

DOI: 10.7652/xjtuxb201412010

融合样本选择与特征选择的 AdaBoost 支持向量机集成算法

杨宏晖, 王芸, 孙进才, 戴健, 李亚安
(西北工业大学航海学院, 710072, 西安)

摘要: 为提高 AdaBoost 分类器集成算法的分类精确度并简化分类系统的复杂度, 提出一种融合样本选择与特征选择的 AdaBoost 支持向量机集成算法(IFSelect-SVME)。该算法在 AdaBoost 算法的每个循环中利用加权免疫克隆样本选择算法进行样本选择, 并用互信息顺序向前特征选择算法进行特征选择, 再利用每个循环优化选择得到的特征样本子集训练个体 SVM 分类器, 并对其进行加权集成, 生成最终的决策系统。对实验所用 9 组 UCI 数据集的仿真结果表明: 与支持向量机集成(SVME)算法相比, IFSelect-SVME 算法的正确分类率有所提高, 且样本数可减少 30.8%~80.0%, 特征数可减少 32.2%~81.5%, 简化了集成结构, 缩短了测试样本的分类时间, 所得到的分类系统具有更好的分类精度。

关键词: 分类器集成; AdaBoost 算法; 支持向量机; 样本选择; 特征选择

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 0253-987X(2014)12-0063-06

An AdaBoost Support Vector Machine Ensemble Method with Integration of Instance Selection and Feature Selection

YANG Honghui, WANG Yun, SUN Jincai, DAI Jian, LI Ya'an
(School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: An AdaBoost support vector machine ensemble (IFSelect-SVME) method with integration of instance selection and feature selection is proposed to improve the classification accuracy of AdaBoost ensemble algorithm and to simplify the complexity of the classification system. The proposed algorithm selects instance subsets via the weighted immune clonal instance selection algorithm in each cycle of AdaBoost algorithm, and the feature subsets are obtained using the mutual information sequential forward feature selection algorithm. Then the individual SVM classifiers are trained by the resulting optimal feature instance subsets in each cycle, and combined via majority vote to generate the final decision system. Simulation results on 9 UCI datasets and comparisons with the AdaBoost SVME algorithm show that the IFSelect-SVME algorithm obtains better classification accuracy with the number of instances decreasing to 30.8%-80.0% and the number of features decreasing to 32.2%-81.5%. The proposed IFSelect-SVME algorithm simplifies the structure of ensemble model, shortens the classification time of test instances, and gives a classification system with better classification accuracy.

收稿日期: 2014-04-29。 作者简介: 杨宏晖(1971—), 女, 博士, 副教授。 基金项目: 国家自然科学基金资助项目(51179157, 60672136); 水下测控技术重点实验室基金资助项目(9140C260505120C26104)。

网络出版时间: 2014-09-09

网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20140909.0840.001.html>

Keywords: classifier ensemble; AdaBoost algorithm; support vector machine; instance selection; feature selection

在统计模式识别领域,分类器集成作为提高分类系统精确度和泛化性能的重要方法^[1],主要通过生成并组合个体分类器来实现分类^[2-4]。但是,分类器集成在实际应用中存在2个主要问题:①训练分类器的样本和特征常含有噪声样本与冗余特征,这导致分类器集成的实际结果与理论值相差甚远;②分类器集成势必增加分类系统的复杂度,导致分类时间过长,限制了分类系统在对在线测试时间有要求领域的应用。近年来,研究者们试图设计出更有效的集成实现方式解决这2个问题。AdaBoost算法^[5-6]作为一种经典的分类器集成算法,通过加大权值来强化对难分样本的学习,提高弱分类器的分类精度,构建更强的集成分类器,但并未去除掉噪声样本和冗余特征的不良影响,且系统仍然具有较高的复杂度;文献[7-8]在AdaBoost算法的基础上进行了改进,引入多视图的概念来构建集成分类器,进一步提高了分类精度,但分类系统仍然不够精简;文献[9]通过动态地选择个体分类器并将其循环集成,大大降低了分类系统的复杂度,但未考虑到训练集中所含的噪声样本和冗余特征;文献[10]将集成思想用在特征选择和分类器设计中,在一定程度上能够去除冗余特征,提高分类精度并节省分类时间,但没有考虑到噪声样本对分类的影响;文献[11]在融合特征选择和分类器集成的基础上又对个体分类器进行了选择,进一步缩小了集成规模,简化了计算复杂度,但噪声样本对分类系统的影响依然存在;文献[12]将样本选择融入AdaBoost分类器集成模型之中来提高集成系统的分类精度,简化集成结构,但又忽略了冗余特征对分类的影响。

针对上述问题,本文提出一种融合样本选择与特征选择的AdaBoost支持向量机^[13]集成算法(AdaBoost SVM ensemble algorithm combined with instance selection and feature selection, IFSelect-SVME)。本文算法在AdaBoost分类器集成框架中融入样本选择^[14-15]和特征选择^[16-17],剔除了训练子集中的噪声样本和冗余特征,提高了集成分类器的学习效率和精度。采用UCI公共数据库中的数据本进行的分类实验结果表明,本文算法能够在降低特征与样本数的同时提高或保持SVME的正确分类率,具有良好的泛化性能。

1 AdaBoost 算法

AdaBoost算法^[5-6]首先采用随机遍历抽样法抽取训练样本集,再用训练样本集训练个体分类器;在每次循环中,根据样本是否被分类正确而赋予不同的加权重值,被错误分类的样本权值较大,被正确分类的样本权值较小;然后根据不同样本的加权重值对原有的训练样本集进行重抽样,重点抽取难分类的样本,直到生成多个个体分类器;最后采用加权多数投票法对个体分类器进行集成。

2 IFSelect-SVME 算法

2.1 IFSelect-SVME 算法的原理

本文IFSelect-SVME算法采用了AdaBoost算法的集成框架,首先采用加权免疫克隆样本选择算法进行样本选择,再使用互信息顺序向前特征选择进行特征选择,以保证在每个循环中只选择最优的样本和特征来训练个体SVM分类器。这样得到的优化子集训练的个体分类器往往具有更高的分类精度,且随着训练集中噪声样本和冗余特征的去除,个体分类器的复杂度也大大降低,由所得到的个体分类器进行加权集成,最终生成的决策系统的分类精度也有所提升。IFSelect-SVME算法原理如图1所示。

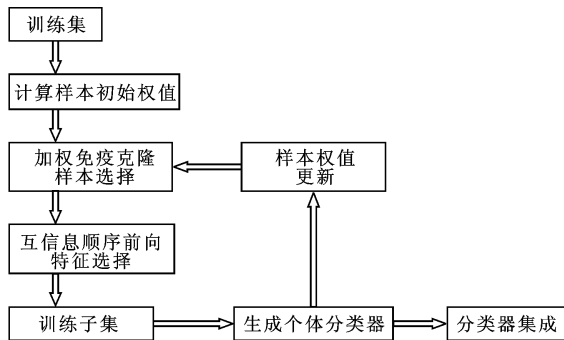


图1 IFSelect-SVME原理图

2.2 加权免疫克隆样本选择算法

在IFSelect-SVME算法中,利用加权免疫克隆样本选择算法进行样本选择^[15]。加权免疫克隆样本选择算法先随机生成初始抗体种群,然后根据样本权值计算抗体的加权亲和度(抗体-抗原加权亲和度与抗体-抗体间亲和度)以及克隆数,再利用克隆

操作、免疫基因操作(海明距离交叉与随机变异)以及克隆选择操作指导种群的进化,循环往复,直到达到终止条件。在该算法中,一个抗体代表一种样本选择方式,抗体-抗原加权亲和度表示某抗体的分类性能,而抗体-抗体间亲和度则表示不同抗体之间的差异程度,体现出抗体群的多样性。每个抗体的克隆数则通过抗体-抗原加权亲和度和抗体-抗体间亲和度来控制。

加权免疫克隆样本选择算法的步骤如下:

Step1 随机生成初始种群。对于某一样本集 $X = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbf{R}^d, y_i \in \{1, \dots, K\}, i = 1, \dots, n\}$, 将样本选择问题编码为二进制基因链 $\Phi = (\phi_1, \phi_2, \dots, \phi_i, \dots, \phi_n)$, 其中 ϕ_i 为个体基因编码, n 为样本数, $\phi_i = 1$ 表示第 i 个样本被选中, $\phi_i = 0$ 表示第 i 个样本未被选中, 随机生成初始抗体种群 O 。

Step2 加权亲和度计算。计算分为以下 2 个步骤。

(1) 抗体-抗原加权亲和度 G_j 计算如下

$$G_j(k) = 1 - E_j(k) \tag{1}$$

式中: $E_j(k) = \sum_{i=1}^n \omega_{y_i \neq \gamma(x_i)}(j)$ 表示第 k 代抗体种群中第 j 个抗体被错误分类的样本权值之和, 其中 $\omega_{y_i \neq \gamma(x_i)}(j)$ 表示第 j 个抗体的基因中某类标 y_i 与测试得到的类标 $\gamma(x_i)$ 不符的样本权值。

(2) 抗体-抗体间亲和度 B_j 计算如下

$$B_j(k) = \min \left\{ \exp \left(\sum_{j \neq q} |\alpha_j(k) - \alpha_q(k)| \right) \right\} \tag{2}$$

式中: $\alpha_j(k)$ 与 $\alpha_q(k)$ 分别表示第 k 代抗体种群中的第 j 个和第 q 个抗体。

Step3 克隆算子操作。操作分为以下 3 个步骤。

(1) 克隆操作: 根据每个抗体进行克隆, 得到克隆种群 O_r ; 用式(3)计算抗体克隆数

$$\lambda_j(k) = \text{int} \left\{ \delta_r B_j(k) G_j(k) / \sum_{q \neq j} G_j(k) \right\} \tag{3}$$

式中: $\lambda_j(k)$ 为第 j 个抗体的克隆数; δ_r 为克隆规模设定值。

(2) 免疫基因操作: 对克隆种群 O_r 以概率 P_s 进行海明距离交叉, 产生交叉种群 O_s , 再对 O_t 按概率 P_t 进行随机变异, 得到变异种群 $O_t^{[13]}$ 。

(3) 克隆选择操作: 比较父代与免疫基因操作后子代抗体种群, 选择出其中抗体-抗原加权亲和度高的一个抗体, 组成新一代抗体种群 O' 。

克隆算子操作保证了在本次循环中, 用优化样本子集训练的个体分类器的分类精度大于等于不做样本选择时训练的个体分类器的分类精度。

Step4 判断。若到达终止条件(终止条件通常为设定的循环次数), 则输出最优解; 否则, 转至 Step2。

2.3 互信息顺序向前特征选择算法

在 IFSelect-SVME 算法中, 用互信息顺序向前特征选择算法进行特征选择。该算法用特征与类别之间的互信息值^[16]来评价特征的分类性能, 并利用顺序向前特征选择方法^[15]来搜索最优特征子集。特征 f 与类别 L 之间的互信息计算公式如下

$$I(L; f) = H(L) - H(L | f) \tag{4}$$

式中: $H(L)$ 是类别 L 的熵; $H(L | f)$ 是在特征 f 条件下类别 L 的熵。互信息值越大, 则表示特征与类别的相关程度越强, 特征包含的分类信息越多, 特征的分类性能也就越强。

2.4 生成分类系统

IFSelect-SVME 算法采用加权多数投票法对每次循环生成的个体 SVM 分类器进行集成, 用每一个个体分类器对样本进行决策, 对自己所预测样本的类别投一票, 最终, 票数最多的类别被确定为这个样本的类别, 直到所有样本的类别被全部预测, 所得到的分类结果即为分类器集成的最终分类结果。

3 实 验

3.1 实验数据

本文利用加州大学用于机器学习的 UCI(Uni-versity of California Irvine, UCI)数据库^[19]中的 9 种数据集对 IFSelect-SVME 算法的性能进行验证实验, 9 种数据集的特征均为数值类型, 包含连续型(continued, C)、离散型(discreted, D)和混合型(mixed, M)特征。数据集的详细信息如表 1 所示。

表 1 实验数据

数据集	特征数	特征类型	样本数	类别数
wine	13	C	178	3
zoo	16	D	101	7
wdbc	30	C	569	2
sonar	60	C	138	2
parkinsons	22	C	195	2
ionosphere	34	C	351	2
Indian-L	10	M	583	2
glass	9	C	214	6
climate	18	C	540	2

3.2 分类实验

采用表 1 中的 9 种数据集对本文的 IFSelect-SVME 算法进行分类实验, 通过分类性能、特征选

择性能、样本选择性能以及分类时间等几个指标对本文 IFSelect-SVME 算法的有效性进行了分析。

3.2.1 IFSelect-SVME 与 SVME 算法的分类性能对比 实验比较了本文 IFSelect-SVME 算法与 SVME 算法的平均正确分类率($\bar{A}$)、样本选择数与特征选择数结果如表 2 所示。

表 2 2 种算法的分类性能对比

数据集	特征选择数		样本选择数		$\bar{A}$	
	SVME	本文算法	SVME	本文算法	SVME	本文算法
wine	13	9	88	25	0.91	0.96
zoo	16	5	50	14	0.94	0.98
wdbc	30	15	200	37	0.92	0.95
sonar	60	20	69	31	0.79	0.81
parkinsons	22	18	60	26	0.79	0.80
ionosphere	34	29	176	78	0.93	0.96
Indian-L	10	2	291	197	0.71	0.73
glass	9	4	107	58	0.52	0.61
climate	18	9	270	128	0.93	0.94

从表 2 中可以看到,对于不同的数据集,本文 IFSelect-SVME 算法选择的特征数与样本数都少于 SVME 算法。对于离散型数据 zoo,IFSelect-SVME 算法在每个循环中平均仅用 31%的特征数和 28%的样本数来训练个体分类器,平均正确分类率比 SVME 算法可提高 4%;对于混合型数据 Indian-L,则平均需用 20%的特征数与 68%的样本数训练个体分类器,正确分类率可提高 2%;本文算法对其他几种连续型数据的正确分类率都有提高,而样本选择数只有 SVME 算法的 18%~54%;特征选择数也只有 SVME 算法的 33%~85%。因此不论对于何种特征类型的数据,IFSelect-SVME 算法均能够提高分类系统的正确分类率,同时大幅度压缩训练样本数与特征数。

3.2.2 样本选择结果 当用加权免疫克隆算法进行样本选择时,每次循环中选择出的样本都是不同的。根据每个样本被选择的次数,定义样本重要性指数,可按式(5)计算

$$V_i = u_i / Z \tag{5}$$

式中: V_i 是第 i 个样本的重要性指数; Z 为训练个体分类器的个数; u_i 是第 i 个样本 Z 次分类器生成中被选择的次数。样本的重要性指数越高,对分类就越有效。用该算法对 9 种数据进行样本选择,得到归一化样本重要性指数。不同样本集的归一化重要性指数如图 2 所示。

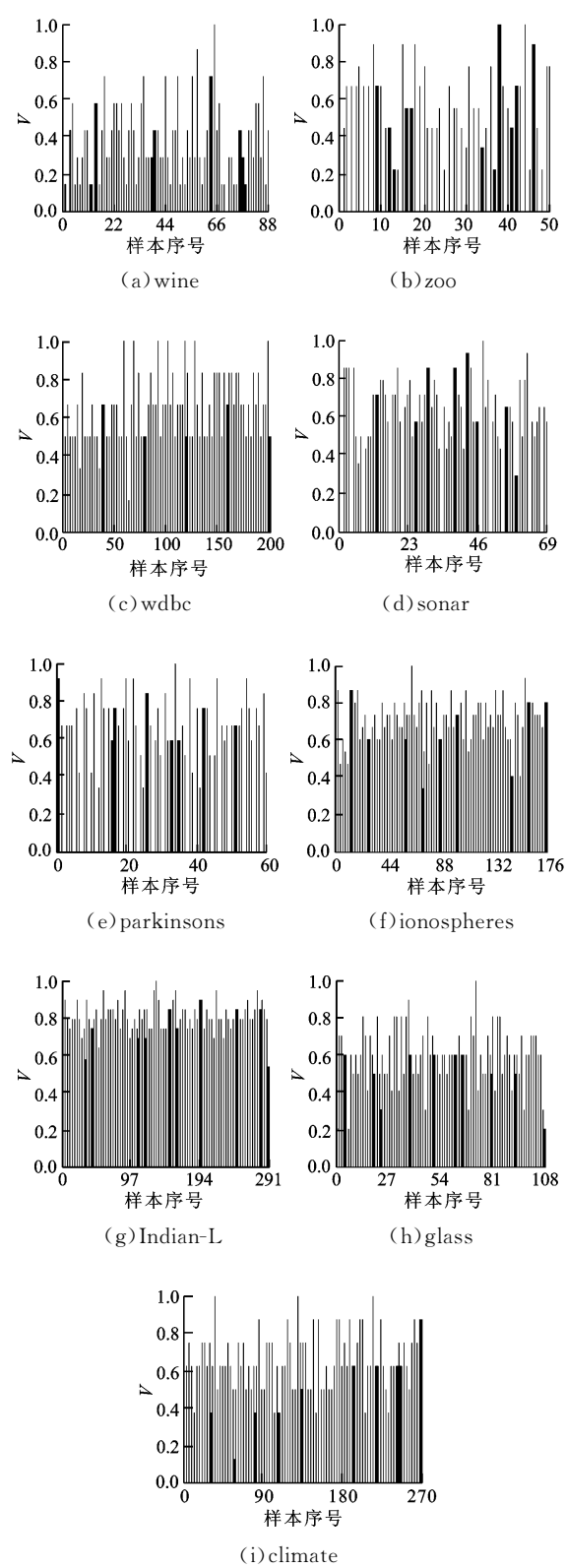


图 2 不同数据集的归一化样本重要性指数

从图 2 可以看出,对于不同数据集,不同样本的归一化重要性指数都有着明显的差别。本文算法只选取了重要性指数高于 0.5 的样本作为样本选择结果。选取重要性指数高的样本来组建最优的样本子

集,就可以训练出分类性能更好的个体分类器。

3.2.3 特征选择结果 根据 IFSelect-SVME 算法中互信息顺序向前特征选择法的原理,可得到选取不同维数特征时对应的平均正确分类率,如图 3 所

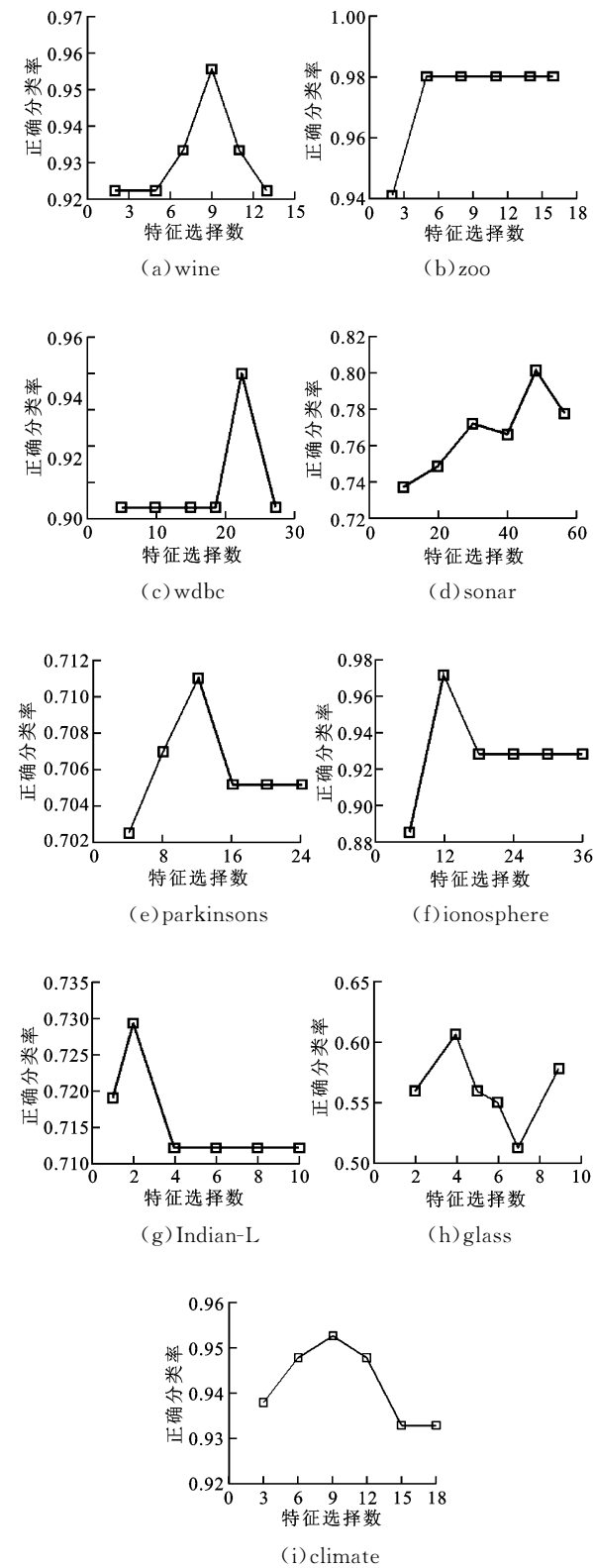


图 3 不同数据集的特征选择数所对应的正确分类率

示。从图 3 可以看出,随着特征选择数的增加,不同数据集的分类率曲线大体呈现先上升,到达某一值后趋于平稳或又有所下降的趋势;glass 与 ionosphere 数据集的分类率曲线上波动,但仍在中间的某一点处取得最大值。这说明,所有的数据集都只需要选取部分维数的特征就可以使得分类率达到最大,过少的特征使得正确分类率不能达到理想的效果,且冗余的特征也会导致分类率变差,说明 IFSelect-SVME 中的特征选择不仅可以降低特征维数,还可以提高分类精度。

3.2.4 IFSelect-SVME 与 SVM 算法的分类时间对比 图 5 给出了 IFSelect-SVME 算法与 SVM 算法训练相同数量分类器的分类时间对比。

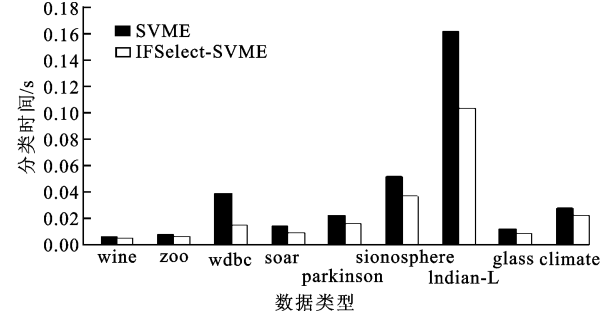


图 4 2 种算法的分类时间对比

从图 4 可以看出,IFSelect-SVME 比 SVM 算法对 9 种数据的分类时间都有所减少,这说明 IFSelect-SVME 算法精简了个体分类器的复杂度,为分类节省了时间。

4 结 论

本文提出了一种融合样本选择和特征选择的 AdaBoost 支持向量机集成算法,利用 AdaBoost 算法构造集成模型,通过样本选择和特征选择优化训练子集来生成个体分类器,并将所得的个体分类器用加权多数投票的方法进行集成,从而得到最终的综合分类器。采用本文算法对 9 种 UCI 数据集进行分类实验,结果表明,本文算法对于多种数据均可有效地指导对最优样本和特征子集的搜索,大幅度减少样本数和特征数,且经过样本选择和特征选择后的分类器集成模型与未经样本选择和特征选择的分类器集成模型相比,具有更好的分类性能,因此本文算法具有一定的普适性和有效性,具有良好的推广价值。

参考文献:

[1] MIKEL G, ALBERTO F, BDURNE B, et al. A re-

- view on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches [J]. *IEEE Transactions on Systems, Man, and Cybernetics: Part C Applications and Reviews*, 2012, 42(4): 463-484.
- [2] 张宏达, 王晓丹, 韩钧, 等. 分类器集成差异性研究 [J]. *系统工程与电子技术*. 2009, 31(12): 3007-3012.
- ZHANG Hongda, WANG Xiaodan, HAN Jun, et al. Survey of diversity researches on classifier ensembles [J]. *Systems Engineering and Electronics*, 2009, 31(12): 3007-3012.
- [3] WINDEATT T. Accuracy/diversity and ensemble MLP classifier design [J]. *IEEE Transactions on Neural Networks*, 2006, 17(11): 1194-1211.
- [4] SUN Shiliang. Local within-class accuracies for weighting individual outputs in multiple classifier systems [J]. *Pattern Recognition Letters*, 2010, 31: 119-124.
- [5] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. *Journal of Computer and System and Science*, 1997, 63(5): 119-139.
- [6] 曹莹, 苗启广, 刘家辰, 等. AdaBoost 算法研究进展与展望 [J]. *自动化学报*, 2013, 39(6): 745-758.
- CAO Ying, MIAO Qiguang, LIU Jiachen, et al. Advance and Prospects of AdaBoost [J]. *Acta Automatica Sinica*, 2013, 39(6): 745-758.
- [7] XU Zhijie, SUN Shiliang. An algorithm on multi-view AdaBoost [C]// *Proceedings of 17th International Conference on Neural Information Processing*. Heidelberg, Germany: Springer, 2010: 355-362.
- [8] XU Zhijie, SUN Shiliang. Multi-source transfer learning with multi-view AdaBoost [C]// *Proceedings of 19th International Conference on Neural Information Processing*. Heidelberg, Germany: Springer, 2012: 332-339.
- [9] 郝红卫, 王志彬, 殷绪成. 分类器的动态选择与循环集成方法 [J]. *自动化学报*, 2011, 37(11): 1290-1295.
- HAO Hongwei, WANG Zhibin, YIN Xucheng. Dynamic selection and circulating combination for multiple classifier systems [J]. *Acta Automatica Sinica*, 2011, 37(11): 1290-1295.
- [10] 孙亮, 韩崇昭, 沈建京, 等. 集成特征选择的广义粗集方法与多分类器融合 [J]. *自动化学报*, 2008, 34(3): 298-304.
- SUN Liang, HAN Chongzhao, SHEN Jianjing, et al. Generalized rough set method for ensemble feature selection and multiple classifier fusion [J]. *Acta Automatica Sinica*, 2008, 34(3): 298-304.
- [11] 马超, 陈西宏, 徐宇亮, 等. 广义领域粗集下的集成特征选择及其选择性集成算法 [J]. *西安交通大学学报*, 2011, 45(6): 34-39.
- MA Chao, CHEN Xihong, XU Yuliang, et al. Ensemble feature selection based on generalized neighborhood rough model and its selective integration [J]. *Journal of Xi'an Jiaotong University*, 2011, 45(6): 34-39.
- [12] GARCIA P N. Constructing ensembles of classifiers by means of weighted instance selection [J]. *IEEE Transactions on Neural Networks*, 2009, 20(2): 258-277.
- [13] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer-Verlag, 2000.
- [14] OLVERA-LOPEZ J A, CARRASCO-CHOA J A, MARTINEZ-TRINIDAD J, et al. A review of instance selection methods [J]. *Artificial Intelligence Review*, 2010, 34(2): 133-143.
- [15] YANG Honghui, ZHOU Xin, WANG Yun, et al. A new adaptive immune clonal algorithm for underwater acoustic target sample selection [C]// *Proceedings of 2013 IEEE International Conference of IEEE Region 10*. Piscataway, NJ, USA: IEEE, 2013: 1-4.
- [16] 杨宏晖, 戴健, 孙进才, 等. 用于水下目标识别的自适应免疫特征选择算法 [J]. *西安交通大学学报*, 2011, 45(12): 28-32.
- YANG Honghui, DAI Jian, SUN Jincai, et al. A new adaption immune feature selection algorithm for underwater acoustic target classification [J]. *Journal of Xi'an Jiaotong University*, 2011, 45(12): 28-32.
- [17] LIU Huan, YU Lei. Toward integrating feature selection algorithms for classification and clustering [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(4): 491-502.
- [18] VINH L T, THANG N D, LEE Y K. An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information [C]// *Proceedings of 2010 10th Annual International Symposium on Applications and the Internet*. Piscataway, NJ, USA: IEEE, 2010: 395-398.
- [19] HETTICH S, BLAKE C, MERZ C. UCI repository of machine learning database [DB/OL]. (1998-07-11) [2014-04-03]. <http://www.ics.uci.edu/mllearn/ML-Repository.html>.

(编辑 刘杨)