分类器的动态选择与循环集成方法

郝红卫 [1] 王志彬 [1] 殷绪成 [1] 陈志强 [2]

摘 要 针对多分类器系统设计中最优子集选择效率低下、集成方法缺乏灵活性等问题, 提出了分类器的动态选择与循环集成方法 (Dynamic selection and circulating combination, DSCC). 该方法利用不同分类器模型之间的互补性, 动态选择出对目标有较高识别率的分类器组合, 使参与集成的分类器数量能够随识别目标的复杂程度而自适应地变化, 并根据可信度实现系统的循环集成. 在手写体数字识别实验中, 与其他常用的分类器选择方法相比, 所提出的方法灵活高效, 识别率更高.

关键词 多分类器系统, 动态选择, 循环集成, 互补指数

DOI 10.3724/SP.J.1004.2011.01290

Dynamic Selection and Circulating Combination for Multiple Classifier Systems

HAO Hong-Wei[1] WANG Zhi-Bin[1] YIN Xu-Cheng[1] CHEN Zhi-Qiang[2]

Abstract In order to deal with the problems of low efficiency and inflexibility for selecting the optimal subset and combining classifiers in multiple classifier systems, a new method of dynamic selection and circulating combination (DSCC) is proposed. This method dynamically selects the optimal subset with high accuracy for combination based on the complementarity of different classification models. The number of classifiers in the selected subset can be adaptively changed according to the complexity of the objects. Circulating combination is realized according to the confidence of classifiers. The experimental results of handwritten digit recognition show that the proposed method is more flexible, efficient and accurate comparing to other classifier selection methods.

Key words Multiple classifier systems, dynamic selection, circulating combination, complementarity factor

多分类器系统是模式识别中一个重要研究方向, 近年来取得了很大进展. 大量理论与实验结果表明, 多分类器系统不但可以提高分类的正确率, 而且可以提高识别系统的泛化能力和鲁棒性[1−3]. 多分类器系统设计的关键在于分类器的选择和集成方法. 一般而言, 多分类器系统首先要构造一定数量的单分类器 (或称候选分类器), 然后从中选择一个最优子集, 再采用某种集成方法对其集成[4].

目前对多分类器系统的研究已不再局限于集成方法的提出和改进, 而更多的是着眼于分类器的选择. 研究结果表明, 所有分类器都参与集成的效果并非最好, 从众多分类器中选择部分互补性强的分类器进行集成可以提高集成的效率并改善其效果[5]. 因此, 分类器的选择方法成为了多分类器系统中的关键问题. 现有的分类器选择方法大部分采用基于集成精度的随机搜索方法. 如何对分类器进行选择以及选择哪个分类器用于集成, 成为目前广大学者研究的重点[6−8]. 文献 [9−10] 利用聚类算法对候选分类器进行聚类, 然后从每个聚类中挑选一个分类器用于集成. 文献 [5, 11−13] 采用遗传算法对分类器子集进行选择. 文献 [14] 通过后向顺序选择方法将那些不能提高集成性能的分类器从初始集中删除. 文献 [15] 提出了一种基于局部精度的动态分类器选择方法. 文献 [16] 则采用自适应 K-近邻规则对分类器进行动态选择. 同时, 大量的分类器选择标准也相继被提出和使用, 如分类器差异性度量、集成精度等. 文献 [17] 对多种分类器差异性度量方法进行了总结、分析, 验证了这些方法的相关性. 文献 [4] 对不同选择标准下的分类器选择方法进行了分析. 我们也在文献 [2] 中对主要的分类器选择方法进行了比较.

上述研究工作是将分类器的选择作为最优化问题来处理, 即根据某种准则函数, 采用各种最优化方法, 寻找一个最优或次优分类器子集, 然后对其进行集成, 使集成后的多分类器系统性能最佳. 一般情况下, 各种最优化方法如遗传算法、分支定界算法等, 需要花费很长的时间, 经过大量的尝试才能找到最优解[2]. 这使得分类器子集的选择需要付出很大的时间代价. 另外, 分类器子集一旦选定就不再变动, 对于任何测试样本都使用该子集进行集成, 系统结

收稿日期 2010-12-03 录用日期 2011-06-13
Manuscript received December 3, 2010; accepted June 13, 2011
国家自然科学基金 (60675006), 国家公益性行业 (气象) 科研专项 (GYHY201106039, GYHY201106047), 中央高校基本科研业务费专项资金 (FRF-BR-10-034B) 资助
Supported by National Natural Science Foundation of China (60675006), the R & D Special Fund for Public Welfare Industry (Meteorology) of China (GYHY201106039, GYHY201106047), and Funds for the Central Universities (FRF-BR-10-034B)
1. 北京科技大学计算机与通信工程学院 北京 100083 2. 广州市劳动保障信息中心 广州 510635
1. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083 2. Guangzhou Labour Security Information Center, Guangzhou 510635

构固定, 缺乏足够的灵活性.

为此, 本文提出了一种分类器子集的选择和集成方法 — 分类器的动态选择与循环集成 (Dynamic selection and circulating combination, DSCC). 该方法能够利用不同分类器模型之间的互补性, 动态选择出对目标有较高识别率的分类器组合, 使参与集成的分类器的数量能够随识别目标的复杂程度而自适应的变化, 并根据可信度实现系统的循环集成. 在手写体数字识别实验中, 与其他常用的分类器选择方法相比, 所提出的方法具有较高的搜索效率, 循环集成的多分类器系统结构灵活, 易于实现, 有较高的识别率和泛化能力.

论文安排如下: 第 1 节定义了互补指数, 详细介绍了分类器动态选择与循环集成算法的过程; 第 2 节给出了算法在两个国际通用手写体数字样本库 MNIST 和 USPS 上的实验结果, 并与其他分类器选择方法进行比较分析; 第 3 节总结了本文的工作.

1 分类器动态选择与循环集成算法

分类器动态选择与循环集成算法可以针对不同待识别目标, 挑选出不同数目的分类器进行集成和识别. 其过程为: 首先构造一定数量的候选分类器, 计算互补指数和整体互补指数. 然后根据整体互补指数对候选分类器进行排序, 并依据可信度, 动态选择出对目标有较高识别率的分类器组合, 构成最优子集. 当所有候选分类器都参与集成仍不能达到初始可信度要求时, 通过循环利用分类器的方法对其集成, 从而得到目标的识别结果. 该算法主要包括两部分: 一是候选分类器排序, 二是分类器子集的动态选择与循环集成.

1.1 候选分类器排序

分类器的互补性是保证多分类器系统具有较高识别率和泛化能力的关键, 是分类器动态选择的依据. 因此, 在进行分类器选择之前, 必须首先解决分类器互补性度量的问题. 为此, 本文提出了一种描述分类器互补性度量的方法 — 互补指数, 其定义如下:

设有两个分类器 E_i 和 E_j, 非空样本集 S 被两个分类器错分的样本集合分别为 S_i 和 S_j, F 为两个分类器的互补指数, $D(S)$ 为 S 中所含样本的个数, 则两分类器之间的互补指数定义如式 (1) 所示:

$$F_{ij}=\frac{D(S_i\cup S_j)-D(S_i\cap S_j)}{D(S)} \tag{1}$$

在上述定义中, F_{ij} 越大, 越说明分类器 E_i 和 E_j 的互补性越强. 反之, 则互补性越弱.

在进行分类器选择时, 除了要考虑分类器之间的互补性外, 还必须考虑待选分类器和已有分类器集合的互补性. 因此, 需要进一步定义描述上述互补性的度量方法 — 整体互补指数.

设有 m 个分类器, 其中已有 n 个分类器入选最优子集, E_k 为当前被考察的分类器, 在 n 个已入选分类器所构成子集中加入 E_k 后, 这 $n+1$ 个分类器的整体互补指数 TF 为

$$TF=\frac{\sum\limits_{i,j=1,i\neq j}^{n+1}F_{ij}}{2\times C_{n+1}^2} \tag{2}$$

根据上述互补性度量方法的定义, 我们设计出一种分类器排序算法 (如算法 1 所示), 其核心思想是根据整体互补指数, 从具有 N 个候选分类器的集合 T 中动态选择分类器并进行排序. 其过程为: 首先选择识别率最高的分类器 c, 放入临时变量 D 中并排在第一位, 其次从剩余的候选分类器集合 T 中选择分类器 e 排在第二位, 选择的标准是它与前面分类器 (临时变量 D 中) 所构成的整体互补指数最大, 然后重复选择 e, 直到所有候选分类器都被选入 D, 各分类器入选的顺序即为排序结果 P.

算法 1. 分类器排序.

Input: 分类器集合 T.

Output: 排序结果 P.

1) 初始化 $D=\phi$, $A=\phi$, $P[0]=N$, $i=1$;
2) 从 T 中选择识别率最高的分类器 c;
3) $D=c$;
4) $T=T-c$;
5) $P[i]=c$;
6) **while** $T\neq\phi$ do
7) $i++$;
8) 从 T 中选择分类器 e 使得 TF 最大;
9) $D=D\cup e$;
10) $T=T-e$;
11) $P[i]=e$;
12) **end while**
13) **return** (P).

按照该算法可以保证在需要 n 个分类器时, 序列 P 中的前 n 个分类器的组合就是互补性最强的分类器子集; 当只需要一个分类器时, 选择的是所有分类器中识别率最高的一个.

1.2 分类器子集的动态选择与循环集成

在分类器排序完成后, 可以根据对可信度的要求进行分类器的动态选择与循环集成 (如算法 2 所示). 当一个或少数几个分类器就能满足识别要求时, 则无需选择更多的分类器; 否则依次添加分类器, 并进行循环集成. 该算法的具体执行过程为:

首先根据识别精度的需要设定初始可信度阈值

θ_0, 令 $\theta = \theta_0$, 然后从已排序的分类器序列 P 中选取第一个分类器对样本 x 进行识别, 当识别结果满足可信度的要求时, 则输出识别结果 R, 无需集成其他分类器; 否则依次选入第 $k\,(k \geq 2)$ 个分类器, 并对所有已入选的分类器按加法规则进行集成, 当满足输出条件 $S_{\max} > k \times \theta$ 时, 输出识别结果 R; 若所有分类器都已选入仍不满足输出条件, 则通过步长 $\Delta\theta$ 降低可信度阈值 $\theta = \theta - \Delta\theta$, 重复上述步骤, 实现循环集成, 直到满足输出条件, 输出识别结果 R.

算法 2. 动态选择与循环集成 (DSCC).

Input: 测试样本 x, 排序结果 P.

Output: 识别结果 R.

1) 初始化 $\theta = \theta_0$, $n = P[0]$, $m = 0$;
2) **while** $\theta \geq 0$ **do**
3) **for** $t = 1$ to n **do**
4) $k = m \times n + t$;
5) 线性集成 k 个分类器: $Y = \sum_{i=1}^{k} y_i$;
6) **if** $S_{\max} > k \times \theta$ **then**
7) **return** (R)
8) **end if**
9) **end for**
10) $m++$;
11) $\theta = \theta - \Delta\theta$;
12) **end while**.

在算法 2 中, y_i 为第 i 个参与集成的分类器的输出结果, Y 是按加法规则集成后的结果, $Y = (O_1, O_2, \cdots, O_c)^t$, c 为类别数, 其中的每一维分量表达了该分量所对应类别在集成后的可信度, $S_{\max}$ 为各分量中的最大值, 所对应的类别即为识别结果. 当满足 $\frac{1}{k}S_{\max} > \theta$ 时, 表明 k 个分类器集成后该类别的可信度超过了设定的可信度阈值, 可以将其所对应的类别作为识别结果输出, 即当满足 $S_{\max} > k \times \theta$ 时, 输出识别结果.

θ 为系统的可信度, 其取值范围为 [0,1], 可信度的初始阈值越大, 参与集成的分类器的个数就越多, 系统的识别精度就会越高, 但效率会降低; 反之, 会降低精度, 提高效率. 可信度初始阈值可根据实际需求和经验来进行选择.

此外, 本文提出的 DSCC 是一种基于可信度计算的方法. 该方法对于输出结果带有类似后验概率的分类器可直接进行选择和集成, 但对于其他输出形式的分类器, 如 SVM、最近邻分类器等, 需要先将其输出值转化到 [0, 1] 上的可信度, 然后再进行选择和集成, 可信度转换方法可依照文献 [18]. 为方便起见, 本文的实验部分采用了 BP 神经网络分类器.

2 实验结果

手写数字识别是一个经典的模式识别问题. 由于类别数小, 使得一些复杂或运算量较大的算法实现起来比较容易. 因此, 数字识别始终是模式识别中各种新方法研究的实验对象, 在算法研究中占据着重要地位[19−20]. 本文使用了国际上通用的手写体数字样本库 — MNIST 库和 USPS 库进行实验. 其中, MNIST 含有 60 000 个训练样本和 10 000 个测试样本, USPS 含有 7 291 个训练样本和 2 007 个测试样本. 我们将 MNIST 数据库中前 50 000 个训练样本分成 10 个子集, 每个子集包含 5 000 个训练样本, 作为单分类器的训练集. 将 MNIST 中剩下的 10 000 个训练样本作为验证集 (Validation set), 用于互补指数的计算及分类器的排序. 用 MNIST 中的 10 000 个测试样本和 USPS 数据库中的全部 9 298 个样本构成两个独立的测试集, 用于测试各单分类器和不同集成方法的性能.

2.1 单分类器的设计

实验中, 我们采用三层结构的 BP 网络, 通过以下两种方式来设计和构造所需的不同单分类器: 1) 选择不同的输入特征; 2) 利用不同的训练样本. 分别使用 $3 \times 5 \times 4 = 60$ 维和 $5 \times 6 \times 4 = 120$ 维两种方向特征[21], 对两种特征分别采用 10 个训练子集进行训练, 得到 20 个不同的 BP 网络分类器. 其中, C1–C10 采用 60 维方向特征, 对应的 BP 神经网络隐节点取 80 个; C11–C20 采用 120 维方向特征, 隐节点取 100 个. 各单分类器的识别率如表 1 所示. 实验中未考虑拒识, 故误识率 = 1 − 识别率.

2.2 分类器动态选择与循环集成

在单分类器设计完成后, 采用验证样本集对上述单分类器进行排序. 根据算法 1, 首先计算分类器之间的互补指数, 然后按照整体互补指数最大的原则对分类器进行排序, 排序结果如表 2 所示.

根据表 2 的排序结果, 对 20 个分类器进行动态选择和循环集成实验, 实验结果如表 3 所示. 为方便对比, 将 20 个分类器中精度最高的 C18 列于表 3 中. 实验中, 可信度初始阈值 $\theta = 0.99$, 步长 $\Delta\theta = 0.05$. 由表 3 可见, 集成后的多分类器系统在两个测试样本集上的识别率分别为 95.93 %、86.35 %. 高于任何一个单分类器的识别率.

此外, 为了进一步对比多分类器系统与单分类器的性能, 我们采用 MNIST 全部 60 000 个训练样本训练出一个单分类器, 该分类器在 MNIST 测试集上的识别率为 95.43 %, 虽然高于 C18 但却低于 DSCC, 进一步说明了 DSCC 的有效性.

表 1 各单分类器在不同样本集上的识别率 (%)
Table 1 Accuracies of individual classifiers on different test sets (%)

分类器	验证集	MNIST 测试集	USPS 测试集
C1	88.43	88.15	68.58
C2	89.77	90.03	74.19
C3	89.71	90.35	71.65
C4	88.86	89.10	73.49
C5	89.78	90.37	71.02
C6	90.42	91.35	71.33
C7	88.89	89.44	67.49
C8	89.15	89.46	63.29
C9	90.73	90.33	74.32
C10	92.01	91.96	74.84
C11	93.00	92.88	79.59
C12	93.48	93.39	84.58
C13	93.72	93.43	80.59
C14	91.06	91.08	82.29
C15	92.91	92.99	81.33
C16	92.47	91.87	79.69
C17	94.03	94.10	82.61
C18	94.86	95.16	84.74
C19	93.39	93.07	81.54
C20	92.31	91.66	81.29

表 2 分类器排序结果
Table 2 Results of classifier sorting

入选顺序 P	分类器	TF
P[1]	C18	0
P[2]	C4	0.105
P[3]	C1	0.104
P[4]	C14	0.103
P[5]	C7	0.102
P[6]	C10	0.1
P[7]	C5	0.099
P[8]	C20	0.097
P[9]	C8	0.096
P[10]	C9	0.094
P[11]	C2	0.093
P[12]	C16	0.092
P[13]	C3	0.091
P[14]	C11	0.090
P[15]	C15	0.089
P[16]	C6	0.088
P[17]	C12	0.087
P[18]	C13	0.086
P[19]	C19	0.085
P[20]	C17	0.083

表 3 DSCC 与单分类器 C18 的性能比较 (%)
Table 3 Comparison of ensemble accuracies with the base classifier C18 (%)

方法	MNIST 测试集	USPS 测试集
C18	95.16	84.74
DSCC	95.93	86.35

在上述实验中, 由于算法在对每一个样本识别时都进行分类器的动态选择和循环集成, 因此, 对每一个样本而言, 参与集成的单分类器个数和最终的可信度各不相同, 这样就克服了已有多分类器集成系统中分类器子集一旦选定就不再变动、缺乏灵活性的缺点. 对于容易识别的样本, 只用一个或少数几个分类器就可以得到满足可信度的识别结果; 而对于难以识别的样本, 则需集成更多的分类器, 甚至重复利用分类器进行识别. 无论是供研究使用的样本库中的样本还是实际应用中的样本, 容易识别的通常占多数, 因此, 多数样本仅需使用少量分类器就能被正确识别, 从而在提高了精度的同时也保证了算法的效率.

2.3 与几种分类器选择算法的比较

为进一步验证本文所提方法的有效性, 我们将该方法与其他常用的分类器选择方法进行对比. 目前, 常用的分类器子集搜索算法有很多, 最典型的有: 穷尽法、顺序前进法 (SFS)、顺序后退法 (SBS)、增 l 减 r 法 (PTA(l, r))、广义增 l 减 r 法 (GPTA(l, r)) 和遗传算法等. 文献 [2] 对各种搜索算法进行了详细的比较. 根据比较结果, 在本实验中, 我们选用穷尽法、SFS 法和遗传算法来进行分类器子集的选择, 并对选择结果按照加法规则进行集成.

实验中, 遗传算法参数设置为: 选用 20 位的二进制字串, 代表 20 个候选分类器, 群体规模为 20, 初始群体通过随机法得到, 最大遗传世代数设为 500, 交叉截断点随机选取, 交叉概率和变异概率分别为 0.9 和 0.01; 分类器动态选择与循环集成算法参数设置为: 可信度初始阈值为 0.99, 步长为 0.001; 实验结果如表 4 所示.

从表 4 的实验结果可以看出, 本文所提出的方法在分类器子集的选择方面不仅比其他几种方法具有更高的效率, 而且具有更高的识别率. 注意到本算法不仅在测试集上而且在验证集上的识别率都超过了穷尽法. 在测试集上识别率超过穷尽法表明本算法具有更好的泛化能力; 在验证集上超过穷尽法则充分说明了循环集成的效果, 因为穷尽法只能选出固定个数的最优子集, 而本算法则能根据样本情况重复利用单分类器. 另外, 从表 4 中还可以发现: 在 MNIST 测试集上, 遗传算法要比 SFS 的效果好, 而

表 4 DSCC 与不同选择方法的性能比较

Table 4 Comparison of ensemble performances with different selection methods

方法	子集搜索时间	分类器个数	验证集 (%)	MNIST 测试集 (%)	USPS 测试集 (%)
穷尽法	10.3 h	4	95.96	95.57	85.99
SFS 方法	8 s	6	95.82	95.42	85.96
遗传算法	37 s	4	95.95	95.60	85.86
DSCC	2 s	−	95.98	95.96	86.42

在 USPS 测试集上则正好相反, 这说明两者的泛化能力不强. 采用上述搜索算法得到的分类器子集一般是针对某个样本集 (验证集) 选出的, 并且一旦选定则不会变化, 以后对任何测试样本集都使用该子集进行集成识别. 这就使某些具有潜在价值的单分类器失去选入该子集的机会, 从而不仅使集成系统缺乏足够的灵活性, 而且效率较低.

然而, 本文所提出的分类器动态选择与循环集成方法则弥补了上述缺陷. 该方法能够根据样本的情况动态选择出所需要的分类器子集并进行循环集成, 从而不仅提高了系统的泛化能力, 而且极大提高了系统的效率.

2.4 参数对算法性能的影响

分类器动态选择与循环集成算法的性能会随算法中可信度初始阈值和步长的改变而变化. 为研究两者之间的关系, 特做如下两组实验:

1) 初始阈值的影响: 采用固定步长 $\Delta\theta = 0.05$, 分别用三个初始阈值 0.99, 0.79 和 0.59 进行实验, 结果如表 5 所示. 由实验结果可知: 当初始阈值减小时, 系统在两个测试样本集上的识别率都在减小, 表明较高的初始阈值会取得较高的识别率.

表 5 初始阈值 θ 对 DSCC 的影响 (%)

Table 5 Influences of initial threshold θ on DSCC (%)

θ_0	MNIST 测试集	USPS 测试集
0.99	95.93	86.35
0.79	95.82	86.22
0.59	95.63	86.13

2) 步长的影响: 采用固定初始阈值 $\theta_0 = 0.99$, 分别用三个步长 0.05, 0.01 和 0.001 进行实验, 结果如表 6 所示. 由实验结果可知: 当步长减小时, 系统在两个测试样本集上的识别率都在增加 (或保持), 表明较小的步长会取得较高的识别率.

表 6 步长 $\Delta\theta$ 对 DSCC 的影响 (%)

Table 6 Influences of step $\Delta\theta$ on DSCC (%)

$\Delta\theta$	MNIST 测试集	USPS 测试集
0.05	95.93	86.35
0.01	95.93	86.39
0.001	95.96	86.42

改变可信度初始阈值和步长会影响集成系统性能的原因在于: 改变初始阈值 θ_0 和步长 $\Delta\theta$ 会改变算法循环集成的次数. 初始阈值 θ_0 越大, 步长 $\Delta\theta$ 越小, 则算法循环的次数越多, 使得参与集成的分类器的个数就越多, 这样通常会使被错分的样本数目减少, 最终提高了系统的整体识别率, 但识别时间也会相对较长; 反之亦然. 由于算法循环次数的增加会降低系统识别的效率, 所以在实际应用中, 应根据识别样本的复杂程度以及对精度的要求, 设置合适的初始阈值和步长, 以兼顾精度和效率.

3 结论

针对目前多分类器系统设计中最优子集选择效率低下、集成方法缺乏灵活性等问题, 本文提出了分类器动态选择与循环集成方法. 在国际通用的两个手写体数字样本库 MNIST 和 USPS 上的实验结果表明: 该方法能充分利用不同分类器之间的互补性, 动态选择出对目标有较高识别率的分类器组合, 使参与集成的分类器数量能够随识别目标的复杂程度而自适应的变化, 而且可以根据可信度的要求, 进行循环集成. 与穷尽法、SFS 法、遗传算法的比较表明, 所提出的方法不仅具有更高的识别率, 而且结构灵活、效率更高. 同时, 本方法还可以通过对可信度初始阈值和步长进行调整, 实现精度和效率的折衷. 在实际应用中, 可以根据不同要求合理设置这两个参数, 兼顾精度和效率.

References

1 Fumera G, Roli F, Serrau A. A theoretical analysis of bagging as a linear combination of classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(7): 1293−1299

2 Hao H W, Liu C L, Sako H. Comparison of genetic algorithm and sequential search methods for classifier subset selection. In: Proceedings of the 7th International Conference on Document Analysis and Recognition. Edinburgh, Scotland: IEEE, 2003. 765−769

3 Brown G. An information theoretic perspective on multiple classifier systems. In: Proceedings of the 8th International Workshop on Multiple Classifier Systems. Reykjavik, Iceland: Springer-Verlag, 2009. 344−353

4 Ruta D, Gabrys B. Classifier selection for majority voting. *Information Fusion*, 2005, **6**(1): 63−81

5 Zhou Z H, Wu J X, Tang W. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 2002, **137**(1−2): 239−263

6 Kang H J, Doermann D. Selection of classifiers for the construction of multiple classifier systems. In: Proceedings of the 8th International Conference on Document Analysis and Recognition. Seoul, Korea: IEEE, 2005. 1194−1198

7 Ko A H R, Sabourin R, Britto A S. From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition*, 2008, **41**(5): 1735−1748

8 Chen L, Kamel M S. A generalized adaptive ensemble generation and aggregation approach for multiple classifier systems. *Pattern Recognition*, 2009, **42**(5): 629−644

9 Liu R, Yuan B. Multiple classifier combination by clustering and selection. *Information Fusion*, 2001, **2**(3): 163−168

10 Li Guo-Zheng, Yang Jie, Kong An-Sheng, Chen Nian-Yi. Clustering algorithm based selective ensemble. *Journal of Fudan University (Natural Science)* , 2004, **43**(5): 689−691 (李国正, 杨杰, 孔安生, 陈念贻. 基于聚类算法的选择性神经网络集成. 复旦学报 (自然科学版), 2004, **43**(5): 689−691)

11 Kim Y W, Oh I S. Classifier ensemble selection using hybrid genetic algorithms. *Pattern Recognition Letters*, 2008, **29**(6): 796−802

12 Santos E M, Sabourin R, Maupin P. Over fitting cautious selection of classifier ensembles with genetic algorithms. *Information Fusion*, 2009, **10**(2): 150−162

13 Jackowski K, Wozniak M. Method of classifier selection using the genetic approach. *Expert Systems*, 2010, **27**(2): 114−128

14 Banfield R E, Hall L O, Bowyer K W, Kegelmeyer W P. Ensemble diversity measures and their application to thinning. *Information Fusion*, 2005, **6**(1): 49−62

15 Didaci L, Giacinto G, Foli F, Marcialis G L. A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern Recognition*, 2005, **38**(11): 2188−2191

16 Didaci L, Giacinto G. Dynamic classifier selection by adaptive K-nearest neighborhood rule. In: Proceedings of the 5th International Workshop on Multiple Classifier Systems. Cagliari, Italy: Springer-Verlag, 2004. 174−183

17 Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensemble and their relationship with the ensemble accuracy. *Machine Learning*, 2003, **51**(2): 181−207

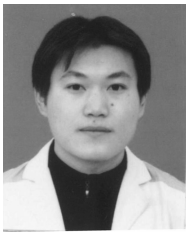
18 Liu C L, Hao H W, Sako H. Confidence transformation for combining classifiers. *Pattern Analysis and Application*, 2004, **7**(1): 2−17

19 Kherallah M, Haddad L, Alimi A M, Mitiche A. On-line handwritten digit recognition based on trajectory and velocity modeling. *Pattern Recognition Letters*, 2008, **29**(5): 580−594

20 Chen Z. Handwritten digits recognition. In: Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition. Las Vegas, USA: CSREA, 2009. 690−694

21 Hao Hong-Wei, Jiang Rong-Rong. Training sample selection method for neural networks based on nearset neighbor rule. *Acta Automatica Sinica*, 2007, **33**(12): 1247−1251 (郝红卫, 蒋蓉蓉. 基于最近邻规则的神经网络训练样本选择方法. 自动化学报, 2007, **33**(12): 1247−1251)

郝红卫 北京科技大学教授. 主要研究方向为图像处理与模式识别. 本文通信作者. E-mail: hhw@ustb.edu.cn
(**HAO Hong-Wei** Professor at University of Science and Technology Beijing. His research interest covers image processing and pattern recognition. Corresponding author of this paper.)



王志彬 北京科技大学博士研究生. 主要研究方向为图像处理与模式识别.
E-mail: wzb1818@yahoo.cn
(**WANG Zhi-Bin** Ph. D. candidate at University of Science and Technology Beijing. His research interest covers image processing and pattern recognition.)

殷绪成 北京科技大学副教授. 主要研究方向为模式识别与机器学习.
E-mail: xuchengyin@ustb.edu.cn
(**YIN Xu-Cheng** Associate professor at University of Science and Technology Beijing. His research interest covers pattern recognition and machine learning.)

陈志强 硕士. 主要研究方向为图像处理与模式识别. E-mail: loaf76@163.com
(**CHEN Zhi-Qiang** Master. His research interest covers image processing and pattern recognition.)