基于约束投影的支持向量机选择性集成

王 磊

(西南财经大学经济信息工程学院 成都 610074) (西南财经大学中国支付体系研究中心 成都 610074)

摘 要 提出两种基于约束投影的支持向量机选择性集成算法。首先利用随机选取的 must-link 和 cannot-link 成对约束集确定投影矩阵,将原始训练样本投影到不同的低维空间训练一组基分类器;然后,分别采用遗传优化和最小化偏离度误差两种选择性集成技术对基分类器进行组合。基于 UCI 数据的实验表明,提出的两种集成算法均能有效提高支持向量机的泛化性能,显著优于 Bagging,Boosting,特征 Bagging 及 LoBag 等集成算法。

关键词 约束投影,选择性集成,支持向量机,分类器

中图法分类号 TP391.4 **文献标识码** A

Constraint Projection-based Support Vector Machines Selective Ensemble Algorithms

WANG Lei

(School of Economics Information Engineering, Southwest University of Finance & Economics, Chengdu 610074, China)

(Research Center of China Payment System, Southwest University of Finance & Economics, Chengdu 610074, China)

Abstract This paper proposed two constraint project-based selective ensemble algorithms of support vector machines. Firstly, projective matrices were determined upon randomly selected must-link and cannot-link constraint sets, with which original training samples were transformed into different representation spaces to learn a group of base classifiers. Then, two selective ensemble techniques of genetic optimization and minimizing deviation errors were utilized to combine base classifiers. Experiments on UCI datasets show that both proposed algorithms improve the generalization performance of support vector machines significantly, which are much better than classical ensemble algorithms, such as Bagging, Boosting, feature Bagging and LoBag.

Keywords Constraint project, Selective ensemble, Support vector machines, Classifiers

集成学习(ensemble learning)技术利用多个基学习器对同一个问题进行学习,可以显著地提高学习系统的泛化能力[1]。最近10年来,集成学习在国内外受到广泛的重视和关注,并被认为是当前机器学习领域的4大研究方向之首[2]。Krogh 等人的研究指出[3],集成学习器的泛化误差等于基学习器的平均泛化误差和平均差异度之差。因此,要增强集成学习器的泛化能力,一方面应尽可能提高单个基学习器的泛化能力,另一方面应尽可能地增大基学习器之间的差异度。大多数经典的集成学习算法都是通过扰动数据集达到增大基学习器之间的差异性的目的,从而获得更优的泛化性能,例如,Bagging,Boosting 和 Arcing 等[4]。

支持向量机(support vector machines,SVM)是一种具有收敛到全局最优、维数不敏感、泛化能力强等优点的新型机器学习方法[5]。直觉上,若采用 SVM 训练一组具有显著差异度的基学习器,则根据文献[3]的解释,将能够获得泛化能力突出的 SVM 集成学习器。然而相比于神经网络、决策树等方法,SVM 是一种稳定的学习技术,直接应用 Bagging 和 Boosting 等算法不能增大基学习器之间的差异度[6],因而它们对于提高 SVM 的泛化性能作用十分有限。针对该问题,采用特殊扰动机制的集成学习算法近几年来受到广泛重视,例如,特征 Bagging 和 LoBag 分别通过扰动特征空间和 SVM 的模型参数来增大差异度[7,8],取得了良好的效果。

本文提出两种新颖的基于约束投影的支持向量机选择性集成算法。首先利用随机选取的成对约束集将训练样本投影到不同的低维空间;然后在投影空间中训练一组 SVM 基分类器;最后,分别通过遗传优化和最小化偏离度误差的策略对基分类器进行选择性集成。基于 UCI 数据集的实验结果表明,本文算法取得的泛化性能十分突出。

1 约束投影

约束投影是处理半监督聚类分析任务时最新出现的技术之一[9]。假设 d 维训练样本集为 $S=\{x_1,x_2,\cdots,x_n \mid \forall x_i \in R^d\}$,它们的类别标记为 $t_i \in \{c_1,c_2,\cdots,c_k\}$,$i=1,2,\cdots,n$。并且令 $\mathcal{M}$ 和 $\mathcal{C}$ 分别表示样本之间的 must-link 和 cannot-link 成对约束关系集,也即

$$\mathcal{M}=\{(x_i,x_j) \mid t_i=t_j, \forall i,j\} \quad (1)$$

$$\mathcal{C}=\{(x_i,x_j) \mid t_i \neq t_j, \forall i,j\} \quad (2)$$

约束投影的实质是寻找一个 $d\times p$ 阶(满足 $d>p$)的投影矩阵 $W=(w_1,w_2,\cdots,w_p)$ 且 $W^{\mathrm{T}}W=I$,将训练集 S 投影到低维空间 R^p,也即 $z_i=W^{\mathrm{T}}x_i$;并且,最大程度上保持原始样本

到稿日期:2008-11-06 返修日期:2009-02-09 本文受国家自然科学基金青年项目(69732010),西南财经大学科学研究基金(QN0806)资助。

王 磊(1978—),男,博士,讲师,主要研究方向为机器学习、模式识别等,E-mail:wanglei_t@swufe.edu.cn。

之间的成对约束关系。也就是说，在投影空间 R^p 中，集合 $\mathcal{M}$ 中的样本对之间的距离最小化，而集合 $\mathcal{C}$ 中的样本对之间的距离最大化。因此，可以通过最大化目标函数 $J(W)$ 来求解投影矩阵 W：

$$J(W)=\frac{1}{2|\mathcal{C}|}\sum_{(x_i,x_j)\in\mathcal{C}}\|W^{\mathrm{T}}x_i-W^{\mathrm{T}}x_j\|^2-\frac{\rho}{2|\mathcal{M}|}\sum_{(x_i,x_j)\in\mathcal{M}}\|W^{\mathrm{T}}x_i-W^{\mathrm{T}}x_j\|^2 \tag{3}$$

其中，$|\mathcal{C}|$ 和 $|\mathcal{M}|$ 分别表示约束集合 $\mathcal{C}$ 和 $\mathcal{M}$ 中的样本对数；而系数 ρ 用于平衡式(3)中的两项对于目标函数值的贡献，可以按式(4)进行取值。

$$\rho=\frac{\frac{1}{|\mathcal{C}|}\sum_{(x_i,x_j)\in\mathcal{C}}\|x_i-x_j\|^2}{\frac{1}{|\mathcal{M}|}\sum_{(x_i,x_j)\in\mathcal{M}}\|x_i-x_j\|^2} \tag{4}$$

分别定义成对约束集 $\mathcal{C}$ 和 $\mathcal{M}$ 的离散度矩阵：

$$D_{\mathcal{C}}=\frac{1}{2|\mathcal{C}|}\sum_{(x_i,x_j)\in C}(x_i-x_j)(x_i-x_j)^{\mathrm{T}} \tag{5}$$

$$D_{\mathcal{M}}=\frac{1}{2\mathcal{M}}\sum_{(x_i,x_j)\in\mathcal{M}}(x_i-x_j)(x_i-x_j)^{\mathrm{T}} \tag{6}$$

则通过简单的线性变换，可以得到优化式(3)的等价问题：

$$J(W)=trace(W^{\mathrm{T}}(D_{\mathcal{C}}-\rho D_{\mathcal{M}})W) \tag{7}$$

显然，式(7)是一个典型的特征值问题。求解 W 等价于求解 $(D_{\mathcal{C}}-\rho D_{\mathcal{M}})$ 的最大的 p 个非负特征值对应的特征向量。不失一般性，令 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$，则 W 取最优值时，满足等式关系：

$$trace(W^{\mathrm{T}}(D_{\mathcal{C}}-\rho D_{\mathcal{M}})W)=\sum_{i=1}^{p}\lambda_i \tag{8}$$

求解出 $W=(w_1,w_2,\cdots,w_p)$ 后，则可以将原始训练集 S 投影到低维空间 R^p，然后在 R^p 中训练 SVM 基学习器。

分析可知，基学习器的性能由投影空间 R^p 中的训练集 $S_W=\{z_1,z_2,\cdots,z_n\,|\,z_i=W^{\mathrm{T}}x_i,\forall x_i\in S\}$ 确定，而 R^p 和投影矩阵 W 完全由成对约束集 $\mathcal{C}$ 和 $\mathcal{M}$ 唯一确定。所以，利用不同约束集训练的 SVM 基学习器之间具有明显的差异度。并且，对于原始训练集 S，样本之间存在 $n(n-1)$ 个成对约束关系，从而最多可以构造出 $2^{n(n-1)}$ 对不同的成对约束集 $\mathcal{C}$ 和 $\mathcal{M}$。

因此，能够很容易地从原始训练集 S 中随机生成 L 对不同的约束集 $\mathcal{C}$ 和 $\mathcal{M}$（假设它们的规模为 $n_{\mathcal{C}}$ 和 $n_{\mathcal{M}}$，且均远小于 $n(n-1)$），并在由它们确定的投影空间中训练 L 个具有明显差异度的 SVM 基学习器。

2 选择性集成

对于 k 分类问题，假设 $\mathcal{F}^1,\mathcal{F}^2,\cdots,\mathcal{F}^L$ 是在投影空间中训练的 L 个 SVM 基分类器，$\beta^1,\beta^2,\cdots,\beta^L\in R$ 是它们的权重因子，满足 $\sum_{i=1}^{L}\beta^i=1$。并令基分类器 $\mathcal{F}^i$ 对于任意样本 x 的判别输出为 k 维向量 y^i，它具有如下的形式：

$$\mathcal{F}^i(x)=y^i=(0,0,1,0,\cdots,0,0)^{\mathrm{T}} \tag{9}$$

其中，y^i 只有第 j 维元素为 1（剩余元素全部为 0），表示基分类器判别样本 x 的类别为第 j 类。

令 $\hat{\mathcal{F}}$ 表示集成分类器，Bagging 和 Boosting 等集成算法通常采用投票（即 $\forall\beta^i=1/L$）或者加权投票的方法将全部 L 个基学习器组合起来获得 $\hat{\mathcal{F}}$，它按如下决策函数对样本 x 所属的类进行判别，也即：

$$\hat{\mathcal{F}}(x)=\arg\max\sum_{i=1}^{L}\beta^i\mathcal{F}^i(x) \tag{10}$$

假设集成分类器 $\hat{\mathcal{F}}$ 对样本 x 进行判别时的误差函数为 $\hat{\epsilon}(x)=\begin{cases}1,\hat{\mathcal{F}}\neq t\\0,\hat{\mathcal{F}}(x)=t\end{cases}$，其中，$t$ 为样本 x 的目标类别。则在样本集 S 上的泛化误差为：

$$\hat{error}=\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}(x_i) \tag{11}$$

最近，Zhou 等人通过严格的理论分析指出，由于随机生成的基学习器之间容易存在相似或者强相关，可以剔除若干基学习器（权重因子 $\beta^i=0$），仅选择性地使用一部分基学习器进行集成（权重因子 $\beta^i\neq0$），并且能够获得泛化性能更优的集成学习器。这种集成策略称为“选择性集成”[10,11]。

本节在前文的工作基础上，提出两种具体的选择性集成算法，实现 SVM 集成分类器的有效训练。

2.1 cp. GASEN 集成算法

文献[10]采用遗传优化技术实现选择性集成，具体措施是：采用遗传优化算法求解出最优的权重因子向量 β^*，然后剔除权重因子小于预设阈值 τ 的基分类器（β^i 重置为 0），而仅将权重因子非 0 的基分类器按式(10)进行集成。

这里，直接将文献[10]给出的选择性集成策略扩展到 SVM 的集成训练过程，得到一种“基于约束投影和遗传优化的选择性集成算法”，简称 cp. GASEN。主要步骤如算法 1。

算法 1 cp. GASEN 算法

输入：训练样本集 S，约束集规模 $n_{\mathcal{C}}$ 和 $n_{\mathcal{M}}$，预设阈值 τ；

Step1：For $i=1,2,\cdots,L$

a) 初始化成对约束集 $\mathcal{C}=\emptyset$ 和 $\mathcal{M}=\emptyset$；

b) 从训练集 S 随机选取样本对 (x_u,x_v)，且 $x_u\neq x_v$；

c) 如果 $t_u=t_v$ 且 $|\mathcal{M}|<n_{\mathcal{M}}$，则将约束对 (x_u,x_v) 添加到 $\mathcal{M}$；如果 $t_u\neq t_v$ 且 $|\mathcal{C}|<n_{\mathcal{C}}$，则将约束对 (x_u,x_v) 添加到 $\mathcal{C}$；

d) 重复步骤 b)和 c)直到满足 $|\mathcal{M}|=n_{\mathcal{M}}$ 和 $|\mathcal{C}|=n_{\mathcal{C}}$；

e) 利用式(7)求解投影矩阵 W，并得到投影空间中的新训练集 $S_W=\{W^{\mathrm{T}}x_u\}_{u=1}^{n}$；

f) 利用 S_W 训练 SVM 基分类器 $\mathcal{F}^i$；

Step2：随机生成规模为 n_{pop} 的权重因子向量 β 的种群，并采用 $h(\beta)=\frac{1}{\hat{error}}$ 作为适应度函数（$\hat{error}$ 由式(11)定义）训练遗传算法，得到最优的权重因子向量 β^*；

Step3：对于任意 $\beta_i^*<\tau$，重置 $\beta_i^*=0$，然后按式(10)得到 SVM 集成分类器 $\hat{\mathcal{F}}$ 的决策函数。

2.2 cp. EDSEN 集成算法

与神经网络学习过程类似，可以采用最小化集成分类器 $\hat{\mathcal{F}}$ 的期望均方误差（MSE）求解最优的权重向量 β^*，即

$$\min_{\beta}E[(\hat{\epsilon}(x))^2]\quad s.t.\quad\sum_{i=1}^{L}\beta^i=1,\ \beta^i\geqslant0 \tag{12}$$

注意到，误差 $\hat{\epsilon}(x)$ 只取 0，1 两个值，并没有区分产生判别误差时 $\hat{\epsilon}$ 的输出值与目标值的偏离程度。因此，重新为集成学习器定义一种“基于偏离度”的误差函数：

$$\hat{\epsilon}_d(x)=\sum_{i=1}^{L}\beta^i\epsilon^i(x) \tag{13}$$

其中，$\epsilon_i(x)=\begin{cases}1,\mathcal{F}^i(x)\neq t\\0,\mathcal{F}^i(x)=t\end{cases}$ 是基分类器 $\mathcal{F}^i$ 的判别误差函数。

分析可知，$\hat{\varepsilon}_d(x)$的值越大，则集成分类器$\hat{\mathcal{F}}$的输出结果越偏离目标值；相反，$\hat{\varepsilon}_d(x)$的值越小，$\hat{\mathcal{F}}$的输出结果越靠近目标值。特别地，若$\hat{\varepsilon}_d(x)=0$，$\hat{\mathcal{F}}$以概率1判别错误；若$\hat{\varepsilon}_d(x)=1$，$\hat{\mathcal{F}}$以概率1判别正确。因此，$\hat{\varepsilon}_d(x)$可以看作$\hat{\varepsilon}(x)$的一般形式。

利用$\hat{\varepsilon}_d(x)$将优化式(12)的目标函数重新写为：

$$E[(\sum_{i=1}^{L}\beta^i\epsilon^i(x))^2]=E[(\sum_{i=1}^{L}\beta^i\epsilon^i(x))(\sum_{j=1}^{L}\beta^j\epsilon^j(x))]=\beta^{\mathrm{T}}Q\beta \tag{14}$$

其中，Q是误差协方差矩阵，有$Q_{i,j}=E[\epsilon^i(x)\cdot\epsilon^j(x)]$。为了便于计算，$Q_{i,j}$可以在单独的验证集上由$\tilde{Q}_{i,j}$估计：

$$\tilde{Q}_{i,j}=\frac{1}{m+1}\sum_{u=1}^{m}\epsilon^i(x_u)\epsilon^j(x_u)$$

验证集可采用boostrap技术从原始训练集S中重复抽样获得，规模为m。因此，将优化问题式(12)改写成：

$$\min_{\beta}\beta^{\mathrm{T}}\tilde{Q}\beta+\delta\|\beta\|^2 \quad s.t.\quad \sum_{i=1}^{L}\beta^i=1,\beta^i\geqslant 0 \tag{15}$$

其中，为了避免过拟合现象的产生，在目标函数中引入正则项$\delta\|\beta\|^2$和正则参数$\delta>0$。

显然，优化式(15)是典型的线性约束二次规划问题。由于正则项的存在，目标函数半正定。根据优化理论，优化问题式(15)的解β^*具有稀疏性的特点，也即，大部分的$\beta_i^*=0$。

因此，可以利用具有稀疏结构特点的β^*对SVM基分类器进行选择性集成。将该方法称为“基于约束投影和最小化偏离度误差的选择性集成算法”，简称cp. EDSEN。主要步骤如算法2所示。

算法2 cp. EDSEN算法

输入：训练样本集S，约束集规模$n_{\mathcal{C}}$和$n_{\mathcal{M}}$，正则参数δ；

Step1：同算法1的Step1；

Step2：按优化式(15)求解最优权重因子向量β^*；

Step3：按式(10)得到SVM集成分类器$\hat{\mathcal{F}}$的决策函数。

3 实验与分析

为了验证本文提出的选择性集成算法的性能，在10个UCI数据集上分别采用cp. GASEN，cp. EDSEN以及Bagging，Boosting，特征Bagging和LoBag训练SVM集成分类器，并比较它们取得的分类器精度。其中，ORHD和PRHD分别表示数据集“optical recognition of handwritten digits”和“pen-Based recognition of handwritten digits”。

实验中，集成规模统一取为$L=50$，成对约束集的规模取值与原数据集相同，即$n_{\mathcal{C}}=n_{\mathcal{M}}=n$；cp. GASEN的阈值$\tau$取为0.01，cp. EDSEN的正则参数取为0.05，特征Bagging每次选用约50%的特征训练基分类器。此外，SVM基分类器均选用RBF核函数，模型参数采用5-fold交叉验证方法选取。

表1统计了上述集成算法在UCI数据集上10次重复实验的平均测试精度(F-Bagging表示“特征Bagging算法”)。表2统计了显著水平为0.05时的双边t检验结果(“W,T,L”分别表示“优于”、“无显著差别”和“差于”的次数)。

表1 集成算法在UCI数据集上获得的测试精度比较(×100%)

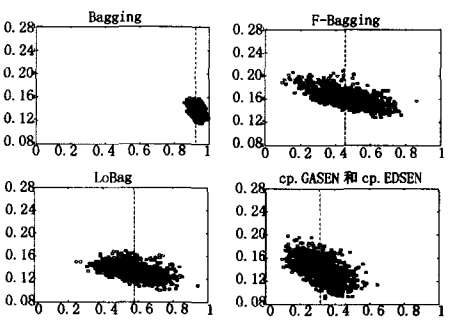
数据集	cp. GASEN	cp. EDSEN	Bagging	Boosting	F-Bagging	LoBag
segment	98.24	98.15	96.46	97.02	97.20	97.61
credit	75.36	75.54	73.33	73.19	73.72	74.70
ionosphere	94.55	94.71	91.47	92.40	92.89	93.84
sonar	91.71	91.42	89.38	89.64	91.26	90.72
heart	86.53	86.78	83.80	84.02	84.45	85.10
vehicle	89.27	89.89	87.65	87.43	88.60	88.73
vowel	96.30	96.35	94.22	95.01	95.37	95.26
waveform	90.42	90.87	88.74	88.95	89.79	89.57
ORHD	99.24	99.32	98.40	98.57	98.82	98.75
PRHD	99.40	99.38	98.51	98.55	98.60	99.00

表1和表2的实验结果表明：在所有数据集上，cp. GASEN和cp. EDSEN算法取得的测试精度均高于剩余4种集成算法。并且，至少在8个数据集上，它们均显著优于Bagging和Boosting算法。当两种算法之间比较时，cp. EDSEN算法取得的精度稍优于cp. GASEN(仅在vehicle上优于后者)。

表2 显著水平为0.05时双边t检验结果(W/T/L)

	cp. GASEN	cp. EDSEN	Bagging	Boosting	F-Bagging	LoBag
cp. GASEN	—	0/9/1	9/1/0	8/2/0	5/5/0	3/7/0
cp. EDSEN	1/9/0	—	9/1/0	9/1/0	7/3/0	4/6/0

为进一步解释本文提出的两种算法在集成SVM分类器时能够取得突出泛化性能的原因，采用kappa统计量来度量不同的集成算法产生的基分类器之间的差异程度[12]，并绘制图1所示的“kappa-误差图”。限于篇幅，图1只给出了Bagging，特征Bagging，LoBag和本文算法在waveform数据集上的实验结果。其中，每一点都对应着任意一对基分类器，且x轴表示它们的kappa值(kappa值越大，它们之间的差异性越小，反之亦然)，y轴表示它们的平均测试误差。



(虚线表示kappa值的平均值)

图1 waveform数据集上的kappa-误差图

显然，cp. GASEN和cp. EDSEN采用约束投影技术产生的基分类器之间具有非常明显的差异度(平均kappa值为0.314)；并且，由于样本间成对约束的存在，使得在投影空间中训练的SVM基分类器具有较小的测试误差。因此，它们能够取得突出的泛化性能。与之比较，Bagging显然不能获得明显的差异度(平均kappa值为0.922)，原因是SVM是一种稳定的机器学习方法[6]。特征Bagging通过扰动特征空间带来的差异程度最接近本文算法，但由于它在训练基分类器时丢弃了样本的部分特征信息，导致单个基分类器的测试误差较高，因而也不能十分有效地提升SVM的泛化性能，如表1所列。

此外，表3给出了cp. GASEN和cp. EDSEN算法在进行选择性集成时，实际使用的基分类器数目占所有基分类器数目的平均百分比。从表3可以看出，cp. GASEN算法使用了51%～68%的基分类器(阈值为0.01时)，cp. EDSEN算法使

(下转第243页)

法的属性个数、规则个数、规则平均长度和规则预测准确度进行了比较，对照实验结果如表1所列。

表1 实验结果比较

属性个数		规则个数		规则平均长度		规则预测准确度	
原属性数	约简属性数	C4.5算法	有判定的极小极大规则学习	C4.5算法	有判定的极小极大规则学习	C4.5算法	有判定的极小极大规则学习
24	15	57	35	6.9	5.4	89.50	93.57
21	17	26	20	4.3	3.6	79.03	80.70
56	39	9	5	1.8	1.7	76.52	81.36
18	13	10	8	4.2	3.7	91.08	94.20

从实验结果来看，在本文中图像情感规则的简化学习算法上所做的创新是成功的。

结束语 本文简述了图像情感语义研究的粗糙集理论及极小极大规则，描述了基于粗糙集理论的属性约简及极小极大规则学习的算法，然后给出了图像情感语义规则抽取及简化流程，最后将规则应用于推理机，配合图像特征与情感语义本体实现图像基于情感语义的检索。其中极小极大规则学习在图像情感语义简化中的应用，在保证规则准确度的基础上，简化了单个规则并减少了规则总数量，为图像情感检索奠定了良好的规则基础。

本文对前人的工作进行了大胆创新，实验结果表明这些创新是较为成功的。但可以预见它们仍然存在着相当大的革新和拓展空间。据此，在认真总结所取得研究成果的同时，下一阶段的研究工作应将规则自动简化机制构建于本系统中。我们乐观地相信它们在图像情感检索中一定会展现出更为优异的表现。

参考文献

[1] 王上飞，王熙法. 图像情感检索研究的进展与展望[J]. 电路与系统学报，2005，10(4)：102-110

[2] 王惠锋，孙正兴. 基于内容的图像检索中的语义处理方法[J]. 中国图像图形学报，2001，6(10)：945-952

[3] Pawlak Z. Rough set[J]. International Journal of Computer and Information Sciences，1982，11(5)：341-356

[4] Mao Xia，Chen Bin，Muta I. Affective property of image and fractal dimension [J]. Chaos Solutions and Fractals，2003，13

[5] Tanaka S，Iwadate Y，Inokuchi S. An attractiveness evaluation model based on the physical features of image regions[C]// IEEE 15th International Conference on Pattern Recognition. Barcelona，Spain，2000：2793-2796

[6] 王上飞，王熙法. 一种双层图像情感检索模型[J]. 系统仿真学报，2004，16(9)：2074-2078

[7] 王伟凝，余英林，张健超. 基于线条方向直方图的图像情感语义分类[J]. 计算机工程，2005，11(31)：7-9

[8] 王军，张庆杰. 极小极大规则学习及在决策树规则简化中的应用[J]. 计算机研究与发展，1998，35(9)：806-809

[9] 孙长嵩，董西国，张健沛. 一个基于粗糙集和决策树的最简分类规则集生成算法[J]. 哈尔滨工程大学学报，2002，23(5)：87-91

(上接第236页)

用了28%～40%的基分类器。这种现象再次验证了Zhou等人的选择性集成策略的有效性[10]，大大降低了集成规模。

表3 本文算法实际使用的基分类器的平均百分比(×100%)

	segment	credit	ionosphere	sonar	heart
cp. GASEN	56.8	67.5	62.5	57.0	51.2
cp. EDSEN	35.1	33.0	37.7	34.9	28.4
	vehicle	vowel	waveform	ORHD	PRHD
cp. GASEN	59.4	64.1	60.9	55.8	61.0
cp. EDSEN	32.3	37.2	39.6	31.5	34.3

结束语 本文提出基于成对约束投影的选择性集成算法。首先，利用随机选取的must-link和cannot-link成对约束集得到投影矩阵，并将原始训练数据在诱导的投影空间中训练SVM基分类器；然后，分别采用遗传优化和最小化偏离度误差两种选择性策略构造集成分类器。基于UCI数据的实验结果表明，本文提出的两种选择性集成算法在训练SVM分类器时取得的泛化性能显著优于Bagging，Boosting，特征Bagging以及LogBag等算法。主要原因之一是它们采用的约束投影技术产生的基分类器之间具有明显的差异性。

参考文献

[1] Hansen L K，Salamon P. Neural Network Ensembles[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence，1990，12(10)：993-1001

[2] Dietterich T G. Machine Learning Research：Four Current Directions [J]. AI Magazine，1997，18(4)：97-136

[3] Krogh A，Vedelsby J. Neural Network Ensembles，Cross Validation，and Active Learning[J]. Advances in Neural Information Processing Systems，1995(7)：231-238

[4] Kuncheva L. Combing Pattern Classifier：Methods and Algorithm[M]. John wiley and Sons，2004

[5] Vapnik V N. 统计学习理论的本质[M]. 张学工，译. 北京：清华大学出版社，2000

[6] Dong Y S，Han K S. A Comparison of Several Ensemble Methods for Text Categorization[C]// IEEE International Conference on Services Computing(SCC'04). 2004：419-422

[7] Valentini G，Dietterich T. 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

[8] Tao D C，Tang X O，et al. Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2006，28(7)：1088-1099

[9] Basu S，Banerjee A，Mooney R J. Active Semi-supervision for Pairwise Constrained Clustering[C]// Proceedings of the SIAM International Conference on Data Mining. 2004：333-344

[10] Zhou Z H，Wu J，Tang W. Ensembling Neural Networks：Many Could Be Better Than All[J]. Artificial Intelligence，2002：137(1/2)：239-263

[11] 朱帮助. 基于特征提取的选择性神经网络集成方法[J]. 计算机科学，2008，35(3)：132-134

[12] Dietterich T. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees：Bagging，Boosting，and Randomization[J]. Machine Learning，2000，40(2)：139-158