选择性集成学习算法的研究

陈 全,赵文辉,李 洁,江雨燕
(安徽工业大学,安徽 马鞍山 243002)

摘 要:通过选择性集成可以获得比单个学习器和全部集成学习更好的学习效果,可以显著地提高学习系统的泛化性能。文中提出一种多层次选择性集成学习算法,即在基分类器中通过多次按权重进行部分选择,形成多个集成分类器,对形成的集成分类器进行再集成,最后通过对个集成分类器多数投票的方式决定算法的输出。针对决策树与神经网络模型在20个标准数据集对集成学习算法 Ada_ens 进行了实验研究,试验证明基于数据的集成学习算法的性能优于基于特征集的集成学习算法的性能,有更好的分类准确率和泛化性能。

关键词:机器学习;集成学习;选择性集成
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2010)02-0087-03

Research of Selective Ensemble Learning Algorithm

CHEN Quan, ZHAO Wen-hui, LI Jie, JIANG Yu-yan
(Anhui University of Technology, Maanshan 243002, China)

Abstract: Through the selective ensemble, the algorithm would be more effective than each single one and better than the algorithm that select all the base classifier, and the algorithm would have effective generalization ability. In this paper, a selective multi-level integrated learning algorithm is presented. In the base classifier by repeatedly carried out by some of the weight of choice, the formation of a number of integrated classifiers, the formation of the integrated classifier re-integration, the final integration of a majority vote classifier algorithm to determine the output. For decision tree and neural network model in 20 data sets the standard learning algorithm of the integrated experimental study of a Ada_ens. Tested based on data integrated learning algorithm is better than the performance of the integrated feature set based on the learning algorithm performance, better classification accuracy and generalization performance.

Key words: machine learning; ensemble learning; selective ensemble

0 引 言

集成学习是将多个不同的单个模型组合成一个模型,其目的是利用这些单个模型之间的差异,来改善模型的泛化性能。选择性集成学习是为了解决某一具体问题而训练出多个分类器,通过某种规则将这些分类器的结果进行整合的学习算法,当学习模型具有较高的正确率且具有差异性时,通过集成学习可以显著地提高学习系统的泛化性能[1]。以往的集成学习方法,如 Adaboost, Bagging,都是将训练出来的所有分类器进行集成。研究发现,选择部分分类器进行集成能得到更好的泛化性能,这种集成思想称为选择性集成[2]。

目前分类问题已经广泛应用于金融、医疗、决策、控制等领域,如银行对是否放贷的决策、对肿瘤恶性和良性的识别等。集成学习是提高分类准确率的重要途径,如果能利用集成学习的思想设计出性能良好的学习算法,将具有很大的经济和现实意义。选择性集成正是通过集成学习提高分类器性能的重要途经。在选择性集成学习的研究中目前国内外研究者主要集中在对基分类器的选择上,大多采用遗传算法来选择较好的基分类器进行集成,然而这对算法的时间复杂度提出了巨大的挑战。

文中在对选择性集成研究的基础上提出了多层次选择性集成,也就是在训练好的基分类器中通过多次按权重进行部分选择,形成多个集成分类器;然后对形成的集成分类器进行再集成,最后通过对各集成分类器多数投票的方式决定算法的输出。研究表明这种学习算法具有良好的准确率和更好的泛化性能。

收稿日期:2009-05-19;**修回日期**:2009-08-17
基金项目:省级教学研究项目(2008jyxm305)
作者介绍:陈 全(1971-),男,硕士,研究方向为机器学习、网络工程、计算机控制。

1 Ada_ens 算法

集成学习的方法首先在训练集上训练出 m 个学习器[$\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_m$],在给出新的样本进行分类预测

时,各分类器首先进行预测,产生 m 个结果[$\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_m$]。最后通过某种规则将这些结果集成起来,如简单的多数投票方式,形成最后的输出 $\hat{y}$。在一种简单的情况下考虑 25 个二元分类器的组合,假设其中每个分类器的误差为 0.35。组合分类器通过对于这些基分类器的预测进行多数表决的方法预测检验样本的类标号。如果所有分类器都是一样的则组合分类器的误差还是 0.35;如果所有的分类器均是独立的则组合分类器的误差为:

$$e_{\text{ensemble}} = \sum_{i=13}^{25} \binom{25}{i} \epsilon^i (1-\epsilon)^{25-i} = 0.06$$

远低于基分类器的误差率。

从上面的说明中能看出,组合分类器的性能优于单个分类器必须满足两个必要的条件:

(1) 基分类器之间应该是相互独立的;

(2) 基分类器应该好于随机猜测分类器,即 $\epsilon_i < 0.5$[3]。目前国内外的研究者的研究主要集中在如何通过加入扰动来产生彼此更加独立的分类器上,最近有研究者着手于对基分类器的选择上,并取得较好的效果。

对于 N 个基分类器构成的选择性集成,通过回归任务进行分析得到的泛化误差公式是:

$$E = \sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{C}_{ij}/N^2 \qquad (1)$$

其中 $\boldsymbol{C}_{ij}$ 表示基分类器之间的相关性。从该式可以看出基分类器之间的相关性越小,则集成分类器的泛化误差越小。从(1) 式不难推导出"坏的" 基分类器的条件:

$$(2N-1)\sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{C}_{ij} < 2N^2 \sum_{\substack{i=1 \\ i\neq k}}^{N} \boldsymbol{C}_{ik} + N^2 E_k \qquad (2)$$

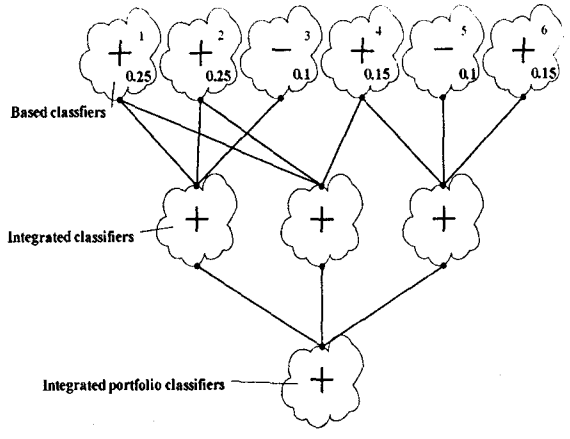
在实际情况中 $\boldsymbol{C}$ 往往是病态矩阵,甚至是不可逆的,因而研究者们多采用遗传算法来选择最佳的基分类器进行集成。而 Ada_ ens 则简单的采用基分类器在训练集上的准确率作为权重进行选择,再通过形成多个中间组合分类器来弥补选择上的欠佳,最后集成这些中间组合分类器的预测结果。这样大大地提高了算法的效率,也提高了算法的泛化性能。表 1 展示了 Ada_ ens 的伪码。

Ada_ ens 先通过构建不同类型的基分类器,如 J48、NaiveBayes、Adaboost. M1 等来提高基分类器之间的独立性,再通过各基分类器在训练数据集上的表现不断调整各基分类器用于集成的权重,按照权重将一定数量的基分类器集成起来形成中间层的组合分类器,如此形成多个中间层的组合分类器,在各组合分类器内部采用简单投票的方式产生输出,最后对这些组合分类器仍采用多数投票的形式产生最后的输出。由于集成学习分类器相对于单个分类器基本上有着更好的分类准确率,与普通的选择性集成相比,Ada_ ens 算法集成是个更高准确率的分类器,因而更容易有好的学习效果。同时由于各分类器中融入了大量的各种基分类器,因而在新的实例的情况下可能有更好的稳定性和泛化性能。

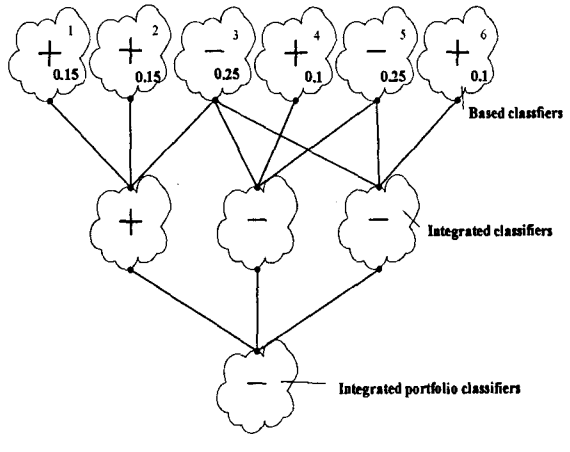
表 1 Ada－ens 集成学习算法

```
[1]  令 D 表示原始训练数据集,k 表示基分类器的个数,T 表示检验数据集,f 表示选择用于组合的分类器个数,h 表示组合分类器的个数
[2]  for i = 1 to k do
[3]  由 D 创建分类器 C1,C2,…,Ck
[4]  计算各分类器在 T 上 errorrate(i),并设置权重 W(i) = 1/(1 - errorrate(i))
[5]  end for
[6]  for j = 1 to h do
[8]  按照各基分类器的 W(i) 来选择 f 个基分类器形成一个组合分类器 F(i)
[9]  end for
[10] for 每一个检验记录 x ∈ T do
[11] F*(x) = vote(f1(x),f2(x)…,fh(x))
[12] end for
```

下面用一个简单的实例演示了 Ada_ ens 的形成过程和解释其可能存在的优点。



a)."+" is the correct classification.



b) "-" is the correct classification.

图 1 Ada－ens 算法比较

在图 1 中通过学习过程假设形成算法 a)、b),在

a)中"+"是某个正确的分类结果,可以看出通过简单的权重选择基分类器123形成了第一个组合分类器,且组合分类器的预测结果为"+",可以看出从基分类器到组合分类器弱化了预测错误分类器的作用而强化了正确分类器的作用。同样的在b)中其展现的是一种比较极端的情况下,如果简单的全部集成必将形成错误的预测。然而Ada_ens通过弱化错误分类和强化正确分类,得到了正确的结果"-"。

2 试验比较

试验程序是基于Weka[4]平台的,数据集试验了20个UCI机器学习数据库中的数据集。表2列出了各数据集的参数。数据集样本实例从57个到2310不等,其中实例在[0,500)的占60%,在[500,1000)的占30%,大于等于1000个实例的样本集占10%。

表2　UCI机器学习数据集参数

name	sizes	attribute	class
balance-scale	625	4	3
breast-cancer	277	9	2
breast-w	683	9	2
colic	368	22	2
credit-a	653	15	2
credit-g	1000	20	2
diabetes	768	8	2
glass	214	9	7
heart-c	296	13	5
heart-h	294	13	5
ionosphere	351	34	2
iris	150	4	3
labor	57	16	2
lymph	148	18	4
segment	2310	19	7
sonar	208	60	2
soybean	562	35	19
vote	232	16	2
vowel	990	13	11
zoo	101	17	7

试验中Ada_ens采用了如J48等10个分类器作为基分类器,每次集成5个分类器形成一个中间组合分类器,反复进行10次,形成10个中间组合分类器,并最终将这10个组合分类器进行集成形成最终分类器。Adaboost和Bagging均用J48作为基分类器。试验采用10次10倍交叉验证(10-Fold Cross Validation)的方法测试算法的准确率。表3展示了Ada_ens算法、普通单分类器(J48[5])、集成分类器(Adaboost[6],Bagging[7])在以上20个UCI数据集上的准确率。表4展示了三种算法的相互比较的结果,其中"win"表示明显好于,"equ"表示两者相当,"loss"表示明显差于。

表3　UCI数据集准确率

name	Ada_ens	Adaboost	Bagging	J48
balance-scale	83.8400%	78.8800%	82.2400%	76.6400%
breast-cancer	72.3776%	69.5804%	73.4266%	75.5245%
breast-w	96.2804%	95.7082%	95.8512%	94.5637%
colic	85.0543%	83.4239%	85.5978%	85.3261%
credit-a	86.9565%	84.2029%	85.3623%	86.0870%
credit-g	75.3000%	69.6000%	74.0000%	70.5000%
diabetes	75.5208%	72.3958%	74.0885%	73.8281%
glass	73.3645%	74.2991%	71.0280%	66.8224%
heart-c	83.4983%	82.1782%	79.2079%	77.5578%
heart-h	82.3129%	78.5714%	78.9116%	80.9524%
ionosphere	93.7322%	93.1624%	93.1624%	91.4530%
iris	94.6667%	93.3333%	95.3333%	96.0000%
labor	82.4561%	89.4737%	84.2105%	73.6842%
lymph	84.4595%	81.0811%	79.0541%	77.0270%
segment	97.1429%	98.4848%	97.4026%	96.9264%
sonar	77.8846%	77.8846%	74.5192%	74.5192%
soybean	94.1435%	92.8258%	93.2650%	91.5081%
vote	96.0920%	95.8621%	96.3218%	96.3218%
vowel	92.2222%	93.3333%	90.4040%	81.5152%
zoo	96.0396%	95.0495%	93.0693%	92.0792%

在表4中可以看出,Ada_ens在16个数据集上明显优于J48,占20个数据集的80%。在14个数据集上明显优于Bagging,占20个数据集的70%。在15个

表4　Ada_ens算法实验结果比较

name	Adaboost	Bagging	J48
balance-scale	win	win	win
breast-cancer	win	loss	loss
breast-w	win	win	win
colic	win	loss	loss
credit-a	win	win	win
credit-g	win	win	win
diabetes	win	win	win
glass	loss	win	win
heart-c	win	win	win
heart-h	win	win	win
ionosphere	win	win	win
iris	win	loss	loss
labor	loss	loss	win
lymph	win	win	win
segment	loss	loss	win
sonar	equ	win	win
soybean	win	win	win
vote	win	loss	loss
vowel	loss	win	win
zoo	win	win	win

(下转第94页)

的旅行距离这两个目标，结合该问题本身的特点，受文献[8]的启发提出用信息素相互独立，但又通过交换信息来协作的混合蚁群算法。当 MMAS－VEI 算法找到可行解以后，通过使用 MMAS－TIME 算法并结合 2－opt 局部搜索来提高可行解的质量，在算法运行过程中信息素被限定在 $[\tau_{min}, \tau_{max}]$ 之间，有利于防止某条路径信息素聚集过高，从而减少计算时间以避免过早收敛。通过使用含 2－opt 和不含 2－opt 局部搜索机制分别对 Solomon 的测试数据进行实验，并与文献[7]中实验做了对比，结果表明了文中算法的有效性。

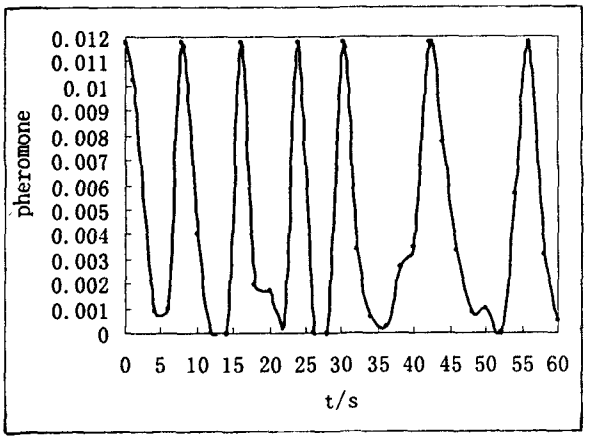


图 2 算法运行时间和信息素的值之间的关系

参考文献：

[1] Dantizig G, Ramser J. The truck dispatching problem[J]. Management Science, 1959, 6(1): 80－91.

[2] Dorigo M. Optimization, Learning and Natural Algorithms [D]. Milan, Italy: Dipartimento di Elettronica, Politecnio di Milano, 1992.

[3] Stützle T, Hoos H H. MAX－MIN Ant System[J]. Future Generation Computer Systems, 2000, 16: 889－914.

[4] Dorigo M, Stützle T. 蚁群优化[M]. 张 军，胡晓敏，译. 北京：清华大学出版社，2007.

[5] 赵传信，张雪东，季一木. 改进的粒子群算法在 VRP 中的应用[J]. 计算机技术与发展，2008，18(6)：240－242.

[6] 刘志硕，申金关. 车辆路径问题的混合蚁群算法设计与实现[J]. 管理科学学报，2007，10(3)：15－22.

[7] 万 旭，林健良，杨晓伟. 改进的最大－最小蚂蚁算法在有时间窗车辆路径问题中的应用[J]. 计算机集成制造系，2005，11(4)：572－576.

[8] Gambardella L M, Taillard E, Agazzi G. MACS－VRPTW: A Multiple Ant Colony System for Vehicle Routing Problems with Time Windows[M]. In: Corne D, Dorigo M, Glover F. New Ideas in Optimization. UK: McGraw－Hill, 1999: 63－76.

[9] Flood M M. The Traveling Salesman Problem[J]. Operations Research, 1956(4): 61－75.

[10] Solomon M. Algorithms for the Vehicle Routing and Scheduling Problem with Time Window Constraints[J]. Operations Research, 1987, 35: 254－365.

[11] 马 良，朱 刚，宁爱兵. 蚁群优化算法[M]. 北京：科学出版社，2008.

(上接第 89 页)

数据集上明显优于 Adaboost 占 75%，1 个数据集上持平占 5%。可以分析出 Ada_ens 在分类准确率、泛化性能、稳定性上都表现出较好的性能。

3 结束语

文中在选择性集成学习的集成上提出了多层次选择性集成算法 Ada_ens，针对决策树与神经网络模型在 20 个标准数据集对集成学习算法 Ada_ens 进行了实验研究，试验证明基于数据的集成学习算法的性能优于基于特征集的集成学习算法的性能，有更好的分类准确率和泛化性能。

下一步工作中着重需要解决的问题：与目前集成学习相似试验的参数均是认为设定的希望通过理论推导或试验分析得出最佳的参数设置和理由；用数学理论佐证为什么选择性集成会更有效[8]，在什么情况下更有效；在真实的情况下测试 Ada_ens 的效果，并依次对其进行改进和完善。

参考文献：

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

[2] 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.

[3] Dietterich T G. Ensemble methods in machine learning[C]// In: Proceedings of the First International Workshop on Multiple Classifier Systems. Cagliari, Italy: [s. n.], 2000: 1－15.

[4] 唐耀华，高静怀. 一种新的选择性支持向量机集成学习算法[J]. 西安交通大学学报，2008(10)：1221－1225.

[5] 蒋望东，林士敏. 基于选择性集成遗传算法的 BNC 结构学习[J]. 计算机辅助工程，2006(3)：46－50.

[6] 王正群，陈世福. 一种主动学习神经网络集成方法[J]. 计算机研究与发展，2005(3)：375－380.

[7] 王建敏，李铁军. 基于神经网络集成学习的智能决策支持系统构建[J]. 电脑知识与技术，2008(9)：2045－2050.

[8] 李 凯，崔丽娟. 集成学习算法的差异性及性能比较[J]. 计算机工程，2008(3)：35－37.