第42卷 第10期 西 安 交 通 大 学 学 报 Vol.42 №10
2008年10月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Oct. 2008

一种新的选择性支持向量机集成学习算法

唐耀华，高静怀，包乾宗
（西安交通大学电子与信息工程学院，710049，西安）

摘要：针对支持向量机(SVM)在应用于集成学习中会失效的问题，提出一种选择性SVM集成学习算法(SE-SVM)，利用$\xi\alpha$误差估计法估计个体SVM泛化性度量，并基于负相关学习理论引入差异性度量，通过递归删除法选择出一组泛化性能优良、相互间差异性大的SVM参与集成学习．基于UCI数据的仿真实验表明，SE-SVM能够平均提高SVM的分类正确率0.4%，比常规的Bagging集成学习方法和负相关集成学习方法的分类正确率分别提高了0.24%和0.16%．

关键词：泛化性度量；集成学习；负相关；支持向量机
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2008)10-1221-05

Novel Selective Support Vector Machine Ensemble Learning Algorithm

TANG Yaohua, GAO Jinghuai, Bao Qianzong
(School of Electronic and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: Focusing on the problem that conventional ensemble learning methods may be invalid when support vector machine (SVM) is used as component learner, a new selective SVM ensemble algorithm is proposed. $\xi\alpha$ estimator is used to estimate the generalization performance of the component SVM, and negative learning theory is used to introduce diversity among component SVMs. A set of component SVMs with high generalization performance and high diversity is selected during ensemble through recursive elimination algorithm. Experimental results on UCI data sets show that compared with single SVM, conventional Bagging ensemble method and negative learning ensemble method, the classification accuracy of selective SVM ensemble is increased on average by 0.4%, 0.24% and 0.16%, respectively.

Keywords: generalization measurement; ensemble learning; negative correlation; support vector machine

尽管集成学习能够有效提高神经网络、决策树等弱学习器的泛化性能，但是对于支持向量机(SVM)这一强学习器是否同样有效，众多学者的研究结果却不尽相同．如Kim等[1]认为简单的Bagging和Boosting算法就能提高分类精度．Valentini[2]基于偏移方差分析法验证了集成学习对SVM有效．相反，Yang[3]以SVM作为成员分类器分析了多种基于差异性最大化的集成学习方法，认为集成分类器分类间隔不会随差异性的增大而单调增大，因此成员分类器差异最大化不能保证集成学习性能提高．

本文在集成误差期望分析的基础上，证明传统集成学习方法是基于经验风险最小化原理，通过最小化集成误差获得的，容易出现过拟合，并提出一种基于泛化性度量与负相差理论相结合的选择性集成算法(SE-SVM)．采用$\xi\alpha$误差估计法评价个体SVM的泛化性能，并利用负相关学习理论选择出泛化性能优良、相互间差异性较大的的一组SVM参与集成．

收稿日期：2007-12-08． 作者简介：唐耀华(1977—)，男，博士生；高静怀(联系人)，男，教授，博士生导师． 基金项目：国家高技术研究发展计划资助项目(2006AA09A102-11)；国家自然科学基金重点资助项目(40730424)．

1 集成学习及其期望误差分析

假设学习任务是利用 N 个学习器通过加权平均方式进行集成学习，各个体学习器 f_i 的学习结果分别被赋予权重 w_i，$i=1,2,\cdots,N$，并满足 $w_i\geqslant 0$，且 $\sum_{i=1}^{N}w_i=1$. 对任一学习样本(x,y)，$x\in\mathbf{R}^d$，$y\in\mathbf{R}$，y 是 x 的期望输出，则集成输出为

$$\hat{f}(x)=\sum_{i=1}^{N}w_i f_i(x) \tag{1}$$

式中：$f_i(x)$是第 i 个个体学习器的输出. 根据集成误差模糊性分解[4]，有

$$e(x)=a(x)-v(x) \tag{2}$$

式中：$e(x)=(\hat{f}(x)-y)^2$ 为集成学习平方误差；$a(x)=\sum_{i=1}^{N}w_i(f_i(x)-y)^2$ 为个体学习器的加权平方误差；$v(x)=\sum_{i=1}^{N}w_i(f_i(x)-\hat{f}(x))^2$ 为个体学习器的差异性度量.

采用负相关学习理论(NC-learning)，通过在个体学习器的加权平方误差中引入一个相关性惩罚函数，可以控制所选个体学习器之间的差异性. 基于负相关学习的集成误差函数定义为

$$e_{NC}(x)=\sum_{i=1}^{N}w_i(f_i(x)-y)^2+\lambda\sum_{i=1}^{N}w_i p_i(x) \tag{3}$$

式中：右边第 2 项为相关性惩罚项；$p_i(x)$为第 i 个个体学习器与其他个体学习器之间的相关性度量

$$p_i(x)=\sum_{j=1,i\neq j}^{N}(f_i(x)-\hat{f}(x))(f_j(x)-\hat{f}(x))=-(f_i(x)-\hat{f}(x))^2 \tag{4}$$

λ 为罚因子，用于控制个体学习器间的相关性度量在集成学习中的作用强度. 将式(4)代入式(3)，得

$$e_{NC}(x)=\sum_{i=1}^{N}w_i(f_i(x)-y)^2-\lambda\sum_{i=1}^{N}w_i(f_i(x)-\hat{f}(x))^2=a(x)-\lambda v(x) \tag{5}$$

显然，当 $\lambda=1$ 时，式(2)与式(5)相同，基于负相关的集成学习方法与集成误差分解是一致的.

假设学习样本 x 按照分布 $p(x)$随机抽取，则 e、a、v 在分布 $p(x)$上对应的期望误差分别为

$$E(e)=\int p(x)e(x)\mathrm{d}x \tag{6}$$

$$E(a)=\int p(x)a(x)\mathrm{d}x \tag{7}$$

$$E(v)=\int p(x)v(x)\mathrm{d}x \tag{8}$$

因此，集成学习的泛化误差为

$$E(e)=E(a)-\lambda E(v) \tag{9}$$

可以得出如下 2 个结论：

(1)由于 $v\geqslant 0$，$E(v)\geqslant 0$，则 $E(e)\leqslant E(a)$，因此集成学习的泛化误差必然小于单个学习器的平均泛化误差，即集成学习性能优于单个学习器学习性能；

(2)个体学习器自身学习性能越好，个体学习器间差异性越大，则集成误差越小，集成效果越好.

因此，为了提高集成学习系统的泛化性能，构建泛化能力强、相互差异大的个体学习器组是集成学习算法实现的目标. 基于集成误差模糊性分解和负相关学习理论的集成学习方法，均通过最小化集成误差 $\min E(e)$选择个体学习器. 然而，在实际应用中，学习样本数量有限，样本概率分布函数 $p(x)$也未知，$\min E(e)$是通过利用训练样本，基于经验风险最小化得到的. 假设 D 为含有 m 个样本的训练数据集，$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，则 e、a、v 在数据集 D 上对应的期望误差分别为

$$E_D(e)=\frac{1}{m}\sum_{j=1}^{m}e_j=\frac{1}{m}\sum_{j=1}^{m}(\hat{f}(x_j)-y_j)^2 \tag{10}$$

$$E_D(a)=\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{N}w_i(f_i(x_j)-y_j)^2 \tag{11}$$

$$E_D(v)=\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{N}w_i(f_i(x_j)-\hat{f}(x_j))^2 \tag{12}$$

则式(9)修正为

$$E_D(e)=E_D(a)-\lambda E_D(v) \tag{13}$$

最小化 $\min E_D(e)$是建立在数据集 D 下基于经验风险最小化原理实现的，容易出现过拟合，影响学习器泛化能力的提高. 这就是集成学习对于神经网络、决策树等弱学习器性能提升明显，但对于 SVM 这种基于结构风险最小化原理的强学习器性能提升效果不明显，甚至有时会失效的原因. 基于上述分析，我们认为对于 SVM 而言，在集成过程中避免使用经验风险最小化原则，直接选择泛化性能优良、相互间差异性较大的个体学习器组将有助于集成学习泛化性能的提高.

2 选择性 SVM 集成学习

给出如下判定函数

$$J=G+\lambda P \tag{14}$$

式中：G 为 SVM 的泛化性度量；P 为 SVM 的差异性度量. 对于第 i 个 SVM，J_i 越大，则该 SVM 的泛化性能就越好，与其他选定的 SVM 差异性就越大，只有那些具有较大 J 值的 SVM 被保留下来参与集成，因此称为选择性支持向量机集成方法.

应用交叉验证法度量 SVM 的泛化性能固然效果较好，但是由于需要对个体 SVM 逐一进行度量，SVM 的训练过程还涉及参数选择，计算量较大. Joachims[5] 针对 SVM 提出了一种有效且计算效率较高的泛化性能估计方法——$\xi\alpha$ 误差估计法，其核心思想是：如果某一样本作为检验样本在应用留一法进行误差估计时验证错误，则该样本满足不等式 $2\alpha_i\Delta^2+\xi_i\geqslant 1$，其中 α、ξ 为利用全部训练样本训练 SVM 所得的解. 满足上述不等式的样本个数 d 即为留一法误差估计错误的上界，该上界可以用来估计 SVM 的泛化性能. $\xi\alpha$ 误差估计法的计算形式为

$$e_{\xi\alpha}=\frac{d}{n}$$

$$d=|\{i:(2\alpha_i\Delta^2+\xi_i)\geqslant 1\}| \tag{15}$$

式中：$\Delta^2=\sup(K(x,x)-K(x,x'))$，其中 $K(\cdot,\cdot)$ 为 SVM 使用的核函数；n 为训练样本数目. 与交叉验证法不同的是，该算法仅需要训练一次 SVM 即可获得估计结果，计算效率高.

为了产生一组具有较大差异的个体 SVM，我们采用 Bagging 方式生成多个训练样本子集，并训练得到一组个体 SVM. 由于采用重采样技术，D_i 中仅含有约 63.2% 的原始训练样本，使得各个体 SVM 之间存在差异. 同时，由于每次训练个体 SVM 时仅使用了部分原始训练样本，因此可以利用训练集 D 来度量个体 SVM 间的差异性.

从现有个体 SVM 组中优选最优学习器组是一个 NP-hard 问题. 本文采用递归删除法，每次删去泛化性能相对较低、与其他个体 SVM 差异性较小的 SVM. 尽管最终结果可能不是最优的 SVM 组合，但该方法计算量小，可操作性强. 为简化讨论，假设所有的个体 SVM 具有相同的权重，各个体 SVM 采用简单平均法集成. 本文 SE-SVM 算法的基本步骤如下.

(1) 训练集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_M,y_M)\}$，初始个体 SVM 规模为 N，SVM 集成规模为 S，罚因子为 λ.

(2) 输入数据预处理，将其归一化为 0 均值、方差为 1 的标准数据.

(3) 基于 Bagging 算法生成 N 个训练集 D_i，训练后生成 N 个支持向量机 S_i，$i=1,2,\cdots,N$，以 f_i 表示 S_i 的预测函数. 初始个体 SVM 集表示为 $F=\{f_1,f_2,\cdots,f_N\}$. 令候选 SVM 集 B 为初始个体 SVM 集，即 $B=F$.

(4) 重复步骤(3)的内容，直至 $|B|=S$.

(4-1) 计算各个体 SVM 的 $\xi\alpha$ 误差估计 $G_i=e_{\xi\alpha}$. 将 G_i 作为 S_i 的泛化性能度量. G_i 越小，S_i 泛化性能越好.

(4-2) 基于负相关学习理论，利用训练集 D 计算各 S_i 对应的相关性惩罚函数 P_i. P_i 越小，表明 S_i 与 B 中其他个体 SVM 的差异性越大

$$P_i=\frac{1}{M}\sum_{j=1}^{M}p_i(x_j)=$$

$$-\frac{1}{M}\sum_{j=1}^{M}(f_i(x_j)-\hat{f}(x_j))^2 \tag{16}$$

(4-3) 计算 S_i 的判定函数值

$$J_i=G_i+\lambda P_i$$

(4-4) 找出对应于最大 J 值的个体 SVM

$$k=\arg\min_i(J_i)$$

(4-5) 删除 S_k，更新候选 SVM 集 B

$$B=B-f_k$$

(5) 算法输出为 SVM 集 B

(6) 集成 SVM 对待预测样本 x 的预测输出为

$$\hat{f}(x)=\frac{1}{|S|}\sum_{f_i\in B}f_i(x) \tag{17}$$

3　仿真实验

采用 UCI 机器学习数据库[6] 中的 5 个数据集进行实验，并将本文算法 SE-SVM 与 Single-SVM、Bagging-SVM 以及文献[7]中所述基于负相关理论的集成学习方法(NC-SVM)进行比较. Single-SVM 利用全部样本训练生成的单个 SVM 进行分类学习. Bagging-SVM 首先利用 Bagging 算法生成多个训练样本子集，获得一组 SVM，然后随机选择一定数量的个体 SVM 进行集成. 本文实验中采用高斯核函数作为 SVM 的核函数，其核参数 σ 和正则化因子 C 的设置根据训练样本按照最优参数搜索算法得到. 在集成过程中，罚因子 λ 从 $\{0,0.1,0.3,0.6,0.8,1,1.5,2\}$ 中分别取值进行实验，其中 $\lambda=0$ 表示忽略 SVM 差异性度量，即个体 SVM 的选择与否完全取决于个体 SVM 自身的泛化性能. 随着 λ 的增大，逐渐增大差异性度量在选择个体 SVM 决策中的权重.

对于每个数据集，由于两类样本数目不平衡，采用等比例随机提取50%的样本作为训练样本，剩余样本作为检验样本进行实验，初始个体SVM的规模$N=100$，将独立运算50次实验取得的检验正确率的平均值作为算法评价指标. 图1给出了针对5个数据集进行实验获得的集成学习系统检验正确率(P)与参与集成学习的个体SVM数目(S)之间的关系. Single-SVM对应的水平线仅表示Single-SVM的检验正确率，与集成SVM的数目无关. 为便于分析，表1列出了3种集成学习方法的最优集成SVM数目及其对应的平均检验正确率.

结合图1与表1可得出以下实验结论：

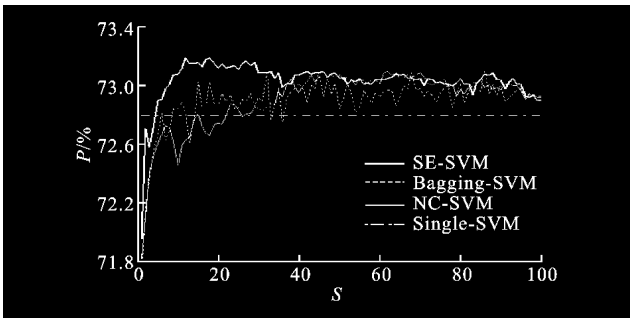
(1)随着参与集成的SVM的数目减少，SE-SVM的学习性能逐渐提高，表明选择部分SVM进行集成获得的学习性能优于集成全部SVM的学习性能，此结论与Zhou等[8]提出的基于人工神经网络的选择性集成学习理论是一致的；

(2)在5个数据集中，SE-SVM取得最优性能时，参与集成的SVM的数目均位于区间[10,30]，这在实际应用中是十分有用的，它为确定集成SVM的数目提供了参考，而Bagging-SVM和NC-SVM的最优集成数目区间难以确定；

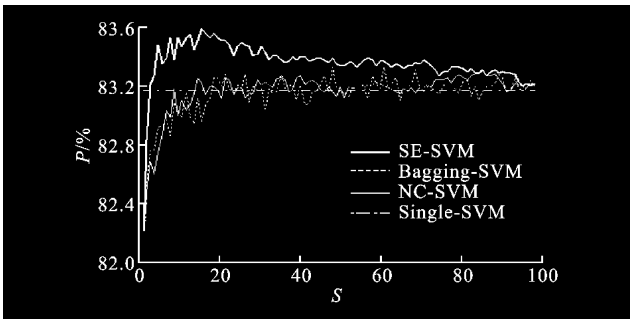
(3)集成学习的根本目的是提高学习系统的泛化性能，与不使用集成学习方法的Single-SVM相比较，SE-SVM能够平均提高SVM的分类正确率0.4%，而Bagging-SVM和NC-SVM只能分别平均提高0.16%和0.24%，因此SE-SVM能够有效地提高集成学习系统的泛化性能，其性能优于常规的Bagging-SVM和NC-SVM集成学习算法.

4 结 论

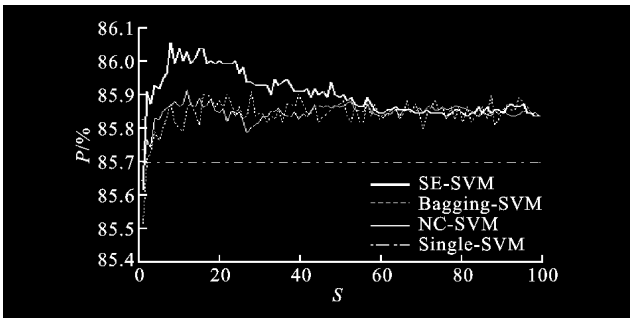
本文基于集成学习的期望误差分析，证明基于集成误差模糊性分解、负相关学习理论的集成学习方法在实际应用中是建立在经验风险最小化基础上，通过最小化训练样本集的集成误差获得的. 这种集成方法容易出现过拟合，对于以结构风险最小化为基础的SVM强学习器，此类集成方法不一定能够提升学习系统的泛化性能. 本文提出一种基于SVM泛化性能度量与负相关学习相结合的选择性集成学习算法SE-SVM，利用$\xi\alpha$误差估计法选择具有较强泛化性能的个体SVM，同时基于负相关学习理论保留具有较大差异的个体SVM组. 实验结果表明，本文算法优于常用的Bagging-SVM和NC-



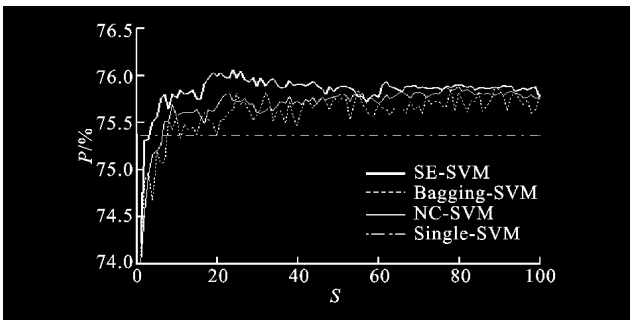
(a)数据集Breast cancer($\lambda=0.1$)



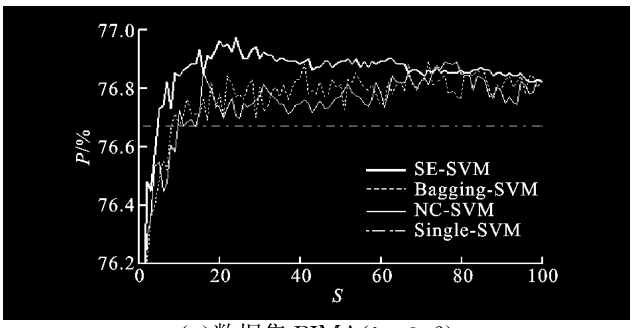
(b)数据集Heart disease($\lambda=0.1$)



(c)数据集Australia credit data($\lambda=0.6$)



(d)数据集German number($\lambda=0.6$)



(e)数据集PIMA($\lambda=0.6$)

图1 集成SVM数目与检验正确率关系的实验结果

表 1　集成学习算法实验结果比较

数据集	最优集成 SVM 数目			P/%			
	Bagging-SVM	NC-SVM	SE-SVM	Single-SVM	Bagging-SVM	NC-SVM	SE-SVM
Breast cancer	40～50	60～70	10～20	72.81±2.43	72.98±2.56	73.09±2.44	73.22±2.44
Heart disease	20～90	80～90	10～20	83.17±2.60	83.17±2.45	83.27±2.32	83.59±2.41
Australia credit data	20～25	10～20	10～15	85.66±1.56	85.85±1.51	85.85±1.57	86.03±1.53
PIMA	60～80	75～85	20～30	76.67±2.05	76.81±2.23	76.85±2.12	76.96±2.11
German number	50～60	70～90	20～30	75.37±1.06	75.68±1.12	75.83±1.15	75.98±1.23
				78.74±1.94	**78.90±1.97**	**78.98±1.92**	**79.16±1.94**

注：黑体数据为平均检验正确率.

SVM 集成学习算法，能够平均提高 SVM 的分类正确率 0.4%，有效提升了学习系统的泛化性能.

对于泛化性和差异性在集成过程中哪个更为重要仍是一个未知的问题，因此集成学习系统泛化性能与个体学习器泛化性能和差异性之间的关系值得进一步研究. 同时，将结构风险最小化理论引入 SVM 集成也是推进 SE-SVM 集成学习算法的另一项研究内容.

参考文献：

[1] KIM H C, PANG S, JE H M, et al. Constructing support vector machine ensemble [J]. Pattern Recognition, 2003, 36(12): 2757-2767.

[2] VALENTINI G. An experimental bias-variance analysis of SVM ensembles based on resampling techniques [J]. IEEE Transaction on Systems, Man and Cybernetics, 2005, 35(6): 1252-1271.

[3] TANG E K, SUGANTHAN P N, YAO X. An analysis of diversity measures [J]. Machine Learning, 2006, 65(1): 247-271.

[4] KROGH A, VEDELSBY J. Neural network ensembles, cross-validation, and active learning [C] // TOURETZKY D, LEEN T. Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 1995: 231-238.

[5] JOACHIMS T. Estimating the generalization performance of a SVM efficiently [C] // Proceedings of ICML 2000. San Francisco, CA, USA: Morgan Kaufniann Publishers Inc., 2000: 431-438.

[6] BLAKE C L, MERZ C J. UCI repository of machine learning databases [DB/OL]. (1998-01-01) [2007-10-01]. http://www.ics.uci.edu/mlearn/MLRepository.html.

[7] ZANDA M, BROWN G, FUMERA G, et al. Ensemble learning in linearly combined classifiers via negative correlation [C] // Proceedings of MCS2007. Berlin, Germany: Springer-Verlag, 2007: 440-449.

[8] ZHOU Zhihua, WU Jianxin, TANG Wei. Ensembling neural networks: many could be better than all [J]. Artificial Intelligence, 2002, 137(1/2): 239-263.

（编辑　刘杨　苗凌）