第23卷 第5期
1997年9月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 23, No. 5
Sep., 1997

# 一种识别手写汉字的多分类器集成方法

肖旭红 戴汝为

（中国科学院自动化研究所 北京 100080）

**摘 要** 根据多信源信息处理与字符识别的经验知识，提出了一个识别手写汉字的多分类器线性集成模型. 这个模型不仅考虑到不同的分类器对不同字符识别能力的不同，而且还考虑了不同的分类器得出的输入字符与参考模板之间相似度的实际大小对判决的影响，及不同分类器提供的候选字符对判决的支持作用，更重要的是提供了一种通过监督学习，利用计算机程序自动计算模型参数的方法，因而实现了一个较好的集成系统. 同时，本文还提供了三个用于集成的分类器，它们集成的结果充分显示了本方法的有效性。

**关键词** 手写汉字识别，综合集成，多分类器集成.

## 1 引言

近年来，多分类器的集成已成为模式识别领域的热门话题，并在模式识别的很多领域，如字符识别领域等取得了较好的应用效果. 与此同时，由钱学森等倡导的综合集成的思想也越来越引起人们的兴趣[1,2]. 区别于普遍意义上的集成或证据组合，综合集成不仅强调多分类器、多证据的叠加或组合，而且更强调人在集成系统中的作用. 人在系统中的作用主要体现在两个方面：一是用计算机模拟人的智能，如对形象思维及抽象思维等的模拟；另一是人对系统的监督、调控作用，如监督学习等；因此，综合集成是集成方法的一种较高境界.

现阶段字符识别多分类器的集成方法可归纳为两类：一类是根据经验的线性加权型，线性权值完全人为确定，因而集成效果因人、因分类器的不同而异；另一类由人指导机器来自动确定系统参数[3,4]，但这些模型均没考虑分类器输出的输入字符与参考模板之间相似度的实际大小对判决的影响，因而没有跳出投票集成的框架，且这些方法勾是以十个数字为研究对象，应用到象汉字这样类别特别多的字符的识别还有困难.

本文提出了一种应用于手写汉字识别的多分类器综合集成的方法. 基本模型为一线性模型，其基本思想是认为各分类器对不同字符的识别能力是不同的. 由分类器输出的不同的候选类别与待识别字符之间必然存在某些相似性，对该字符的识别起着一定的支持作用，并认为它们之间的关系可以用一个线性模型来近似，而模型的参数可通过监督学习和人的经验，利用计算机程序确定. 相对于以前提出的分类器线性叠加集成系统，本文方法的优越性在于提出了一种通过监督学习自动确定模型参数的方法，并考虑到了不同的

1） 国家八六三高技术计划和自然科学基金资助项目.

收稿日期 1995-12-14

分类器得出的输入字符与参考模板之间相似度的实际大小对判决的影响及各候选字对判决的支持作用.实验结果充分显示了本文方法的高效性和优越性.

## 2 多分类器集成模型

为了便于说明,首先对有关的表示进行定义,假设

1)输入字符用 $x$ 表示;

2)字符类别用 $C_m$ 表示,其中 $m=1,2,\cdots,M$;

3)不同的分类器用 $e_i(i=1,2,\cdots,I)$来区分.

单分类器的输出可以有多种形式,本文是以输入字符与其对应类别模板之间的相似度作为输出,并且对每一个输入字符,依次输出它与可能的 $N(N\leqslant M)$个候选类别模板之间的相似度.对分类器 $e_i$ 与输入 $x$,其 $N$ 个输出分别用 $m_i(x,j,k)(k=1,2,\cdots,N)$表示.其中 $j$ 代表候选字符类别标号,$k$ 代表候选顺序标号,$k$ 越小,对应的 $m_i(x,j,k)$越大.为了保证集成的有效性,各分类器的输出度量尺度应该一致,因此必须对相似性度量进行尺度规一化,本文采用如下变换:

$$m_i(x,j,k)\leftarrow m_i(x,j,k)/m_i(x,j,1). \qquad k=1,2,\cdots,N$$

集成模型主要基于以下几条经验知识:

1)各分类器对各字符的分类能力是不同的;

2)由于手写字符的书写形变较大,同一字符的不同次书写与候选模板之间的相似度也不同;

3)对输入字符 $x$,分类器 $e_i$ 输出的较低阶次候选类别(如首候选、第二候选等)一般来说与 $x$ 有较大的相似性(误识情况除外),而较高阶次的候选类别必须与 $x$ 相似性较小.

基于以上经验知识,作以下假设:

1)分类器 $e_i$ 输出的各阶候选对集成决策的支持量是不同的,一般来说,候选阶次越高,即候选顺序标号 $k$ 越大,支持作用越小.令 $W_k$ 代表第 $k$ 候选的支持因子,$k$ 越大,$W_k$ 越小,$W_k$ 可有多种形式,如

$$W_k=e^{-\alpha(k-\delta)} \text{ 或 } W_k=1.0-\beta\times k,$$

其中 $\alpha,\delta,\beta$ 为非负常数,因而 $W_k\leqslant 1.0$.

2)分类器 $e_i$ 的第 $k$ 个候选(类别为 $C_i$)对集成判决 $x\in C_j$ 的支持作用 $A_i(x,j,k,l)$可用下面经验公式表示:

$$A_i(x,j,k,l)=m_i(x,l,k)\times R_i(j,l)\times W_k, \tag{1}$$

其中 $R_i(j,l)$表示分类器 $e_i$ 将类别 $C_j$ 中的样本判为属于类别 $C_i$ 的可能性.(1)式表明:分类器 $e_i$ 将类别 $C_j$ 中的样本判为属于类别 $C_i$ 的可能性越大,对判决 $x\in C_j$ 的支持作用越大;同时,输入 $x$ 与类别 $C_l$ 的模板越相似.对判决 $x\in C_j$ 的支持作用也越大.

3)集成系统输出的输入 $x$ 与输出类别 $C_j$ 间的总相似度 $M(x,j)$可用各 $A_i(x,j,k,l)$线性组合来表示,即

$$M(x,j)=\sum_{i=1}^{I}\sum_{k=1}^{N}A_i(x,j,k,l). \tag{2}$$

因此集成系统模型可用图1表示

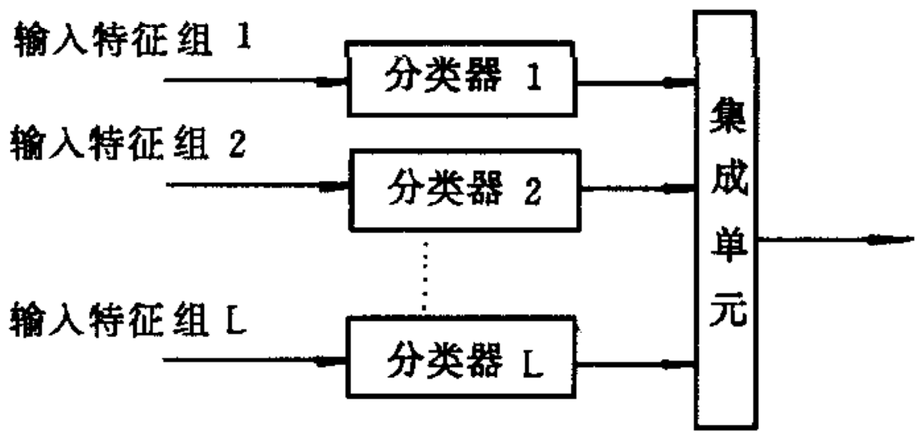


图1　综合集成系统结构框图

每一单分类器分别输出 $N$ 个候选字符的标号及相似度,因此由 I 个单分类器构成的集成变换网络共有 $L\times N$ 项输入$\{m_i(x,j,k)\,|\,i=1,2,\cdots,L,k=1,2,\cdots,N\}$,分别对应 $L\times N$ 个候选字符,这些候选字符分别由 $L$ 个分类器提供,因此其对应的字符类别有可能是相同的,但总候选类别数不会少于 $N$). 集成的作用就是以各分类器的输出作为输入,通过公式(2)所述变换,得到输入与这些候选字符模板间的一组新的相似性度量 $M(x,j)$,若 $M(x,l)=\max M(x,j)$,则 $x\in C_l$,集成系统输出标号 $l$.

## 3　$R_i(j,l)$的近似与计算

$R_i(j,l)$的选择应满足:1)$C_j$ 类的字符与 $C_i$ 类的字符越相似,$R_i(j,l)$越大. 2)当 $j=l$ 时,$R_i(j,l)$取最大值.

### 3.1　识别矩阵

通过监督学习,对每一字符类别 $C_j$,不难得到以下矩阵[3]:

$$R_j=\begin{bmatrix} r_{11}^j & r_{12}^j & \cdots & r_{1M}^j \\ r_{21}^{j1} & r_{22}^j & \cdots & r_{2M}^j \\ \vdots & \vdots & \vdots & \vdots \\ r_{I1}^j & r_{I2}^j & \cdots & r_{IM}^j \end{bmatrix}, \tag{3}$$

其中每一行对应一个分类器,每一列对应一个字符类别,元素 $r_{il}^j$1)当 $j=1$时表示分类器 $e_i$ 对字符类 $C_j$ 的识别率;2)当 $j\neq l$ 时表示分类器 $e_i$ 将类 $C_j$ 中的字符误识为属于类别 $C_i$ 的误识率. 容易看出,元素 $r_{il}^j$满足 $R_i(j,l)$所要求的条件,可以作为对 $R_i(j,l)$近似.

### 3.2　用 $r_{il}^j$近似 $R_i(j,l)$

由于各学习样本书写的规范程度不一样,各类字符的结构也不同,就可能对有的字符类,所有的分类器识别率都较高;而对另一些字符类,所有的分类器识别率都较低. 不难看出,若直接令 $R_i(j,l)=r_{il}^j$,根据公式(1),(2),必然倾向于把输入 $x$ 识别成各分类器具有较高识别率的候选类别. 因此 $R_i(j,l)$还必须满足对各类字符具有相同的度量尺度,这里选择

$$R_i(j,l)=r_{jl}^j/r_{\max}^j, \tag{4}$$

其中 $r_{\max}^j=\max\limits_i r_{ij}^j$,这就保证对每一字符类别 $C_j$,$\max\limits_i R_i(j,j)=1.0$,即对各类字符度量尺度一致.

由于对一特定的输入字符类别,分类器总是倾向于把它误识别成与它比较类似的较固定的几个字符类别,实际应用时,为了节省存储空间及减少运算时间,单分类器输出的

候选类别数 $N$ 远远小于字符类别总数 $M$；由于只有少数几个与输入 $x$ 很类似的字符对 $x$ 的识别有支持作用，而与输入 $x$ 所属类别很不相似的字符类对肯定判决的支持量应为负，对每一字符类别 $C_j$，只存储少数 $R_i(j,l)$，比如，只保留具有较大值的 $L$ 个 $R_i(j,l)$，其余的被赋为0(无作用)或一个很小的负数(负作用).

## 4 单分类器介绍

### 4.1 基于多层外围结构的分类器(分类器1)

分别从上、下、左、右四个方向对细化字符图象扫描，扫描线第一次碰到的黑象素被当成字符的最外围结构，第二次碰到的黑象素被当成字符的次外围结构，每一方向的结构被非均匀地分成8份，图2.$a$，图2.$c$ 所示分别为细化字符“啊”及其最外围、次外围结构.

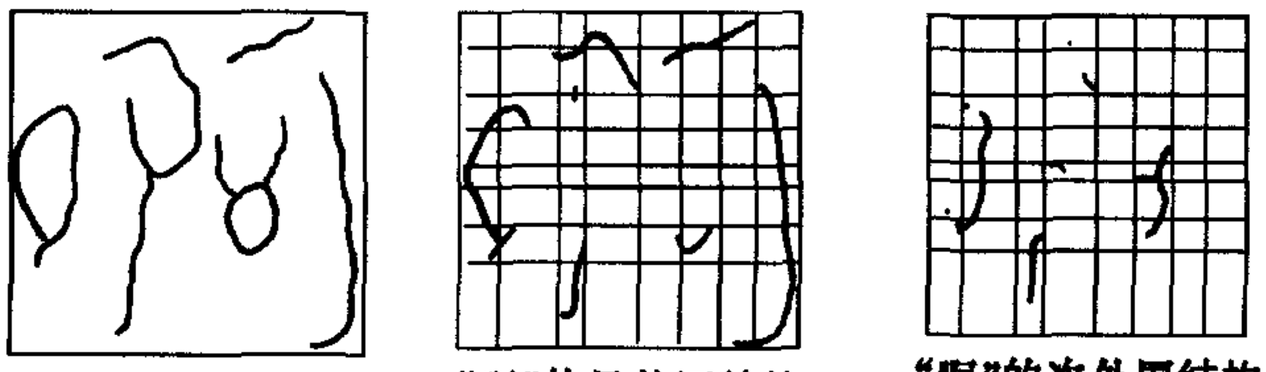


图2 细化字符“啊”及其层次外围结构

对外围结构中的每一黑象素点，抽取以下特征：

1)四方向属性$(P_1,P_2,P_3,P_4)$：假设 $l_m=(m=1,2,\cdots,8)$分别表示从该图象点沿上、下、左、右及四对角所定义的八方向到笔划边界的游长，则

$$P_i=\frac{l_i+l_{i+4}}{\sqrt{\sum_{j=1}^{4}(l_j+l_{j+4})^2}};$$

2)该点到扫描线起点的距离 $d$

这样，对每一个黑象素点，抽取了五维特征：$(P_1,P_2,P_3,P_4,d)$.

被非均匀分割的每一子区域被当成一个子结构，子结构的各特征为其内各黑象素点对应特征的统计平均值. 设子结构 $S$ 由 $n$ 个黑象素点组成. 其特征分别为$(P_1^j,P_2^j,P_3^j,P_4^j,d^j)(j=1,2,\cdots,n)$，则 $S$ 的特征为

$$P_i^S=\sum_{j=1}^{n}P_i^j/n \qquad (i=1,2,3,4),$$

$$D^S=\sum_{j=1}^{n}d^j/n.$$

由于每一方向外围结构被分为八个子结构，因此，最外、次外两层外围结构共被分成8×4×2=64个子结构. 假设输入 $x$ 的外围子结构集用$\{S_1,S_2,\cdots,S_{64}\}$表示，其候选类别模板的子结构集用$\{r_1,r_2,\cdots,r_{64}\}$表示，则输入与候选类别模板之间的距离为

$$d(s,r) = \sum_{i=1}^{64} \mathrm{dis}p(s_i,r_i) * (|D^{S_i} - D^{r_i}| + 8)/8.0,$$

其中　　$$\mathrm{dis}p(s_i,r_i) = \sum_{j=1}^{4} |P_j^{S_i} - P_j^{r_i}|,$$

$d(s,r)$越大，输入与该候选模板之间的相似性越小.

### 4.2　基于四角结构的分类器(分类器2)

分别从四对角方向对细化字符图象扫描，图3所示为图2.$a$ 中所示“啊”的四角结构. 每一对角区域的结构被均匀地分成六等份，每一份被当成一个子结构，总共可得到6×4＝24个子结构，基于四角结构的分类器对每一子结构，抽取与4.1.1节的外围结构分类器类似的特征，并利用类似的判决准则.

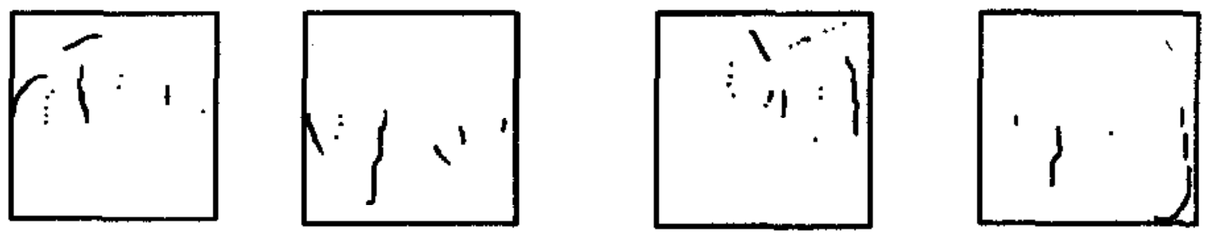

图3　沿四对角方向扫描所得字符“啊”的四角结构

### 4.3　基于局部笔画方向的分类器(分类器3)

这一分类器所抽取的特征类似于文献[5]中的 $LDC$，即对细化字符图象的每一黑象素点，抽取4.2.1中所描述的四方向属性，每一图象被分成8×8＝64个矩形区域，每个矩形区域为一子结构，其特征为区域内黑象素点对应特征的统计平均值；与文献[5]所不同的是每一矩形区域是根据输入字符结构非均匀划分而成. 相应的距离准则为

$$d(s,r) = \sum_{i=1}^{64} \mathrm{dis}p(s_i,r_i).$$

## 5　实验结果

为了检验本文所述方法的性能，将三个分类器的输出结果加以综合集成，并对中国科学院自动化所收集的4$M$ 手写字符样本库[6]中的3755个一级国标进行了学习和测试(每字用100个样本优类样本对单分类器进行学习，20个良类样本用于集成变换参数的学习，另外每字分别用5个优类及良类样本进行测试)，实验结果见表1.

表 1

| 分类器<br>识别率(%) | 分类器1 | 分类器2 | 分类器3 | 综合集成 |
|---|---|---|---|---|
| 优类样本 | 83.4 | 80.1 | 78.9 | 89.9 |
| 良类样本 | 75.28 | 74.36 | 68.19 | 81.7 |

为了比较各参数对集成系统的影响，进行了以下实验，实验结果示于表2.

### 5.1　分类器输出的多候选对判别的支持作用

$a$)　$N$＝1，即每个分类器只输出首候选，对应于表(2)中 $T_1$；

*b*) $N=5$,即每个分类器输出5个候选,对应于表(2)中 $T_2$.

5.2 输入与候选之间相似度的实际大小对判决的影响

*a*) 考虑分类器输出的输入与候选之间相似度的实际大小对判决的影响,仍取 $N=5$,则与5.1中 *b*)为同一实验.

*b*) 不考虑分类器输出的输入与候选之间相似度的实际大小对判决的影响,即令2节中 $m_i(x,j,k)$为常数,如令 $m_i(x,j,k)=1$,对应于表中 $T_3$.

5.3 既不考虑其它候选对判决的支持作用($N=1$),也不考虑分类器输出的输入与候选之间相似度的实际大小对判决的影响($m_i(x,j,k)$为常数).

5.4 采用下面的线性叠加模型(图4),线性权值根据经验直接人为设定,结果对应于 $T_5$.

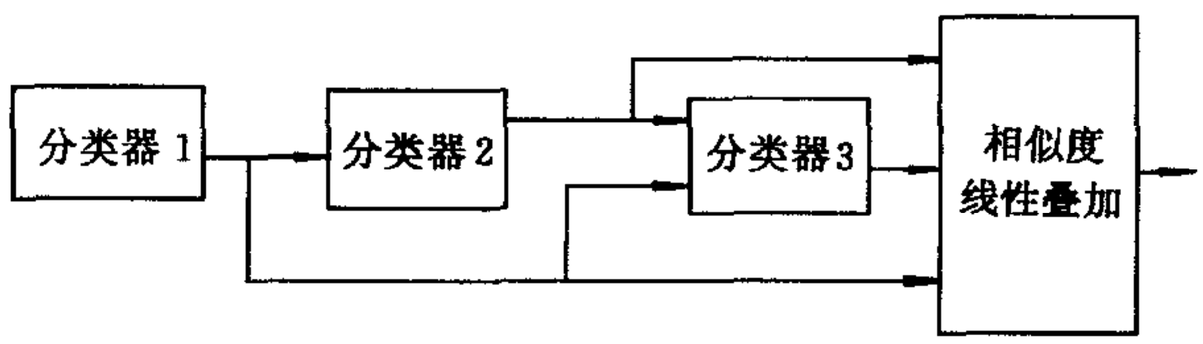


图4 多分类器线性集成模型

表 2

| 实验 | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|---|---|---|---|---|---|
| 识别率(%) | 77.5 | 81.7 | 76.86 | 76.24 | 80.02 |

实验结果表明,分类器输出的多候选及输入与候选之间相似度的实际大小对判决都会有一定的作用,输入与候选之间相似度的实际大小对判决的作用尤其重要,若不考虑输入与候选之间相似度的实际大小对判决的作用,则多候选的作用也不明显($T_3\approx T_4$). 从实验结果看,直接根据经验线性加权的结果也较好. 这主要是因为三分类器所抽取的特征及所用的判决准则都是同一个度量等级,可以简单叠加;除此之外,这种简单叠加的一重大缺陷是实验结果在很大程度上取决于实验者人为所取的参数,没有普遍意义.

## 6 结语

本文基于人的经验,提出了一个识别手写汉字的多分类器综合集成的线性模型,并提供了一种通过监督学习,自动决定模型参数的方法. 通过对三个手写字符分类器的综合集成,充分体现了本文方法的有效性. 当然,实验只能证明本文提出的模型及参数确定方案的可行性,关于更合理的模型及更有效参数确定方法还有待进一步研究.

## 参 考 文 献

[1] 钱学森,于景元,戴汝为. 一个科学新领域——开放的复杂巨系统及其方法论. 自然杂志,1990,**13**(1):3—10.

[2] 戴汝为等. 智能系统的综合集成. 杭州:浙江科技出版社,1995.

[3] Xu L, Krzyzak A, Suen C Y. Method of combining multiple classifiers and their application to handwriting recognition. *IEEE Trans. SMC*, 1990, **22**(3):418—435.

[4] 郝红卫，戴汝为. 一种综合集成方法及其在字符识别中的应用. 模式识别与人工智能，1996，**9**(1)：10－20.

[5] Teruo Akiyama, Norihiro Hagita. Automated entry system for printed documents. *Pattern Recognition*, 1990, **23**(11)：1141－1154.

[6] Tai J W (Dai R W), Liu Y J, Zhang L Q. A new approach for feature extraction and feature selection of handwritten Chinese character recognition, from pixels to features Ⅲ：fronties in handwriting recognition, Elsevier Science Publishers B. V, 1992, 479－489.

# A METASYNTHETIC APPROACH FOR HANDWRITTEN CHINESE CHARACTER RECOGNITION

XIAO XUHONG　　DAI RUWEI

(*Institute of Automation, Chinese Academy of Sciences, Beijing* 100080)

**Abstract** As a high-level integration approach, metasynthesis has drawn much attention. In this paper, a metasynthetic approach for combination of multiple classifiers is proposed. As a first step, a linear integration model is built. In this model, not only the degree of the similarity between the input and its ranked candidates are appled, but also the supporting effects of multiple candidates are taken into account, and their contributions to the integration result is expressed by a linear function. Then, an algorithm for automatically calculating the arguments of the model through supervised learning is provided. Experiments on metasynthesis of three individual classifiers for handwritten Chinese character recognition are given to demonstrate the effectiveness of our method.

**Key words** Handwritten Chinese character recognition, metasynthesis, multiple classifier integration.

**肖旭红**　1992年毕业于中国科技大学无线电电子学系，1997年获中国科学院自动化所博士学位. 主要研究领域为模式识别，人工智能，图象处理等.

**戴汝为**　主要从事人工智能，模式识别等研究. 1980—1982在美国普渡大学电机系作访问学者，现任中国科学院自动化所研究员，博士生导师，中国科学院院士，《模式识别与人工智能》主编，并任清华大学、汕头大学等近三十所高校的兼职或名誉教授(请见本刊19卷5期).