一种基于 AdaBoost 的 SVM 分类器

王晓丹[1]， 孙东延[1,2]， 郑春颖[1]， 张宏达[1]，赵学军[1]
(1.空军工程大学 导弹学院，陕西 三原 713800；2.西安电子科技大学 电子工程学院，陕西 西安 710071)

摘 要：针对 AdaBoost 的分量分类器的分类精度和差异性互为矛盾、以至于该矛盾的存在降低了 AdaBoost 算法的分类精度和泛化性的问题，提出了一种变 σ - AdaBoostRBFSVM 算法，通过根据训练样本调整各个分量分类器的核函数参数值，使分量分类器在精度和差异性之间达到一定的平衡，从而提高了集成分类器的分类精度和泛化性。对标准数据集的分类实验结果表明了算法的有效性。

关键词：支持向量机；AdaBoost 算法；分类器
中图分类号：TP391 **文献标识码**：A **文章编号**：1009 - 3516(2006)06 - 0054 - 04

Boosting 是在 PAC 学习问题框架模型下提出的一种提高任意给定弱分类器分类精度的方法[1]，在解决实际问题时有一个重大的缺陷，即要求事先知道弱学习器学习正确率的下限，但这在实际问题中难以做到。AdaBoost 即自适应提升(AdaBoost，Adaptive Boosting)算法[2]，解决了早期 Boosting 算法很多实践上的困难，不需要预先知道弱学习器学习正确率的下限，可以很容易应用到实际问题中。用支持向量机(SVM，Support Vector Machine)作为 AdaBoost 的分量分类器得到的集成分类器的分类性能、泛化性能如何等是近年来倍受研究者们关注的问题[3-6]。本文对如何用 AdaBoost 算法提升 RBFSVM(以径向基函数为核函数的 SVM)的分类精度进行了研究，基于分量分类器分类精度和差异性间的平衡，提出了一种变 σ - AdaBoostRBFSVM 算法，并用标准数据集进行了实验。

1 AdaBoost 及分析

AdaBoost 算法的基本过程是：依次训练一组分量分类器，其中每个分量分类器的训练集都是选择由其它分量分类器给出的“最富信息”(most informative)的样本组成，最后用线性加权集成这些分量分类器，从而得出最终判决结果。其中，“最富信息”样本的选取方法：每个训练样本都被赋予一个权重，表明它被某个分量分类器选入训练集的概率。如果某个样本被当前弱分类器准确分类，那么它的权重就会被降低，则在构造下一个分量分类器的训练集时，它被选中的概率就被降低；相反，如果某个样本没有被正确分类，则它的权重就相应被提高，它入选下一个分量分类器的训练集的概率被提升。通过这种方式，Adaboost 能够“聚焦于”那些比较困难(容易出现错分)的样本。

在具体实现上，令每个训练样本的初始权重都相等，对于第 t 次迭代操作，需要根据第 $t-1$ 次训练得到的样本权重来选取新的训练样本集，进而训练分类器 C_t。然后，用分类器 C_t 对整个样本集进行测试，提高被它错分样本的权重，同时降低可以被正确分类样本的权重。之后，权重更新过的样本集被用来训练下一个分类器 C_{t+1}，整个训练过程如此迭代进行，直到满足结束条件为止。

假设 $\boldsymbol{x}_i$ 和 y_i 表示原始样本集中的样本和它们的类标记，$w_t(i)$ 表示第 t 次迭代时样本 $\boldsymbol{x}_i$ 的权重分布，h_t 是分量分类器 C_t 的决策函数，$h_t(\boldsymbol{x}_i)$ 是由分量分类器 C_t 给出的对样本 $\boldsymbol{x}_i$ 的类标记(+1 或 - 1)。AdaBoost 算法描述如下：

收稿日期：2005 - 12 - 28
基金项目：陕西省自然科学研究计划项目(2004F36)
作者简介：王晓丹(1966 -)，女，陕西汉中人，教授，博士生导师，主要从事智能信息处理、模式识别等研究.

第 1 步　输入:一组有标记的训练样本集 $D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$,$x_i\in X,y_i\in Y=\{-1,+1\}$ 。弱学习算法,迭代次数 T。

第 2 步　初始化:初始化各样本对应的权值: $w_1(i)=1/n,i=1,2,\cdots,n$ 。

第 3 步　For $t=1,2,\cdots,T$

1)在加权训练样本集上用弱学习算法训练弱分类器 C_t得到 h_t,或训练使用按照 $w_t(i)$ 采样的弱分类器 C_t 得到 h_t;

2) 计算 C_t的训练误差 ε_t:$\varepsilon_t=\sum_{i=1}^{n}w_t(i)$,$y_i\neq h_t(x_i)$,即 ε_t相当于错分样本的权值 w_t之和;

3) 设置弱分类器 C_t的权值:$\alpha_t=\frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$;

4) 更新训练样本的权值:$w_{t+1}(i)=\frac{w_t(i)\exp\{-\alpha_t y_i h_t(x_i)\}}{Z_t}=\frac{w_t(i)}{Z_t}\times\begin{cases}e^{-\alpha_t} & if\ y_i=h_t(x_i)\\ e^{\alpha_t} & if\ y_i\neq h_t(x_i)\end{cases}$其中 Z_t 为归一化系数,使 $\sum_{i=1}^{n}w_{t+1}(i)=1$;

第 4 步　输出:总体分类器的判决函数值:$H(x)=\mathrm{sign}[\sum_{t=1}^{T}a_t h_t(x)]$ 。

AdaBoost 方法中,错分样本权值靠因子 $(1-\varepsilon_t)/\varepsilon_t$获得提升,从而增加了错分样本的总权值(设 $\varepsilon_t>0.5$)。增加错分样本的权值并减少被正确分类样本权值的结果是,更高权值的样本对训练中的分类器影响更大,因此使分类器更关注错分样本,这些错分样本通常是最靠近决策边界的样本。

在 T个分量分类器的分类错误率都满足 $\varepsilon_t<0.5$ 的情况下,令 $\gamma_t=0.5-\varepsilon_t$,则 H 的分类错误率上界理论上为[2] $\frac{1}{n}|\{i:H(x_i)\neq y_i\}|\leqslant\prod_{t=1}^{T}\sqrt{1-4\gamma_t^2}\leqslant\exp(-2\sum_{t=1}^{T}\gamma_t^2)$ 。可以看出,随着分量分类器数目 T 的增加,H 的分类错误率指数级下降。因此,只要每个分量分类器都是弱分类器(分类正确率稍大于 50% ,即比随机猜测略好),AdaBoost 就能提升其分类正确率。如果 T 足够大,总体分类器的训练误差就能够任意小,而且当 T 非常大时,过拟合现象很少发生。

2　一种基于 AdaBoost 的 SVM 分类器

对于不稳定的分类器,使用 AdaBoost 算法可以改善其分类准确率。如果分类器是稳定的,即训练数据集中的变化只在分类器上引起很小的变化,则 AdaBoost 对性能改善做出的贡献通常将很小。

差异性是影响集成分类器泛化性能的重要因素[7]。AdaBoost 分量分类器的精度和它的差异性互为矛盾,即两个分量分类器的精度越高,它们之间的差异性就越少。因此,只有当精度和差异性达到某种平衡时,AdaBoost 才能表现出较好的性能。但 AdaBoost 本身对这一问题仍没有一个有效的解决措施。

RBFSVM 有高斯宽度 σ 和规则化参数 C 两个参数,任一个的改变都导致分类器性能的改变。通过选择合适的 C 和 σ 可以有效避免过拟合。对 RBFSVM 的性能分析发现[3],C 值过小,分类器学习能力不好,但当 C 在一个合适的范围内取值时,RBFSVM 的性能可以简单地通过调整 σ 值改变,且 σ 对分类器的影响更大。

在用RBFSVM作为AdaBoost的分量分类器时,如果对所有RBFSVM都取相同σ 值,将会引起以下问题:相对过大的 σ 产生的 RBFSVM 性能过弱,即它们的分类精度都小于50%,因此不能满足 AdaBoost 中对分量分类器的要求;另一方面,相对小的 σ 常常使得 RBFSVM 已经很健壮,分量分类器的错误高度相关、差异性小,而使增强它们可能变得无效。更重要的是, σ 太小甚至可能使 RBFSVM 对训练样本过拟合。因此,需要为各分量分类器找到合适的 σ 。由于对于 RBFSVM,在选择合适的参数 C 时,用训练样本集的标准差作为高斯宽度 σ ,可以获得较高的分类精度,本文通过将训练每个分量分类器的样本集的标准差作为该分量分类器的 σ 值,以控制分量分类器的分类精度,避免了参数 σ 在所有分量分类器中取值相同带来的上述问题,从而得到一种基于 AdaBoost 的 SVM 分类器,以下称为变 σ - AdaBoostRBFSVM 算法,算法描述如下:

第 1 步　输入:一组有标记的训练样本集 $D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$,$x_i\in X=\{-1,+1\}$ 。弱学习器为 RBFSVM,迭代次数 T。

第 2 步 初始化:初始化各样本对应的权值:$w_1(i) = 1/n$,$i = 1,2,\cdots,n$ 。

第 3 步 For $t = 1,2,\cdots,T$

1) 按照 $w_t(i)$ 在 D 中采样,得到训练弱学习器 C_t的训练样本集 d_t;

2) 计算 d_t的标准差 σ :sqrt(mean(var(d_t))) ;

3) 以 d_t为训练样本集,σ 为参数训练弱学习器 C_t得到 h_t, C_t 为以 σ 为参数的 RBFSVM;

4) 计算 C_t的训练误差 ε_t: $\varepsilon_t = \sum_{i=1}^{n} w_t(i)$,$y_i \neq h_t(\boldsymbol{x}_i)$,即 ε_t相当于错分样本的权值 w_t之和;

5) 设置弱分类器 C_t的权值:$\alpha_t = \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$;

6) 更新训练样本的权值:$w_{t+1}(i) = \frac{w_t(i)\exp\{-\alpha_t y_i h_t(\boldsymbol{x}_i)\}}{Z_t} = \frac{w_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t}, & if\ y_i = h_t(\boldsymbol{x}_i) \\ e^{\alpha_t}, & if\ y_i \neq h_t(x_i) \end{cases}$ 其中 Z_t 为归一化系数,使 $\sum_{i=1}^{n} w_t(i) = 1$;

第 4 步 输出:总体分类器的判决函数值:$H(\boldsymbol{x}) = \text{sign}[\sum_{i=1}^{n} a_t h_t(\boldsymbol{x})]$ 。

3 实验及结果

采用标准数据集,对变 σ - AdaBoostRBFSVM 的性能与 SVM(采用径向基核函数)及 AdaBoostSVM(用固定参数 σ 值的 RBFSVM 作为 AdaBoost 分量分类器)的性能进行实验比较。SVM 使用的是 Steve Gunn SVM Toolbox。标准数据集采用的是 Westontoynonliner 数据集和 Wine 数据集。Westontoynonliner 数据集,共 1 000 个样本,样本维数为 52,样本分为 2 类;Wine 数据集,共有 178 个样本,样本维数为 13,样本分为 3 类,以第一类作为正类,其它两类作为负类进行实验。变 σ - AdaBoostRBFSVM、AdaBoostSVM 及 SVM 在相同条件下进行实验,以比较 3 种算法的分类性能。分量分类器数量 T 取 10,惩罚参数 C = 1 000。

对 Westontoynonliner 数据集,实验选取训练样本集的大小分别为 50,150,200,300,500,随机抽取数据集中 128 个样本作为测试样本集。对于 SVM 和 AdaBoostSVM,核函数参数 σ 取 12。对于变 σ - AdaBoostRBFSVM 和 AdaBoostSVM,分别选取训练样本集的一部分(1/2 - 1/10)来训练分量分类器,取 3 次实验分类正确率的平均值作为结果。实验结果如图 1 所示。图中,横坐标为训练样本数,纵坐标为分类正确率,Ada - SVM 指 AdaBoostSVM,改进 Ada - SVM 指变 σ - AdaBoostRBFSVM。

对 Wine 数据集,实验选取训练样本集的大小分别为 50,80,100,130,150,随机抽取数据集中 79 个样本作为测试样本集。对于 SVM 和 AdaBoostSVM,核函数参数 σ 分别取 2,6,12,取分类结果的平均值作为结果分类正确率。对于变 σ - AdaBoostRBFSVM 和 AdaBoostSVM,分别选取训练样本集的一部分(1/2 - 1/8)来训练分量分类器,取 3 次实验分类正确率的平均值作为结果,实验结果如图 2 所示。

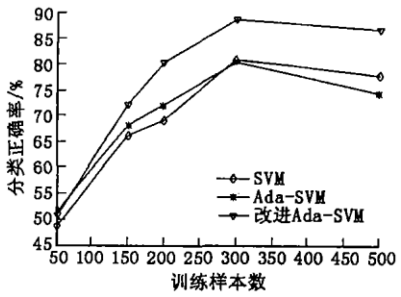

图1 对 Westontoynonliner 数据集的实验结果

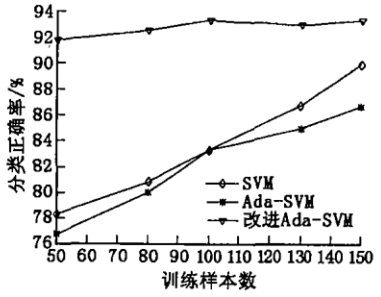

图 2 对 Wine 数据集的实验结果

从图 1 和图 2 可知:AdaBoostSVM 相对于 SVM 来说,分类精度没有明显改善,而变 σ - AdaBoost RBFSVM 则较明显地提高了分类精度;对于所用 Wine 数据集,由于实验是将其第一类作为正类,其它两类作为负类进行,存在样本分布的不均衡(正类 59 个样本,负类 119 个样本),由图 2 可知变 σ - AdaBoostRBFSVM 对

于提高非均衡数据集的分类精度更有效。

所以,变 σ - AdaBoostRBFSVM 与单个 SVM 分类器相比,具有 SVM 模型选择容易的优点;与用固定参数值的 RBFSVM 作为 AdaBoost 分量分类器的 AdaBoostSVM 算法相比,具有泛化性能良好的优点。

4　结束语

本文在分析 AdaBoost 的性能与分量分类器性能间的关系的基础上,对如何用 AdaBoost 算法提升 RBFSVM 的分类精度进行了研究。针对 AdaBoost 的分量分类器的分类精度和它的差异性互为矛盾、以至于该矛盾的存在降低了 AdaBoost 算法整体的分类精度和泛化性的问题,提出了一种变 σ - AdaBoostRBFSVM 算法,它通过根据训练样本调整各个分量分类器的核函数参数 σ 值,使 AdaBoost 的每个分量分类器都达到一个适当的分类精度,在精度和差异性之间达到一定的平衡,从而提高了集成分类器的分类精度和泛化性。对标准数据集的分类实验表明了该方法的有效性,实验同时表明了该算法对于提高非均衡数据集的分类精度具有显著的效果。

参考文献:

[1] Schapire R E. The Strength of Weak Learnability[J]. Machine Learning, 1990,5(2):197-227.

[2] Freund Y, Schapire R E. A Decision-theoretic Generalization of Online Learning and an Application to Boosting[J]. Journal of Computer and System Sciences, 1997,55(1):119-139.

[3] Valentini G, Dietterich T G.. Bias-variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods[J]. Journal of Machine Learning Research. 2004, 5:725-775.

[4] Pavlov D, Mao J, and Dom B. Scaling-up Support Vector Machines Using Boosting Algorithm[A]. Sanfeliu A, Villanueva J J, Vanrell M, Alquezar R, Jain A K, Kittler J(eds.). Proc. of the 15th International Conference on Pattern Recognition[C]. Los Alamitos: IEEE Computer Society Press, 2000,2: 2219-2222.

[5] Kim H C, Pang S, Je H M, et al. Constructing Support Vector Machine Ensemble. Pattern Recognition, 2003,36(12):2757-2767.

[6] 王晓丹, 王积勤. 支持向量机研究与应用[J]. 空军工程大学学报(自然科学版), 2004, 5(3): 49-55.

[7] Ludmila I K, Christopher J W. Measures of Diversity in Classifier Ensembles and their Relationship with the Ensemble Accuracy [J]. Machine Learning, 2003,51(2):181-207.

(编辑:田新华)

A Combined SVM Classifier Based on AdaBoost

WANG Xiao-dan[1], SUN Dong-yan[1,2], ZHENG Chun-ying[1], ZHANG Hong-da[1], ZHAO Xue-jun

(1. The Missile Institute, Air Force Engineering University, Sanyuan 713800, Shaanxi, China; 2. School of Electronic Engineering, Xidian University, Xi'an 710071, Shaanxi, China)

Abstract: The relation between the performance of AdaBoost and that of component classifiers is analyzed, and the approach of improving the classification performance of RBFSVM is studied. There is an inconsistency between the accuracy and the diversity of component classifiers, and the inconsistency affects the generalization performance of the algorithm. A new variable δ - AdaBoostSVM is proposed by adjusting the kernel function parameter of the component classifier based on the distribution of training samples, and it improves the classification performance by making a balance between the accuracy and diversity of component classifiers. Experimental results indicate the effectiveness of the proposed algorithm.

Key words: support vector machine; AdaBoost Algorithm; classifier