

一种改进的支持向量机 BS-SVM

郭亚琴¹, 王正群²

(1 紫琅职业技术学院, 江苏 南通 226002; 2 扬州大学 信息工程学院, 江苏 扬州 225009)

摘 要: 提出了一种改进的 SVM: BS-SVM, 它先对训练样本进行分类, 根据每个样本到模式类样本均值的距离, 将训练样本分为三种: 好样本、差样本、边界样本, 然后用边界样本训练得到分类器。实验表明, BS-SVM 相比 SVM 在分类正确率、分类速度以及使用的样本规模上都表现出了一定的优越性。

关键词: 支持向量机; 训练样本; 样本分类; 边界样本

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2010)06-0054-03

An Improved SVM: BS-SVM

GUO Ya-qin¹, WANG Zheng-qun²

(1 Zilang Vocational Technical College, Nantong 226002, China;

2 School of Information Engineering, Yangzhou University, Yangzhou 225009, China)

Abstract: A support vector machine constructs an optimal hyperplane from a small set of samples near the boundary. This makes it sensitive to these specific samples and tends to result in machines either too complex with poor generalization ability or too imprecise with high training error, depending on the kernel parameters. SVM focuses on the samples near the boundary in training time, and those samples intermixed in another class are usually no good to improve the classifier's performance, instead they may greatly increase the burden of computation. In order to improve the generalization ability we present an improved SVM: BS-SVM. It first classifies the training set. According to the distance between the sample and the mean sample, the training sample is classified three classes: good sample, poor sample and boundary sample, then trains the SVM with boundary sample. Experimental results show that BS-SVM is better than SVM in speed and accuracy of classification.

Key words: SVM; training sample; sample classification; boundary sample

1 引言

支持向量机 SVM(Support Vector Machine)以其泛化能力强著称, 并因此得到了人们的青睐。它仅考虑类的边界情况, 用较少的向量(支持向量)来分类。一方面, 它要使得两类边界之间的宽度最大(分类边界将居于两类边界的正中间); 另一方面, 它还要使得错分的代价不要太高。两方面权衡的结果, 就是要使得下面的式子取得最小值: $1/h + l(e)$, 其中 h 表示两类边界之间的宽度, $l(e)$ 表示错分带来的损失。

争取最大的边界宽度是为了保证分类器具有较强的泛化能力。同时这里的“最大”是有条件的, 就是不能付出太多的错分代价, SVM 是这两种要求折中的结果。在训练分类器时, SVM 的着眼点在于两类的交界部分, 那些混杂在另一类中的点往往无助于提高分类器的性能, 反而会大大增加训练器的计算负担, 同时它们的存在还可能造成过学习, 使泛化能力减弱。文献[1]中提出了 NN-SVM 方法, 文献[2]中提出了 NSVM。文中提出了一种新的训练 SVM 的方法 BS-SVM。

收稿日期: 2009-05-25; 修回日期: 2009-07-25

基金项目: 国家自然科学基金项目(60875004); 江苏省高校自然科学基金基础研究项目(07KJB520133); 紫琅职业技术学院科研项目(科研 2008003)

2 SVM

SVM 支持向量机可用于模式分类、回归分析、主成分分析等^[3,4]. 下面以模式分类为例来介绍支持向量机的含义.

给定一组训练数据 $(x_1, y_1), \dots, (x_l, y_l), x_i \in R^n, y_i \in \{-1, +1\}, i = 1, 2, \dots, l$. 我们要寻找一个分类规则 $f(x)$, 使它能对未知类别的新样本(新样本与训练样本独立同分布)作尽可能正确的划分. 支持向量机用于分类问题其实就是寻找一个最优分类超平面, 把此平面作为分类决策面. 同时它还通过引进核函数巧妙地解决了在将低维空间向量映射到高维空间向量时带来的“维数灾难”问题^[5-6].

2.1 最优分类超平面

在训练集线性可分情形, SVM 就是要构造一个最优超平面

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

这个超平面既要满足下面的约束条件:

$$y_i[(w \cdot x_i) + b] \geq 1, i = 1, 2, \dots, n \quad (2)$$

同时还要使下面的函数取得最小值:

$$\Phi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \cdot w) \quad (3)$$

通过求解最优化问题可得最优超平面的形式如下:

$$\sum_{SV} y_i \alpha_i^0 (x \cdot x_i) + b_0 = 0 \quad (4)$$

式中, SV 表示支持向量, α_i^0 是拉格朗日乘子.

在训练集线性不可分时, 我们引进松弛因子 $\xi_i \geq 0$ 及惩罚函数 C , 这时需要做的是在约束条件 $y_i((w \cdot x_i) + b) \geq 1 - \xi_i, i = 1, 2, \dots, n$ 下最小化函数:

$$\Phi(\xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \quad (5)$$

类似可得最优超平面, 有了最优超平面, 可以得到分类函数:

$$f(x) = \text{sgn}[\sum_{SV} y_i \alpha_i^0 (x \cdot x_i) + b_0] \quad (6)$$

2.2 核函数

支持向量机的特点之一在于核函数的引入, 低维空间向量集往往难于划分, 因此, 自然想把它们映射到高维空间, 但随之而来的是计算复杂度的增加, 核函数巧妙地解决了这个问题.

若函数 $K(x, y)$ 满足 Mercer 条件^[2], 则 $K(x, y) = \Phi(x)\Phi(y)$, 其中 Φ 表示某个映射. 这样, 只要选取适当的核函数就可以得到对应高维空间的分类函数

$$f(x) = \text{sgn}[\sum_{SV} y_i \alpha_i^0 K(x, x_i) + b_0] \quad (7)$$

式中, $\Phi(x), \Phi(y)$ 是比 x, y 更高维的向量, 由于 $K(x, y)$ 只涉及 x, y , 因此计算没有涉及高维运算. 我们将分类函数类型为式(7)的学习机称为支持向量机.

3 BS-SVM

尽管支持向量机追求的目标是较强的泛化能力, 但相对于具体的样本集, 也可能出现过学习的问题, 如两类样本集混叠较严重时, SVM 的决策面可能由于过分复杂反而降低了其泛化能力. 基于以上考虑, 文中提出了一种改进 SVM: BS-SVM. 取全部训练样本进行训练, 得到全部模式的模型, 然后对全部训练样本进行识别. 在识别训练样本时, 先计算每个模式类的样本均值 $\mu_j (j = 1, 2, \dots, c)$, 训练样本到样本 x_i 均值的距离定义如下:

$$d_i^j = \|x_i - \mu_j\|, i = 1, 2, \dots, l; j = 1, 2, \dots, c \quad (8)$$

根据式(8)定义的距离, 计算样本属于模式类的概率 P_i^j , 定义如下:

$$P_i^j = \frac{1}{c-1} \left[1 - \frac{d_i^j}{\sum_{k=1}^c d_i^k} \right] \quad (9)$$

保存每个样本属于模式类的概率, $p_i^j (i = 1, 2, \dots, l; j = 1, 2, \dots, c)$, 按照概率值由大到小顺序排序, 即 $P_i^1 > P_i^2 > \dots > P_i^c$, 设 x_i 正确候选字的位置为 n , 即 P_i^n 表示在期望模型下的识别概率, 然后按照式(10)计算相对概率值 RP :

$$RP = \begin{cases} P_i^1 - P_i^2, n = 1 \\ P_i^n - P_i^1, n \neq 1 \end{cases} \quad (10)$$

当 $n = 1$, 表示正确候选字的概率最大, 该样本能够被正确识别, 即 $RP > 0$, 且 RP 越大, 样本质量就越好, 就越不容易与其他类样本混淆. 当 $n \neq 1$, 表示该样本正确候选字的概率不是最大的, 该样本被错误识别, n 越大, 就意味着这个样本容易与更多的模式类混淆. 我们将 RP 值位于 0 附近的样本称为边界样本 (boundary sample), 将 $RP > 0$ 的非边界样本称为好样本, 将 $RP < 0$ 的非边界样本称为差样本. 将训练样本分类后, 再用边界样本训练 SVM 的方法称为 BS-SVM^[7].

当然, 上述优点的获得也是付出了一定的代价: 那就是对训练样本分类需要额外的时间, 但是这点代价与上述任何一种收益相比都是微不足道的. 因

为更高的正确率、更快的分类速度以及训练更大的样本集是我们追求的首要目标。

4 实验结果

4.1 实验数据

数据 1, ringnorm, 该数据集是 Leo Breiman 生成的用于两类划分的样本集, 每一类都是取自一个 20 维的多变量正态分布, 类 1 的均值为 0, 方差为单位元的四倍; 类 2 的均值为 $(a, a, \dots, a)$, 方差为单位元, 其中 $a = 2/\sqrt{20}$. 在实验中通过截取某 k 维得到相应 k 维正态分布的样本集。

数据 2, waveform, 此数据集是 3 类 21 维的特征向量集, 在实验中把第二类作为一类, 其他的两类作为另一类。

数据 3, SVM 工具包自带的数据集 iris, 它是两种植物的 4 维特征向量集^[8]。

4.2 实验结果及分析

实验一:

在实验过程中, 对于数据集 2 和 3, 我们采用了 5 倍交叉验证技术, 将每个训练集随机分成 5 组, 每次用其中 2 组作为测试集, 其余的 3 组作为训练集, 实验是 5 次实验结果的平均值。对于数据 1, 我们取前 4 000 个作为训练集, 后 3 400 个作为测试集, 由于数据集 1 本身是随机排列的, 向量的各分量之间也是独立的, 因此这样选取是合理的。实验过程中, 核函数使用径向基函数, 其中 $\sigma = 0.5$, 惩罚参数 $C = 100$ 。

采用好样本、差样本、边界样本训练 SVM, 比较了在这 3 个数据集上的识别结果, 结果见表 1。

表 1 三种样本的实验结果

数据集	好样本	差样本	边界样本
ringnorm	76.3	72.2	80.3
waveform	67.7	59.9	68.0
iris	90.5	66.7	92.9

从表 1 可以看出, 在这三种样本中, 边界样本具有最高的识别率, 实验结果表明用边界样本训练 SVM 是合理的。

实验二:

实验中, 核函数使用径向基函数, 其中惩罚参数 $C = 100$. 我们重点对正态分布样本集进行了训练, 实验结果如图 1~图 4 所示, 这些图都是相对于数据集 1 实验的结果, 数据集 1 为 7 400 个 20 维向量的集合, 每个向量还附带一个类标。我们通常截取 20 维向量的前 K 个分量构成 K 维向量, 取前 N 个

样本作为训练集, (图中横坐标对应 K 或 N 的具体取值), 后 3 400 个样本作为测试集。

实验结果表明, BS-SVM 分类正确率有了明显的提高, BS-SVM 的支持向量数大大减少了, 从而分类时间也大大减少了用于训练 SVM 的边界样本, 也有较大程度的减少。

可以看出, BS-SVM 与 SVM 相比都具有很好的性能。因此, 在样本集太大(受硬件条件限制)无法直接训练的情况下, 我们可以采用文中提到的 BS-SVM 方法。

5 结束语

改进的支持向量机——BS-SVM, 先对训练样本进行分类, 根据每个样本到模式类样本均值的距离, 将训练样本分为三种: 好样本、差样本、边界样本, 然后用边界样本训练得到分类器。实验表明, 相比 SVM, BS-SVM 分类正确率更高、分类速度更快、能训练的样本集更大。可见 BS-SVM 是解决大样本训练 SVM 的有效途径, 同时还可以增强 SVM 的泛化能力。

参考文献:

- [1] 李红莲, 王春花, 袁保宗. 一种改进的支持向量机 NN-SVM[J]. 计算机学报, 2003, 26(8): 1015-1020.
- [2] Yao Yong, Zhao Hui, Liu Zhijing. NSVM decision tree for multi-class classification[J]. Chinese Journal of Electronics, 2008, 17(4): 627-629.
- [3] 叶菲, 罗景青, 俞志富. 一种改进的并行处理 SVM 学习算法[J]. 微电子学与计算机, 2009, 26(2): 40-43.
- [4] 冯佳, 宋胜利, 王荔, 等. 一种新的 SVM 多层增量学习方法 HISVML[J]. 微电子学与计算机, 2009, 26(5): 216-222.
- [5] 张永, 迟忠先, 米滢. 一类直接构造的模糊多类支持向量分类器[J]. 计算机工程与应用, 2008, 44(8): 12-15.
- [6] 刘斌, 张楠. 基于 LS2SVM 的在线文本识别方法[J]. 微电子学与计算机, 2009, 26(3): 192-199.
- [7] 方景龙, 陈铄, 潘志庚, 等. 复杂分类问题支持向量机的简化[J]. 电子学报, 2007, 11(5): 78-82.
- [8] Hsu C, Lin C. A comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks, 2002, 13(2): 415-425.

作者简介:

郭亚琴 女, (1981-), 硕士研究生, 助教. 研究方向为模式识别、人工智能等。

王正群 男, (1965-), 博士后, 副教授. 研究方向为机器学习、神经网络集成等。