

大数据中边界向量调节熵函数支持向量机研究

林 蔚, 刘 婷, 吕伟国

(哈尔滨工程大学 理学院, 黑龙江 哈尔滨 150001)

摘 要: 当训练集的规模很大时, 一般的支持向量机的学习过程需要占用大量的内存, 寻优速度缓慢, 不利于实际应用. 提出了一种预抽取支持向量的支持向量机调节熵函数法. 首先, 利用两凸包相对边界向量方法预抽取出边界向量; 然后, 利用支持向量机调节熵函数法来训练预抽取的边界向量. 实验表明, 采用这种方法来训练样本集不仅降低了学习的代价, 还提高了分类速度.

关键词: 数据挖掘; 支持向量机; 调节熵函数; 预抽取支持向量

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2016)08-0149-04

Support Vector Machine Based on Boundary Vectors Adjustable Entropy Function

LIN Wei, LIU Ting, LV Wei-guo

(College of Science, Harbin Engineering University, Harbin 150001, China)

Abstract: When the size of the training set is large, learning process in general support vector machines take a lot of memory, optimizing slow, is not conducive to practical application. This paper presents a boundary vector-based SVM adjustable entropy function method. First, we use two methods convex hull boundary vectors relative pre-extracted boundary vectors; Then, to train pre-drawn boundary vectors in SVM adjustable entropy method. Experiments show that this method not only reduces the training sample set the price of learning, but also improve the classification rate.

Key words: data mining; support vector machine; adjustable entropy; pre-extraction support vector machine

1 引 言

数据挖掘(Data Mining, DM)是从大量的不完全的、嘈杂的、模糊的随机数据库的过程中提取隐含的、未知的、潜在利用信息和知识^[1-3]. 熟为人知的数据挖掘方法有遗传算法、决策树法、神经网络、统计分析方法、粗糙集法、支持向量机等^[4-5].

文献[6]提出的支持向量机(Support Vector Machine, SVM)已经成为解决分类、回归和其他统计学习问题的流行方法之一. 其中, SVM 的训练需要解一个与问题样本数目同样多参数的凸二次规划问题, 特别是对于大规模数据集的训练存在着一定的困难. 现有的大多数典型算法都是基于序贯输入

样本的方式或分解训练集, 如 SMD 算法、块算法和分解算法等. 这些算法虽然在提高效率的同时还节省了内存空间, 但是在设计和实验方面都比较复杂, 不易于实现.

在 1990 年后人们发掘了一种新的数值优化方法——极大熵方法^[1-7]. 其在解决非线性规划和极值等问题时, 具有收敛速度快, 易于计算机实现, 数值稳定性好等优势. 但是进一步的研究表明, 极大熵函数法还存在着一定的不足, 当参数 P 趋于无穷大时, 容易产生数值溢出现象. 后来带调节因子的熵函数法的补充了熵函数法的不足, 参数 p 的值无需取得很大就可以快速趋于问题的精确解.

本文提出了一种预抽取支持向量的调节熵函数

法. 先利用相对边界向量方法预抽取出边界向量; 在利用支持向量分类机 (Support Vector Classification, SVC) 调节熵函数法来训练预抽取的边界向量, 从而为 SVC 求解最优化问题开辟了一条新路径.

2 支持向量机

建立在统计学习理论和结构风险最小化原理基础上而产生的 SVM, 其基本原理是找到一个满足分类要求的最优分类超平面, 并且找的超平面要在保证分类精度的同时也可以满足超平面两侧空白的区域最大化.

设样本 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l) \in R^n$ 线性可分, 类别标号为 $y_i \in \{-1, 1\}$, $g(x) = \omega \cdot x + b$ 为 n 维空间线性判别函数一般形式, 分类面方程为

$$\omega \cdot x + b = 0 \quad (1)$$

权向量和分类阈值分别为 ω 和 b , 所有样本被分类面正确的分类, 需满足:

$$y_i[(\omega \cdot x) + b] - 1 \geq 0, i = 1, 2, \dots, n \quad (2)$$

成立的样本点是支持向量的条件是满足上述条件且使得 $\|\omega\|^2$ 达到最小的分类面, 优化问题的最优分类函数:

$$f(x) = \text{sgn}\left\{\sum_{i=1}^n a_i^* y_i (x_i \cdot x) + b^*\right\} \quad (3)$$

3 预抽取边界向量

3.1 线性可分情况下

定义 1 (距离) 给定两个不同样本 $x_1, x_2 \in R^N$, x_i 表示 n 维样本的第 i 维, 则两个样本间距离可表示为:

$$d(x_1, x_2) = \|x_1 - x_2\| = \sqrt{\sum_i (x_1^i - x_2^i)^2} \quad (4)$$

定义 2 (边界向量) 我们定义边界向量为位于模式边界上的向量.

假设有两类样本 $x_{1i}, x_{2j}, i = 1, 2, \dots, l, j = 1, 2, \dots, m, d(x_{1i}, x_{2j})$ 为两类样本中不同点的距离, 每一个 i 值 $\text{mind}(x_{1i}, x_{2j}), i = 1, 2, \dots, l$ 所对应的向量为第一类样本中的一个边界向量. 因此, 当 i 或 j 取遍所有值, 可以获得包含全部支持向量的两类样本的边界向量集合, 并且边界向量的数量仅占总样本数量的一小部分.

3.2 非线性可分情况下

非线性情况下, 利用一个非线性核函数 Φ 将原

问题映射到线性可分的高维空间 H 中.

定义 3 (距离) 给定两个原输入空间中的不同样本 $x_1, x_2 \in R^N$, 二者之间的距离可以表示为 $\|x_1 - x_2\|$, 其在高维特征空间中的距离可以表示为 $d(x_1, x_2) = \|\Phi(x_1) - \Phi(x_2)\|$

$$= \sqrt{k((x_1, x_2)) - 2k(x_1, x_2) + k(x_1, x_2)} \quad (5)$$

由定义 3 可知样本在非线性高维空间中线性可分, 利用预抽取边界向量方法找出仅占总样本数量比例很小的边界向量作为训练集训练 SVC, 算法的训练速度将有很大幅度的提高.

预抽取边界向量的实现步骤如下:

设两类样本集合为 $A, B; x_{Ai} \in A, y_{Bj} \in B, i = 1, 2, \dots, l, j = 1, 2, \dots, m;$

(1) 求出 $\text{mind}(x_{Ai}, y_{Bj}), j = 1, \dots, m;$

(2) 将 $i = 1, 2, \dots, l$ 取遍, 找到 A 样本集中所有样本的最小值和 B 样本集中对应的点, 作为预抽取的样本集合 $A_1;$

(3) 求出 $\text{mind}(x_{Ai}, y_{Bj}), i = 1, 2, \dots, l;$

(4) 将 $j = 1, \dots, m$ 取遍, 找到样本集 B 中所有样本的最小值和其对应的样本集 A 中的样本点, 作为预抽取的样本集合 $B_1;$

(5) 对 A_1 和 B_1 样本集进行唯一化处理, 得到的样本集作为预抽取训练样本集合.

4 SVC 的调节熵函数法

4.1 无约束 SVC 模型

训练样本 $\{(x_1, y_1), \dots, (x_l, y_l)\}, x_i \in R, y_i \in R$, 是 k 维的, l 为总样本个数, $x_1 = 1$ 或 -1 是该模型类别, SVC 要求解二次规划问题:

$$\begin{aligned} \min & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t. } & y_i(\omega \cdot x_i - b) + \xi_i \geq 1 \\ & \xi_i \geq 0, i = 1, 2, \dots, l. \end{aligned} \quad (6)$$

最优分类面为 $\omega \cdot x - b = 0$.

令 $r_i = r_i(\omega, b) = 1 - y_i(\omega \cdot x_i - b)$ 在式(6)中的约束条件为 $\xi_i \geq r_i, \xi_i \geq 0$ 而目标函数又线性极小化 ξ_i , 则必有 $\xi_i = \max\{0, r_i\}, i = 1, 2, \dots, l$, 于是式(6)与如下含有 $k+l$ 个变量的无约束不可微规划等价.

$$\min F(\omega, b) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^l \max\{0, r_i\} \quad (7)$$

$$\text{令 } \|\xi\|_1 = \sum_{i=1}^l \max\{0, r_i\}, \|\xi\|_\infty = \sum_{i=1}^l \max$$

$\{0, r_1, r_2, \dots, r_l\}$ 由向量范数等价定理在有限维线性空间上的性质可得到如下不等式:

$$\frac{1}{l} \|\xi\|_1 \leq \|\xi\|_\infty \leq \|\xi\|_1$$

于是规划(7)同下式无约束问题等价:

$$\min_{\omega, b} \Phi(\omega, b) = \frac{1}{2} \|\omega\|^2 + C \max\{0, r_1, r_2, \dots, r_l\} \quad (8)$$

4.2 SVC 的调节熵函数法

对于极值问题 $\min_{x \in R^k} \max_{1 \leq i \leq l} f_i(x)$, 连续可微函数 $f_i(x) \in R^k, l \geq 2$ 的正整数, 通过极大熵函数推出的可微函数 $F_p(x) = \frac{1}{p} \ln(\sum_{i=1}^l \exp(pf_i(x)))$ 称为极大熵函数, 并用该可微函数逼近最大值函数 $F(x) = \max_{1 \leq i \leq l} f_i(x)$, 只有当参数 $p \rightarrow \infty$ 时, $F_p(x)$ 的解才能趋于 $F(x)$ 的精确解, 因此, 带调节因子的熵函数法避免了熵函数法数值溢出的不足。

调节熵函数的定义如下:

$$F(p, \mu) = \frac{1}{p} \ln(\sum_{i=1}^l \mu_i \exp(pf_i(x))) \quad (9)$$

式中, $\mu \in \Delta = \{\mu \in R^l, \sigma \leq \mu_i \leq \frac{1}{l+1}, i = 1, 2, \dots, l, \sum_{i=1}^l \mu_i = 1\}$, $p > 0$, P 和 μ 均为参数, σ 为一个给定非负常数, 根据调节熵函数定义(7)可以等价于如下无约束问题:

$$\min_{\omega, b} \Phi_{\langle p, \mu \rangle}(\omega, b) = \frac{1}{2} \|\omega\|^2 + \frac{c}{p} \ln(\sum_{i=1}^l \mu_i \exp(pr_i)) \quad (10)$$

其中, $r_0(\omega, b) \equiv 0, p > 0, \mu \in \Delta$ 为参数。

易知 $\Phi_{\langle p, \mu \rangle}(\omega, b)$ 为任意阶可微凸函数, 所以式(10)中的任一局部最优解也是全局最优解。

定义 $\Delta_I = \{\mu \in \Delta \mid \text{对 } i \notin I(\omega, b), \mu_i = 0\}$,

$$R(\omega, b) = \max_{1 \leq i \leq l} r_i(\omega, b),$$

$$I(\omega, b) = \{i \mid r_i(\omega, b), i = 0, 1, \dots, l\}.$$

下述定理根据极大熵函数性质得到。

定理 1 设 $\mu \in \Delta, \mu \neq 0$, 则有:

$$C \ln_{i \in I} \mu_{\max} / p \leq \Phi_{\langle p, \mu \rangle}(\omega, b) - \Phi(\omega, b) \leq 0$$

其中 $\mu_{\max} = \max\{\mu_i, i \in I(\omega, b)\}$, 当 $p \rightarrow \infty$ 时 $\Phi_{\langle p, \mu \rangle}(\omega, b)$ 一致收敛于 $\Phi(\omega, b)$ 。

定理 2 对于 $(\omega, b) \subset R^k \times R, p \in R^+$ 有 $\Phi_{\langle p, \mu \rangle}(\omega, b) = \Phi(\omega, b)$, 当且仅当 $\mu \in \Delta_i$ 。

根据上述内容, 可以同时调节 p 和 μ , 使

$\Phi_{\langle p, \mu \rangle}(\omega, b)$ 快速收敛到 $\Phi(\omega, b)$ 。为了避免 $\exp(pr_i)$ 溢出的数值计算问题, 当 p 充分大后, 通过调节 μ 的值使 $\Phi_{\langle p, \mu \rangle}(\omega, b)$ 收敛到 $\Phi(\omega, b)$ 。下面给出调节 μ 的方法。

设 $p > 0, \mu \in \Delta, (\omega, b)$ 是 $\Phi_{\langle p, \mu \rangle}(\omega, b)$ 的极小解, 得到 $\nabla \Phi_{\langle p, \mu \rangle}(\omega, b)$, 即

$$\nabla \Phi_{\langle p, \mu \rangle}(\omega, b) = \begin{pmatrix} \omega \\ 0 \end{pmatrix} + C \sum_{i=1}^l \frac{\mu_i \exp(pr_i)}{\sum_{j=0}^l \mu_j \exp(pr_j)} y_i \begin{pmatrix} -x_i \\ 1 \end{pmatrix} = 0 \quad (11)$$

$$\sigma_i = \frac{\mu_i \exp(pr_i)}{\sum_{j=0}^l \mu_j \exp(pr_j)}, i = 0, 1, \dots, l \quad (12)$$

则有 $\lambda_i \in \Delta$ 且满足

$$\begin{pmatrix} \omega \\ 0 \end{pmatrix} + C \sum_{i=1}^l \sigma_i y_i \begin{pmatrix} -x_i \\ 1 \end{pmatrix} = 0 \quad (13)$$

所以下一个调节因子的值为式(11)定义的 σ , 可以得出 $\Phi_{\langle p, \mu \rangle}(\omega, b)$ 的 Hessian 矩阵为

$$\nabla^2 \Phi_{\langle p, \mu \rangle}(\omega, b) = \begin{bmatrix} \mathbf{I}_k & 0 \\ 0 & 0 \end{bmatrix} + Cp \sum_{i=1}^l \sigma_i \begin{bmatrix} x_i x_i^T & -x_i \\ -x_i^T & 1 \end{bmatrix} - Cp \cdot \left(\sum_{i=1}^l \sigma_i y_i \begin{pmatrix} -x_i \\ 1 \end{pmatrix} \right) \cdot \left(\sum_{i=1}^l \sigma_i y_i \begin{pmatrix} -x_i^T & 1 \end{pmatrix} \right) \quad (14)$$

式中, $\mathbf{I}_k$ 为 k 阶单位矩阵, σ_i 如(12)定义。

4.3 基于边界向量的调节熵函数的 SVC

将预抽取出的边界向量作为训练集, 二次无约束规划问题(10)的 Newton 算法如下:

算法: 输入 $\{(x_i, y_i)\}_{i=1}^l$, 初值 $\mu_i^0 = \frac{1}{l+1}, (i =$

$0, 1, \dots, l), k = 0$,

$$I_0 = (\omega_0, b_0) \in R^{k+1}, \varepsilon_1 > 0, \varepsilon_2 > 0, p_0 = 1.$$

Step1: 计算 $g_k = \nabla \Phi_{\langle p_k, \mu_k \rangle}(z_k)$, 若 $\|g_k\| \leq \varepsilon_1$ 转 Step4;

$$\text{Step2: 计算 } H_k = \nabla^2 \Phi_{\langle p_k, \mu_k \rangle}(z_k), d_k = -H_k^{-1} g_k;$$

$$\text{Step3: 一维搜索 } \min_{\lambda > 0} \Phi_{\langle p_k, \mu_k \rangle}(z_k + \lambda d_k) \text{ 得 } \lambda_k,$$

$$z_{k+1} = z_k + \lambda_k d_k, k = k + 1 \text{ 转 Step1};$$

$$\text{Step4: } z^*(p_k) = z_k, \text{ 若 } \|z^*(p_k) - z^*(p_{k-1})\|$$

$$\geq \varepsilon_2, \text{ 则计算 } p_{k+1} = 10p_k, \mu_i^{k+1} = \frac{\mu_i^k \exp(pk r_i(z_k))}{\sum_{j=0}^l \mu_j^k \exp(pk r_j(z_k))},$$

$i = 0, 1, \dots, l. k = k + 1$ 转 Step1, 否则转下一步;

Step5: 输出 $(\omega^*, b^*) = z_k$, 算法结束。

根据 Newton 算法二次收敛性, 为满足 ε_1 约束

仅需 p 的取值较小和几部迭代,从而使得上述算法二次收敛.

算法运行时除了原始数据的存储外,本方法只涉及了 $k+l$ 阶方阵的求逆运算和存储,对于线性分类问题一般有 k 远远小于 l ,之前的算法需要处理 $l \times l$ 阶矩阵.所以,本文提出的预抽取支持向量的 SVC 的调节熵函数法,无论是在计算量或者在存储量上都较小.

4.4 数值实验

下面通过两个实验验证该算法有效性,实验都是在 MATLAB7.0 下编写,分别采用了二次规划算法和二次收敛的 Newton 算法求解的无约束规划问题.

实验 1 人工数据分类实验

当 $p=10$ 时,算法收敛情况如图 1 所示,结果与二次规划算法完全一致且二者的最优分类面几乎重合,图中最优分类超平面以 SH 表示,边界支持向量以 SV 表示.

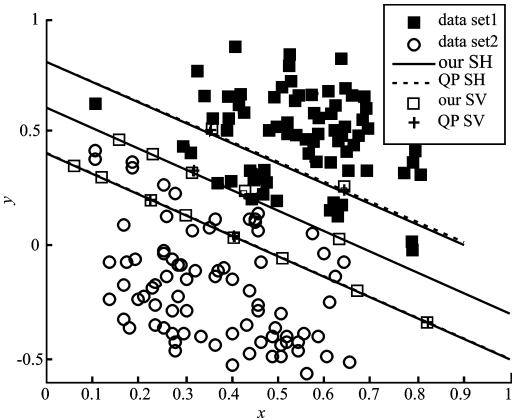


图 1 人工数据分类

实验 2 实际数据分类

样本数据如表 1 所示.

表 1 实验二数据

样本	x_1	x_2	Y
1	1	1	-1
2	3	3	1
3	1	3	1
4	3	1	-1
5	2	2.5	1
6	3	2.5	-1
7	4	3	-1

取正则化参数 $c=2$,零向量做为初始值,当 $p=10$ 时,利用本文提出的算法得到的最优分类超平

面(OUR SH) 如图 2 所示,斜率和截距分别为 0.5 和 1.5,得到的结果是分类问题的精确解.

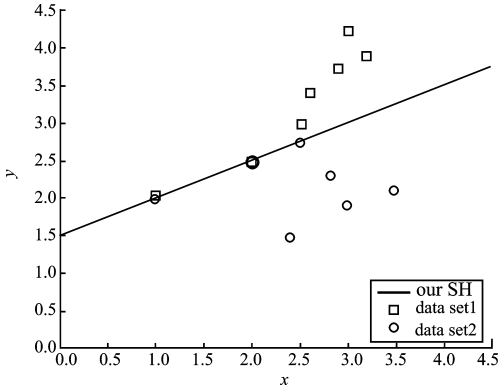


图 2 实际数据分类

实验结果分析:本文实验结果说明不需要将参数 p 的值取的过大得到的最优解也可以与文献[8]结果一致.

5 结束语

本文提出的一种预抽取边界向量的 SVC 调节熵函数刚好可以满足 SVC 的不足,将预抽取边界向量与 SVC 调节熵函数方法相结合.通过人工数据实验和实际数据测试实验表明,保证分类精度的同时克服了 SVC 在训练集很大时的不足,本文提出的算法明显优于其他 SVC 算法.

参考文献:

[1] O L Mangasarian, Mathematical programming in data mining[J]. Data Mining and Knowledge Discovery, 1997,2(1):183-201.

[2] 张敏. 云计算环境下的并行数据挖掘策略研究[D]. 南京:南京邮电大学,2011:1-60.

[3] 郑盼丽,戴壮红. 自动数据挖掘算法[J]. 计算机系统应用,2012,21(11):218-221,193.

[4] 邬伟杰. 基于支持向量机的指纹识别技术与实现方法研究[D]. 上海:华东理工大学,2015:1-50.

[5] 于航. 基于支持向量机模型的股指期货高频交易策略研究[D]. 北京:北京理工大学,2015:1-48.

[6] V N Vapnik. The Nature of Statistical Learning Theory [M]. New York: Springer, 2000:25-46.

[7] 李兴斯. 一类不可微优化问题的有效解法[J]. 中国科学(A 辑),1994,24(4):371-377.

[8] 陶卿,孙德敏,范劲松,等. 基于闭凸包收缩的最大边缘线性分类器[J]. 软件学报,2002,13(3): 404-409.

带相关处理器,采用基于快速傅里叶变换的单通道窄带相关处理计算方法,结合双门限判决接收算法,实现了对窄带信号的盲分离算法改进.研究表明,采用本文算法能有效提高对窄带信号的盲分离检测性能,信号的恢复和特征检测效果较好,提高了信道增益,实现信道均衡.

参考文献:

- [1] Zibullevsky M, Pearlmutter B A. Blind source separation by sparse decomposition in a signal dictionary[J]. Neural computation, 2001, 13(4):863-882.
- [2] Li Y Q, Cichocki A, Amari S. Analysis of sparse representation and blind source separation [J]. Neural Computation, 2004, 16(6):1193-1234.
- [3] He Z S, Xie S L, Fu Y L. Sparse representation and blind source separation of ill-posed mixtures[J]. Science in China Series F: Information Sciences, 2006, 49(5): 639-652.
- [4] Bofill P, Zibulevsky M. Underdetermined blind source separation using sparse representations [J]. Signal Process, 2001, 81(11):2353-2362.
- [5] Xie S L, Liu Y, Yang J M, et al. Time-Frequency Approach to Underdetermined Blind Source Separation [J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(2):306-315.
- [6] 靳晓艳, 周希元, 张琬琳. 多径衰落信道中基于自适应 MCMC 的调制识别[J]. 北京邮电大学学报, 2014, 37(1):31-34.
- [7] 韩慧鹏, 梁红, 胡旭娟. 自适应 IIR 陷波器在信号检测中的应用[J]. 弹箭与制导学报, 2008, 28(2): 315-317.
- [8] Xie S L, Liu Y, Yang J M, et al. Time-frequency approach to underdetermined blind source separation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(2):306-315.
- [9] 饶雨泰, 杨凡. 网络入侵扰动下的网络失稳控制方法研究[J]. 科技通报, 2014, 30(1): 185-188.
- [10] 黎峰, 吴春明. 基于能量管理的网络入侵防波动控制方法研究[J]. 计算机仿真, 2013, 30(12): 45-48, 335.
- [11] Alfaro V M, Vilanovab R. Robust tuning of 2DoF five-parameter PID controllers for inverse response controlled processes[J]. Journal of Process Control, 2013, 23(4): 453-462.
- [12] 杨明, 郝亮, 徐殿国. 基于自适应陷波滤波器的在线机械谐振抑制[J]. 哈尔滨工业大学学报, 2014, 46(4): 63-69.
- [13] 王瑞, 马艳. 基于分数阶傅里叶变换的线性调频脉冲信号波达方向估计[J]. 兵工学报, 2014, 35(3): 421-427.
- [14] 欧世峰, 高颖, 赵晓晖. 自适应组合型盲源分离算法及其优化方案[J]. 电子与信息学报, 2011, 33(5): 1243-1247.
- [15] 张子木. 物联网接入技术研究与系统设计[J]. 电子设计工程, 2016, 24(2):112-120.

作者简介:

陈 婧 女, (1984-), 硕士, 讲师. 研究方向为软件工程、嵌入式数据存储. E-mail: chenjing38495@163. com.

(上接第 152 页)

作者简介:

刘 婷 女, (1989-), 硕士研究生. 研究方向为数据挖掘. E-mail: liuting1126@126. com.

林 蔚 女, (1964-), 博士, 副教授. 研究方向为无线传感器网络、数据挖掘.

吕伟国 男, (1990-), 硕士研究生. 研究方向为数据挖掘.