

基于错分样本的 AdaBoost 支持向量预选取算法

叶 菲, 罗 军

(解放军陆军军官学院, 安徽 合肥 230031)

摘 要: 支持向量机在大样本情况下训练速度慢, 支持向量预选取可以解决这个问题. AdaBoost 算法重点关注错分样本, 而错分样本一般都处于分类边界, 支持向量就由分类边界样本构成. 因此, 提出基于错分样本的 AdaBoost 支持向量预选取算法, 该算法通过 AdaBoost 提升过程, 使得越是容易被错分的样本权值越大, 从而实现支持向量的预选取, 通过仿真实验验证了算法的有效性.

关键词: AdaBoost; 错分样本; SVM; 支持向量预选取

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2013)04-0050-03

The Support Vector Pre-extracting Method Based on Error Samples AdaBoost Algorithm

YE Fei, LUO Jun

(Army Officer Academy, PLA, Hefei 230031, China)

Abstract: SVM has slow training speed in the condition of a great deal of samples, and support vector Pre-extracting can solve this problem. Adaboost algorithm emphasizes error samples, and error samples are boundary samples which can construct support vector. So the support vector pre-extracting method based on error samples AdaBoost algorithm is proposed. Through AdaBoost upgrading process the algorithm gets error samples a big weight, and the support vector pre-extracting carries out, in the end the simulation experiments validate the algorithm is effective.

Key words: AdaBoost; error sample; SVM; Support Vector Pre-extracting

1 引言

支持向量机(SVM)是基于统计学习理论而提出的一种小样本学习方法, 在解决非线性及高维输入空间等分类问题中表现出许多特有的优势. 但是当样本数量较多时, 训练要花费大量的时间, 这成为制约支持向量机应用的主要问题. 解决的途径之一就是在不影响支持向量机性能的前提下, 预先从已知类标号的样本集中选取一部分样本作为训练样本, 从而减小训练样本的数目^[1]. 本文利用 AdaBoost 训练算法标记错分样本进行 SVM 训练样本的预选取.

2 SVM 训练样本预选取原理分析

SVM 的基本思想用图 1 所示的二维情况来说

明. 图中实心点和空心点分别代表两类样本, H 为分类超平面, H_1 、 H_2 分别为过类中距超平面最近的样本且平行于超平面的直线, 它们之间的距离叫作分类间隔(margin). 最优分类平面就是要求超平面不但能将两类正确分开(即训练错误率为 0), 且要使分类间隔最大. 分类超平面的方程为:

$$W^T \cdot x + b = 0 \quad (1)$$

对分类超平面进行归一化, 使得对线性可分的样本集 $\{(x_i, y_i)\}$, $i=1, 2, \dots, n$, $x_i \in R^N$, $y_i \in \{+1, -1\}$, 满足:

$$y_i[(w^T \cdot x_i) + b] - 1 \geq 0 \quad i=1, 2, \dots, n \quad (2)$$

此时分类间隔等于 $\frac{2}{\|w\|}$, 使间隔最大等价于使 $\|w\|$ 最小, 故满足 (2) 式且使 $\|w\|$ 最小的分类面就是最优分类面. 其中 H_1 和 H_2 上的训练样

本使式(2)中的等号成立,支撑了最优分类面,被称作支持向量^[2-3]。

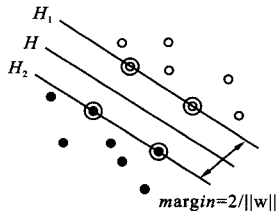


图1 最优分类面示意图

SVM的核心问题就是利用训练样本获取支持向量。从图1可以看出,支持向量位于分类的边界,是最容易被错分的样本,而那些离分类边界较远的样本则不可能成为支持向量。因此,预选取算法就是要选出位于分类边界的样本作为训练样本,从而减少训练样本数,提高支持向量机的训练效率,同时不影响 SVM 的分类性能。

3 基于错分样本的 AdaBoost 预选取算法

AdaBoost 是一种典型的集成学习方法,思想如下:给定训练集 $\{(x_i, y_i)\}, i=1, 2, \dots, n, x_i \in X, X$ 表示实例空间, $y_i \in Y, Y$ 为标签集。 D 为定义在训练集上的分布,为 (x_i, y_i) 的分布权值, P 为定义在 $X \times Y$ 上的分布,用于有放回的随机抽取 (x_i, y_i) 。 AdaBoost 根据权值分布 D ,反复调用弱学习器 T 次,每一次都根据 P 对训练集进行有放回随机抽样,生成新的训练集。第 t 回迭代时样本上的分布权值为 $D_t(i)$ (初始时所有样本 (x_i, y_i) 的分布权值为 $\frac{1}{n}$)。

在每一回迭代中都将增加错分样本的分布权值,以使弱分类器在下一轮迭代中对它们集中学习,从而生成合适的弱分类器 $h_t: X \rightarrow Y, h_t$ 的目标就是在 $D_t(i)$ 上的分类错误率 ϵ_t 最小。这样, T 次迭代后,用加权多数投票把 T 个弱分类器组合生成最终分类器 H 。只要弱分类器的分类正确率比随机猜测略好, AdaBoost 就能提升其分类正确率^[4-5]。

可以看出, AdaBoost 算法中随着迭代次数的增加,那些权值越大的样本就是越容易被错分的样本,也就是处于分类边界的样本^[6-7]。本文就是基于此思想,提出基于错分样本的 AdaBoost 样本预选取算法。为简便起见,本文只考虑二类分类问题,多类分类问题可以用级联的二类分类问题来实现。假设训练集 $S = \{(x_i, y_i)\}, i=1, 2, \dots, n, x_i \in R^N, y_i \in \{+1, -1\}$ 。预选取算法的具体过程如下:

(1) 令 $t=1$, 对于每一个样本 (x_i, y_i) , 初始化样

本分布权值 $d_t(i) = d_1(i) = \frac{1}{n}$, 样本抽取概率 $p_t(i) = p_1(i) = \frac{1}{n}$;

(2) 根据样本抽取概率 $p_t(i)$ 对样本集进行抽样产生训练集 D_t ;

(3) 在 D_t 上采用决策树作为分类器基本学习算法,训练出弱分类器 $h_t: X \rightarrow \{+1, -1\}$, 用 h_t 对原训练集 S 中的所有样本分类,将分类错误的样本放入错误样本集 R_t ;

(4) 计算 h_t 的错误率为 $\epsilon_t = \sum_{h_t(x_i) \neq y_i} d_t(i)$, 令 $\beta_t = \frac{\epsilon_t}{1 - \epsilon_t}$, 采用如下公式更新权值向量和样本抽取概率。

$$d_{t+1}(i) = d_t(i) \beta_t^{\frac{1}{2}(1 + |y_i - h_t(x_i)|)} \quad (3)$$

$$p_{t+1}(i) = \frac{d_{t+1}(i)}{\sum_{i=1}^n d_{t+1}(i)} \quad (4)$$

(5) 如果 $t < T$, 令 $t = t + 1$, 根据更新后的样本抽取概率 $p_{t+1}(i)$ 从 $S - R_t$ (S 中除去被上一个分类器分类错误的样本) 的样本中进行重采样,把重采样集合连同 R_t 一起作为新的训练集 D_t , 返回步骤(4); 否则, 进入下一步;

(6) 假设实际需求的训练样本数目为 M , 则从样本集中选出分布权值最大的 M 个样本, 即为预选选出的 SVM 训练样本。

该算法在每一次提升过程中,都把上一次分类器分类错误的样本直接加入作为下一个分类器的训练样本,这样使算法可以更为关注错分的样本。

4 仿真实验

仿真1: 二类分类的例子

随机产生两类均匀分布的非线性可分数据, 分别用星号和加号表示, 如图2所示。利用基于错分样本的 AdaBoost 支持向量预选取算法, 图中加圈号

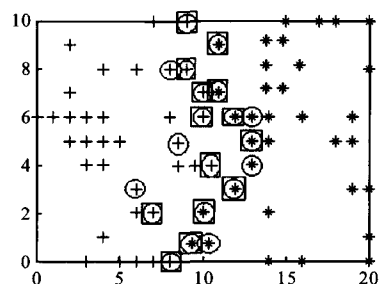


图2 非线性可分的二类分类支持向量预选取

的样本是预选取的边界样本,加矩形的是训练得到的支持向量.表 1 是支持向量机在没有经过支持向量预选取、经过支持向量预选取两种不同条件下的分类结果,从表中可以看出支持向量机在两种条件下的分类效果大致相同,但是经过支持向量的预选取可以减少运算量.

表 1 非线性可分二类分类时 SVM 的分类性能比较

SVM 算法	训练样本	边界样本	支持向量	训练时间	检验样本	识别率
未加支持向量预选取	80	未选	18	24.7548s	100	93.75%
增加支持向量预选取	80	20	14	8.3713s	100	90%

仿真 2:多类分类的例子

设定五部雷达辐射源信号,其参数设置如表 2 所示.每部雷达信号仿真 40 个样本,其中 20 个样本作为训练样本,20 个样本作为测试样本,并在仿真过程中加入了 1%以内的误差.将多类分类问题转换为级联的二类分类问题来实现.图 3 是雷达 1 和雷达 2 在 AdaBoost 算法中构建出的决策树分类器.表 3 是支持向量机在没有经过支持向量预选取、经过支持向量预选取两种不同条件下的分类结果,从表中可以看出支持向量机在两种条件下的分类效果大致相同,但是经过支持向量的预选取可以减少运算量.

表 2 仿真雷达辐射源信号参数设置

雷达	RF/MHz	PRI/ms	PW/us	训练样本	测试样本
1	25 35 45	64	10	20	20
2	32 42	54	30	20	20
3	20	40 60	12	20	20
4	25	30	15	20	20
5	30	35 55	20	20	20

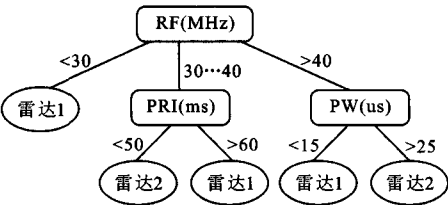


图 3 雷达 1 和 2 在 AdaBoost 算法中构建的决策树分类器

表 3 多类分类时支持向量机的分类性能比较

SVM 算法	训练样本	边界样本	支持向量	训练时间	检验样本	识别率
未加支持向量预选取	100	未选	32	195.8329	100	93.75%
增加支持向量预选取	100	67	28	101.2374s	100	90%

5 结束语

本文提出基于错分样本的 AdaBoost 支持向量预选取算法.从仿真结果可以看出,该支持向量预选取算法可以选出位于分类边界的样本作为训练样本,而减少训练样本的数目,提高支持向量机的训练效率.

参考文献:

[1] 李青,焦李成,周伟达.基于向量投影的支持向量预选取[J].计算机学报,2005,28(2):145—152.

[2] 潘继斌.最优分类超平面的几何意义[J].湖北师范学院学报:自然科学版,2005,25(1):33—35.

[3] 张威,魏冬生.基于 Adaboost 与支持向量机的人脸特征提取[J].微电子学与计算机,2007,24(5):69—72.

[4] 方义秋,卢道兵,葛君伟.一种改进 Adaboost 算法的方法[J].微电子学与计算机,2010,27(7):40—43.

[5] Stan Z Li, ZhenQiu Zhang. FloatBoost learning and statistical face detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(9):1112—1123.

[6] 贺灏,贺建飙.一种基于分布式计算的 Adaboost 训练算法[J].计算机应用与软件,2010,27(2):258—260.

[7] 李闯,丁晓青,吴佑寿.一种改进的 Adaboost 算法—AD AdaBoost[J].计算机学报,2007,30(1):103—109.

作者简介:

叶菲女,(1980—),博士,讲师.研究方向为雷达信号处理、智能信息处理.

罗军男,(1977—),博士,讲师.研究方向为空间信息处理.