第36卷 第14期 计 算 机 工 程 2010年7月
Vol.36 No.14 Computer Engineering July 2010
·人工智能及识别技术· 文章编号：1000—3428(2010)14—0163—03 文献标识码：A 中图分类号：TP391.4

一种分类器选择方法

牛 鹏，魏 维，李峻金，郭建国
(西安通信学院研究生管理大队，西安 710106)

摘 要：在按照“测试-选择”方法设计多分类器系统时，从超量生成的候选分类器集中选取一个最优子集是关键环节之一。基于此，定义一个组合适宜度概念，提出一种新的分类器选择方法。将该方法用于高光谱遥感数据分类实验中，并从具有27个候选的分类器集中挑选子集。实验结果表明，该方法在选择效率和识别精度方面具有优势，能保证所选子集的泛化能力。

关键词：组合适宜度；分类器选择；高光谱数据

Classifier Selection Method

NIU Peng, WEI Wei, LI Jun-jin, GUO Jian-guo
(Management Unit of Graduate Student, Xi'an Communications Institute, Xi'an 710106)

【Abstract】 Selecting an optimal subset of classifiers from overproduced candidates is a key step when the "test-and-select" method is employed to design a multiple classifier system. This paper therefore defines a new concept——Degree of Combination Fitness(DCF) and presents a new DCF-based classifier selection method. In its application to experiments of hyperspectral data classification, the new method selects a subset from a pool of twenty-seven candidate classifiers. Results show that in comparison to other popular methods, the proposed method has advantages both in efficiency and recognition accuracy. Besides, it can guarantee the generalization ability of the selected set as well.

【Key words】 Degree of Combination Fitness(DCF); classifier selection; hyperspectral data

1 概述

多分类器系统是模式识别领域的一个发展方向，近年来取得很大进展。设计多分类器系统时采用测试-选择方法[1]是一种常用的解决方案。其基本思想是首先超量生成一个很大的候选分类器集，然后从中选择一个最优的分类器子集。

分类器子集并非越大越好。分类器过多不仅增加计算复杂度，分类准则上的差异还可能降低系统的识别率。关于分类器最优子集的选择方法已有很多研究，大致可以分为2类：基于不精准多样性的搜索法和直接搜索法[2]。前者选出的分类器子集的多样性在某种标准下达到最好，后者则要求选出的分类器子集在验证集上的分类精度最高。文献[3]提出了2种启发式选择方法：第1种方法(后面用BestN表示)要固定分类器子集的个数 N，然后按照分类精度由大及小从候选分类器集中选取前 N 个；另外一种方法(后面用BestClass表示)从每一种类型的分类器中选取分类精度最高的单分类器构成分类器子集。文献[4]根据多样性先将分类器聚类，然后根据与其他类别多样性度量平均距离最大的原则，从各类中选取一个成员，构成分类器子集。文献[5]提出一种利用聚类算法去除冗余分类器的选择方法，该方法选取每簇中靠近聚类中心的模型构成分类器子集。文献[6]对基于不同的多样性标准的分类器选择做了实验比较。文献[7]提出一种基于IWCECR差异性度量指标的启发式搜索方法。文献[8]采用遗传算法选择分类器。

基于不精准多样性的搜索法在搜索时主要考虑分类器集多样性，减少了计算量但不能保证选出的分类器子集的精度；直接搜索法以分类器子集在验证集上的分类精度最高为准则，但一般伴随着巨大的搜索空间和复杂的计算量，而且不一定能保证分类器子集在测试集上的泛化能力。

本文提出了一种新的搜索算法，在构建分类器子集时既考虑了分类器集的多样性，又考虑了单分类器的精确性。为了描述本文的算法，提出了一个新的概念：组合适宜度。实验表明，该算法在搜索效率、分类精度和泛化能力方面都具有优势。

2 基于组合适宜度的选择方法

2.1 成员分类器的选择要求

在构建多分类器系统时，总是希望成员分类器的精度要高而且分类器集的多样性要好。成员分类器的精度太低，组合出的多分类器系统的精度也不会很高；分类器的分类结果应具有多样性，如果成员分类器对相同的样本产生同样的分类错误，那么组合分类器系统的精度将不可能有任何提高。

文献[9]以回归学习的集成推导出重要的集成学习的泛化误差公式，这个公式对于分类器的组合有着同样的意义。对于 N 个学习机，它们的组合误差公式如下：

$$E=\overline{E}-\overline{A}$$

其中，$\overline{E}$ 为 N 个学习机的绝对误差的加权平均；$\overline{A}$ 为 N 个学习机相对于组合的误差的加权平均。$\overline{E}$ 指示出学习机固有的误差，$\overline{A}$ 指示出这些学习机之间的差异。该式表明要获得好的组合效果就需要降低成员学习机的误差并增加学习机间的差异。基于这种考虑，本文提出了一种新的基于组合适宜度的搜索算法。

2.2 组合适宜度

度量一个分类器集的多样性有多种方法[6]，主要分为成对度量法和非成对度量法2类，输出相关系数是成对度量方

作者简介：牛 鹏(1985－)，男，硕士研究生，主研方向：模式识别；魏 维，教授、博士；李峻金，硕士研究生；郭建国，学士

收稿日期：2010-01-20 **E-mail：**jdnp@tom.com

法的一种。给定单分类器 C_i 和 C_j，设样本集 S 中 C_i 与 C_j 同时正确分类的样本数为 a，C_i 正确而 C_j 错误分类的样本数为 b，C_i 错误而 C_j 正确分类的样本数为 c，2 个分类器同时分类错误的样本数为 d，则 2 个分类器的输出相关系数表示为

$$\rho_{ij}=\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \quad -1\le\rho_{ij}\le 1$$

当 2 个分类器正确分类的样本与错误分类的样本相同时 ρ 为 1，若 ρ 为负值表明 2 个分类器错分的样本不一样。一个分类器集的多样性可以用两两单分类器之间的输出相关系数的平均值来度量，平均值越小，分类器集的多样性越好。

为了描述本文的算法，引入一个新的概念组合适宜度(Degree of Combination Fitness, DCF)来表征一个单分类器与一个分类器集的组合适宜程度，定义如下：

$$D=e^{R}/e^{C\bar{\rho}}$$

其中，R 表示单分类器在验证集上的总体分类精度；$\bar{\rho}$ 表示单分类器与分类器集中每个分类器的输出相关系数的平均值，C 是大于 0 的常数。当 R 越大 $\bar{\rho}$ 越小时，单分类器与分类器集的组合适宜度就越大。

2.3 算法实现

本文算法的具体实现如下：

输入 候选分类器集 T

输出 选出的分类器子集 S_o

```
初始化：N=T 中分类器的个数;S[1, 2, …,N]= φ ;
计算 T 中每个分类器 C 在验证集上的分类精度;
计算 T 中两两分类器之间的 ρ ;
找出分类精度最大的分类器 Cm;
S1={Cm};T=T-S1;
for i=1:(N-1)
max=0;
    for 对于 T 中的每一个分类器 C
        计算 C 与 Si 的组合适宜度 D;
        if D>max
           max=D;
           G=C;
        end if
    end for
 Si+1=Si∪G;
 T=T-G;
end for
mmax=0;
for i=1:N
   计算 Si 在验证集上的总体分类精度 OA;
    if   OA>mmax
         mmax=OA;
         So=Si;
    end if
end for
return So;
```

本算法无需迭代，一次搜索即可完成，大大减少了计算量。组合适宜度的定义既考虑了成员分类器的精度，又考虑了分类器集的多样性，而且算法也考虑了组合分类器的分类精度。实验表明，与其他算法相比，本算法既能保证所选分类器集的分类精度，又能保证组合分类器的泛化能力，而且效率较高。

3 实验结果

本文实验是在高光谱遥感数据上进行的，原始图像 92AV3C(ftp://ftp.ecn.purdue.edu/ biehl/MultiSpec)取自 1992 年 6 月拍摄的美国印第安纳州西北部遥感试验区的一部分，它包含 220 个波段。首先去除含噪声较大的 20 个波段，然后用自适应波段选择的方法降为 90 维，并选取像素数目较多的 9 类作为研究对象。各个样本集中的像素随机选取，训练集、验证集、测试集包含像素数分别为 4 674, 900, 3 771。

实验使用 27 个成员分类器：4 种结构的 BP 神经网络，在不同的初始参数条件下每种结构训练 3 次共 12 个分类器($C1$~$C12$)；12 个初始参数不同的 PNN 分类器($C13$~$C24$)；3 个 KNN(k=7,9,11)分类器($C25$~$C27$)。组合分类器采用相对多数投票法。各分类器在不同样本集上分类精度如表 1 所示。实验中机器配置为：Celeorn 2.4 GHz，内存 1 GB，操作系统 Windows XP SP2。实验环境为 Matlab7.01。

表 1　27 个分类器在不同样本集上的分类精度

分类器	验证集/(%)	测试集/(%)
$C1$	77.11	79.16
$C2$	75.33	77.70
$C3$	74.56	78.25
$C4$	75.22	78.92
$C5$	78.78	80.48
$C6$	76.67	79.18
$C7$	69.78	75.07
$C8$	65.22	69.61
$C9$	53.78	63.01
$C10$	72.22	75.66
$C11$	74.00	77.43
$C12$	58.56	68.10
$C13$	80.56	78.65
$C14$	81.67	79.95
$C15$	81.78	80.11
$C16$	81.89	80.14
$C17$	82.11	80.43
$C18$	82.33	80.80
$C19$	82.11	81.44
$C20$	82.67	81.84
$C21$	82.44	82.23
$C22$	82.11	82.07
$C23$	82.00	81.94
$C24$	81.22	81.68
$C25$	79.22	80.91
$C26$	79.33	80.64
$C27$	78.44	79.90

实验将本文算法与几种常见的分类器子集选择方法进行了比较。常见的搜索算法有：BestN，BestClass，前向搜索法，后向搜索法，遗传算法，聚类选择法，穷举法以及文献[7]中的启发式算法等。聚类选择算法一般先将分类器聚类，然后从每簇中选取一个代表参与组合，这种方法可以去除相似度高的冗余个体，保留差异度大的个体，但缺乏对单个分类器精度的考虑。穷举法是种全局最优解的搜索法，对包含 N 个分类器候选分类器集其搜索次数将达到 2^N-1 次。在本文的机器配置和实验环境中，穷举法的搜索时间预估为 2 000 多个小时，而且穷举法是一种直接搜索法，在验证集上得到的最优选择在测试集上不一定仍然是最优选择，因此，本文未给出穷举法的实验结果。遗传算法是一种经典次优搜索法，也是一种直接搜索法，本文遗传算法参数设置为：选用 27 位的二进制字符串(1 代表参与组合，0 代表不参与)，初始群体个数为 30，群体代数设置为 500，变异率为 0.01，目标函数为分类器子集在验证集上的总体分类精度的相反数。实验结果见表 2，其中，前向搜索 1、前向搜索 2 分别表示从最好和随机的分类器开始搜索。从表 2 的实验结果可以看出，验证集上和测试集上的分类精度并不是简单的正比关系，这与组合分类器的泛化能力有关。

表 2　组合分类器的分类精度

选择方法		CPU 运行时间/s	挑选出的分类器个数	验证集/(%)	测试集/(%)
遗传算法		92.0	14	83.67	82.15
前向搜索法 1		5.0	2	83.56	81.89
前向搜索法 2		9.0	4	83.67	82.21
启发式算法[7]		70.0	4	83.67	82.29
BestClass		0.2	6	79.00	81.15
聚类选择[4]		74.0	6	83.44	82.82
BestN		0.2	3	82.67	81.86
		0.2	5	82.11	81.41
		0.2	7	82.56	81.78
		0.2	9	82.00	81.44
本文算法	C=0.00	39.0	16	83.11	81.94
	C=0.20	39.0	6	83.89	82.37
	C=0.25	40.0	8	83.44	82.79
	C=0.30	39.0	6	83.56	82.60
	C=0.50	39.0	6	83.56	82.63
	C=0.60	39.0	6	83.33	82.87
	C=1.00	39.0	1	82.67	81.84
	C=10.00	39.0	1	82.67	81.84

在 C 取值合适时(如 0.30)，本文算法在构造分类器子集时能兼顾分类器集的多样性和成员分类器的精确性，组合分类器在验证集和测试集上的分类精度均比较高；当 C 取值较大时(如 10.00)，构造分类器子集时只考虑成员分类器之间的多样性，组合效果比较差；当 C 取值较小时(如 0.00)，构造分类器子集时只考虑成员分类器的精确性，组合效果也不理想。在 C 取适当的值时，本文的算法在分类精度、选择效率和泛化能力等多方面均具有一定的优势。

4　结束语

本文提出了一种基于组合适宜度的分类器选择方法，该方法在搜索时综合考虑了分类器集的多样性和成员分类器的精度以及组合分类器的精度。与其他常见的算法相比，本文的方法在选择效率和识别精度等方面均具有优势，而且能够保证泛化能力。

针对本文方法，下一步的研究工作将放在如何根据待分类的数据的特性来确定合适的参数 C，以便使该方法具有更广泛的应用价值。

参考文献

[1] Sharkey A J C, Sharkey N E, Gerecke U, et al. The “Test and Select” Approach to Ensemble Combination[C]//Proceedings of the 1st International Workshop on Multiple Classifier Systems. [S. 1.]: Springer-Verlag, 2000: 30-44.

[2] Ruta D, Gabrys B. Application of the Evolutionary Algorithms for Classifier Selection in Multiple Classifier Systems with Majority Voting[C]//Proceedings of MCS’01. [S. 1.]: Springer-Verlag, 2001: 399-408.

[3] Partridge D, Yates W B. Engineering Multiversion Neural-net Systems[J]. Neural Computation, 1996, 8(9): 869-893.

[4] Giacinto G, Roli F. An Approach to the Automatic Design of Multiple Classifier Systems[J]. Pattern Recognition Letters, 2001, 22(1): 25-33.

[5] 李国正, 杨　杰, 孔安生, 等. 基于聚类算法的选择性神经网络集成[J]. 复旦大学学报, 2004, 43(5): 689-691.

[6] Aksela M. Comparison of Classifier Selection Methods for Improving Committee Performance, Multiple Classifier Systems[Z]. (2003-10-02). http://www.cis.hut.fi/projects/hcr/aksela_mcs2003. pdf.

[7] 郝红卫, 陈志强. 一种新的启发式分类器选择方法[J]. 计算机工程, 2008, 34(2): 206-208.

[8] Sirlantzis K, Fairhurst M C, Hoque M S. Genetic Algorithms for Multi-classifier System Configuration: A Case Study in Character Recognition[C]//Proceedings of MCS’01. [S. 1.]: Springer-Verlag, 2001: 99-108.

[9] Krogh A, Vedelsby J. Neural Network Ensembles, Cross Validation and Avtive Learning[C]//Proceedings of ANIPS’95. [S. 1.]: MIT Press, 1995: 231-238.

编辑　索书志

(上接第 162 页)

5　结束语

本文采用最大熵分割法和肤色模型对人脸图像进行人眼定位。仿真实验表明，本方法可以准确、快速地定位人眼，对于不同的光照、背景、头部偏转旋转、面部表情等都具有良好的鲁棒性。但由于作为约束条件的肤色模型易受到背景中类肤色信息的影响，对检测出的人眼候选区域进行准确定位时存在一定的影响。如何改进约束条件，对人眼候选区域更好地进行判断、定位将是下一步的研究工作。

参考文献

[1] Sobottka K, Pitas I. A Fully Automatic Approach to Facial Feature Detection and Tracking[C]//Proc. of the 1st International Conference on Audio and Video Based Biometric Person Authentication. London, UK: Springer-Verlag, 1997: 77-84.

[2] 张娜娜, 马　燕. 基于灰度投影函数的眼睛定位方法[J]. 计算机工程, 2006, 32(10): 193-195.

[3] Yuille A L, Hallinan P W, Coben D S. Feature Extraction from Faces Using Deformable Templates[C]//Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Press, 1989: 104-109.

[4] Heisele B, Serre T, Pontil M, et al. Categorization by Learning and Combining Object Parts[C]//Proc. of Advances in Neural Information Processing System. Vancouver, Canada: [s. n.], 2002: 1239-1245.

[5] Deng J Y, Lai Feipei. Region-based Template Deformation and Masking for Eye-feature Extraction and Description[J]. Pattern Recognition, 1997, 30(3): 403-419.

[6] Yang Peng, Du Bo, Shan Shiguang, et al. A Novel Pupil Localization Method Based on Gabor Eye Model and Radial Symmetry Operator[C]//Proc. of International Conference on Image Processing. [S. l.]: IEEE Press, 2004: 67-70.

[7] 李　嵩, 刘党辉, 沈兰荪. 基于 Gabor 变换的人眼定位方法[J]. 测控技术, 2006, 25(5): 27-29.

[8] 彭进业, 俞卞章, 王大凯, 等. 多尺度对称变换及其应用于定位人脸特征点[J]. 电子学报, 2002, 30(3): 363-366.

[9] 胡步发, 王　忠. 基于对称变换与自评估的人眼定位新方法[J]. 电路与系统学报, 2006, 11(1): 133-137.

[10] 李粉兰, 徐可欣. 一种应用于人脸正面图像的眼睛自动定位算法[J]. 光学精密工程, 2006, 14(2): 320-326.

[11] 周德龙, 潘　泉, 张洪才, 等. 最大熵阈值处理算法[J]. 软件学报, 2001, 12(9): 1420-1422.

[12] Hsu R L. Face Detection in Color Images[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002, 24(5): 696-706.

[13] Soille P. 形态学图像分析原理与应用[M]. 2 版. 王小鹏, 译. 北京: 清华大学出版社, 2008.

编辑　顾逸斐