ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 446－454
DOI:10. 16337/j. 1004-9037. 2018. 03. 007

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

极速非线性判别分析网络

谢群辉　陈松灿

(南京航空航天大学计算机科学与技术学院,南京,211106)

摘　要:由于线性判别分析仅是线性方法,难以有效应对非线性问题,而对其非线性化是解决这一问题的关键途径。非线性化判别方法主要包括神经网络和核化方法。神经网络判别分析方法虽然继承了神经网络所具有的自适应、分布存储、并行处理和非线性映射等优点,但也遗传了其训练速度慢且易陷入局部最小值缺点;而核线性判别分析方法虽能获得全局最优解析解,但因受制于隐节点数目(等于样本个数),当数据规模大时,计算成本变大。本文受随机映射启发,对神经网络判别分析方法进行极速化改造,实现了一种极速非线性判别分析方法,兼具神经网络的自适应性和全局最优解的快速性。最后在UCI真实数据集上的实验表明,极速非线性判别分析方法具有更优的分类性能。

关键词:线性判别分析;神经网络;核判别分析;极速化

中图分类号:TP183　　**文献标志码**:A

Extreme Nonlinear Discriminant Analysis Network

Xie Qunhui, Chen Songcan

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China)

Abstract: As the linear discriminant analysis (LDA) is just a linear method and is difficult to effectively deal with nonlinear problems, non-linearizing LDA is a crucial strategy to enable it to solve such nonlinear problems. Nonlinear LDA is mainly based on two strategies, neural networks and kernelization. A representative of the former strategy is the neural network discriminant analysis (NNDA). Athough NNDA inherits the advantages such as self-adaption, parallel processing, distributed storing and nonlinear mapping of neural networks, its training is quite time-consuming and likely to get trapped in local minimum. While the representative of the latter strategy is the kernel linear discriminant analysis (KLDA). Although KLDA can obtain a global optimal analytical solution, its computational cost is rather high, due to the fact that the number of hidden nodes of KLDA is equal to the size of training samples, especially in large scale scenarios. Inspired by the idea of random map, a novel extreme nonlinear discriminant analysis (ENDA) is proposed by reconstructing NNDA via extreme learning strategy in this paper. ENDA shares both the self-adaption of NNDA and the efficient computation of global optimal solution of KLDA. Finally, experimental results on UCI datasets demonstrate the superiority of ENDA over KLDA and NNDA in classification accuracy.

Key words: linear discriminant analysis; neural network; kernel discriminant analysis; speedup

---

基金项目:国家自然科学基金(61472186)资助项目;中国博士后科学基金(20133218110032)资助项目。
收稿日期:2016-09-07;**修订日期**:2016-10-08

引 言

线性判别分析(Linear discriminant analysis,LDA)是一种流行的有监督数据降维和可视化工具,广泛用于模式识别和机器学习等各个领域[1-2];它最大化类间散度的同时最小化类内散度,提取数据特征,将高维投影到2～3维即获得可视化,能更直观理解和探索潜在的数据结构[3],有利于其后续的分类算法获得更好的泛化性能。判别分析可分为线性型和非线性型两大类,上述提及的LDA[4]是线性型的代表之一,而传统的非线性判别方法主要包括神经网络和核方法,如核LDA(Kernel linear discriminant analysis,KLDA)[5],流形判别分析[6],(深度)神经网络[7-8]和深度判别分析[9]等。其中KLDA采用核函数将输入样本映射到高维特征空间,而后利用LDA进行线性分类。而神经网络(Neural network,NN)则通过多层结构的非线性变换将输入映射到输出空间以实现非线性判别。尽管NN缺乏核方法那样的简洁表示方式,但NN所具有的分布并行特点和其对连续函数的万能逼近性使其获得了广泛应用。事实上,基于NN的判别分析可追溯到1995年Mao和Jain的先驱性工作[10],即神经网络判别分析方法(Neural networks discriminant analysis,NNDA),该工作促发了众多非线性化分类和可视化研究。目前该文被引数已超680次,并在1996年获得了IEEE TNN期刊年度最佳论文奖[11]。然而,与其他多层前向网络训练类似,NNDA存在优化收敛速度慢、易陷入局部最小、且因网络结构复杂易导致过拟合等问题。1999年由Scholkopft等人[12]提出的KLDA,利用核技巧避免了NNDA复杂的优化。相对NNDA,其判别投影仅需通过求解一广义特征值方程即可获得解析解,不仅快速而且学习性能优良,因而受到广泛关注和应用,该工作目前的引用数已超2 240次。

然而传统的KLDA在学习过程中存在数据的可伸缩性问题[13],当数据量增加到一定规模后,算法所学得的判别方向复杂度与训练样本数成线性增长,这在常规计算资源上已难以胜任相应的学习。相比KLDA的这种复杂性,本文所提出的ENDA算法本身只线性于隐节点数,而独立于训练样本数,在现实应用中,隐节点数通常比训练样本数要小很多。因此,当实验数据规模变大时,在保证分类性能的前提下,ENDA计算成本比KLDA更小。另一方面,近年来的众多研究表明[14],对传统NN(如卷积网络)的深化学习能大大提升图像分类[8]、语音识别[15]和自然语言处理[16]等的识别性能,由此也促发了对NNDA向深度化学习[17]的研究,在对NNDA的跟踪研究发现,此类研究主要有两大趋势:(1)网络变大变深的深度学习;(2)限于常规计算资源的加速化(例如典型ELM)极速学习。然而深度学习的成功往往需付出高昂的代价,是因为:(1)深度学习需要学习大量参数,样本少了很易"过拟合",数据大小成为其性能提升的关键[18];(2)模型复杂化需要庞大的计算资源和巨大的时间开销。而"没有免费午餐定理"告诉人们:算法的优劣必须针对具体的学习任务,其有效性必须考虑"偏好"问题,结果是:(1)时间开销大和计算资源缺乏,即难以在常规计算资源下完成深度学习;(2)所需要的大数据的标记涉及昂贵的人力和物力等问题。综上所述,当数据规模达不到深度学习要求的情况下(在机器学习UCI储存数据库公布的348个数据集当中,296个数据集的主要样本例数集中在50 000个以下,约占85%),如何在常用PC计算资源下进行更有效的学习变得更有意义和更加紧迫,所以笔者更偏重于NN加速化的研究。在保证分类性能的前提下,ENDA能迅速处理这类常规数据的分类任务,与传统NNDA不同是因为其利用了快速、可靠的随机映射。

最新研究表明[19]使用随机投影的简单NN与人类学习具有很大相关性和相似鲁棒性。因此也为10年前采用此原理的极速学习机(Extreme learning machine,ELM)提供了认知原理上的解释,尽管ELM被提出以来得到了广泛关注,并已在特征学习、分类、回归和聚类[20]等方面获得了一系列拓展,但就笔者所知,还未有对NNDA相应的ELM改造。本文基于这一事实,对NNDA进行极速化,构建出一种极速非线性判别分析方法(Extreme nonlinear discriminant analysis,ENDA),使其兼具NNDA的万能逼近能力和KLDA能解析获得全局最优解的快速性。最后在UCI机器学习库真实数据集上进行实验,

结果显示 ENDA 比 KLDA 和 NNDA 具有更优的分类性能。

1 极速学习机模型

极限学习机 ELM[21]是一种特殊单隐层前向神经网络(Single-hidden layer feedforward neural networks, SLFN),由 Huang 等人于 2004 年提出,目前已获得了 1 180 多次引用。不同于 SLFN 传统梯度下降学习算法,ELM 随机产生输入层到隐层权重和偏置,克服 SLFN 反复迭代计算导致的收敛慢、且不能保证全局最优解等问题。在优化隐节点和输出节点的权重上,ELM 采用正则化最小二乘法快速求得闭合解[22]。ELM 由输入层,隐层和输出层 3 层网络组成,其中隐节点常用非线性激活函数。典型的非线性激活函数包括 Sigmoid 函数、高斯函数和径向基函数。不失一般性,这里采用式(1)中的 Sigmoid 函数作为隐层神经元的激活函数

$$g(x)=\frac{1}{1+\exp(-x)} \tag{1}$$

ELM 不仅有速度上优势,更重要的是理论上也证明了其具有与 SLFN 同样的对非线性分段连续函数的万能逼近能力[23]。其学习过程可视为两步:(1)确定网络 NN 隐层的神经元数,随机设置输入权重和偏差;(2)确定网络权重。在特征空间中通过使训练误差平方和最小解析求得输出权重的最小范数解,达到优化输出权重的目的。ELM 除了快速外还能实现对不同类型数据的分析,并已渐渐成为一种新型的快速学习范式。

现设有 N 个样本的训练数据集 $\{\boldsymbol{x}_j,\boldsymbol{t}_j\}_{j=1}^{N}$,其中 $\boldsymbol{x}_j=[x_{j1},x_{j2},\cdots,x_{jn}]^{\mathrm{T}}\in\mathbf{R}^n$ 为 n 维特征输入向量,$\boldsymbol{t}_j=[t_{j1},t_{j2},\cdots,t_{jm}]^{\mathrm{T}}\in\mathbf{R}^m$ 为对应的输出目标向量。第一步按式(2)执行,其中输入层与隐层间的权重和偏置分别从 $(-1,1)$和$(0,1)$上随机均匀产生,所得隐层的输出为

$$h_i(\boldsymbol{x}_j)=g(\boldsymbol{w}_i^{(1)\mathrm{T}}\boldsymbol{x}_j+b_i) \tag{2}$$

式中:$\boldsymbol{w}_i^{(1)}=[w_{i1}^{(1)},w_{i2}^{(1)},\cdots,w_{in}^{(1)}]^{\mathrm{T}}(i=1,\cdots,L,j=1,\cdots,N)$ 为输入层连接第 i 个隐层单元输入权值,b_i 为偏置。定义 $\boldsymbol{H}$ 为 L 维的隐层映射特征空间表示

$$\boldsymbol{H}(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_L,b_1,b_2,\cdots,b_L,\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)=[\boldsymbol{h}(\boldsymbol{x}_1),\boldsymbol{h}(\boldsymbol{x}_2),\cdots,\boldsymbol{h}(\boldsymbol{x}_N)]^{\mathrm{T}}\in\mathbf{R}^{N\times L} \tag{3}$$

式中:$\boldsymbol{h}(\boldsymbol{x}_j)=[g(\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_j+b_1),g(\boldsymbol{w}_2^{\mathrm{T}}\boldsymbol{x}_j+b_2),\cdots,g(\boldsymbol{w}_L^{\mathrm{T}}\boldsymbol{x}_j+b_L)]$。第二步,通过最小化总训练误差平方和准则式(4),即可获得优化输出权重。

$$\min_{B}L_{\mathrm{ELM}}=\frac{1}{2}\|\boldsymbol{B}\|^2+\frac{C}{2}\sum_{j=1}^{N}\|\boldsymbol{\xi}_j\|^2 \tag{4}$$

$$\text{s.t.}\quad \boldsymbol{h}(\boldsymbol{x}_j)\boldsymbol{B}=\boldsymbol{t}_j^{\mathrm{T}}-\boldsymbol{\xi}_j^{\mathrm{T}}\quad j=1,\cdots,N$$

式中:$\boldsymbol{\xi}_j$ 为第 j 个训练模式的误差,$\boldsymbol{B}=[\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_L]^{\mathrm{T}}\in\mathbf{R}^{L\times m}$,$C>0$ 为惩罚系数。而 ELM 的输出方程为

$$f(\boldsymbol{x}_j)=\sum_{i=1}^{L}\boldsymbol{\beta}_i g(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}_j+b_i)=\boldsymbol{h}(\boldsymbol{x}_j)\boldsymbol{B} \tag{5}$$

式中 $\boldsymbol{\beta}_i=[\beta_{i1},\beta_{i2},\cdots,\beta_{im}]^{\mathrm{T}}$ 为输出权重。因此最终优化问题式(4)可重写为

$$\min_{B}\boldsymbol{L}_{\mathrm{ELM}}=\frac{1}{2}\|\boldsymbol{B}\|^2+\frac{C}{2}\sum_{j=1}^{N}\|\boldsymbol{T}-\boldsymbol{H}\boldsymbol{B}\|^2 \tag{6}$$

式中 $\boldsymbol{T}=[\boldsymbol{t}_1,\cdots,\boldsymbol{t}_N]^{\mathrm{T}}\in\mathbf{R}^{N\times m}$。借助正则化最小二乘法求解获得全局最优的权阵 $\boldsymbol{B}$ 为

$$\boldsymbol{B}=\left(\frac{\boldsymbol{I}}{C}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T} \tag{7}$$

式中 $\boldsymbol{I}$ 为单位矩阵。利用式(7)可快速计算 $\boldsymbol{B}$[22]。ELM 从理论和经验上获得了性能保证[24]。借助其思想,拟对由 Mao 和 Jain 所提出的 NNDA 进行极速化改造,建立 ENDA 模型。

2 极速非线性判别分析网络

Mao 等人的 NNDA 属于浅层网络,其目标是学习非线性降维,但其遗传了 SLFN 的训练慢、易陷于局部最优的缺点。同时基于深度学习思想实现的多层 ELM[25-26],能够实现对海量数据的建模,尽管能获得比深度网络相对快的训练,但对本文所要处理的数据规模仍显得"大材小用",偏离了传统 ELM 简单易解的特性。本文目的在于极速化 NNDA ,使 ENDA 在常规计算设施上能处理比 NNDA 更大规模的数据集,同时继承了 ELM 能解析求解的优点。ENDA 网络结构如图 1 所示,具体分两步:(1)随机生成 NN 的输入层连接权重和偏置,构成了一个随机映射,结果使模型具有简单快速性;(2)视隐层输出为新形成的训练数据,用 LDA 准则优化进输出层连接权重,同时可用于可视化。同 ELM 求解一样,不仅无局部最小,并且能快速解析求得全局闭合最优解,最后形成一个新的判别特征空间,而后用所获特征对目标进行分类。由于采用了随机非线性变换和后续的判别优化,使 ENDA 对数据有着自适应性和期望的分类性能。

ENDA 算法第一部分实现原始特征 $\boldsymbol{X}=[\boldsymbol{x}_1^{\mathrm{T}},\boldsymbol{x}_1^{\mathrm{T}},\cdots,\boldsymbol{x}_N^{\mathrm{T}}]\in\mathbf{R}^{N\times n}]$ 到非线性特征 $\boldsymbol{H}=[\boldsymbol{h}_1^{\mathrm{T}},\boldsymbol{h}_2^{\mathrm{T}},\cdots,\boldsymbol{h}_N^{\mathrm{T}}]\in\mathbf{R}^{N\times L}$ 的极速随机映射;第二部分通过使用 LDA 对 $\boldsymbol{H}$ 进行降维。LDA 作为一种有效的降维和分类方法,具有以下优点:(1)可作为可视化工具;(2)通过获取最佳投影方向问题转化为求解最大特征值的问题,易求得分类面法方向的解析解。在多分类任务中,假定一个有包含 c 类的样本,LDA 降维实际是从 L 维空间向 P 维空间投影,设 $P\leqslant c-1$。对 $\boldsymbol{H}$ 经过式(9)投影后得到新样本 $\hat{\boldsymbol{H}}=[\hat{\boldsymbol{h}}_1^{\mathrm{T}},\cdots,\hat{\boldsymbol{h}}_N^{\mathrm{T}}]\in\mathbf{R}^{N\times P}$。设

$$\boldsymbol{W}=[\boldsymbol{w}_1^{(2)},\ \boldsymbol{w}_2^{(2)},\cdots,\boldsymbol{w}_P^{(2)}] \tag{8}$$

$$\hat{\boldsymbol{H}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H} \tag{9}$$

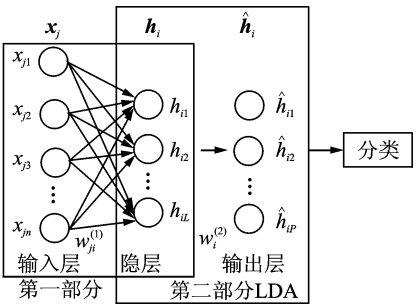

图 1 ENDA 网络结构图

Fig. 1 Structure of ENDA Network

式中 $\boldsymbol{w}_i^{(2)}$ 为 LDA 的投影权重。样本总类内散度矩阵 $\boldsymbol{S}_b$ 和类间散度矩阵 $\boldsymbol{S}_w$ 分别定义为

$$\boldsymbol{S}_b=\sum_{i=1}^{c}n_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^{\mathrm{T}} \tag{10}$$

$$\boldsymbol{S}_w=\sum_{i=1}^{c}\sum_{\hat{\boldsymbol{h}}\in\hat{\boldsymbol{H}}_i}(\hat{\boldsymbol{h}}-\boldsymbol{\mu}_i)(\hat{\boldsymbol{h}}-\boldsymbol{\mu}_i)^{\mathrm{T}} \tag{11}$$

式中:$\hat{\boldsymbol{H}}_i$ 表示 $\hat{\boldsymbol{H}}$ 中属于第 i 类的样本,$\boldsymbol{\mu}_i=\dfrac{1}{n_i}\sum\limits_{\hat{\boldsymbol{h}}\in\hat{\boldsymbol{H}}_i}\hat{\boldsymbol{h}}$,$\boldsymbol{\mu}=\dfrac{1}{N}\sum\limits_{i=1}^{c}n_i\boldsymbol{\mu}_i$,$n_i$ 为第 i 类样本数。构建 LDA 优化目标函数如下

$$\max_{\boldsymbol{W}}\boldsymbol{J}(\boldsymbol{W})=\frac{|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W}|}{|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W}|} \tag{12}$$

然后对其进行判别分析求解,最大化 $\boldsymbol{J}(\boldsymbol{W})$ 等价于求解如下广义特征问题,即

$$\boldsymbol{S}_b\boldsymbol{W}=\lambda\boldsymbol{S}_w\boldsymbol{W} \tag{13}$$

通过求解式(13)可计算出 LDA 投影矩阵 $\boldsymbol{W}$,它由 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 的 $c-1$ 个最大广义特征值所对应的特征向量组成,此 $\boldsymbol{W}$ 即为最佳的判别投影矩阵。通过以上两部分计算可得到整个 ENDA 投影权重,完整的 ENDA 算法如算法 1 所示。

算法 1 ENDA 算法

输入:样本集 $N=\{(\boldsymbol{x}_j,\boldsymbol{t}_j)\mid\boldsymbol{x}_j\in\mathbf{R}^n,\boldsymbol{t}_j\in\mathbf{R}^m\}_{j=1}^{N}$,激活函数 $g(\cdot)$,隐层节点数 L。

输出:ENDA 输出 $\hat{\boldsymbol{h}}$。

步骤 1 随机生成 ENDA 第一部分中的权重 $\boldsymbol{w}_i^{(1)}$ 和偏置 $b_i,i=1,\cdots,L$;

步骤 2 计算 ENDA 的隐层输出 h_i（通过式(2)）；

步骤 3 对隐层输出 h_i 进行判别分析，通过式(13)计算出投影权重 $\boldsymbol{W}$，得到 ENDA 的输出 $\hat{\boldsymbol{h}}$。

3 仿真及性能分析

3.1 设置

实验环境如下：MATLAB2013a，Intel(R) Core™ i5-3470 CPU @3.20 GHz，16.0 GB 内存，64 位 Win10 操作系统。如表 1 所示，实验所用数据为 UCI(http://archive.ics.uci.edu/ml/)数据集。实验前需对数据进行了归一化，然后将处理后的数据作为 ENDA 的输入进行训练。为了验证模型的有效性，将 ENDA 算法与 NNDA，KLDA，ELM 算法进行了比较分析，为公平起见，NNDA，ELM 与 ENDA 的隐层节点参数相同，统一使用 K 近邻算法(K-Nearest Neighbor)，分别对 NNDA，KLDA 和 ENDA 进行分类。实验过程中参数 K 采用 10 折交叉验证。

表 1 UCI 数据集

Tab. 1 UCI datasets

Dataset	Type	Instance	Feature	Classe
Wine	Real	178	13	3
Segment	Real	2 310	19	7
Waveform3	Real	5 000	21	3
Pen-digits	Integer	10 992	16	10
Letter	Integer	20 000	16	26
Sensorless	Real	58 509	49	11

按照 Mao 等人提出 NNDA 方法实验，采用的 NN 包含两个隐层的误差反向传播算法(Error back propagation，BP)网络，输入层节点个数对应特征维数，输出层神经元个数等于类别数。通过反向传播学习算法训练前馈神经网络两个隐层的权重，并且固定最后隐层的节点数作为要投影空间的维度。NN 训练参数设置如下：学习率为 0.7，迭代次数 5 000 次，动量因子为 0.3，最小误差为 0.001。KLDA 所采用的高斯映射核函数的类型为 $K(x,y)=\exp(-\|x-y\|^2/2\delta^2)$，其中映射函数控制参数 δ 由集合 $\{0.01,0.1,1,10,100,500,1\ 000\}$ 取值。

3.2 可视化

为验证 ENDA 模型的非线性特征学习能力，实验中随机抽取 UCI 的 4 组数据集通过模型投影到二维空间进行可视化，结果如图 2～5 所示。

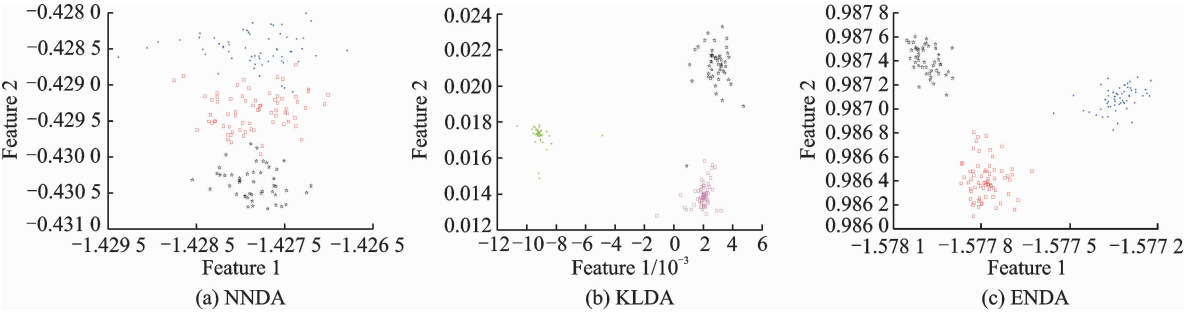

图 2 Wine 数据集的可视化

Fig. 2 Wine dataset visualized by three algorithms

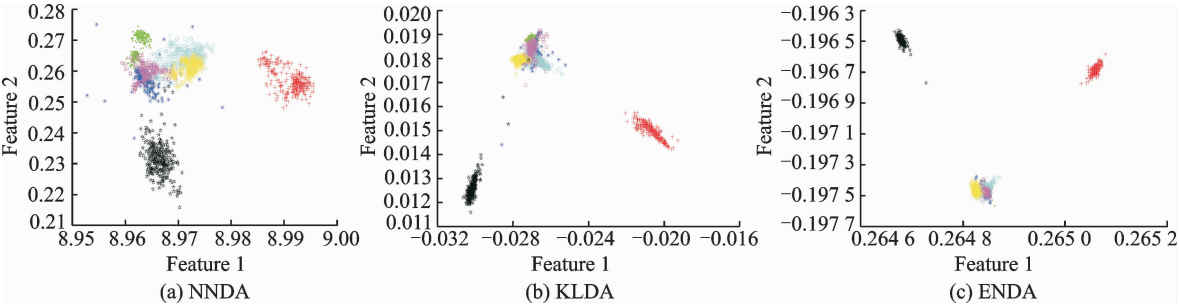

图 3 Segment 数据集的可视化

Fig. 3 Segment dataset visualized by three algorithms

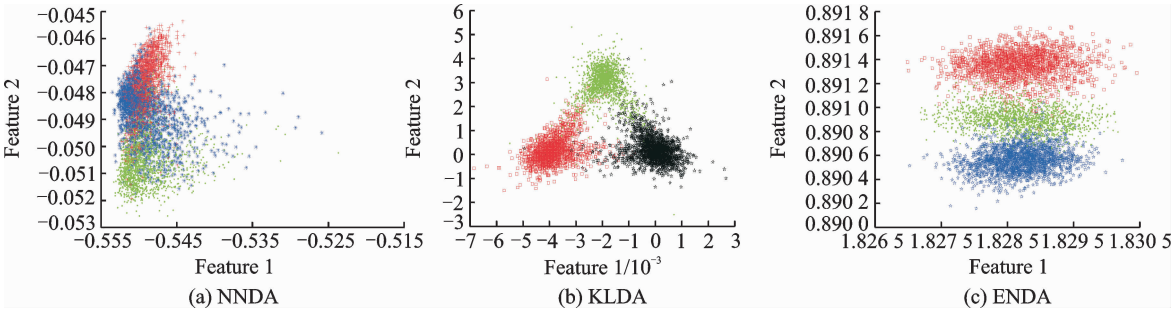

图 4 Waveform3 数据集的可视化

Fig. 4 Waveform3 dataset visualized by three algorithms

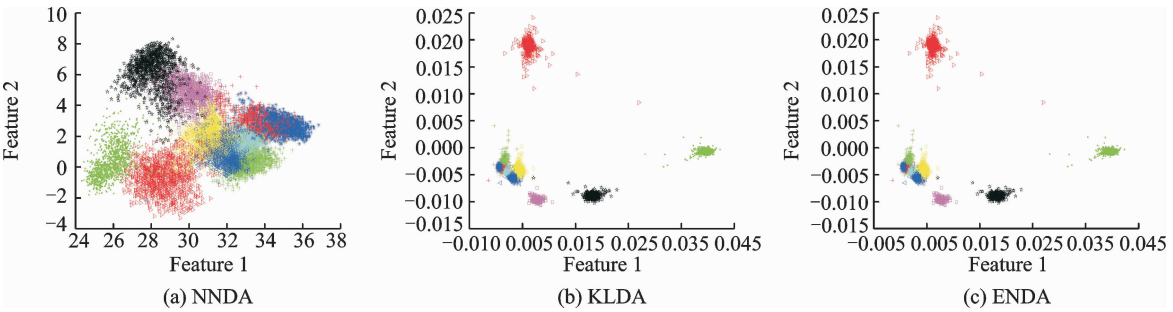

图 5 Pen-Digits 数据集的可视化

Fig. 5 Pen-Digits dataset visualized by three algorithms

由图 2～5 可以观察到样本分别经过 NNDA,ENDA 和 KLDA 方法投影二维投影空间后的效果:数据经过投影之后都具有最大的分离度,其中 ENDA 和 KLDA 投影分类效果更直观有效。同时实验结果表明对数据线性不可分问题,非线性变换是一个强大的方法。

在隐节点参数设置相同情况下,NNDA 不仅花费的训练时间更长,而且投影出来效果分离程度并不明显,如图 4,5 所示,其原因有:(1)NNDA 最后隐层到输出层训练权重并没有充分利用;(2)NNDA 为避免扭曲严重,以线性函数替代 Sigmoid 函数。在图 4,5 中,随着数据规模和特征属性逐渐复杂情况下,ENDA,KLDA 比 NNDA 分类投影后数据分开,抽取特征更明显。ENDA 与 KLDA 投影效果差别不大,但 KLDA 的不足表现在对数据集规模较大样本,计算时间和空间复杂度变大。综上所述 ENDA 可作为一种极速且稳定的可视化工具,其原理简单直观,通过可视化可以更加了解数据的内在特性,同时在

判别分析时提取最大特征，不仅降低数据中的不相关和冗余信息，同时有利于数据的后续分类。

3.3 性能分析

对 KLDA，NNDA，ELM 和 ENDA 这 4 种相关算法进行实验对比，NNDA，ELM 和 ENDA 隐层节点设置为相同参数，采用统计测试集数据分类的错误率和时间进行对比分析。实验结果为 10 次测试结果平均值和标准差，如表 2 所示。其中错误率公式为

$$\mathrm{ErrorRate}=\frac{\mathrm{NumError}}{\mathrm{NumAll}} \tag{14}$$

表 2 错误率和耗时比较

Tab. 2 Comparisons of classification error rate and training time

Dataset		Wine (L=80)	Segment (L=300)	Waveform (L=3 000)	Pen-digits (L=1 100)	Letter (L=1 100)	Sensorless (L=5 000)
KLDA	错误率	0.053±0.05	**0.026 ±0.02**	0.159±0.02	0.004±0.002	0.047±0.005	*
	耗时/s	0.546 9	45.281 3	465.859 4	4 863.7	25 891	
NNDA	错误率	0.065±0.09	0.038±0.02	*	0.016±0.003	0.051±0.006	*
	耗时/s	8.859 6	4 939.8		12 646	12 639	
ELM	错误率	0.026±0.04	0.045±0.01	0.168±0.01	0.004±0.004	0.060±0.003	0.352±0.002
	耗时/s	**0.009 4**	**0.206 3**	**5.567 2**	**7.862 5**	**13.670 3**	**30.354 4**
ENDA	错误率	**0.015 ±0.05**	0.027±0.01	**0.095 ±0.13**	**0.002 ±0.001**	**0.022 ±0.005**	**0.213 ±0.003**
	耗时/s	0.640 6	3.203 1	244.468 8	17.878 1	971.375 0	1.321 3×10³

表 2 中 * 代表溢出，性能较好结果用粗体标出，虽然 ELM 是 4 种方法里最快速的，但 ENDA 的分类性能相比 ELM 得到很大提升。从综合情况看，当数据规模逐渐变大的情况下，ENDA 算法始终表现出较好分类精度和快速性。分析原因如下：(1)实验中发现 NNDA 需要训练多层神经网络权重和偏置，导致学习效率低、开销大；NNDA 双隐层网络结构致使参数增多，而 ENDA 通过随机映射简化了模型结构，只需要优化 LDA 层权重，就可以达到比 NNDA 更好的效果。(2)NNDA 不仅时间花费巨大，而性能并没有得到改善，从表 2 中 ENDA 和 NNDA 分类结果可以看出，ENDA 性能更优。由于 NNDA 在最后一个隐层到输出层训练出的权重并没发挥训练作用，而且 NNDA 为了防止变形严重用线性函数替代了 Sigmoid 函数，导致学习效果不理想。

从表 2 中耗时结果可看出，ENDA 比 KLDA 计算更快速。随着数据集规模增大，除了 Wine 数据集外，ENDA 的表现明显，准确性更高。这是因为 KLDA 隐节点数与训练样本数呈线性关系(计算复杂度为 N^3)，需要计算 $N\times N$ 大小的核矩阵，而 ENDA 只与隐节点数的大小呈线性关系，只需计算 $N\times L$ 大小的核矩阵，而隐节点参数 L 通常远小于 N，当数据规模变大时，KLDA 计算成本较高甚至溢出，而 ENDA 则表现出更好的可伸缩性。

综上所述，ENDA 比 NNDA，KLDA 更具有极速性，这充分表明随机映射在不影响分类性能前提下大幅度降低了 NN 的复杂性。与 KLDA 不同的是，ENDA，NNDA，ELM 都充分发挥了 NN 的万能逼近能力，而 ENDA 不仅强化了随机映射，而且能保证全局最优解析解，使学习更加迅速和鲁棒。

4 结束语

本文探讨了 LDA 三种非线性化方法。受随机映射启发，将 NNDA 进行改造，提出 ENDA 算法，避免了 NNDA 需要迭代的调整权重和易陷入局部最优等问题，并且具有 KLDA 的全局最优特性，同时又避免了 KLDA 对样本依赖的影响，使其更具极速性和鲁棒性。

由于 ENDA 随机设置权重和偏置带来的不稳定性，很容易联想到集成的学习方法，利用对样本数

据进行重采样技术[27-28]加强分类器泛化性能,对于同质或异质个体学习器集成学习,可使整个模型决策更加智能化。通常对产生的个体学习器进行集成的学习算法涉及稀疏修剪或多样性的度量相关[29]。多样性度量能确保信息完整性,而稀疏化建立原则是在不使用更多的个体学习器的情况下就能达到剪枝的目的。Yin 等人[30]提出同时结合稀疏正则化和多样性度量这两种方法用于凸二次规划求解,极大地改进了集成泛化性能,并有利于大规模数据的并行分布式表示[31-32],具有很好的扩展性,这将是下一步研究内容。

参考文献:

[1] 杜海顺, 张平, 张帆. 一种基于双向 2DMSD 的人脸识别方法[J]. 数据采集与处理, 2010, 25(3): 369-372.
Du Haishun, Zhang Ping, Zhang Fan. Face recognition method based on bidirectional two-dimensional maximum scatter difference[J]. Journal of Data Acquisition and Processing, 2010, 25(3): 369-372.

[2] 牛璐璐, 陈松灿, 俞璐. 线性判别分析中两种空间信息嵌入方法之比较[J]. 计算机科学, 2014, 41(2): 49-54.
Niu Lulu, Chen Songcan, Yu Lu. Comparison between two approaches of embedding spatial information into linear discriminant analysis[J]. Computer Science, 2014, 41(2): 49-54.

[3] Zhang X Y, Huang K, Liu C L. Feature transformation with class conditional decorrelation[C]//2013 IEEE 13th International Conference on Data Mining. Dallas, Texas:IEEE,2013:887-896.

[4] Izenman A J. Linear discriminant analysis[M]. New York:Springer, 2013: 237-280.

[5] Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach[J]. Neural Computation, 2000, 12(10): 2385-2404.

[6] Wang R, Chen X. Manifold discriminant analysis[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009. Anchorage, Alaska: IEEE, 2009: 429-436.

[7] Stuhlsatz A, Lippel J, Zielke T. Feature extraction with deep neural networks by a generalized discriminant analysis[J]. Neural Networks and Learning Systems, IEEE Transactions on, 2012, 23(4): 596-608.

[8] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[9] Wu L, Shen C, Hengel A V D. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person reidentification[J]. Pattem Recognition, 2016,65:238-250.

[10] Mao J, Jain A K. Artificial neural networks for feature extraction and multivariate data projection[J]. Neural Networks, IEEE Transactions on, 1995, 6(2): 296-317.

[11] Hassoun M H. 1996 IEEE transactions on neural networks outstanding paper award[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1996(4): 802-802.

[12] Scholkopft B, Mullert K R. Fisher discriminant analysis with kernels[J]. Neural Networks for Signal Processing IX, 1999, 1:1.

[13] Cai D, He X, Han J. Speed up kernel discriminant analysis[J]. The VLDB Journal, 2011, 20(1): 21-33.

[14] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv Preprint arXiv, 2013:1312.6229.

[15] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[16] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[17] Huang G B, Wang D H, Lan Y. Extreme learning machines:A survey[J]. International Journal of Machine Learning and Cybernetics, 2011, 2(2): 107-122.

[18] Hu G, Peng X, Yang Y, et al. Frankenstein: Learning deep face representations using small data[J]. IEEE Trans Image Process,2017,27(1):293-303.

[19] Arriaga R I, Rutter D, Cakmak M, et al. Visual categorization with random projection[J]. Neural Computation, 2015, 27(10): 2132.

[20] 刘金勇, 郑恩辉, 陆慧娟. 基于聚类和微粒群优化的基因选择方法[J]. 数据采集与处理, 2014, 29(1):83-89.
Liu Jinyong, Zheng Enhui, Lu Huijuan. Gene selection based on clustering method and particle swarm optimazition[J]. Journal of Data Acquisition and Processing, 2014, 29(1):83-89.

[21] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: A new learning scheme of feedforward neural networks[C]// Proc IJCNN. Budapest, Hungary:[s. n.],2004(2):985-990.

[22] Tang J, Deng C, Huang G B. Extreme learning machine for multilayer perceptron[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 27(4): 809-821.

[23] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1): 489-501.

[24] Huang G B. What are extreme learning machines? Filling the gap between frank rosenblatt's dream and John von neumann's puzzle[J]. Cognitive Computation, 2015, 7(3): 263-278.

[25] Tang J, Deng C, Huang G B. Extreme learning machine for multilayer perceptron[J]. IEEE Transactions on Neural Networks & Learning Systems, 2016, 27(4):809-821.

[26] Kasun L L C, Zhou H, Huang G B, et al. Representational learning with ELMs for big data[J]. IEEE Intelligent Systems, 2013, 28(6): 31-34.

[27] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[28] Schapire R E, Freund Y, Bartlett P, et al. Boosting the margin: A new explanation for the effectiveness of voting methods [J]. The Annals of Statistics, 1998, 26(5): 1651-1686.

[29] Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy [J]. Machine Learning, 2003, 51(2): 181-207.

[30] Yin X C, Huang K, Yang C, et al. Convex ensemble learning with sparsity and diversity[J]. Information Fusion, 2014, 20: 49-59.

[31] Wang H, He Q, Shang T, et al. Extreme learning machine ensemble classifier for large-scale data[C]//Proceedings of ELM-2014. Singapore: Springer International Publishing, 2015: 151-161.

[32] Van Heeswijk M, Miche Y, Oja E, et al. GPU-accelerated and parallelized ELM ensembles for large-scale regression[J]. Neurocomputing, 2011, 74(16): 2430-2437.

作者简介：

谢群辉(1984-)，男，硕士研究生，研究方向：人工智能和机器学习，E-mail：xiequnhui@163.com。

陈松灿(1962-)，博士，教授，博士生导师，研究方向：人工智能、机器学习和数据挖掘等领域，E-mail：s.chen@nuaa.edu.cn。

（编辑：夏道家）