

# 正则化流形信息极端学习机

刘德山, 楚永贺, 闫德勤

(辽宁师范大学计算机与信息技术学院, 辽宁 大连 116081)

**摘 要:** 基于流形学习的思想和理论方法, 提出刻画流形信息的正则化的极端学习机(MELM)算法。该算法利用流形信息刻画数据的几何结构和判别信息, 克服 ELM 在有限样本上学习不充分的问题; 能够有效提取数据样本的判别信息避免数据样本信息重叠; 利用最大边准则有效解决类间散度矩阵和类内散度矩阵的奇异问题。为验证所提方法的有效性, 实验使用普遍应用的图像数据, 将 MELM 与 ELM 以及相关最新算法 RAFELM、GELM 进行识别率和计算效率的对比。实验结果表明, 该算法能够显著提高 ELM 的分类准确率和泛化能力, 并且优于其他相关算法。

**关键词:** 极端学习机; 几何结构; 流形信息; 机器学习

**中图分类号:** TP18

**文献标识码:** A

## Regularized manifold information extreme learning machine

LIU De-shan, CHU Yong-he, YAN De-qin

(College of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China)

**Abstract:** By exploiting the thought of manifold learning and its theoretical method, a regularized manifold information extreme learning machine algorithm aimed to depict and fully utilize manifold information was proposed. The proposed algorithm exploited the geometry and discrimination manifold information of data to perform network of ELM. The proposed algorithm could overcome the problem of the overlap of information. Singular problems of inter-class and within-class were solved effectively by using maximum margin criterion. The problem of inadequate learning with limited samples was solved. In order to demonstrate the effectiveness, comparative experiments with ELM and the related update algorithms RAFELM, GELM were conducted using the commonly used image data. Experimental results show that the proposed algorithm can significantly improve the generalization performance of ELM and outperforms the related update algorithms.

**Key words:** extreme learning machine, geometry, manifold information, machine learning

### 1 引言

近年来, 极端学习机(ELM, extreme learning machine)<sup>[1,2]</sup>的研究成为一个热点领域。ELM 随机产生隐层节点的输入权值和偏置值, 所有参数中仅有输出权值需经过分析确定, 从而将传统的神经网络的求解过程转化为一个线性模型。文献[3, 4]通过分析指出 ELM 在随机产生隐层节点的输入权值和偏置值, 分析确定输出权值保持了 SLFN(single-hidden layer feedforward network)的通用逼近能力, 与基于

梯度思想的传统神经网络方法相比具有更高的效率, 同时能够得到一个全局最优解。文献[5~7]指出 ELM 与支持向量机(SVM, support vector machine)<sup>[8]</sup>及其改进算法相比具有更好的泛化能力。从 2006 年至今, ELM 在理论和应用这 2 个方面取得了很大的进展。文献[9]进一步证实了 ELM 的通用逼近能力, 文献[10]在局部泛化误差模型下进一步探究了 ELM 的泛化能力, 对于具体问题的特殊需求已在 ELM 算法的基础上提出了各种改进。例如将 ELM 应用于在线连续数据学习<sup>[11-13]</sup>、不平衡数据学习<sup>[14]</sup>、噪声及

收稿日期: 2016-04-12; 修回日期: 2016-10-10

基金项目: 国家自然科学基金资助项目(No.61105085); 辽宁省教育厅基金资助项目(No.L2014427)

**Foundation Items:** The National Natural Science Foundation of China (No.61105085), Liaoning Provincial Department of Education Project(No.L2014427)

缺失数据<sup>[15,16]</sup>、人脸识别<sup>[17]</sup>等。近年来,研究者将并行技术应用到 ELM 中,使 ELM 在处理大数据和实时推理时具有更好的稳定性。

ELM 在模式识别领域得到了广泛的应用,然而在有监督的学习中,由于数据样本的有限,使 ELM 学习不充分而限制了 ELM 的泛化能力。为了克服 ELM 所存在的上述缺陷,文献[17]提出了一种基于图判别信息正则化的极端学习机(GELM, discriminative graph regularized extreme learning machine), GELM 结合数据样本的判别信息,构造拉普拉斯特征映射(LE, Laplacian eigenmap)<sup>[18]</sup>结构,将其作为正则项引入到 ELM 中。该方法不仅继承了 ELM 的优点,并在一定程度上克服了 ELM 在学习过程中训练不充分及小样本问题,同时,在学习过程中考虑到数据的类内几何结构及判别信息。

GELM 利用 LE 获取了数据局部几何结构信息,但却无法对数据的全局几何结构信息进行有效地获取。从聚类角度看,不同类的数据样本位于不同的流形结构中,利用局部映射(如 LE)算法将数据映射到低维空间时会出现信息重叠的现象, GELM 没有考虑到数据样本间的差异判别信息,在一定程度上限制了 GELM 的泛化能力。基于以上分析,本文提出一种基于流形信息正则化极端学习机(MELM, manifold information regularized extreme learning machine), MELM 充分考虑到数据样本的几何特征及其所蕴含的判别信息,通过最大化类间散度矩阵和最小化类内散度矩阵,起到优化极端学习机输出权值的作用,从而提高 ELM 的泛化能力。

MELM 方法的创新在于以下 3 点。

1) 将局部保持投影 (LPP, locality preserving projection)算法思想<sup>[19]</sup>引入 ELM 中,有效获取数据的全局和局部几何结构信息,克服 ELM 在有限样本上学习不充分的问题。

2) 利用数据样本的判别信息将类间散度矩阵和类内散度矩阵引入 ELM 中,有效避免数据样本信息重叠的现象。

3) 利用最大边准则(MMC, maximum margin criterion)<sup>[20]</sup>方法有效解决类间散度矩阵和类内散度矩阵的奇异问题。

通过与 ELM、RAFELM<sup>[21]</sup>、GELM<sup>[17]</sup>算法的比较,实验结果表明,本文所提算法有较好的识别性能和计算效率,优于对比算法。

## 2 ELM

对于  $N$  个不同的样本  $(x_j, t_j)$ , 可表示为矩阵  $X = (x_1, x_2, \dots, x_N)^T \in R^{D \times N}$ , 其中, 向量  $t_j = (t_{j1}, t_{j2}, \dots, t_{jm})^T \in R^m$ , 具有  $L$  个隐层节点的激活函数为  $g(x)$  的 ELM 模型为

$$\sum_{i=1}^L \beta_i g(\mathbf{a}_i x_j + b_i) = t_j \quad (1)$$

其中,  $j=1, 2, \dots, N$ ,  $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{im})$  为连接第  $i$  个隐层节点与输入节点的输入权值向量,  $\beta_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{im})$  为连接第  $i$  个隐层节点与输出节点的输出权值向量,  $b_i$  为第  $i$  个隐层节点的偏置值,  $\mathbf{a}_i \cdot x_j$  表示  $\mathbf{a}_i$  和  $x_j$  的内积,  $t_j = (t_{j1}, t_{j2}, \dots, t_{jm})^T \in R^m$  为对应于样本  $x_j$  的期望输出向量。对所有数据样本进行整合, 式(1)可以改写为

$$H\beta = T \quad (2)$$

其中,  $H$  是网络隐层节点输出矩阵,  $\beta$  为输出权值矩阵,  $T$  为期望输出矩阵。

$$H = \begin{pmatrix} g(\mathbf{a}_1 \cdot x_1 + b_1) & \cdots & g(\mathbf{a}_N \cdot x_1 + b_N) \\ \vdots & \ddots & \vdots \\ g(\mathbf{a}_1 \cdot x_N + b_1) & \cdots & g(\mathbf{a}_N \cdot x_N + b_N) \end{pmatrix}_{N \times L} \quad (3)$$

$$\beta = \begin{pmatrix} \beta_1^T \\ \vdots \\ \beta_L^T \end{pmatrix}_{L \times m}, T = \begin{pmatrix} t_1^T \\ \vdots \\ t_N^T \end{pmatrix} = \begin{pmatrix} t_{11} \cdots t_{1m} \\ \vdots \\ t_{N1} \cdots t_{Nm} \end{pmatrix} \quad (4)$$

当隐层节点个数和训练样本个数相同时 (即  $L = N$ ), 本文通过式(2)直接求矩阵  $H$  的逆矩阵得到最优的输出权值矩阵  $\beta$ , 但大多数情况下隐层节点的个数远小于训练样本的个数 (即  $L \ll N$ ), 此时, 矩阵  $H$  为奇异矩阵, 本文利用最小二乘解的方法对式(2)进行求解。

$$\hat{\beta} = \arg \min_{\beta} \|H\beta - T\|^2 = H^+ T \quad (5)$$

其中,  $H^+$  为矩阵  $H$  的广义逆矩阵。

为了提高传统 ELM 的稳定性和泛化能力, Huang<sup>[22]</sup>提出了等式优化约束的 ELM (equality optimization constraints based ELM), 等式优化约束的 ELM 优化式子的特点是不仅最小化训练误差  $\xi$ , 同时最小化输出权值  $\beta$ , 因此, 等式优化约束的 ELM 目标式子可写为

$$\min_w \frac{1}{2} \|\beta\|^2 + \frac{1}{2} C \sum_{i=1}^N \xi_i^2$$

$$\text{s.t. } h(x_i)\beta = t_i^T - \xi_i^T, \quad i = 1, 2, \dots, N \quad (6)$$

其中,  $\xi_i = (\xi_{i1}, \dots, \xi_{im})^T$  为对应于样本  $x_i$  的训练误差向量,  $C$  为惩罚参数。式(6)的求解可通过拉格朗日方法转化为无条件最优化问题进行求解。

### 3 MELM

为了克服 ELM 训练过程中学习不充分以及出现的数据样本信息重叠问题, 在本节提出一种基于流形信息正则化极端学习机算法。

#### 3.1 局部保持投影算法思想的引入

局部保持投影算法构造了数据样本的邻接图矩阵, 通过投影变换将数据样本映射到低维空间, LPP 作为 LE 算法的线性近似, 保持了 LE 算法良好的局部性质, 又能够获取数据样本全局的几何结构信息。因此, 将 LPP 算法引入到 ELM 优化模型中, 以获取数据的局部及全局几何结构信息, 从而提高 ELM 的泛化能力。

对于给定的数据集矩阵  $X = (X_1, X_2, \dots, X_M) \in R^{D \times N}$ ,  $M$  为数据样本的类别数,  $x_i^m$  为第  $m$  类第  $i$  个样本, 第  $m$  类数据样本个数为  $n_m$ , 数据样本总数为  $N$ , 则 LPP 的优化可表示为

$$J = \sum_{m=1}^M \sum_{i,j=1}^{n_m} \left| \mathbf{u}^T x_i^m - \mathbf{u}^T x_j^m \right|^2 W_{ij}^m \quad (7)$$

其中,  $\mathbf{u}^T$  为投影变换矩阵,  $W_{ij}^m$  为邻接权值矩阵, 可采用  $k$  近邻法对  $W_{ij}^m$  进行定义。

$$W_{ij}^m = \begin{cases} \exp\left(-\frac{\|x_i^m - x_j^m\|^2}{2}\right), & \\ x_i^m \in N_k(x_j^m) \text{ 或 } x_j^m \in N_k(x_i^m), & \\ \text{并且 } x_i^m \text{ 和 } x_j^m \text{ 属于同一类} & \\ 0, \text{ 其他} & \end{cases} \quad (8)$$

对式(7)进行推导为

$$\begin{aligned} J &= \sum_{m=1}^M \sum_{i,j=1}^{n_m} (\mathbf{u}^T x_i^m - \mathbf{u}^T x_j^m)^2 W_{ij}^m \\ &= \sum_{m=1}^M \left( \sum_{i=1}^{n_m} \mathbf{u}^T x_i^m D_{ii}^m (x_i^m)^T \mathbf{u} - \sum_{i,j=1}^{n_m} \mathbf{u}^T x_i^m W_{ij}^m (x_j^m)^T \mathbf{u} \right) \\ &= \sum_{m=1}^M \mathbf{u}^T X_m (D_m - W_m) X_m^T \mathbf{u} = \mathbf{u}^T X (D - W) X^T \mathbf{u} \\ &= \mathbf{u}^T XLX^T \mathbf{u} \end{aligned} \quad (9)$$

其中,  $D_{ii}^m = \sum_j W_{ij}^m$ ,  $L = D - W$  为拉普拉斯矩阵。

LPP 的优化目标可改写为

$$\begin{aligned} \min \mathbf{u}^T XLX^T \mathbf{u} \\ \mathbf{u}^T XDX^T \mathbf{u} = 1 \end{aligned} \quad (10)$$

LPP 通过式(10)寻找最佳的投影方向保持数据的内在几何结构, 能够很好地获取数据的内在几何结构信息。

#### 3.2 MELM 的提出

定义 1 令

$$Z_m = X_m (D_m - W_m) X_m^T \quad (11)$$

数据样本的类内离散度矩阵  $S_B$  以及数据样本的类间离散度矩阵  $S_W$  分别定义为

$$S_B = \sum_{m=1}^M Z_m \quad (12)$$

$$S_W = \sum_{i=1}^M (\mathbf{u}^i - \mathbf{u})(\mathbf{u}^i - \mathbf{u})^T \quad (13)$$

式(11)中  $D_m$  是一个对角矩阵, 其对角线上的元素为  $(D_m)_{ii} = \sum_j W_{ij}^m$ ,  $W^m$  为第  $m$  类样本的邻接权值矩阵, 可定义为式(8)。

上述定义中,  $X_m$  为第  $m$  类所有样本构成的样本矩阵,  $M$  为数据样本的类别个数,  $\mathbf{u}^i$  为数据样本的类内样本均值向量,  $\mathbf{u}$  为数据样本的总体均值向量, 类间离散度矩阵和类内离散度矩阵分别体现了输入数据样本空间的分布特征和判别信息。

定义 2 矩阵  $S = S_B - (1 - \lambda)S_W$  ( $0 \leq \lambda \leq 1$ ) 称为信息差距阵,  $\lambda$  为大于零的常量。

上述定义中, 参数  $\lambda$  起调节类内判别信息和类间判别信息的作用, 当参数  $\lambda$  增大时偏向于类内信息; 反之, 加大惩罚类间差异信息。因此, 在适当的  $\lambda$  下, 利用信息差别矩阵能够较好地利用数据蕴含的判别信息, 增强了模式分类的能力。

构造 MELM 的优化模型为

$$\begin{aligned} \min_W \text{Tr} \left( \frac{1}{2} W^T S W \right) + \frac{1}{2} C \sum_{i=1}^N \varepsilon_i^2 \\ \text{s.t.} \sum_{i=1}^L w_i f(\mathbf{a}_i \cdot \mathbf{x} + b_i) - \mathbf{t}_j^T = \varepsilon_i^T, \quad j = 1, 2, \dots, N \end{aligned} \quad (14)$$

其中,  $W$  为输出权值矩阵,  $\boldsymbol{\varepsilon}$  为训练误差向量,  $C$  为惩罚参数,  $f(x)$  为激活函数。

依据最优化条件(KKT)理论和文献[22]对 ELM 模型的推导过程, ELM 训练模型等价于解决一个双重优化问题。

$$L = \text{Tr} \left( \frac{1}{2} W^T S W \right) + \frac{1}{2} C \sum_{i=1}^N \varepsilon_i^2 + \sum_{i=1}^N \sum_{j=1}^m \alpha_{ij} (f(x_i) w_j - \mathbf{t}_{ij} + \varepsilon_{ij}) \quad (15)$$

其中,  $w_j$  为第  $j$  个输出节点的权值向量,  $W=[w_1, w_2, \dots, w_m]$ , 依据文献[22]对 ELM 模型的推导过程, 本文可以有相应的 KKT 最优条件如下

$$\frac{\partial L}{\partial w_j} = 0 \rightarrow w_j = \sum_{i=1}^N \alpha_{ij} f(x_i)^T \rightarrow WS = F^T \alpha \quad (16)$$

$$\frac{\partial L}{\partial \epsilon_i} = 0 \rightarrow \alpha_i = C \epsilon_i, i = 1, \dots, N \quad (17)$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow f(x_i)W - t_i^T + \epsilon_i^T = 0 \quad (18)$$

其中, 向量  $\alpha_i = [\alpha_{i1}, \dots, \alpha_{im}]^T$ ,  $\alpha = [\alpha_1, \dots, \alpha_N]$ ,  $F$  为隐层节点输出矩阵。

1) 当训练样本个数远大于隐层节点个数时, 即  $N \gg L$ , 通过式(16)和式(17)求得

$$\alpha = (F^T)^+ WS \quad (19)$$

$$\epsilon_i = \frac{\alpha_i}{C} \quad (20)$$

其中,  $(F^T)^+$  是  $F^T$  的广义逆矩阵。

由式(19)和式(20)可以得到

$$\epsilon = \frac{(F^T)^+ WS}{C} \quad (21)$$

将式(16)和式(21)代入式(18), 可以得到

$$FW - T + \frac{(F^T)^+ WS}{C} = 0 \rightarrow W = \left( F^T F + \frac{S}{C} \right)^{-1} F^T T \quad (22)$$

由式(22)得出 MELM 的输出函数为

$$g(x) = f(x)W = f(x) \left( F^T F + \frac{S}{C} \right)^{-1} F^T T \quad (23)$$

2) 当训练样本个数远小于隐层节点个数时即  $N \ll L$ , 由式(16)和式(17), 可以得到

$$W = F^T \alpha S^{-1} \quad (24)$$

$$\epsilon_i = \frac{\alpha_i}{C} \quad (25)$$

将式(24)和式(25)代入式(18), 可以得到

$$FF^T \alpha S^{-1} - T + \frac{\alpha}{C} = 0 \rightarrow \left( FF^T + \frac{S}{C} \right) \alpha = ST \rightarrow W = F^T \left( F^T F + \frac{S}{C} \right)^{-1} ST \quad (26)$$

由式(26)MELM 的输出函数为

$$g(x) = f(x)W = f(x)F^T \left( F^T F + \frac{S}{C} \right)^{-1} ST \quad (27)$$

给定训练样本向量  $X = (x_1, x_2, \dots, x_N)$  和训练样本的期望输出矩阵  $T = (t_1, t_2, \dots, t_D) \in R^{D \times N}$ , 激活函数为  $f(x)$ , 隐层节点数为  $L$  (隐层节点个数等于人脸图像数据的维数), MELM 算法总结步骤如下。

1) 初始化训练样本集, 利用 LPP 优化式(10)得到数据的投影。

2) 根据定义 1 计算  $S_B$  和  $S_W$ , 根据定义 2 计算  $S$ 。

3) 随机指定网络输入权值向量  $a_i$  和偏置值  $b_i$ 。

4) 通过激活函数计算隐层节点输出矩阵  $F$ 。

5) 计算输出权值,  $W = F^T \left( F^T F + \frac{S}{C} \right)^{-1} ST$  或

$$W = \left( F^T F + \frac{S}{C} \right)^{-1} F^T T。$$

由于本文所提出的算法需要引入类内散度矩阵和类间散度矩阵, 因此, 与 ELM 算法相比具有较高的空间复杂度, 尤其在高维数据样本中此种情况更为明显, 为了提高本文算法的执行效率, 在测试高维数据样本时使用 PCA 降维对数据进行降维处理。

## 4 实验结果及分析

为了验证所提出算法的有效性, 本文从 2 个方面进行实验, 首先将 MELM 与 ELM、RAFELM<sup>[21]</sup>, GELM<sup>[17]</sup>进行对比, 其次将 MELM 与卷积神经网络算法 (CNN) 和深度信念神经网络算法 (DBN)<sup>[23]</sup>进行对比。

实验数据使用广泛应用的图像数据集。部分数据集如图 1 所示。不同图像数据在实验中的设置如表 1 所示。

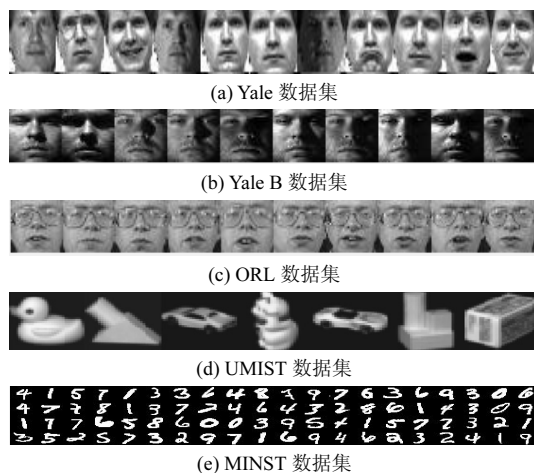


图 1 不同人脸图像的训练集

数据集	维数	样本总数	每类样本数	类别数
Yale	1 024	165	11	15
Yale B	1 024	2 414	55	38
ORL	1 024	400	10	40
UMIST	10 305	575	18	20
MINST	784	70 000	7 000	10

Yale 数据集包含了 15 人的 165 张图像，每个人包含 11 张图像，图像显示了每个人快乐、悲伤、正常等不同情况下的状态，图像大小为  $32 \times 32$  像素。Yale B 数据集包含了 38 人的 2 414 张图像，每个人包含 55 张图像，图像显示了每个人快乐、悲伤、正常等不同情况下的状态，图像大小为  $32 \times 32$  像素。ORL 数据集包含了 40 人的 400 张图像，每个人包含 10 张图像，本文选取了不同时期、不同光照条件下面部表情的图像，图像大小为  $32 \times 32$  像素。UMIST 数据集包含了 20 人的 575 张图像，选取了不同时期、不同光照条件下面部表情的图像，图像大小为  $112 \times 92$  像素。MINST 数据集

为手写数字图像，包含 70 000 张图像，图像大小为  $28 \times 28$  像素。

#### 4.1 ELM 相关算法比较

本节将 MELM 与 ELM、RAFELM<sup>[21]</sup>、GELM<sup>[17]</sup> 进行对比。在 Yale、ORL 实验中本文随机选取训练集个数为  $K=\{3, 4, 5, 6\}$ ，剩余部分为测试集；在 Yale B 实验中随机选取训练集个数为  $K=\{25, 30, 35, 40\}$ ，剩余部分为测试集。为了实验的公平性，4 种不同 ELM 算法均使用相同的惩罚参数  $C=2^{-8}$ ，对于 Yale 数据集，MELM 算法中调节参数设置为  $\lambda=0.4$ ；对于 ORL 数据集，MELM 算法中调节参数设置为  $\lambda=0.5$ ；对于 Yale B 数据集，MELM 算法中调节参数设置为  $\lambda=0.3$ 。MELM 与 ELM、RAFELM、GELM 的识别率曲线如图 2~图 4 所示。

图 2~图 4 给出了 4 种不同的 ELM 算法在 Yale、Yale B、ORL 数据上的识别率曲线。可以看出，本文所提方法 MELM 的识别率曲线显著高于其他算法的识别率，其原因在于 MELM 不仅考虑到同类数据样本间的判别信息，同时，还考虑到不同类数据样本间的差异判别信息，有效避免了运用 GELM

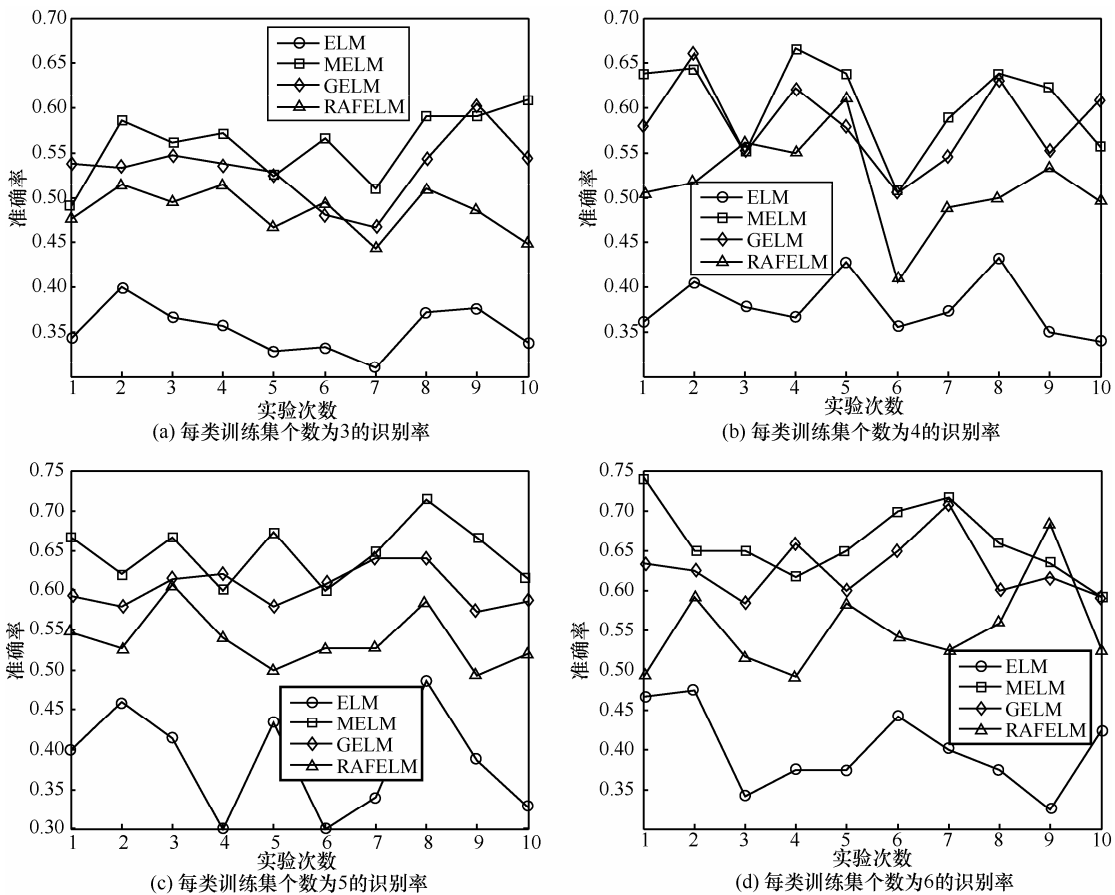


图 2 MELM、ELM、RAFELM、GELM 算法在 Yale 数据集上的识别率

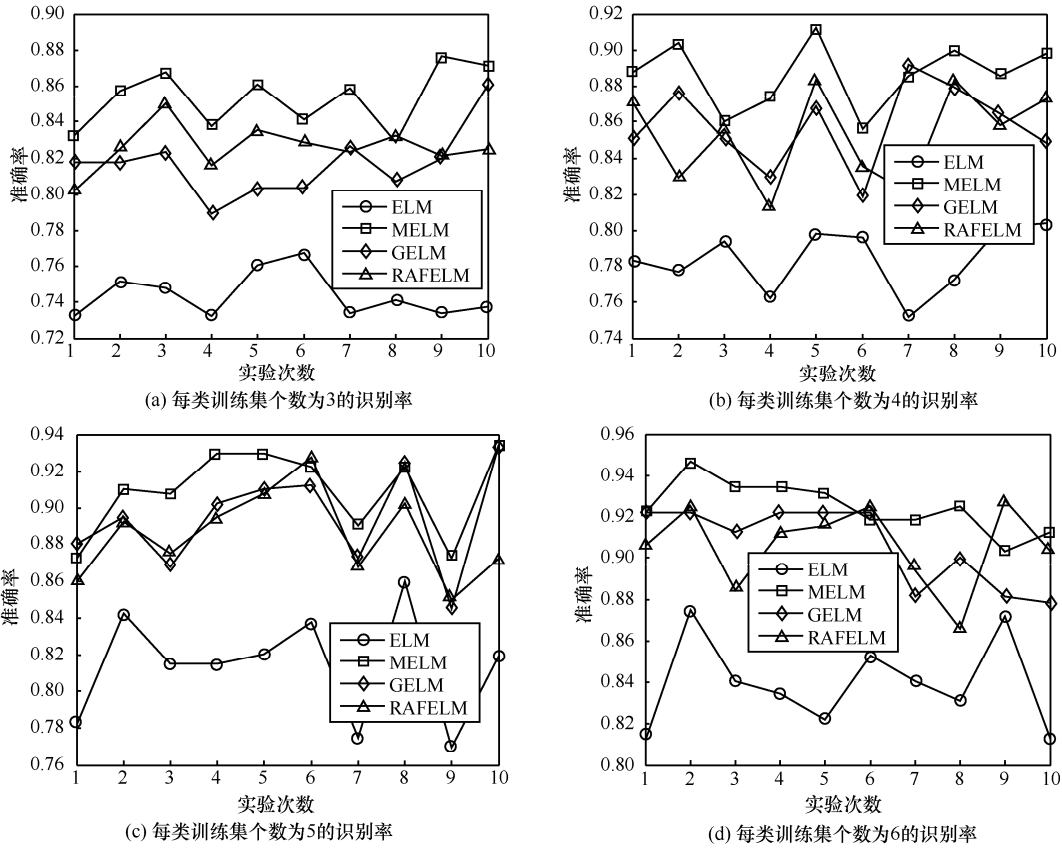


图 3 MELM、ELM、RAFELM、GELM 算法在 ORL 数据集上的识别率

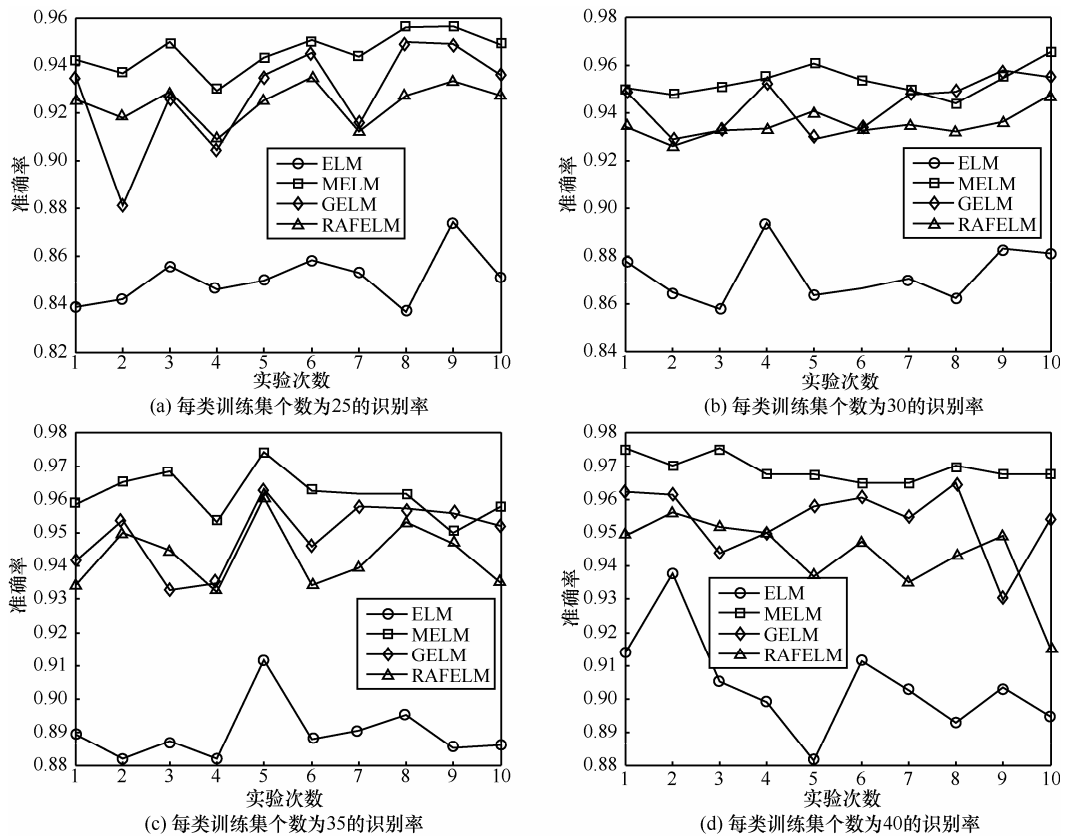


图 4 MELM、ELM、RAFELM、GELM 算法在 Yale B 数据集上的识别率

方法时所造成数据样本信息重叠的现象。MELM 算法中融入 LPP 算法技术, 使该算法能够有效获取数据样本的局部和全局的几何结构信息, 从而提高 ELM 的分类性能和泛化能力。从图 2~图 4 中也可以看出, GELM 算法的识别率也取得了良好的效果, 因为 GELM 算法考虑到同类数据样本具有相似性, 并且将这种相似性利用拉普拉斯特征映射结构进行表示。这也从侧面证明了 MELM 算法理论思路的正确性, 也就是说如何有效提取样本的内在信息与关联信息是关键。虽然 RAFELM 算法的识别率高于 ELM, 但明显低于 MELM、GELM。这是由于 RAFELM 仅仅是从激活函数的角度改进了 ELM 算法, 而没有考虑到数据样本的几何特征和数据蕴含的判别信息。图 5 给出了不同 ELM 算法在 PCA 降维算法下的识别率曲线, 由图 5 可以看出 MELM 算法取得了比较好的识别效果。

为了更好地显示出不同 ELM 模型的平均性能, 使用直方图表示不同 ELM 模型的平均识别率和标准偏差, 如图 6 所示。其中, 3 train 为每类训练集个数为 3 的识别率, 下同。

#### 4.2 不同 ELM 算法时间复杂度分析

为了评价本文所提算法 MELM 的计算效率,

本文对 ELM、RAFELM、GELM、MELM 的时间复杂度进行分析。对于 ELM 算法输出权值  $W = \left( HH^T + \frac{I}{C} \right)^{-1} HT^T$ , 在求解输出权值  $W$  时本文需要计算一个  $L \times L$  的矩阵  $HH^T$  ( $L$  为隐层节点个数) 的逆矩阵, ELM、RAFELM、GELM 和 MELM 有相同的时间复杂度, RAFELM 的输出权值与 ELM 的输出权值相同为  $W = \left( HH^T + \frac{I}{C} \right)^{-1} HT^T$ , MELM 的输出权值形式为  $W = \left( HH^T + \frac{S}{C} \right)^{-1} HT^T$ , GELM<sup>[13]</sup> 的输出权值为  $W = (HH^T + \lambda_1 H L H^T + \lambda_2 I)^{-1} HT^T$ , 这些算法都需要计算一个  $L \times L$  的矩阵  $HH^T$  的逆矩阵, 因此, ELM、RAFELM、GELM 和 MELM 的时间复杂度都为  $O(L^3)$ 。

为了定量地评价 MELM 的计算效率, 将 ELM、RAFELM、GELM、MELM 在图像数据集上进行实验。MELM 与 ELM、GELM 均采用 Sigmoid 函数作为激活函数, RAFELM 采用自身改进的激活函数, 4 种不同 ELM 算法隐层节点个数均设置为 200, 使用

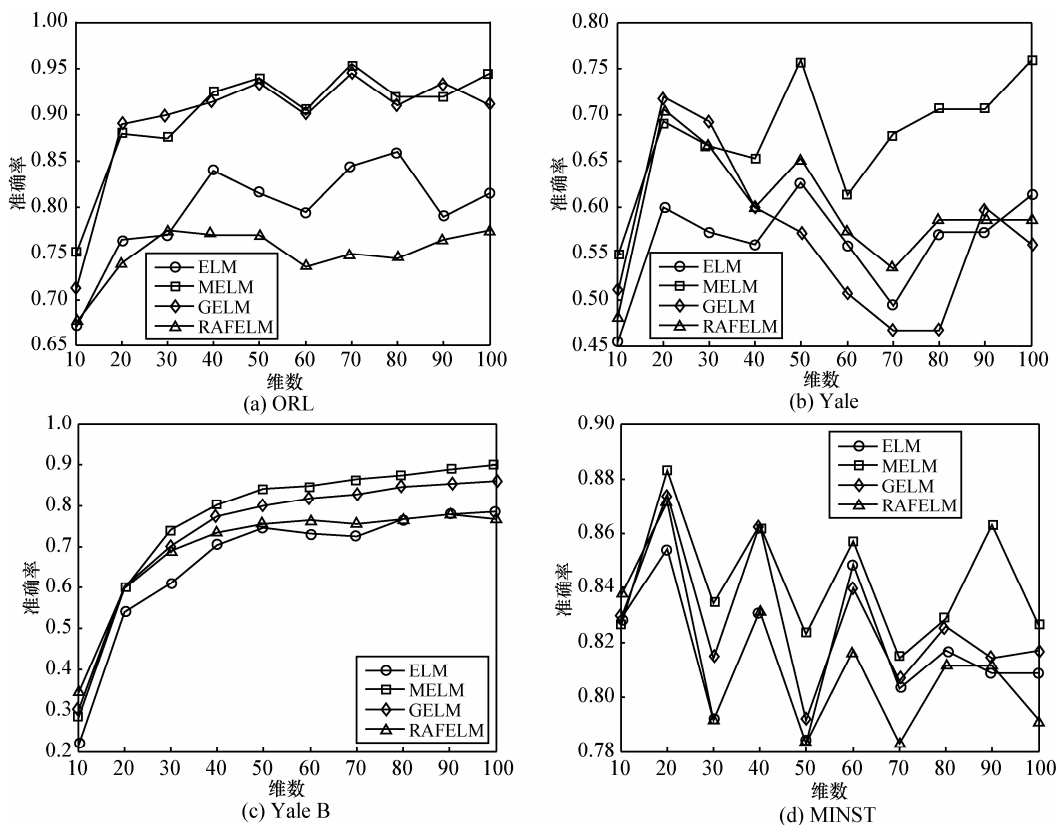


图 5 MELM、ELM、RAFELM、GELM 算法在 PCA 降维算法下的识别率

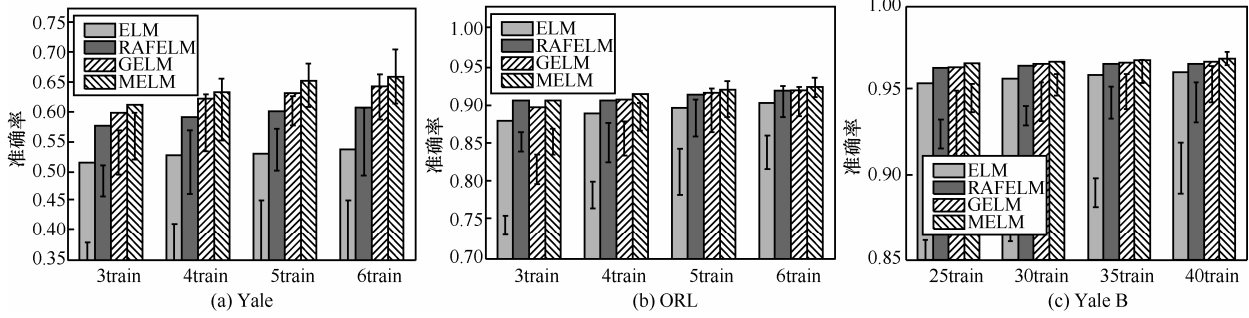


图 6 不同 ELM 算法在不同数据集上的统计

相同的惩罚参数  $C = 2^{-8}$ ，对于 Yale 数据集 MELM 算法中调节参数设置为  $\lambda = 0.4$ ，对于 UMIST 数据集 MELM 算法中调节参数设置为  $\lambda = 0$ ，对于 ORL 数据集 MELM 算法中调节参数设置为  $\lambda = 0.5$ ，对于 Yale B 数据集 MELM 算法中调节参数设置为  $\lambda = 0.3$ 。实验平台信息为惠普工作站处理器：Intel(R)Xeon(R) CPU E5-1603 0 @2.80 GHz。内存：8.00 GB。系统类型：Windows 7 64 位操作系统，Matlab 2010b。图 7 给出了 4 种不同 ELM 算法在图像数据集上的运行时间曲线，表 2 给出了 4 种不同

ELM 算法在图像数据集上具体的平均运行时间。

表 2 不同 ELM 算法在图像数据集上的平均训练时间/s

数据集	ELM	RAFELM	GELM	MELM
Yale	0.025 3	0.234 0	0.670 8	0.276 9
ORL	0.062 4	0.456 3	1.721 8	0.522 6
Yale B	0.197 6	3.918 2	34.268 2	0.798 2
UMIST	0.052 0	0.512 2	1.814 8	0.270 4

由图 7 及表 2 给出的 4 种不同的 ELM 算法在图像数据集上的运行时间的比较，可以看出 MELM

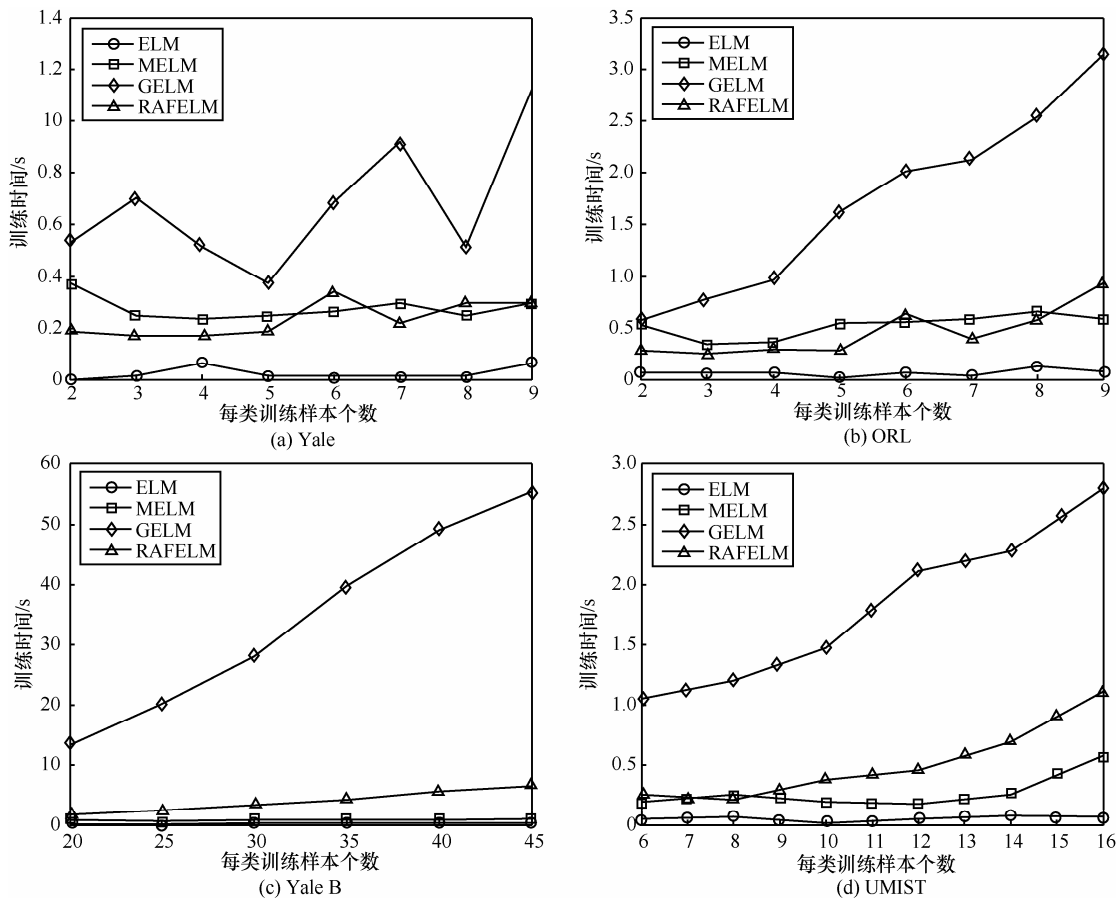


图 7 不同 ELM 算法在图像数据集上的运行时间

的运行时间高于 ELM 算法的运行时间，这是由于 MELM 将类内散度矩阵和类间散度矩阵引入到 ELM 模型中势必增加运行时间。可以看出 GELM 的运行时间远远高于其他 3 种算法，在数据集 Yale B、UMIST 上 MELM 的平均运行时间低于 RAFELM 和 GELM 的平均运行时间，综合以上分析，对于 MELM、GELM、RAFELM 时间复杂度都为  $O(L^3)$  的情况下，MELM 表现出了良好的计算效率。

### 4.3 MELM 与相关神经网络算法比较

为了进一步证明本文算法的有效性，将 MELM 与卷积神经网络算法和深度信念神经网络算法进行对比。当训练样本较少时，DBN 算法和 CNN 算法需要迭代多次才会取得较好的识别效果。在 ORL、Yale、UMIST 数据集上实验时，由于这些数据都为小数据样本，为此将 DBN 算法和 CNN 算法的迭代次数设置为 500，根据文献[23,24]推荐的参数范围，将 DBN 算法的学习率设置为 0.05，CNN 算法的学习率设置为 1，对于 MINST 数据将 DBN 算法和 CNN 算法的迭代次数设置为 1，DBN 算法的学习率设置为 0.05，CNN 算法的学习率设置为 1。

MELM 与 DBN、CNN 的识别率曲线如图 8 所示，DBN、CNN、MELM 算法在人脸数据集上的识别率如表 3 所示。

表 3 DBN、CNN、MELM 算法在人脸数据集上的识别率

数据集	DBN		CNN		MELM	
	最大识别率	平均识别率	最大识别率	平均识别率	最大识别率	平均识别率
Yale	80.00%	46.00%	70.00%	52.26%	86.67%	69.95%
ORL	51.50%	44.76%	25.00%	11.30%	95.00%	86.06%
UMIST	85.00%	57.27%	80.00%	26.82%	98.75%	83.40%
MINST	90.60%	81.32%	67.46%	35.17%	89.48%	85.78%

由图 8 及表 3 可知，MELM 在 4 个不同的人脸数据集上取得了较好的识别效果，并且显著高于 DBN 及 CNN 算法的识别率。由图 8 可以看出 DBN 和 CNN 算法在数据集 Yale、ORL、UMIST 的识别率低于 MELM 算法。还可以看出 DBN、CNN 的识别率维持在比较低的水平，这是由于 Yale、ORL、UMIST 为小数据集，在小数据集上 DBN、CNN 并没有取得较好的识别效果，图 8 中 DBN 算法在 MINST 数据集上识别率取得了较好的效果，CNN

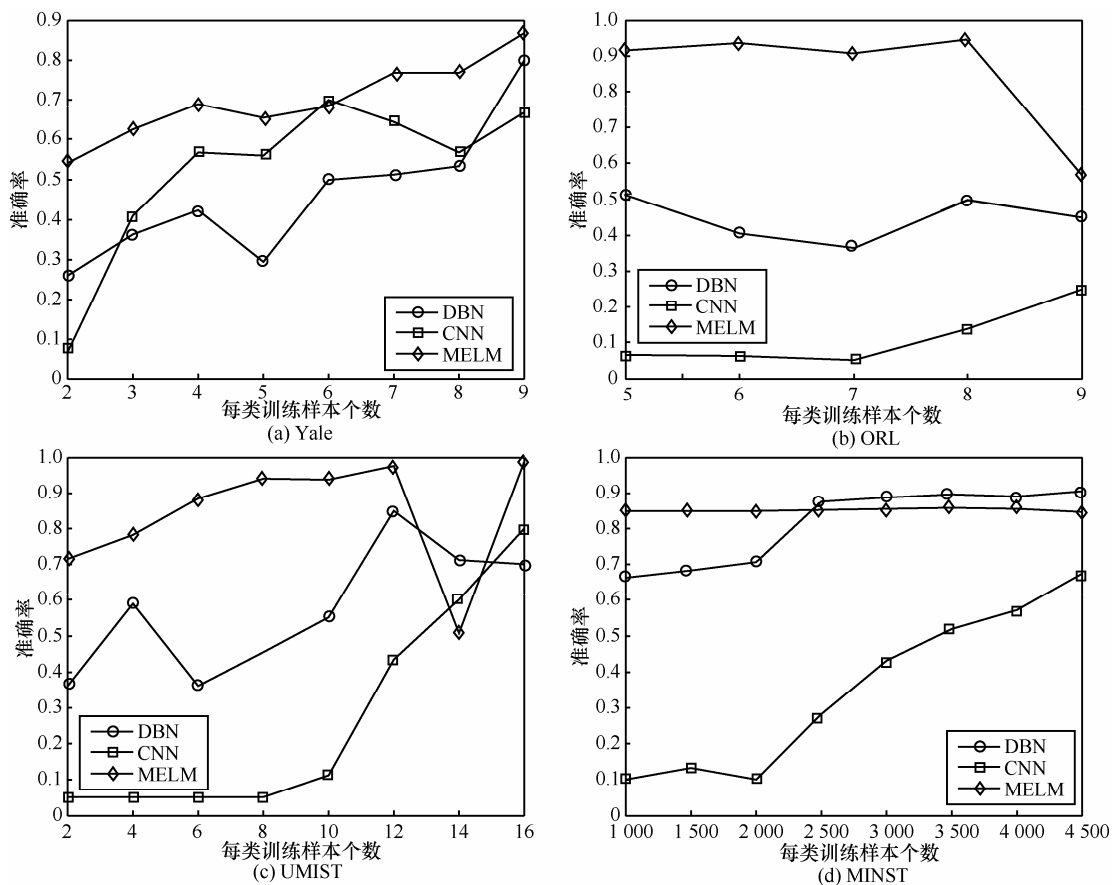


图 8 DBN、CNN、MELM 在人脸数据集上的识别率

在 4 个数据集中均没有取得较好的识别效果。从表 3 可以看出 MELM 算法的识别率高于 DBN 算法和 CNN 算法的识别率，因此，MELM 在模式识别中可以作为有效的分类方法。

#### 4.4 MELM 算法的参数分析

在本节实验中主要是测试参数  $\lambda$  对 MELM 性能的影响。实验过程中，Yale、ORL 数据集随机选取每类训练样本个数为 5，剩余部分为测试集。在 Yale B 实验中随机选取每类训练样本个数为 25，剩余部分为测试集。MELM 采用 Sigmoid 函数作为激活函数，隐层节点个数设置为图像数据的维数，惩罚参数  $C = 2^{-8}$ ，实验结果如图 9 及表 4 所示。

由图 9 可以看出， $\lambda$  的取值不同会产生不同的

效果，由表 4 可知，对于 Yale 数据集最优的  $\lambda$  为 0.4，ORL 数据集最优的  $\lambda$  为 0.5，Yale B 数据集最优的  $\lambda$  为 0.3，UMIST 数据集最优的  $\lambda$  为 0。实验表明，可以通过改变调节参数  $\lambda$  的不同取值来改善 MELM 的分类性能。

#### 5 结束语

极端学习机的发展得益于其在各个领域的广泛应用，也促使了研究领域的深刻性。本文结合流形学习思想技术，在对极端学习机的泛化能力和信息提取能力研究的基础上，提出一种新的极端学习机算法 (MELM)。该算法能够有效获取数据的全局和局部几何信息，充分考虑到数据的类内和类间

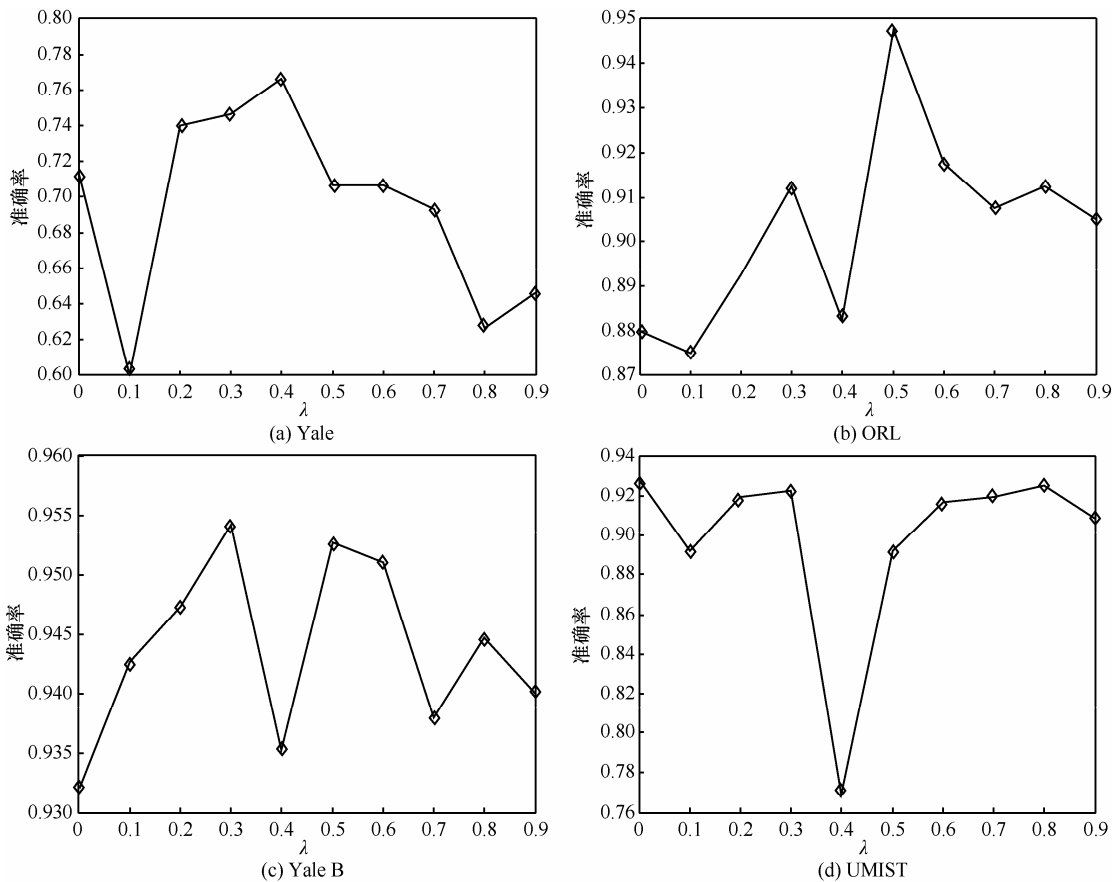


图 9 在不同的图像数据集上参数  $\lambda$  对 MELM 算法的影响

表 4 不同的图像数据集上，参数  $\lambda$  变化对 MELM 算法识别率的影响

数据集	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$
Yale	71.33%	60.00%	74.00%	74.67%	76.67%	70.67%	70.67%	69.33%	62.67%	64.67%
ORL	88.00%	87.50%	89.25%	91.25%	88.25%	94.75%	91.75%	90.75%	91.25%	90.50%
Yale B	93.21%	94.26%	94.74%	95.42%	93.53%	95.26%	95.11%	93.79%	94.47%	94.00%
UMIST	92.78%	89.17%	91.94%	92.22%	76.94%	89.17%	91.67%	91.94%	92.50%	90.83%

信息的自然属性, 对不同结构的数据矩阵计算的奇异性问题, 使算法有很好的稳定性和泛化能力。使用通用的数据集 Yale、Yale B、ORL 将本文所提出算法与相关最新算法和神经网络算法做了对比实验。对相关算法进行了复杂度分析, 在参数变化下对新算法性能的表现也进行了实验分析。实验和分析结果证明了所提出算法的有效性和先进性。

### 参考文献:

- [1] TANG J, DENG C, HUANG G B. Extreme learning machine for multilayer perception[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, (99): 1-13.
- [2] HUANG G B, ZHOU H M, DING X J, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(2): 513-529.
- [3] HUANG G B, CHEN L, SIEW C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes[J]. IEEE Trans Neural Networks, 2006, 17(4): 879-892.
- [4] ZHANG R, LAN Y, HUANG G B, et al. Universal approximation of extreme learning machine with adaptive growth of hidden nodes[J]. IEEE Trans Neural Networks Learn Systems, 2012, 23(2): 365-371.
- [5] FERNÁNDEZ-DELGADO M, CERNADAS E, BARRO S, et al. Direct kernel perceptron (DKP): ultra-fast kernel ELM-based classification with noniterative closed-form weight calculation[J]. Neural Networks, 2014, 50: 60-71.
- [6] HUANG G, SONG S J. Semi-supervised and unsupervised extreme learning machines[J]. IEEE Transactions on Cybernetics, 2014, 44(12): 2405-2417.
- [7] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70: 489-501.
- [8] VAPNIK V N. Statistical learning theory[J]. Encyclopedia of the Sciences of Learning, 2010, 41(4): 3185-3185.
- [9] HUANG G B, ZHOU H M, DING X J, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(2): 513-529.
- [10] WANG X Z, SHAO Q Y, MIAO Q, et al. Architecture selection for networks trained with extreme learning machine using localized generalization error model[J]. Neurocomputing, 2013, 102: 3-9.
- [11] ZHAO J W, WANG Z H, PARK D S. Online sequential extreme learning machine with forgetting mechanism[J]. Neurocomputing, 2012, 87: 79-89.
- [12] RONG H J, HUANG G B, SUNDARARAJAN N, et al. Online sequential fuzzy extreme learning machine for function approximation and classification problems[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(4): 1067-1072.
- [13] LIANG N Y, HUANG G B, SARATCHANDRAN P, et al. A fast and accurate online sequential learning algorithm for feedforward networks[J]. IEEE Transactions on Neural Networks, 2006, 17(6): 1411-1423.
- [14] ZONG W W, HUANG G B, CHEN Y. Weighted extreme learning machine for imbalance learning[J]. Neurocomputing, 2013, 101: 229-242.
- [15] YU Q, MICHE Y, EIROLA E, et al. Regularized extreme learning machine for regression with missing data[J]. Neurocomputing, 2013, 102: 45-51.
- [16] MAN Z H, WANG D H, CAO Z W, et al. Robust single-hidden layer feedforward network-based pattern classifier[J]. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23 (12): 1974-1986.
- [17] PENG Y, LU B L. Discriminative graph regularized extreme learning machine and its application to face recognition[J]. Neurocomputing, 2015, 149: 340-353.
- [18] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396.
- [19] HE X, NIYOGI P. Locality preserving projections[J]. Advances in Neural Information Processing Systems, 2004, 17:153-160.
- [20] LI H, JIANG T, ZHANG K. Efficient robust feature extraction by maximum margin criterion[J]. Advances in Neural Information Processing Systems, 2003, 16:71-78
- [21] LIU S, FENG L, XIAO Y. Robust activation function and its application: semi-supervised kernel extreme learning method[J]. Neurocomputing, 2014, 144: 318-328.
- [22] HUANG G B. An insight into extreme learning machines: random neurons, random features and kernels[J]. Cognitive Computation, 2014, 6: 376-390.
- [23] HINTON G E, OSINDERO S. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

### 作者简介:



刘德山 (1970-), 男, 辽宁辽阳人, 辽宁师范大学副教授, 主要研究方向为模式识别、数据挖掘。



楚永贺 (1989-), 男, 河南郑州人, 辽宁师范大学硕士生, 主要研究方向为数据降维、机器学习等。



闫德勤 (1962-), 男, 辽宁沈阳人, 辽宁师范大学教授, 主要研究方向为模式识别、机器学习等。