.渤海大学 信息科学与工程学院, 辽宁 锦州 121000)

摘要: 快速的支持向量机分类算法——FCSVM对支持向量集采用变换的方式,用支持向量集的子集代替全部支持向量进行分类计算,在保证不损失分类精度的前提下使得分类速度较传统SVM算法有较大提高.为了获得最小的支持向量子集,同时避免支持向量的移动,对FCSVM算法进行了改进.采用二分法优化分类函数中的支持向量数,给出了变换矩阵存在的充要条件及构造方法,减少了计算量.实验结果表明,改进的快速分类算法较大幅度地减少了计算复杂度,提高了分类速度,尤其在训练集规模庞大、支持向量数量较多的情况下,效果更加明显.

关键词: 支持向量机; 快速算法; 分类; FCSVM; 改进
中图分类号: TP18 **文献标识码:** A

0 引言

支持向量机 SVM(support vector machines)是一种基于统计学习理论的机器学习方法,其核心内容是在1992到1995年间提出的,目前仍处在不断发展阶段^[1-2].支持向量机可用于模式识别、回归分析和主成分分析等^[3-6].在SVM训练方面的主要成果有Platt的SMO(sequential minimal optimization)方法^[7]、Osuna等的Chunking方法^[8]和Joachims的SVM light方法^[9]等,这些方法都是针对训练过程的,不涉及分类过程.在训练SVM的过程中使用了所有样本,因此对分类速度的提高没有影响.Lee等提出了一种减少SVM训练时间、加快训练速度的方法RSVM^[10](reduced support vector machines).该方法在训练过程中不是采用所有样本,而是随机地选取其中的一个子集进行训练,通过削减训练规模而达到加快训练速度的目的.同时,由于产生的支持向量数量随之减少,也在一定程度上加快了分类速度,但是由于损失了部分支持向量分类精度有所下降,特别是当支持向量数量较多时,其分类精度会有较大程度的下降.

Burges提出一种提高分类速度的方法^[11],该方法在分类函数中不使用支持向量,而使用一个约简的向量集.与标准SVM不同的是该向量集中的向量既不是训练样本也不是支持向量,而是经过变换的特殊向量.该方法取得了一定的效果,但是在寻找这个约简的向量集的过程中要付出巨大的计算代价,以至于不能广泛地在实际中使用.文献[12]提出了一种快速的支持向量机分类算法FCSVM(fast classification for support vector machines),该方法通过变换方式减少分类函数中的支持向量,从而达到提高分类速度的目的.但由该算法的实现过程可知,它没能使分类函数中的支持向量减少到最少程度,因而分类速度也没有提高到最大幅度.本文就此对该分类算法(FCSVM)进行改进.即采用二分法在支持向量集中选择一个最小的满足精度要求的子集作为分类函数的计算主体,同时给出变换矩阵存在的充要条件和构造方法,以降低构造变换矩阵的计算复杂度,避免支持向量的移动;在保证分类精度的条件下,使分类函数中的支持向量数达到最少,提高分类速度.

收稿日期: 2005-08-20; 修回日期: 2007-01-05.
基金项目: 公安部资助项目(20032129001).
作者简介: 秦玉平^{*}(1965-),男,博士生,教授, E-mail: qlq888888@sina.com; 王秀坤(1945-),女,教授,博士生导师.
©1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1 快速的支持向量机分类算法 (FCSVM)

快速的支持向量机分类算法 (FCSVM)^[12] 的思想是通过适当变换特征空间中的内积矩阵, 进一步变换分类函数的形式, 以减少分类函数中的支持向量, 提高分类速度, 同时保留其他支持向量的信息, 使全部支持向量的信息保留在分类函数中, 以保证在不损失分类精度的前提下通过减少支持向量提高分类速度. 减少分类函数中支持向量的方法是先将支持向量集 S (支持向量数为 N) 随机划分为两个初始子集 P (m 个支持向量, 分类函数中保留的支持向量) 和 Q (n 个支持向量, 通过变换用前 m 个支持向量表示), 构造变换矩阵, 如果不满足精度要求, 则从 Q 中转移一个支持向量到 P 中, 更新子集 P 和 Q , 构造变换矩阵, 重复执行此操作, 直至满足精度要求为止. 若初始划分就满足精度要求, 该方法并没有继续判断是否有更少的支持向量也满足精度要求, 因此, 该方法不能使分类函数中的支持向量数达到最少, 分类速度不能达到最快.

2 改进的快速分类算法

为了快速地获得满足精度的最小支持向量子集, 本文采用二分法优化分类函数中的支持向量, 同时给出变换矩阵的构造方法. 改进的快速分类算法描述如下:

设有 l 个 d 维样本 x_i 构成样本集 X , 即 $X = \{x_i | x_i \in X, i = 1, \dots, l\}$, 令样本属于二类, 分别标识为 $+1$ 和 -1 , 即类标识集 $Y = \{y_i | y_i \in \{+1, -1\}\}$. 训练后产生 N 个支持向量 s_j , 它们构成支持向量集 S , 即 $S = \{s_j | s_j \in S, j = 1, \dots, N\}$. 利用二分法将 S 划分为 2 个支持向量子集 P 和 Q , 分别包含 m 和 n 个支持向量, 即 $P = \{s_k | s_k \in S, k = 1, \dots, m\}$, $Q = \{s_k | s_k \in S, k = m+1, \dots, N\}$. 将所有样本 x_i 和支持向量 s_j 通过某个非线性函数 H 映射到一个高维特征空间 H , 即 $H: X \rightarrow H$. 样本 x_i 和支持向量 s_j 在该高维空间中的映像分别为 $H(x_i)$ 和 $H(s_j)$.

对于内积 $H(x_i)H(s_j)$ 而言, 在 H 为非线性的情况下直接计算是非常困难的, 此时可以选择一个满足 Mercer 条件的核函数 $K(x_i, s_j) = H(x_i)H(s_j)$, 通过计算原样本空间中 x_i 和 s_j 之间的内积而间接得到高维特征空间 H 中的内积

$H(x_i)H(s_j)$, 在此令 $K_{ij} = H(x_i)H(s_j)$, 则 X 关于 P 和 Q 在特征空间 H 中的内积矩阵 K_m 和 K_n 如式 (1) 和式 (2) 所示:

$$K_m = \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1(l-1)} & K_{1l} \\ K_{21} & K_{22} & \cdots & K_{2(l-1)} & K_{2l} \\ \vdots & \vdots & & \vdots & \vdots \\ K_{(m-1)1} & K_{(m-1)2} & \cdots & K_{(m-1)(l-1)} & K_{(m-1)l} \\ K_{m1} & K_{m2} & \cdots & K_{m(l-1)} & K_{ml} \end{pmatrix} \quad (1)$$

$$K_n = \begin{pmatrix} K_{(m+1)1} & K_{(m+1)2} & \cdots & K_{(m+1)(l-1)} & K_{(m+1)l} \\ K_{(m+2)1} & K_{(m+2)2} & \cdots & K_{(m+2)(l-1)} & K_{(m+2)l} \\ \vdots & \vdots & & \vdots & \vdots \\ K_{(N-1)1} & K_{(N-1)2} & \cdots & K_{(N-1)(l-1)} & K_{(N-1)l} \\ K_{N1} & K_{N2} & \cdots & K_{N(l-1)} & K_{Nl} \end{pmatrix} \quad (2)$$

对于支持向量集 S , 设 SVM 的分类函数为符号函数 $f(x)$, 其具体形式如式 (3) 所示:

$$f(x) = \operatorname{sgn} \left(\sum_{j=1}^N \mathcal{T}_j y_j K(x, s_j) + b \right) \quad (3)$$

其中 N 是支持向量数, $\mathcal{T}_j$ 是拉格朗日乘子, 且满足 $\mathcal{T}_j > 0$, b 是阈值.

根据支持向量子集 P 和 Q , 对式 (3) 的求和项进行分解^[12], 得到式 (4):

$$\sum_{j=1}^N \mathcal{T}_j y_j K(x, s_j) = \sum_{j=1}^m \mathcal{T}_j y_j (x, s_j) + \sum_{j=m+1}^N \mathcal{T}_j y_j (x, s_j) \quad (4)$$

构造一个如式 (5) 所示的变换矩阵 W^T , 使得式 (6) 成立:

$$W^T = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1(m-1)} & W_{1m} \\ W_{21} & W_{22} & \cdots & W_{2(m-1)} & W_{2m} \\ \vdots & \vdots & & \vdots & \vdots \\ W_{(n-1)1} & W_{(n-1)2} & \cdots & W_{(n-1)(m-1)} & W_{(n-1)m} \\ W_{n1} & W_{n2} & \cdots & W_{n(m-1)} & W_{nm} \end{pmatrix} \quad (5)$$

$$K_n = W^T K_m \quad (6)$$

对于给定的 K_n 和 K_m , 式 (6) 中 W^T 存在的充分必要条件是

$$K_n K_m^{(1)} K_m = K_n \quad (7)$$

其中, $K_m^{(1)} \in K_m \{1\}$, 当式 (7) 有解时, 其通解为

$$W^T = K_n K_m^{(1)} + y - y K_m K_m^{(1)} \quad (y \text{ 为任意矩阵}) \quad (8)$$

这里 $K_m \{1\}$ 表示 K_m 的 $\{1\}$ -逆.

利用变换矩阵 W^T , 将式 (4) 转化为如式 (9)

所示的形式:

$$\sum_{j=1}^N T_j y_j K(\mathbf{x}, \mathbf{s}_j) = (\mathbf{A}_m^T + \mathbf{A}^T \mathbf{W}^T) \mathbf{K}_t \quad (9)$$

其中 $\mathbf{A}_m^T$ 、 $\mathbf{A}^T$ 和 $\mathbf{K}_t$ 分别如式 (10)、(11) 和 (12) 所示:

$$\mathbf{A}_m^T = (T_1 y_1 \quad T_2 y_2 \quad \cdots \quad T_m y_m) \quad (10)$$

$$\mathbf{A}^T = (T_{m+1} y_{m+1} \quad T_{m+2} y_{m+2} \quad \cdots \quad T_N y_N) \quad (11)$$

$$\mathbf{K}_t = (K_{t1} \quad K_{t2} \quad \cdots \quad K_{tm})^T \quad (12)$$

由式 (3) 和式 (9) 得到精简的分类函数为

$$f(x) = \text{sgn}((\mathbf{A}_m^T + \mathbf{A}^T \mathbf{W}^T) \mathbf{K} + \mathbf{b}) \quad (13)$$

为使分类函数式 (13) 的分类速度达到最快, 必须在保证精度要求的条件下使其包含的支持向量数最少. 解决式 (14) 的优化问题便可达到此目的.

$$\begin{aligned} \min \quad & m \\ \text{s. t.} \quad & \mathbf{K}_m = \mathbf{W}^T \mathbf{K}_m \\ & \|\Theta' - \Theta\| \leq X \text{ (为分类误差)} \end{aligned} \quad (14)$$

其中

$$\begin{aligned} \Theta &= (\theta_1 \quad \theta_2 \quad \cdots \quad \theta_l), \Theta' = (\theta'_1 \quad \theta'_2 \quad \cdots \quad \theta'_l), \\ \theta_i &= \sum_{j=1}^N T_j y_j K(\mathbf{x}_i, \mathbf{s}_j), \theta'_i = (\mathbf{A}_m^T + \mathbf{A}^T \mathbf{W}^T) \mathbf{K}_m \end{aligned} \quad (15)$$

式 (14) 的求解步骤如下:

- 步骤 1 $low = 1, high = N$;
- 步骤 2 如果 $low \geq high, m = low$, 进入步骤 7; 否则, $m = [(low + high) / 2]$;
- 步骤 3 按照式 (1) 和 (2) 分别计算出 $\mathbf{K}_m$ 和 $\mathbf{K}_n$;
- 步骤 4 按照式 (7) 判定变换矩阵 $\mathbf{W}^T$ 的存在性, 若 $\mathbf{W}^T$ 不存在, $low = m$, 进入步骤 2;
- 步骤 5 按照式 (8) 计算 $\mathbf{W}^T$, 对所有测试样本, 根据式 (15) 得到 Θ 和 Θ' ;
- 步骤 6 如果 $\|\Theta' - \Theta\| > X, low = m$; 否则, $high = m$, 进入步骤 2;
- 步骤 7 循环终止.

3 改进的快速分类算法分析

由优化问题的求解步骤可知, 改进的快速分类算法采用二分查找法优化分类函数中的支持向量数, 即使找到满足精度要求的支持向量集 P , 还要继续查找是否有更少的满足精度要求的支持向量集 P 存在. 因此, 改进的算法能使分类函数中的支持向量数达到最少, 从而使分类器速度达到最大幅度. 特别在训练集规模庞大或者支持向量

数量较多的情况下, 其速度优势更加明显. 这些从下面的实验结果中也可以看到. 另外, 改进前的分类算法, 若初始划分不满足精度要求, 则从子集 Q 向子集 P 转移支持向量, 每转移一个支持向量, 就要构造一次变换矩阵, 因此, 构造变换矩阵的计算复杂度为 $O(N)$. 改进后分类算法, 不仅避免了支持向量的移动, 而且构造变换矩阵的计算复杂度为 $O(\log_2 N)$. 由此可知, 改进后的算法提高了支持向量子集 P 的优化速度. 由于该算法对其他 n 个支持向量不是采取简单抛弃的方式, 而是通过变换矩阵 $\mathbf{W}^T$ 将其利用起来, 在 $\mathbf{W}^T$ 中保留了这些支持向量的信息, 因此在减少计算量、提高速度的情况下保证了分类器精度.

4 实验结果及分析

实验所用的数据集中包含了 2 000 篇刑事审讯文档, 按文档内容分为两类. 将这些文档分成两个部分, 一部分作为训练集 (1 200 篇), 另一部分作为测试集 (800 篇). 用 SMO 算法训练后得到的支持向量数为 122. 实验过程中采用了径向基函数 (RBF) 作为分类器的核函数, 其中 SVM 参数为 $C = 100$, RBF 参数为 $\gamma = 110$. 表 1 是重复 10 次提取支持向量子集和对测试集进行分类的实验结果.

表 1 实验结果

Tab. 1 Experiment results					
实验 次序	选取支持向量数		损失精度 /%		改进后较改进 前速度提高 /%
	改进前	改进后	改进前	改进后	
1	74	68	0	0	8.2
2	68	68	0	0	0
3	87	68	0	0	11.6
4	79	68	0	0	13.9
5	68	68	0	0	0
6	68	68	0	0	0
7	120	68	0	0	43.7
8	68	68	0	0	0
9	68	68	0	0	0
10	98	68	0	0	30.4

从表 1 中可以看到, 在保证不损失精度的前提下, 改进后的算法使分类函数中的支持向量数达到最少, 分类速度平均能够提高 10.78%.

5 结 论

本文对快速的支持向量机分类算法 (FCSVM) 进行了改进, 改进的算法采用二分法在支持向量集中提取了一个最小的满足精度要求

的子集作为分类函数的计算主体,降低了求解支持向量集和构造变换矩阵的时间复杂度,使分类函数中的支持向量数在保证精度要求的前提下减少到最小,从而使得计算量减少到最少,使分类速度提高到最大幅度.由于分类函数中隐式地使用了被剔除的支持向量,分类函数事实上保留了所有支持向量的信息,因此分类精度没有损失.特别是当训练集规模较大或者支持向量数目较大时,效果更加明显.

参考文献:

[1] ZHAN G Xue-gong. Introduction to statistical learning theory and support vector machines [J]. *Acta Autom Sinica*, 2000, 26(1): 32-42

[2] 许建华, 张学工, 李衍达. 支持向量机的新发展 [J]. *控制与决策*, 2004, 19(5): 481-484

[3] VAPNIK V N. *Statistical Learning Theory 2nd ed* [M]. New York: Springer-Verlag, 1999

[4] LASKOV P. Feasible direction decomposition algorithms for training support vector machines [J]. *Machine Learning*, 2002, 46(1/3): 315-349

[5] 曾文华, 马健. 支持向量机增量学习的算法与应用 [J]. *计算机集成制造系统—CIMS*, 2003, 9(12): 144-148

[6] COLLOBERT R, BENGIO S. SVM torch a

support vector machine for large-scale regression and classification problems [J]. *J Machine Learning Res*, 2001, 1: 143-160

[7] PLATT J C. Fast training of support vector machines using sequential minimal optimization [M] // SCHOLKOPF B, BURGESS C, SMOLA A J. *Advances in Kernel Methods Support Vector Learning*. Cambridge: MIT Press, 1999: 185-208

[8] KEERTHI S, GILBERT E. Convergence of a generalized SMO algorithm for SVM classifier design [J]. *Machine Learning*, 2002, 46(1/3): 351-360

[9] LIN Chih-jen. On the convergence of the decomposition method for support vector machines [J]. *IEEE Trans on Neural Networks*, 2001, 12(6): 1288-1298

[10] LEE Y J, MANGASARIAN O L. RSV M: Reduced support vector machines [R]. Wisconsin Data Mining Institute, Computer Sciences Department, University of Wisconsin, 2000

[11] YANG M H, AHUJA N. A geometric approach to train support vector machines [C] // *Proceeding of CVPR 2000*. Hilton Head Island [s n], 2000: 430-437

[12] 刘向东, 陈兆乾. 一种快速支持向量机分类算法的研究 [J]. *计算机研究与发展*, 2004, 41(9): 1327-1332

Research on an improved fast classification algorithm of support vector machines

QIN Yu-ping^{* 1,2}, WANG Xiu-kun¹

(1.School of Electr. and Inf. Eng., Dalian Univ. of Technol., Dalian 116024, China;
2.College of Inf. Sci. and Eng., Bohai Univ., Jinzhou 121000, China)

Abstract The fast classification for support vector machines (FCSVM) performs transformation on the full set of support vectors, a subset of support vectors, which contains fewer support vectors, is used in classification. The speed of classification is much faster than that of conventional SVM under the condition that the precision of classification does not decline. In order to obtain minimal subset of support vectors and avoid movement of support vectors, the improvement on the method is presented. The minimal subset of support vectors is gotten by dichotomy. The condition of sufficient and necessary transform matrix existing and its construction method are given, the amount of computation is reduced. The experimental results show that the improved algorithm can remarkably reduce the computation complexity and the speed of classification is the fastest, especially in the case of large number of support vectors.

Key words support vector machine; fast algorithm; classification; FCSVM; improvement

©1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. <http://www.cnki.net>