DOI:10.11830/ISSN.1000-5013.201611103

采用负相关学习的 SVM 集成算法

洪铭,汪鸿翔,刘晓芳,柳培忠

(华侨大学 工学院,福建 泉州 362021)

摘要: 为了平衡集成学习中多样性与准确性之间的关系,并提高决策分类器的泛化能力,提出一种基于负相关学习和 AdaBoost 算法的支持向量机(SVM)集成学习方法.将负相关学习理论融合到 AdaBoost-SVM 的训练过程中,利用负相关学习理论计算基分类器间的相关性,并根据相关性的值自适应调整基分类器的权重,进而得到加权后的决策分类器.在 UCI 数据集中进行仿真,结果表明:相较于传统的负相关集成学习算法和 AdaBoost-SVM 算法,所提出的方法分类准确率更高,泛化能力更好.

关键词: 负相关学习;误差-分歧分解;AdaBoost-SVM;集成学习;分类器

中图分类号: TP 391 **文献标志码:** A **文章编号:** 1000-5013(2018)06-0942-05

SVM Ensembles Algorithm Using Negative Correlation Learning

HONG Ming, WANG Hongxiang, LIU Xiaofang, LIU Peizhong

(College of Engineering, Huaqiao University, Quanzhou 362021, China)

Abstract: In order to balance the relationship between diversity and accuracy in ensemble learning, and improve the generalization ability of decision classifier, a new support vector machine (SVM) ensemble learning method based on negative correlation learning and AdaBoost algorithm is proposed. The negative correlation learning theory is integrated into the training process of AdaBoost-SVM, and the correlation between the base classifiers is calculated by using the negative correlation learning theory. Furthermore, the weight of the base classifier is adjusted adaptively according to the correlation value. The simulation results of UCI dataset show that compared with the traditional negative correlation ensemble learning algorithm and AdaBoost-SVM algorithm, the proposed method can get higher classification accuracy and better generalization ability.

Keywords: negative correlation learning; error-ambiguilty decomposition; AdaBoost-SVM; ensemble learning; classifier

集成学习是通过构建并结合多个学习器来完成学习任务[1].随着集成学习技术的快速发展,各种集成学习算法被广泛应用于工程、生物、医学、图像处理和计算机视觉等领域[2-5].虽然集成学习器的预测效果显著优于单个学习器,但随着基学习器数目增多,所需的计算和存储开销也逐渐增加,基学习器之间的差异性难以保证[6-9].Zhou 等[2]提出了选择性集成的概念,剔除一些精度不高和作用不大的基分类器进行集成能获得更好的效果[10].AdaBoost 是一种有效的集成学习方法.它使用权重更新的方法对难训练的样本赋予更高的权重以训练出一系列基学习器[11],即基学习器的差异性主要通过样本扰动实现.但是,将稳定的学习算法,如支持向量机(SVM),作为 AdaBoost 的基分类器时,基分类器之间通常

收稿日期: 2016-11-03

通信作者: 柳培忠(1976-),男,讲师,博士,主要从事计算机视觉、机器学习、嵌入式系统的研究.E-mail:pzliu@hqu.edu.cn.

基金项目: 国家自然科学基金资助项目(61203242);福建省泉州市科技计划项目(2014Z113,2014Z103);华侨大学研究生科研创新能力培育计划资助项目(1400422003)

存在较大的相关性和冗余性，融合后的决策分类器容易造成过拟合[11]. 而集成学习器的分类精度主要由基分类器的准确性和多样性决定，基分类器的准确性越高、多样性越大，则集成效果越好[12]. 因而有必要使用 SVM 等稳定的分类器作为基分类器[13-14]. 针对 AdaBoost 算法的上述缺陷，本文将基于负相关学习(NCL)的相关性惩罚项引入 SVM 的集成学习过程中，以提高集成系统的精确度和泛化能力.

1　负相关学习

负相关学习最早应用于神经网络集成[15]，它的理论来源于误差-方差分解和分歧分解[16-18]，它对每一个基分类器显式地添加一个相关性惩罚项，保证集成系统的多样性.

给定训练集$\{x_i, y_i\}_{i=1}^N$，NCL 融合 T 个基分类器 $h_t(x)$构建集成系统 $H(x)=\frac{1}{T}\sum_{t=1}^{T}h_i(x_i)$.

每一个基学习器的误差 e_t 定义为

$$e_t=\sum_{i=1}^{N}(h_t(x_i)-y_i)^2+\lambda p_t. \tag{1}$$

式(1)中：e_t 为第 t 个基学习器的训练误差；p_t 为相关性惩罚函数；λ 为惩罚项 p_t 的权重参数，用于协调训练误差与惩罚项之间的关系.

由式(1)可知：当$\lambda=0$ 时，相关性惩罚项都是 0，每个基分类器都是独立地训练；随着 λ 的增大，集成系统越来越多地偏重于惩罚项，基分类器相互之间并不完全独立，每个基分类器的误差都要受到自身和其他分类器分类结果的影响. 由于这种影响是负相关的，从而可以训练出一些差异性较大的分类器.

相关性惩罚函数 p_t 为

$$\begin{aligned}p_t&=\sum_{i=1}^{N}\{(h_t(x_i)-H(x_i))\sum_{k\neq t}^{T}(h_k(x_i)-H(x_i))\}=\\&\sum_{i=1}^{N}\{(h_t(x_i)-H(x_i))(\sum_{k\neq t}^{T}h_k(x_i)-(T-1)\times H(x_i))\}=\\&\sum_{i=1}^{N}\{(h_t(x_i)-H(x_i))(T\times H(x_i)-h_t(x_i)-(T-1)\times H(x_i))\}=\\&-\sum_{i=1}^{N}(h_t(x_i)-H(x_i))^2.\end{aligned} \tag{2}$$

由式(2)可知：p_t 的幅值$|p_t|$越小，表明第 t 个基学习器与集成系统的差异性越小. 因而，采用$|p_t|$计算相关性惩罚值.

2　NCAB-SVM 算法

集成系统的性能主要由基学习器的准确性和多样性决定，其泛化误差与多样性、准确性的关系为

$$E=\bar{E}-D. \tag{3}$$

式(3)中：E 为集成系统的泛化误差；$\bar{E}$ 为基分类器的平均错误率；D 为基分类器之间的多样性. 因此，在提高基分类器准确性的同时，保证分类器间的多样性，便可减小集成系统的泛化误差. 然而，多样性和准确性本身是一个矛盾体，即增加多样性的同时，准确性一般会降低，如何在二者之间找到一个平衡是集成学习研究的重点.

文中基于负相关学习理论与 AdaBoost 算法提出一种新的 SVM 集成学习算法(NCAB-SVM). 通过在 AdaBoost 中引入负相关学习的惩罚项，协调集成系统的多样性和准确性，并选用 RBF 核 SVM (RBFSVM)作为 AdaBoost 的基分类器. RBFSVM 主要包含 C，σ 两个参数. 其中，参数 C 控制模型的复杂度和训练误差；σ 为高斯核的带宽. 由 RBFSVM 的性能分析[14]可知：相对于 C，RBFSVM 的性能主要由核参数 σ 决定. 对于一个给定范围的 C 值，RBFSVM 的性能仅随着 σ 的改变而改变. 当 σ 值很大时，RBFSVM 的分类精度通常小于 50%；当 σ 值很小时，分类精度较高，但是分类结果与高度相关，很难得到好的集成效果. 因此，通过固定参数 C，使用步长 σ_s 更新 σ，以获得一系列不同性能的 RBFSVM 基分

类器. NCAB-SVM 算法的主要流程如下.

输入:给定训练数据集$\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_m,y_m)\}$;

过程:

步骤 1 初始化数据权重 $D_1(x_i)=1/m$,惩罚项 $p_1(x_i)=1$,初始化 $\sigma=\sigma_m$,σ 的最小值 $\sigma_{\min}$,σ 的更新步长 σ_s,惩罚项阈值 DIV,迭代次数 $t=2$;

步骤 2 根据输入样本训练一个 RBFSVM,记作 h_1,则有 $H(x)=h_1$;

步骤 3 do while($\sigma>\sigma_{\min}$);

步骤 4 根据输入样本训练一个 RBFSVM,记作 h_t;

步骤 5 计算训练误差:$\varepsilon_t=\sum_{i=1}^{m}D_t(x_i),y_i\neq h_t(x_i)$;

步骤 6 计算每一个样本 x_i 的惩罚值:$p_t(x_i)=\sum_{i=1}^{N}(h_t(x_i)-H(x_i))^2$;

步骤 7 if $1/m\sum_{i=1}^{m}p_t(x_i)<\mathrm{DIV}\parallel\varepsilon_t>0.5;\sigma=\sigma-\sigma_s$;continue;

步骤 8 根据误差和惩罚项计算 h_i 的权重 $\alpha=\frac{1}{2}\lg\frac{\sum_{i=1}^{m}(p_t(x_i))^{\lambda}D_t(x_i),y_i=h_t(x_i)}{\sum_{i=1}^{m}(p_t(x_i))^{\lambda}D_t(x_i),y_i\neq h_t(x_i)}$;

步骤 9 更新权重 D_t,得到新的权重 $D_{t+1}(x_i)=\frac{D_t(x_i)\exp(-\alpha_t h_t(x_i)y_i)}{Z_t}$;

步骤 10 更新集成分类器 $H(x)=\mathrm{sign}\sum_{t=1}^{T}\alpha_t h_t(x)$;

步骤 11 更新迭代次数 $t=t+1$;

输出:决策分类器 $H(x)=\mathrm{sign}\sum_{t=1}^{T}\alpha_t h_t(x)$.

由此可知:NCAB-SVM 算法在集成时进行权重更新,而不是在数据层或算法层进行. 原始的 Bagging 或 Boosting 方法,仅通过数据扰动或参数扰动实现多样性,但这并不能保证生成的基分类器彼此不相关. 提出的 NCAB-SVM 算法通过设置阈值的方式自动删除相关性和误差太大的基分类器(步骤7). 同时,对具有不同相关性和训练误差的分类器赋予不同的权重. 最后,集成得到的决策分类器的输出将由这个权重值决定.

NCAB-SVM 算法在步骤 8 中同时使用相关性惩罚项和精确度计算分类器的权重. 其中,λ 控制惩罚项的强度,以协调多样性和精确性. 由步骤 9 可知,当基分类器与集成系统的差异很小或准确率较低时,该基分类器被赋予的权重将很小,对集成系统的影响也很小.

综上所述,NCAB-SVM 算法具有以下 2 个优点:1) 使用 RBFSVM 作为基分类器,通过自动调整核参数获取一系列准确率较高的基分类器;2) 算法基于分类器之间的相关性和准确性进行加权,在增加多样性的同时,能保证基分类器的准确性,最终提高了集成系统的性能.

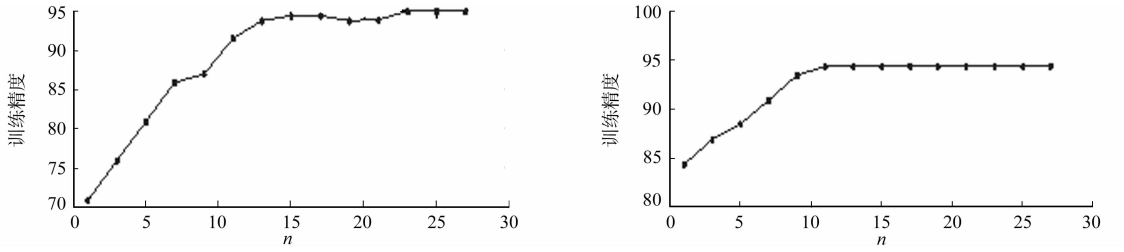
3 实验结果与分析

为了验证 NCAB-SVM 算法的性能,使用 UCI 数据库[19]的 10 个数据集进行试验,并与集成学习方法 SVM,AdaBoost-SVM[16],Bagging 和基于负相关学习的相关性数据修正学习(NCCD)算法[20]进行比较. 从收敛性分析、分类误差和多样性等 3 个角度对文中算法进行验证. 对于 NCAB-SVM 算法,设置相关参数:σ_{ini}为 20;$\sigma_{\min}$为 -20;DIV 为 0.8.

3.1 算法的收敛性分析

由于文中算法的迭代次数主要由 RBFSVM 的高斯核宽度 σ 的更新步长 σ_s 决定. 因此,通过调整 σ_s 来调整迭代次数. Ionosphere 和 Soybean 两个数据集的训练精度与迭代次数(n)的关系,如图 1 所示. 由图 1 可知:随着迭代次数的增加,训练精度逐渐提高,最后趋于稳定;当迭代次数等于 20 时,精度已经达

到稳定. 因此，试验中步长 σ_s 设为 2.



(a) Ionosphere 数据集　　(b) Soybean 数据集

图 1　不同数据集上迭代次数与训练精度的关系

Fig. 1　Relationship between number of iterations and training accuracy in different data sets

3.2　泛化误差的比较

采用十折交叉验证的方法估计泛化误差，取平均值作为最后分类器的训练误差. AdaBoost-SVM 算法和 Bagging 算法的迭代次数均为 50 次，NCCD 算法的迭代次数为 20. 实验结果如表 1 所示.

由表 1 可知：相比其他算法，NCAB-SVM 算法在大部分数据集都获得了最小的分类误差，表明文中算法不仅改进了 AdaBoost-SVM 算法的性能，而且相比其他负相关学习算法，其引入 SVM 作为集成系统的基分类器可以提高集成系统的性能；基于样本扰动的 Bagging 方法相对于纯 SVM 的分类精度没有显著改善. 因此，对存在冗余或无用信息的基分类器进行选择性裁剪是有必要的.

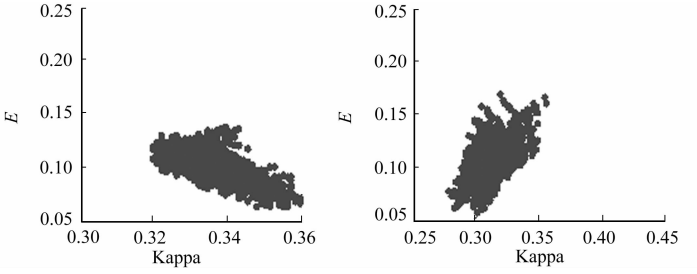
表 1　不同算法分类误差的比较

Tab. 1　Comparison of classification errors of different algorithms

数据集	SVM	AdaBoost-SVM	Bagging	NCCD	NCAB-SVM
Promoter	27.454±4.815	23.905±3.801	24.091±4.321	15.000±9.103	17.727±4.520
Sonar	14.227±1.106	12.250±1.103	12.381±3.241	22.857±15.681	10.952±2.300
Ionosphere	6.250±0.235	4.580±1.706	5.211±0.950	4.584±2.523	4.047±0.950
House-votes-84	3.007±0.935	2.965±0.647	3.218±0.726	4.023±1.817	2.528±0.484
Breast-w	21.270±2.002	12.666±2.542	19.055±1.502	18.358±1.454	15.671±0.853
Pima	29.780±2.143	26.592±2.143	26.900±1.744	25.071±2.081	24.025±1.960
German	26.189±2.351	23.780±1.544	25.130±1.437	49.221±17.987	24.900±1.776
Hypothyroid	29.780±2.143	15.235±0.862	26.900±1.744	18.673±2.064	12.270±0.157
Soybean-large	11.244±1.795	8.316±1.250	10.665±1.281	7.952±0.714	4.080±1.124
Insurance	16.067±0.435	9.828±0.935	13.240±0.912	7.219±2.391	6.450±0.177

3.3　多样性比较

为验证文中算法相对于 AdaBoost-SVM 算法引入相关性惩罚项后，集成系统基分类器的多样性的改变，使用 Kappa-Error 图[21]分析基分类器间的差异性. 对 AdaBoost-SVM 和 NCAB-SVM 使用 Ionosphere 数据集训练的 Kappa-Error 图，如图 2 所示.



(a) AdaBoost-SVM 算法　　(b) NCAB-SVM 算法

图 2　Ionosphere 数据集上的 Kappa-Error 图

Fig. 2　Kappa-Error figure on Ionosphere dataset

由图 2 可知：相比于 AdaBoost-SVM，NCAB-SVM 具有较低的分类误差和较高的差异性. 为更明确地表示 NCAB-SVM 与 AdaBoost-SVM 算法的差异，使用计算各个基分类器的 Kappa 统计值和平均分类错误率之积求和再平均的方法进行定量分析，两种算法对 Ionosphere 数据集的计算结果分别为 0.320 4，0.337 6，可得 NCAB-SVM 算法的分类效果更好.

4　结束语

提出一种基于负相关学习和 AdaBoost 算法的 SVM 集成学习方法. 将负相关学习理论融合到 Ad-

aBoost-SVM的训练过程中,利用负相关学习理论计算基分类器间的相关性,并根据相关性的值自适应调整基分类器的权重,进而得到加权后的决策分类器.算法使用RBFSVM作为基分类器,通过自动调整核参数获取一系列准确率较高的基分类器.同时,基于分类器之间的相关性和准确性进行加权,在增加多样性的同时,不降低基分类器的准确性,最终提高了集成系统的性能.

参考文献:

[1] LEBANON G,LAFFERTY J. Boosting and maximum likelihood for exponential models[C]// Advances in Neural Information Processing Systems. Vancouver:Neural Information Processing Systems,2002:447-454.

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

[3] 倪志伟,张琛,倪丽萍. 基于萤火虫群优化算法的选择性集成雾霾天气预测方法[J]. 模式识别与人工智能,2016,29(2):143-153. DOI:10.16451/j.cnki.issn1003-6059.201602006.

[4] TRAN V T,TEMPEL S,ZERATH B,*et al*. MiRBoost: Boosting support vector machines for microRNA precursor classification[J]. RNA,2015,21(5):775-785. DOI:10.1261/rna.043612.113.

[5] XU Jian,TANG Liang,LI Tao. System situation ticket identification using SVMs ensemble[J]. Expert Systems with Applications,2016,60:130-140. DOI:10.1016/j.eswa.2016.04.017.

[6] 张春霞,张讲社. 选择性集成学习算法综述[J]. 计算机学报,2011,34(8):1399-1410.

[7] MAO Shasha,JIAO Licheng,XIONG Lin,*et al*. Greedy optimization classifiers ensemble based diversity[J]. Pattern Recognition,2011,44(6):1245-1261. DOI:10.1016/j.patcog.2010.11.007.

[8] LAZAREVIC A,OBRADOVIC Z. Effective pruning of neural network classifier ensembles[J]. International Joint Conference on Neural Networks. Washington DC:IEEE Press,2001:796-801. DOI:10.1109/IJCNN.2001.939461.

[9] MARTINEZ-MUNEZ G,SUAREZ A,*et al*. Using boosting to prune bagging ensembles[J]. Pattern Recognition Letters,2007,28(1):156-165.

[10] FREUND Y,SCHAPIRE R E. Experiments with a new boosting algorithm[C]// Proceedings of the Thirteenth International Conference on International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,1996:148-156.

[11] 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.

[12] 姚旭,王晓丹,张玉玺,等. 基于AdaBoost和匹配追踪的选择性集成算法[J]. 控制与决策,2014(2):208-214. DOI:10.13195/j.kzyjc.2012.1472.

[13] CHANG Tiantian,LIU Hongwei,ZHOU Shuisheng. Large scale classification with local diversity AdaBoost SVM algorithm[J]. Journal of Systems Engineering and Electronics,2009,20(6):1344-1350.

[14] LI Xuchun,WANG Lei,SUNG E. AdaBoost with SVM-based component classifiers[J]. Engineering Applications of Artificial Intelligence,2008,21(5):785-795. DOI:10.1016/j.engappai.2007.07.001.

[15] LI Leijun,ZOU Bo,HU Qinghua,*et al*. Dynamic classifier ensemble using classification confidence[J]. Neurocomputing,2013,99(99):581-591.

[16] MARGINEANTU D D,DIETTERICH T G. Pruning adaptive boosting[C]// Fourteenth International Conference on Machine Learning. San Francisco:Morgan Kaufmann Publishers Inc,1997:211-218.

[17] 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(3):725-775.

[18] UEDA N,NAKANO R. Generalization error of ensemble estimators[C]// IEEE International Conference on Neural Networks. Washington DC:IEEE Press,1996:90-95. DOI:10.1109/ICNN.1996.548872.

[19] ASUNCION A,NEWMAN D. UCI machine learning repository[EB/OL]. [2016-11-03]. http://archive.ics.uci.edu/ml/index.php.

[20] CHAN Z S H,KASABPV N. A preliminary study on negative correlation learning via correlation-corrected data (NCCD)[J]. Neural Processing Letters,2005,21(3):207-214. DOI:10.1007/s11063-005-1084-6.

[21] KUNCHEVA L I,WHITAKER C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy[J]. Machine Learning,2003,51(2):181-207. DOI:10.1023/A:1022859003006.

(**责任编辑:** 钱筠 **英文审校:** 吴逢铁)