一种基于AP的仿生模式识别方法

丁 杰[1] 杨静宇[2]
(国网电力科学研究院信通所 南京 211106)[1] (南京理工大学计算机学院 南京 210094)[2]

摘 要 提出了一种基于仿射传播聚类(Affinity Propagation Clustering,AP Clustering)和仿生模式识别理论(Biomimetic Pattern Recognition,BPR)的识别方法。该方法通过AP聚类选择代表训练样本,依据仿生模式识别理论构建并划分样本空间,通过计算待识样本到各特征子空间的相对距离,根据其所处空间进行分类识别。在因空间重叠造成拒识的情况下,通过计算基于类条件的后验概率对样本进行相对区别。在Concordia大学CENPARMI手写体数字库与南京理工大学手写金额库上进行了实验,结果表明,该方法在识别率方面优于传统的分类器。

关键词 AP聚类,仿生模式识别,后验概率,类条件置信变换,手写体数字识别
中图法分类号 TP391 **文献标识码** A

AP Clustering Based Biomimetic Pattern Recognition

DING Jie[1] YANG Jing-yu[2]
(Research Institute of Information Technology and Communication, State Grid Electric Power Research Institute, Nanjing 211106, China)[1]
(Department of Computer, Nanjing University of Science and Technology, Nanjing 210094, China)[2]

Abstract A classify based on AP Clustering and biomimetic pattern recognition was proposed. It can relatively classify the samples by calculating the distance to the relative subspace. The training sample space was constructed by the AP algorithm and bionic pattern recognition theory. The posterior probabilities based on the class condition were estimated to reduce the reject rate caused by the space overlapping with low misclassification. Experiments were performed with Concordia University CENPARMI's handwritten digit database and Nanjing University of Science and Technology's handwritten amount database. Experimental results indicate that the proposed classifier has a higher recognition rate than the traditional classifiers.

Keywords Affinity propagation clustering, Biomimetic pattern recognition, Posterior probability, Class-conditional confidence transformation, Handwritten digit recognition

1 引言

仿生模式识别[1]从人类认知过程出发,以高维空间几何为理论基础,将空间中简单的点的划分转换为点的覆盖过程,克服了传统模式识别将复杂的识别过程完全归纳为类别划分的弊端,已用于人脸确认[2]和人脸识别[3]等研究,显示了识别效果的优越性。仿生模式识别以拓扑空间中对高维流形进行覆盖为研究对象,即在高维空间中离开一条首尾相接的空间曲线的最小距离小于某一定值的所有点的集合,空间曲线可由筛选的样本点序列彼此相连的若干直线段逼近。然而对于大多数字符集,特别是非限制手写体字符,训练样本集中存在两方面问题:

(1)书写风格的差异造成字符模式的不稳定,训练集中不可能包含所有书写风格的样本,大规模的训练样本中存在冗余。

(2)训练样本集中存在噪声干扰,表现为特征空间中游离的样本点。

目前大多数方法[5-7,16]关注于特征提取、分类器设计和分类器组合,忽视了对训练样本集的构建。AP聚类算法[8]具有有效地构建代表样本集的效果,能够去除样本空间的冗余和离群样本干扰。本文采用AP聚类筛选代表样本点构建样本空间。在分类识别阶段,针对两类拒识情况:(1)因阈值较小而导致待识样本处于所有特征子空间之外,(2)因增大阈值出现空间重叠,依据待识样本与特征子空间的相对距离做置信评估,根据各类模式散布计算后验概率,实现样本的相对划分,提高识别率。

2 样本空间构成

2.1 代表样本集构建

Frey和Dueck[8]提出一种简单有效的基于消息传递的数据聚类方法,即Affinity Propagation(AP)方法,该方法能构建有效的代表训练样本集合,它将所有样本数据同时作为初始代表样本,每个样本数据被当作网络中的一个结点,然后AP算法递归地沿着网络的边缘传输实值信号,直到生成一个代表样本集及相应的聚类。代表样本集使得所有样本与其代表样本的相似度总和最大。

到稿日期:2010-06-10 返修日期:2010-09-01 本文受国家自然科学基金(60632050),国家"863"计划项目(2006AA01Z119)资助。
丁 杰(1983-),男,博士,研究员,主要研究方向为OCR相关技术,E-mail:dingjie_star@163.com;杨静宇(1941-),男,博士,教授,主要研究方向为模式识别与智能系统。

聚类样本 x_i 和 x_k 间的相似性度量 $s(i,k)$ 通过计算二者间的负欧式距离得到，即：

$$s(i,k)=-\|x_i-x_k\|^2 \tag{1}$$

依据样本间相似度构建相似度矩阵并设定参考值之后，数据样本间的两类消息（可靠度、有效度）分别采用不同的机制不断地更新，二者可认为是对数似然比。

可靠度 $r(i,k)$：反映了 x_k 作为 x_i 的代表样本的可信度，其更新规则如下：

$$r(i,k)\leftarrow s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \tag{2}$$

有效度 $a(i,k)$：反映了 x_i 选择 x_k 作为其代表样本的累计可信度，通过收集来自样本数据的信息而确定是否每一个候选代表样本都是个合适的代表样本。$a(i,k)$ 的更新规则如下：

$$a(i,k)\leftarrow\min\{0,r(k,k)+\sum_{i'\notin\{i,k\}}\min\{0,r(i',k)\}\} \tag{3}$$

AP 算法通过组合可靠度和有效度来确定代表样本，二者仅需要在数据对之间传输。对于样本 x_i，当 $a(i,k)+r(i,k)$ 值最大化时，x_k 即为 x_i 的代表样本；当局部 $a(i,k)+r(i,k)$ 值保持不变时，消息传递过程将停止。

对一些非限制手写体字符，由于书写风格的不同造成字符模式不稳定，要获得较高的识别率，往往需要大量的训练样本。然而在实际中，训练集不可能包含所有书写风格的样本，并且大规模训练样本中的许多样本对识别来说都是冗余的，因此需要重新构建样本空间。

设 c 类模式 $\{\omega_1,\omega_2,\cdots,\omega_c\}$，模式类 ω_i 有 n_i 个训练样本，记为 $X_{i,k}(k=1,2,\cdots,n_i)$，基于随机分组的思想，将模式类 ω_i 中的训练样本按顺序分割为 m_i 个样本数目相等的子集，对每个子集采用 AP 方法来获得子代表样本，记为 $\hat{X}_{i,k}(k=1,2,\cdots,m_i)$，合并各子集的代表样本，构建模式类 ω_i 的代表样本集 $\hat{X}_i$。

2.2 子空间的构成

仿生模式识别认为，同类样本在模式空间中是连续的，即特征空间中同类样本点集 A 满足：

$$\forall x,y\in A,\forall\varepsilon>0,\exists B\subset A \tag{4}$$

$$B=\{x_m\mid x_1=x,x_n=y,d\{x_m,x_{m+1}\}<\varepsilon,1\leqslant m\leqslant n-1\}$$

$d(x,y)$ 表示样本点 x,y 间的距离度量。区别于传统模式识别以不同类样本在特征空间的最佳划分为目标，仿生模式识别以一类样本在特征空间的分布的最佳覆盖作为目标。

特征空间 R^n 中模式类 ω_i 的最佳覆盖近似于 ω_i 的代表样本集与 n 维超球体的拓扑乘积，即以 $\hat{X}_i$ 中代表样本 $\hat{X}_{i,k}$ $(k=1,2,\cdots,m_i)$ 为球心、常数 r 为半径的 m_i 个 n 维超球体的并。

构造代表样本集 $\hat{X}_i$ 中代表样本序列 $Seq(\hat{X}_i)=\{\hat{Y}_{i,1},\hat{Y}_{i,2},\cdots,\hat{Y}_{i,m_i}\}$ 如下：

$$\begin{cases}\hat{Y}_{i,1}=\hat{X}_{i,1}\\ A_{k-1}=\hat{X}_i-\{\hat{Y}_{i,1},\hat{Y}_{i,2},\cdots,\hat{Y}_{i,k-1}\}(1<k\leqslant m_i)\\ \hat{Y}_{i,k}=\hat{X}_{i,l},l=\arg(\min_{\hat{X}\in A_{k-1}}(d(\hat{Y}_{i,k-1},\hat{X})))\end{cases} \tag{5}$$

可以证明 $Seq(\hat{X}_i)$ 满足如下条件：

$$\exists\varepsilon>0,d(\hat{Y}_{i,k},\hat{Y}_{i,k+1})\leqslant\varepsilon<d(\hat{Y}_{i,k-1},\hat{Y}_{i,k+1}) \tag{6}$$

则特征空间 R^n 中对模式类 ω_i 近似覆盖的子空间 Ω_i 由以下公式给出：

$$\begin{cases}\Omega_{i,j}=\{x\mid d(x,y)\leqslant r_i,y\in\Lambda_{i,j},x\in R^n\}\\ \Lambda_{i,j}=\{y\mid y=\alpha\hat{Y}_{i,j}+(1-\alpha)\hat{Y}_{i,(j+1)\bmod m_i},\alpha\in[0,1]\}\\ \Omega_i=\bigcup_{j=1}^{m_i}\Omega_{i,j}\end{cases} \tag{7}$$

式中，r_i 为模式类 ω_i 的最小类间距离，计算如下：

$$r_i=\min(\min_{j\neq i}\|x-y\|),x\in w_i,y\in w_j \tag{8}$$

3 基于类条件的空间覆盖识别过程

对待识样本 X，其特征为 x，定义其到各类子空间 $\Omega_i(1\leqslant i\leqslant c)$ 覆盖区的最小距离 ρ_i 如下：

$$\begin{cases}\rho_i=\min_{j=1}^{m_i}(\|x-\Lambda_{i,j}\|)\\ \Lambda_{i,j}=\{y\mid y=\alpha\hat{Y}_{i,j}+(1-\alpha)\hat{Y}_{i,(j+1)\bmod m_i},\alpha\in[0,1]\}\end{cases} \tag{9}$$

设样本 X 最近邻类为 ω_i，即：

$$\rho_i=\min_{1\leqslant k\leqslant c}(\rho_k) \tag{10}$$

与前面所定阈值 r_i 相比，若满足 $\rho_i<r_i$ 且 $\forall k,k\neq i$ 有 $\rho_k>r_k$，则认为 X 属于子空间 Ω_i。由于噪声的存在和高维空间的不可预料性，识别过程在以下情形产生拒识：

(1) X 越出子空间覆盖范围，即样本 X 不属于任何模式类对应的子空间。

(2) 子空间可能会出现重叠，使得样本 X 处于多个子空间的交叉重叠处。

后验概率估计是模式分类器组合方法的基础，后验概率被认为集中在待识样本的最近邻类与次近邻类上[10]，因此，考虑待识样本的次近邻类是必要的。设 X 的次近邻类为 ω_j，即：

$$\rho_j=\min_{1\leqslant k\leqslant c,k\neq i}(\rho_k) \tag{11}$$

ρ_i 与 ρ_j 差别越大，决策结果越可靠、越可信，反之则越不可靠、越不可信。1NN 分类器的置信度[9]定义为：

$$Conf(X)=1-\frac{\rho_i}{\rho_j} \tag{12}$$

一般地，特征空间中各个模式类的散布矩阵是彼此不等的。假设 X 的最近邻类 ω_i 和次近邻类 ω_j 分别具有散布矩阵 Σ_i 和 Σ_j（见图 1），则上述 1NN 分类器对 $X\in\omega_i$ 的置信度不高，而此时认为 $X\in\omega_i$ 的后验概率大于 $X\in\omega_j$ 的后验概率是合理的。因此，后验概率不仅与置信度 $Conf(X)$ 有关，而且与模式类的分布有关，形式如下：

$$P(\omega_i\mid X)=\frac{P(C(X)=\omega_i,X\in\omega_i\mid Conf(X))}{P(C(X)=\omega_i\mid Conf(X))} \tag{13}$$

假设对每一个模式类 ω_k，都存在一个置信变换函数 $h_k(\cdot)(1\leqslant k\leqslant c)$，可将置信区间分成若干连续区间，后验概率通过区间端点插值估计[10]，从而样本 X 具有最近邻类 ω_i 和次近邻类 ω_j 时，后验概率估计为：

$$\begin{cases}P(\omega_i\mid X)=h_i(Conf(X))\\ P(\omega_j\mid X)=1-P(\omega_i\mid X)\\ P(\omega_k\mid X)=0\quad k\neq i,k\neq j,1\leqslant k\leqslant c\end{cases} \tag{14}$$

待识样本 X 具有多个特征 $x_1,x_2,\cdots,x_T$，对特征 $x_t(1\leqslant t\leqslant T)$，使用 K 个分类器识别决策的后验概率 $P_k(\omega_i\mid x_t)(1\leqslant k\leqslant K)$。Kittler 加法规则[11]给出的多分类器组合的后验

概率式(15)、判别 $X\in\omega_\zeta$ 规则式(16)如下：

$$P(\omega_i|X)=\frac{1}{KT}\sum_{t=1}^{T}\sum_{k=1}^{K}P_k(\omega_i|x_t) \tag{15}$$

$$\zeta=\arg\left(\max_{1\leqslant k\leqslant c}(P(\omega_k|X))\right) \tag{16}$$

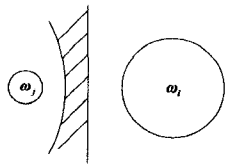

图 1　具有不同散布矩阵的两类

4　实验结果及分析

4.1　实验 1

数据集使用了扩充的 NUST 手写金额库。NUST 手写金额字符库包括了手写数字库和手写汉字库，是对数千张银行实际使用的支票经过定位、分割以及二值化等前期处理后，在大小写金额区域上采样，生成的一个较大规模的手写体字库，扫描分辨率为 200DPI。手写体数字库共有 30000 幅字符图像，其中 20000 幅字符图像（每类 2000 个样本）作为训练集，其余 10000 幅（每类 1000 个样本）作为测试集。一组 NUST 样本示例见图 2。在字符集上提取如下 5 组特征：

特征 1　笔画归正图像的点阵特征。在水平和垂直方向上对图像进行伸缩变换，并对图像作单侧的腐蚀和膨胀运算。将字符大小和笔画宽度变换为固定量值，划分图像为固定单位并统计和排列各单位的前景点数。

特征 2　笔画归正图像的模糊笔画方向统计特征。对图像作线性规范化处理并抽取基于模糊笔画方向的统计特征[12]。

特征 3　非线性归正图像的模糊笔画方向统计特征。对图像作非线性规范化处理并抽取基于模糊笔画方向的统计特征[12]。

特征 4　链码特征。对归一化的字符图像采用 8 方向链码跟踪并描述轮廓[13]。

特征 5　梯度特征。归一化图像上各像素的梯度，将梯度向量按链码方向分解成梯度方向平面，对各平面使用高斯模板作卷积，生成特征向量[13]。

表 1　NUST 手写数字库上的组合识别结果

方法	子集数	单特征分类器错误率(%)					组合错误率(%)
		特征 1	特征 2	特征 3	特征 4	特征 5	
AP	10	10.31	10.42	10.12	10.79	9.87	9.46
	20	9.56	9.43	8.54	9.01	8.72	8.25
	40	7.31	7.56	7.11	6.97	6.86	6.43
	80	6.78	6.61	6.52	6.71	6.64	6.31
	100	6.13	5.97	5.68	6.56	5.23	5.02
	200	6.22	6.03	5.92	6.43	5.46	5.27
AP+BPR	10	5.11	4.96	4.68	4.44	4.12	3.83
	20	4.37	4.21	3.66	3.46	3.26	3.08
	40	3.22	2.63	2.40	2.43	2.04	1.38
	80	2.59	1.89	1.58	2.12	1.68	1.87
	100	2.09	1.77	1.32	2.22	1.66	0.76
	200	2.29	1.57	1.42	2.18	1.34	0.88

采用随机分组的思想将 NUST 数字金额库的训练集划分成若干子集，子集数从 10 到 200，再基于 AP 形成代表样本集。用本文方法，进行基于上述 5 组特征的单分类器组合识别实验，分类器为最近邻距离分类器，识别错误率见表 1。针对 NUST 数据集，AP 和 AP+BPR 方法基于 5 类特征的随机分组与最终识别结果关系见图 3。从图中可以看出，随着子集数目的增加，字符的识别率呈上升趋势，在子集数目达到 100 左右时趋于稳定。

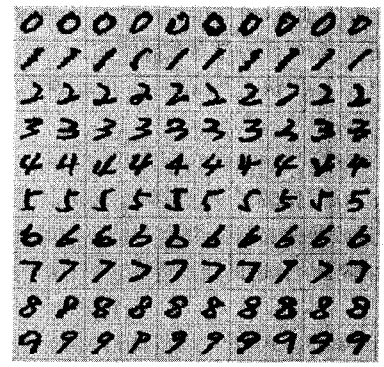

图 2　NUST 手写体数字库样本示例

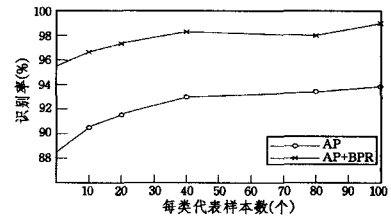

图 3　NUST 库样本分组与识别率关系图

4.2　实验 2

为了验证本文所提方法的有效性，在 Concordia 大学的 CENPARMI 库上将本方法与其他分类器进行了比较[14,15]。CENPARMI 数据集包含了 6000 幅在 USPS 的信封图像上采集到的数字字符图像，扫描分辨率为 166DPI；其中 4000 幅数字图像（每类 400 个样本）指定为训练集，一组 CENPARMI 样本示例见图 4。

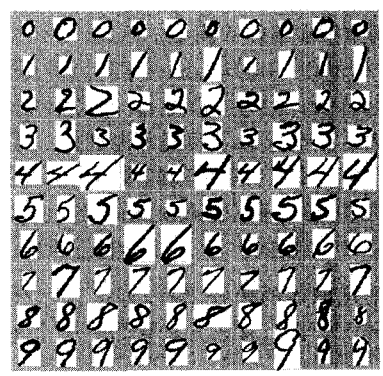

图 4　CENPARMI 手写体数字库样本示例

文献[14]针对支持向量机问题提出了基于先验知识的支持向量变换，从而提高了特征空间中的分类面稳定性和可靠性。文献[15]基于分步解决（divide and conquer）的原则，提出了一种通用的分类器学习框架，它有效地降低了多层分类器的复杂度，在小样本集上同样取得了较好的分类识别效果。实验中将本文算法与上述方法进行了比较，比较实验中划分的子集数为 100，比较识别的统计结果见表 2，其中各项指标定义如下：

（下转第 230 页）

和变异对种群多样性影响的内在机制如何等，通过这些内在机制的研究，找出提升种群多样的策略，都是下一步需要进行研究的工作。

参 考 文 献

[1] Kennedy J, Eberhart R C. Particle swarm optimization [C]// Proceedings of IEEE International Conference on Neural Networks. Piscataway, USA, 1995: 1942-1948

[2] Clerc M, Kennedy J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space [J]. IEEE Transactions on Evolutionary Computation, 2002, 2(6): 58-73

[3] Peram T, Veeramachaneni K, Mohan C K. Fitness-distance-ratio based particle swarm optimization[C]//Proceedings of the IEEE Swarm Intelligence Symposium. Indianapolis, Indiana, 2003: 174-181

[4] Mendes R, Kennedy J, Neves J. The fully informed particle swarm: Simpler, maybe better [J]. IEEE Transactions on Evolutionary Computation, 2004, 7(8): 204-210

[5] van den Bergh F, Engelbrecht A P. A cooperative approach to particle swarm optimization [J]. IEEE Transactions on Evolutionary Computation, 2004, 6(8): 225-239

[6] 刘衍民，赵庆祯. 一种基于动态邻居和变异因子的粒子群算法[J]. 控制与决策，2010，25(7)：968-974

[7] Stacey A, Jancic M, Grundy I. Particle swarm optimization with mutation[C]// Proceeding of IEEE Congress on Evolutionary Computation. Canberra, Australia, 2003: 1425-1430

[8] Deb K, Goyal M. A combined genetic adaptive search(GeneAS) for engineering design [J]. Computer Science and Informatics, 1996, 26(4): 30-45

[9] Martinez S Z, Coello C. Hybridizing an evolutionary algorithm with mathematical programming techniques for multi-objective optimization [C] // Proceedings of Genetic and Evolutionary Computation Conference. New York, USA, 2008: 769-770

[10] Shi X H, Liang Y C, Lee H P, et al. Particle swarm optimization-based algorithms for TSP and generalized TSP [J]. Information Processing Letters, 2007, 103(5): 169-176

[11] Deb K, Agrawal R B. Simulated binary crossover for continuous search space [J]. Complex Systems, 1995, 9(3): 115-148

(上接第 226 页)

$$\text{错误率}=\frac{N_e}{N}\times 100\% \tag{17}$$

$$\text{拒识率}=\frac{N_r}{N}\times 100\% \tag{18}$$

$$\text{可靠性}=\frac{N-N_e-N_r}{N-N_r}\times 100\% \tag{19}$$

式中，N_e 为错误识别的样本数，N_r 为拒识的样本数，N 为样本总数。从表 2 数据可以看出本文提出的基于 AP 的仿生模式识别方法对训练集进行筛选时，使用了原训练集中部分代表样本作为样本空间的特征点集，并以此构建特征子空间。识别过程依据子空间的相对划分完成，获得了较为理想的识别结果，识别结果明显优于其它分类器，证明本文方法能去除离群样本干扰，构建有效的代表样本集，同时能够保持字符模式的稳定性。

表 2 CENPARMI 库上的比较识别结果

样本数	分类器	错误率(%)	拒识率(%)	可靠性(%)
6000	K nearest neighbors	2.71	7.28	97.29
	Virtual SVM[14]	1.28	0	98.72
	Local Learning Framework[15]	1.86	0	98.14
	AP+BPR	0.86	1.45	99.14

结束语 本文提出了一种基于 AP 的仿生模式识别方法。从理论上讲，识别算法在训练集规模越大时越好，但同时所需的训练时间也越多，限制了其在计算、存储资源有限的特定系统中的应用。基于空间覆盖的识别方法，在小样本识别方面具有传统识别方法不可替代的优越性，而 AP 方法在样本筛选方面取得了较好的结果，类条件后验概率估计也在一定程度上提高了识别率。但在样本空间的构建和降低时间复杂度方面还需作进一步的研究。

参 考 文 献

[1] 王守觉. 仿生模式识别(拓扑模式识别)——一种模式识别新模型的理论与应用[J]. 电子学报，2002，30(10)：1417-1420

[2] 王守觉，徐健，王宪保，等. 基于仿生模式识别的多镜头人脸身份确认系统研究[J]. 电子学报，2003，31(1)：1-5

[3] 王守觉，曲延锋，李卫军，等. 基于仿生模式识别与传统模式识别的人脸识别效果比较研究[J]. 电子学报，2004，33(7)：1057-1061

[4] 高庆吉，王晓华，赵为平. 对粘连和缺损数字串分割的研究[J]. 模式识别与人工智能，2000，13(1)：99-102

[5] Teow L N, Loe K F. Robust vision-based features and classification schemes for off-line handwritten digit recognition[J]. Pattern Recognition, 2002, 35(3): 2355-2364

[6] 荆晓远，杨静宇. 基于相关性和有效互补性分析的多分类器组合方法[J]. 自动化学报，2000，26(6)：741-747

[7] Liu C L, Nakashima K, Sako H, et al. Handwritten digit recognition: benchmarking of state-of-the-art techniques [J]. Pattern Recognition, 2003, 36: 2271-2285

[8] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976

[9] Smith S J, Bourgoin M O, Sims K, et al. Handwritten character classification using nearest neighbor in large databases[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(9): 915-919

[10] 娄震，金忠，杨静宇. 基于类条件置信变换的后验概率估计方法[J]. 计算机学报，2005，28(1)：18-24

[11] Kittle J, Hatef M, Duin R P W. On combining classifiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(3): 226-239

[12] 王正群，叶晖，孙兴华，等. 基于模糊方向特征的手写体汉字识别[J]. 模式识别与人工智能，2001，14(3)：317-320

[13] Cheng L L, Nakashima K, Sako H, et al. Handwritten digit recognition: investigation of normalization and feature extraction techniques[J]. Pattern Recognition, 2004, 37(2): 265-279

[14] Scholkoph B, Burges C, Vapnik V. Incorporating invariances in support vector learning machines[C]// The International Conference on Artificial Neural Networks. Springer Lecture Notes in Computer Science, 1996, 1112: 47-52

[15] Dong J X, Krzyzak A, Suen C Y. Local learning framework for handwritten character recognition[J]. Engineering Applications of Artificial Intelligence, 2002, 15(2): 151-159

[16] Wierer J, Boston N. A handwritten digit recognition algorithm using two-dimensional Hidden Markov Models for feature extraction[C]// The 32nd IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu, Hawaii, USA, 2007: 505-508