

基于相关性和有效互补性分析 的多分类器组合方法¹⁾

荆晓远 杨静宇

(南京理工大学计算机系模式识别和机器视觉实验室 南京 210094)

(E-mail: yangjy@mail. njust. edu. cn)

摘要 定义了分类器组合中的相关向量和有效互补性的概念,并提出了一种新的组合准则,即最大有效互补准则. 对人脸图象作正交小波变换,得到它在不同频带上的四个子图象,然后分别提取奇异值特征. 实验表明,这四组特征之间以及相应的分类结果之间的相关性都较小,组合结果明显优于原始图象的奇异值特征的分类效果,并优于常用的组合方法——计分法的效果.

关键词 小波变换, 人脸识别, 奇异值特征, 多分类器组合, 相关性, 有效互补性.

COMBINING CLASSIFIERS BASED ON ANALYSIS OF CORRELATION AND EFFECTIVE SUPPLEMENT

JING Xiaoyuan YANG Jingyu

(Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094)

Abstract We define the correlative vector and effective supplement for classifiers combination, and bring forward a new combination rule, ie, maximal effective supplement rule. We do orthogonal wavelet transform of a face image, get its four sub-images of different frequency bands, then respectively extract their singular value features. We find in experiment that the correlation within these four feature groups and the correlation of their sorted results are all small, and that the combination results are obviously superior to the classification results of singular value features of the initial images and superior to the commonly used mark-counting combing method. We combine the classifiers with linear weights and use genetic algorithm to train the confident weights of every classifier.

Key words Wavelet transform, face recognition, singular value feature, classifiers combination, correlation, effective supplement.

1) 国家自然科学基金(69672013)和国家教委博士点基金资助项目.

收稿日期 1998-08-03 收修改稿日期 1999-01-25

1 引言

近年来,多分类器的组合方法已成为模式识别领域的前沿研究课题,并在模式识别的很多应用方面,如字符识别、目标识别等领域取得了较好的应用效果.在已提出的许多方法中,常用的有传统的择多判决法(如投票表法、计分法等),根据后验概率的线性加权法,贝叶斯估计,证据推理法,模糊推理法,以及将分类结果作为一种新的输入特征的神经网络组合方法.另外,还有一些组合的策略,如多级分类方法.

选择合适的组合准则是分类器组合的首要问题.常见的准则如采用少数服从多数规则,最小错误率规则等.分类器组合的方法虽然不少,但是到目前为止仍然不太成熟,在分类器的组合中一个重要的观点是,分类器之间存在的差异对最后的组合结果非常有用,因为差异本身就是构成互补的必要条件.因此,如何有效地利用这种差异性,设计出合理、通用的组合准则是该领域研究中的重点问题,同时也是难点问题.需要指出的是,通常模式识别理论中的分类器概念是指分类的方法.而在多分类器组合中所采用的分类器概念,一般是指选用不同的特征和不同的分类方法进行模式分类,得到的多种分类结果,它包含的范围更加广泛,本文采用这种广义上的分类器概念.

2 基于小波变换的人脸图象的奇异值特征提取

人脸图象识别是一项有着重要的理论研究价值与应用价值和极具挑战性的课题,综述^[1]比较全面地阐述了已有的各项研究成果,它指出代数特征反映了图象的内在属性,是一种本质特征.文献[2]用它来作人脸图象的分类和识别,取得了一定的效果.但是在以前的工作中,很少有针对代数特征去做图象的处理,而是直接对原始图象矩阵做各种代数变换,来提取代数特征;并且对于人脸图象而言,同一个人的不同人脸样本变化不大.当人脸的姿态和表情等发生较大变化时,实验表明奇异值特征的分类效果不是很理想.因此,我们除了选择合适的分类器以外,需要对人脸图象做合适的处理.

在实验中,我们从南京理工大学模式识别与机器视觉实验室的人脸图象库 FDB603 中抽取了18个人的脸部图象.在建立人脸图象库时,我们让每个人坐在固定的位置上,摄像机与人脸的相对位置和光照条件略有变动,脸部表情和姿态等有较大变化.每人拍12次,得到大小为 64×64 ,256级灰度的图象,这样共有 $18 \times 12 = 216$ 个样本.图1给出了其中一个人的样本图象集.

小波分析由于其优良的时频域局部性能而成为有力的信号分析处理工具.我们选择由多尺度分析获得的标准正交小波基,如具有紧支集的 Daubechies 正交小波基,来对 $64 \times 64 \times 8$ 灰度级的人脸图象做离散小波变换,其中 Daubechies 正交小波的滤波器系数为 $h(n) = \{0.542, 0.307, -0.035, -0.078, 0.023, -0.030, 0.012, -0.013, 0.006, 0.006, -0.003, -0.002\}$.这样就得到四个子图象(包括低分辨率子图象,水平方向和垂直方向子图象,对角方向子图象).对四个 32×32 大小的子图象分别计算奇异值特征,得到四个32维的特征向量.由于它们的特征维数相等,采用下式来表示特征向量之间的相关性

$$R_{xy} = |(E(\mathbf{x}^T \cdot \mathbf{y}) - E(\mathbf{x})^T \cdot E(\mathbf{y}))|, \quad (1)$$



图1

其中 \mathbf{x} 和 \mathbf{y} 是两个相同维数的随机向量. 这样, 就得到四个子图象的奇异值特征的相关程度, 归一化到 $[0, 1]$ 区间. 实验表明, 低频子图象与三个高频子图象之间相关性较大, 而三个高频子图象之间相关性较小. 另外, 我们也提取了原始图象矩阵的傅氏频谱的幅值特征. 不论我们的实验结果, 还是文献[3]的结果都表明, 傅氏频谱的幅值特征具有良好的分类性能. 在实验中, 我们对原始图象矩阵做快速离散傅立叶变换, 得到频谱的幅值矩阵.

实验也表明, 奇异值特征和频谱特征之间相关性较小. 但是我们认为, 特征之间的相关性并不能代表分类结果之间的相关性. 因为对于相同的特征集, 用不同的分类方法进行分类, 所得到的分类结果之间的相关程度通常不一样. 因此, 对于多分类器组合而言, 除了选用相关性较小的特征来分类以外, 有必要进一步分析分类结果之间的相关性大小. 下面, 我们从分类结果的角度来分析.

3 分类器的相关性分析

为了便于说明, 先对有关的表示进行定义. 假设: 1) 类别数用 c 表示; 2) 分类器的个数用 k 表示; 3) 每个分类器的分类结果用 $\mathbf{e}_i (i=1, 2, \dots, k)$ 来区分, 且 $\mathbf{e}_i = (e_{i_1}, e_{i_2}, \dots, e_{i_c})$; 4) 输入样本用 \mathbf{x}, \mathbf{y} 表示, $\mathbf{x} = (x_1, x_2, \dots, x_k), \mathbf{y} = (y_1, y_2, \dots, y_k)$.

模式识别理论中评价向量之间的相关程度有多种方法, 主要有基于向量之间距离的和基于向量之间夹角的度量方法. 但作者认为, 这些方法用于分类结果之间的相关性度量并不合适. 理由是分类器产生的结果就是要找到最有可能属于的候选类别. 我们将分类结果向量的各分量按从大到小或从小到大的顺序依次排列(根据实际情况而定), 第一个分量对应于第一个候选类别, 依此类推. 如果两个分类结果向量 \mathbf{e}_i 和 \mathbf{e}_j 的第一个分量对应着不同的候选类别, 那么它们的相关性比较就没有意义, 即该分量的相关程度为 0. 如果 \mathbf{e}_i 和 \mathbf{e}_j 的第一个分量对应着相同的候选类别, 且大小相等(当分量值做了归一化之后), 则相关程度为 1. 所以相关值落在 $[0, 1]$ 之间.

用 r_t 来表示 \mathbf{e}_i 和 \mathbf{e}_j 的第 t 个分量之间的相关值, $t=1, 2, \dots, c$, 定义如下:

$$r_t = 1 - \left| \frac{e_{i_t}}{\sum_{s=1}^c e_{i_s}} - \frac{e_{j_t}}{\sum_{s=1}^c e_{j_s}} \right| \left/ \left(\frac{e_{i_t}}{\sum_{s=1}^c e_{i_s}} + \frac{e_{j_t}}{\sum_{s=1}^c e_{j_s}} \right) \right., \quad (2)$$

其中 $e_{i_t} / \sum_{s=1}^c e_{i_s}$ 表示对分量做归一化操作, 分量之间的差除以分量之间的和表示了它们的相对差异程度, 而不是绝对的差异程度. 其好处在于将度量尺度归一化, 即对于分量值

较大的场合和分量值较小的场合使用统一的衡量标准.

文献[4]认为存在以下经验知识:对输入样本 x , 分类器 e_i 输出的较低阶次候选类别(如第一、第二候选等)一般与 x 有较大的相关性(误识情况除外);而较高阶次的候选类别与 x 的相关性较小. 在这里,我们采用类似的方法:分类结果向量 e_i 输出的各阶候选对相关程度的支持量是不同的,即候选阶次越高,支持作用越小. 令 w_t 表示第 t 阶候选的支持因子, w_t 可有多种形式,如 $w_t = e^{-\alpha(t-\delta)}$ 或 $w_t = 1.0 - \beta \times t$, 其中 α, δ, β 为非负常数.

在实验中,采用如下形式效果较好

$$w_t = e^{-t}/e^{-1}, \quad (3)$$

这里 $\alpha=1, \delta=0, t=1, 2, \dots, c$, 除以 e^{-1} 表示对 w_t 做归一化. w_t 的前10个分量依次为 $w_t = (0.6321, 0.2325, 0.0855, 0.0315, 0.0116, 0.0043, 0.0016, 0.0006, 0.0002, 0.0001, \dots)$.

可见,第一、第二候选所占权重很大,尤其是第一个候选分量.

定义1. 分类器 e_i 和 e_j 的相关程度 $R_{i,j}$ 定义为

$$R_{i,j} = \sum_{t=1}^c w_t \cdot r_t, \quad (4)$$

于是就得到测试样本 x 的各分类器之间的相关值,用相关矩阵 R_x 表示如下:

$$R_x = \begin{bmatrix} 1 & R_{1,2} & \cdots & R_{1,k} \\ R_{2,1} & 1 & \cdots & R_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ R_{k,1} & R_{k,2} & \cdots & 1 \end{bmatrix}. \quad (5)$$

显然, R'_x 是一个实对称矩阵,为计算方便起见,取它的上三角部分,将其表示为向量的形式.

定义2. 分类器组合的相关向量 R_x 定义为

$$R_x = [R_{1,2}, R_{1,3}, \dots, R_{1,k}, R_{2,3}, R_{2,4}, \dots, R_{2,k}, \dots, R_{k-1,k}]. \quad (6)$$

4 基于有效互补性的多分类器组合方法

互补信息虽然描述了同一对象的不同方面,但并不是对分类器的组合都有益处的,只有能够提高组合效果的互补信息才是有用的. 我们使用有效互补性的概念来表示分类器之间有利于组合的相互作用. 首先分析样本正确分类的程度. 众所周知,最小错误率是最常用的判别规则,具有最直接的实际分类意义. 但作者认为,用最小错误率准则来表示组合结果的正确与否,是一种硬性的评价方法. 如果有两个组合结果 A 和 B , 它们的第一个候选类别相同,最小错误率准则就认为它们同样正确或者同样错误,不妨设为都正确. 但是,若 A 的第一候选与第二候选非常接近,而 B 相差较大,则显然 B 的正确程度要高于 A . 所以,如果使得组合过程中样本权值的训练更倾向于 B , 即采用不确定性的方法来表示正确的程度,那么所得到的权值参数抗噪声能力会更强,也就是容错性和推广能力更高.

下面给出一种评价方法. 设组合结果中,第一候选分量为 x_1 , 第二候选分量为 x_2 , 分量的平均值为 \bar{x} , 客观上正确的候选分量为 x_t , 其中 $t=1, 2, \dots, c$, c 为类别数. 不妨设 $x_1 > x_2 \geq \dots \geq x_c$, 正确程度为 G , 其中 $x_1 > x_2$, 否则 $G=0$. 对以下两种情况:

1)若 $t=1$,即组合结果正确,则如果第一候选分量和第二候选分量之差,相比于它和分量的平均值之差越大,则正确的程度越高,有

$$G = (x_1 - x_2)/(x_1 - \bar{x}). \quad (7)$$

显然, $G>0$.

2)若 $t \neq 1$,即组合结果不正确,则 $G=0$.

我们可以用相关向量来表示分类器之间的不相关程度.设 R_x 的最大值为

$$R_0 = [1,1,\dots,1], \quad (8)$$

分类器之间的整体互补度可用 N 表示为

$$N = \frac{\|R_0 - R_x\|}{\|R_0\|}. \quad (9)$$

定义3. 分类器之间的有效互补程度为

$$E = G \cdot N^a. \quad (10)$$

定义3说明,有效互补性是由正确程度 G 和整体互补度 N 决定的,其中 a 为 $[0,1]$ 之间的一个小数,用它来权衡 G 和 N 之间对于 E 的相对重要性,在实验中取 a 为 0.5.有效互补性是描述分类器之间的相对作用,而不是单个分类器的某种性能.

在日常生活中,人们对于各种意见进行综合决策时,最常用的思路就是针对每个具体问题,给出各种意见的权值.相比无反馈的分类器组合算法而言,用有监督的学习方法(如遗传算法)训练权值参数,可以得到更好的识别结果.本文用二进制编码来表示权值参数,并且提出了一种指导训练的组合格则.

定义4. 最大有效互补准则是指对于所有训练样本而言,加权后的各分类器之间有效互补程度的总和 J 越大越好,即

$$J = \max\left(\sum_{i=1}^n E_i\right) = \max\left(\sum_{i=1}^n (G_i \cdot N_i^a)\right), \quad (11)$$

其中 n 代表训练样本的总数.因此,使用 J 来表示染色体的适应度值.

对每个分类器赋予一个可信度权值,通过遗传算法训练求得.这样得到的一组权值反映了分类器之间整体上的有效互补程度,具有统计意义.并且,对每一个训练错误的样本再单独进行训练,得到这个样本点所对应的一组权值.对于测试样本而言,当它和整体训练出错的样本的最小距离小于某一阈值时,就选择最接近的那个样本所对应的权值进行组合,否则就使用整体训练得到的权值进行组合.这样做的好处在于,既考虑到训练样本集的统计性质,又顾及到个别样本的情况.在实验中,距离用欧氏距离度量,阈值取为 0.05.

5 实验结果

在模式识别的应用中,以下三种分类方法被广泛使用:1)对特征直接做 K 近邻分类,这里采用最近邻分类法;2)对特征做最佳鉴别变换后,再用最近邻方法进行分类;3)对特征用 BP 神经网络进行分类.分别用 $C1 \sim C3$ 来表示.将现有特征用 $F0 \sim F5$ 来表示,其中 $F0$ 对应原始图象的奇异值特征, $F1$ 、 $F2$ 、 $F3$ 和 $F4$ 分别对应低频子图象、水平方向子图象、垂直方向子图象和对角方向子图象的奇异值特征, $F5$ 对应原始图象频谱的幅值特征.

为了比较和其它组合算法的实验效果,本文采用一种常用的组合算法——计分法作为对比标准.在实验中使用的计分法是将每一种类别在各个分类结果中的排名相加,例如第一名得1分,第 k 名则得 k 分,从而得到该类别融合后的得分总和,取总分最小的类别作为计分法的选择结果.对于加权组合方法而言,取每个人图像的第1个到第3个样本来设计分类器,并取他的第4个到第6个样本来训练分类器的加权参数,然后用第7个到第12个样本来测试组合结果.这样测试样本总数为样本的类别数乘以每个人的测试样本个数,即有 $18 \times 6 = 108$ 个.识别率则是指分类正确的测试样本个数除以测试样本总数的百分比.由于计分法不需要训练有关参数,所以为了公平地比较组合算法的性能和显示单个分类器的识别率,对于计分法和各分类器而言,取每个人图像的第1个到第6个样本来设计分类器,然后用第7个到第12个样本来测试组合结果和分类结果.

实验显示,对于整体权值的训练代数一般不超过20代,而对于出错的样本的训练一般不超过4代,故计算量不大.我们在奔腾166微机上,训练耗时一般在20分钟以内.

1)对于小波变换子图象的奇异值特征,采用同一种分类方法,得到的四个分类器的组合结果见表1所示.

表1

多分类器的组合	四个分类器				计分法	本文方法	原始图象 (F0-C1)
	F1-C1	F2-C1	F3-C1	F4-C1			
识别准确率 ($\times 100\%$)	71.29	49.07	51.85	43.51	73.14	85.18	75.00

整体训练得到的权值为

$$W_{F1-C1} = 0.7197, W_{F2-C1} = 0.1084, W_{F3-C1} = 0.0068, W_{F4-C1} = 0.2100.$$

2)对于原始图象的奇异值特征分别用三种分类方法进行分类,对于原始图象的频谱特征用最近邻方法进行分类,然后对所得到的四个分类器作组合.如表2所示.

表2

多分类器的组合	四个分类器				计分法	本文方法
	F5-C1	F0-C1	F0-C2	F0-C3		
识别准确率 ($\times 100\%$)	87.96	75.00	71.29	72.22	80.56	94.44

整体训练得到的权值为

$$W_{F5-C1} = 0.9004, W_{F0-C1} = 0.0056, W_{F0-C2} = 0.0010, W_{F0-C3} = 0.0849.$$

本文的组合方法肯定还有许多不足之处需要进一步的改进,希望能够起到抛砖引玉的作用,以推动此研究领域的发展.

参 考 文 献

- 1 Rama Chellappa *et al.* Human and machine recognition of faces; a survey. In: Proc. IEEE, 1995, **83**(5):705~740
- 2 Hong Z Q. Algebraic feature extraction of image for recognition. *Pattern Recognition*, 1991, **24**(2):211~219.
- 3 Shigeru Akamatsu *et al.* A robust face identification scheme——KL expansion of an invariant feature space. *SPIE*, 1991, **1607**:71~83
- 4 肖旭红,戴汝为.一种识别手写汉字的分类器组合方法.自动化学报,1997, **23**(5):621~627

荆晓远 1971年生,南京理工大学计算机系模式识别与智能控制专业博士研究生,研究方向为图象处理、模式识别、神经网络、信息融合等。

杨静宇 1941年生,南京理工大学信息学院院长、教授、博士生导师。目前主要研究方向为图象处理、模式识别、计算机视觉、智能机器人等。已出版著作6部,发表论文100多篇,获省部级以上科技奖11项。

第二届全国“技术过程故障诊断与安全性”学术会议 (中国 SAFEPROCESS' 2001)征文通知

经中国自动化学会批准,第二届全国“技术过程故障诊断与安全性”学术会议将于2001年9月下旬在上海举行。

一、征文范围

化工与石油化工、电力、冶金、航空航天、电子、机械、交通运输等行业中,涉及实时监测控制、故障检测与诊断、系统可靠性与安全性、安全控制与管理、容错控制等方面的新理论、新方法、新技术和新的应用成果。

会议论文集将由中文核心刊物正式出版。

二、征文要求

- 1) 全文不超过7000字;
- 2) A4单页打印稿两份;
- 3) 论文经审稿录用后,再按会议论文要求的格式,用 Word 97文稿编排。

三、征文截稿日期:2001年3月31日

录用通知发出日期:2001年4月30日

正式论文截稿日期:2001年5月31日

来稿请寄:

200135 上海 浦东大道1550号 上海海运学院科研处 蔡荣 先生收。

有关会议具体情况,请通过 E-mail:safe2001@shmtu.edu.cn 联系。

主办单位:中国自动化学会 技术过程的故障诊断与安全性专业委员会。

承办单位:上海海运学院(代章)。

2000年8月1日