三种判别分析方法在元音库上的分类①

潘志方，杨 峰，邵和鸿

(温州医学院 信息与工程学院，温州 325035)

摘 要：分别用降秩线性判别分析(RRLDA)、降秩二次判别分析（RRQDA）和主成分分析+线性判别分析(PCA+LDA)三种模型对数据进行了分析，并在元音测试数据集上进行了测试。分别画出了这三种模型的误分类率曲线，画出了 RRLDA 和 PCA+LDA 分别降至二维后的最优分类面。从实验结果中可以发现，RRLDA 模型的实验结果优于 PCA+LDA 模型，而 RRQDA 的误分类率相当的高，这是因为 PCA 在降维过程中仅仅要求数据分散，而忽略了数据的类内和类间的信息。同时，曲线提示 RRLDA 在子空间的维数取 2 时具有最好的泛化能力，PCA+LDA 在子空间的维数取 4 时具有最好的泛化能力，RRQDA 在第 10 维才有最好的检验误差率。

关键词：元音；类；降秩；判别分析；主成分分析

Classification About Vowel Database Using Discriminant Analysis Methods

PAN Zhi-Fang, YANG Feng, SHAO He-Hong

(School of Information & Engineering, Wenzhou Medical College, Wenzhou 325035, China)

Abstract: The paper analyzes vowel data using reduced-rank linear discriminant analysis (RRLDA), reduced-rank quadratic discriminant analysis (RRQDA) and principal component analysis plus linear discriminant analysis (PCA+LDA). Then it drew some curves of false classification about the three model. A curved surface of the best classification has drawn for RRLDA and PCA+LDA after reduced rank to two dimensions. From the result, it can be conclude that RRLDA is good than PCA+LDA. The false classification of RRQDA is considerably big, because PCA ignores the information of classification about data and only disperses data during reducing rank. Simultaneously these curves prompts RRLDA owning the best generalizing ability when its dimension is 2 in subspace, and PCA+LDA owning the best generalizing ability when its dimension is 4 in subspace, and RRQDA owning the best verify error rate in tenth dimension.

Keywords: vowel; classification; reduced-rank; discriminant analysis; PCA

1 引言

语言是人类最重要的交流工具，随着社会的不断发展，各种各样的机器参与了人类的生产活动和社会活动，因此改善人和机器之间的关系，使人对机器的操纵更加便利就显得越来越重要。随着电子计算机和人工智能机器的广泛应用，人们发现，人和机器之间最好的通信方式是语言通信，而语音是语言的声学表现形式。这样就需要进行语音识别和语音合成，这是语音信号处理技术的主要研究内容。

语音识别的研究始于 20 世纪 50 年代。语音识别技术的根本目的是研究出一种具有听觉功能的机器，使机器能接受人类的语音，理解人的意图。由于语音识别本身所固有的难度，人们提出了各种限制条件下的研究任务，并由此产生了不同的研究领域。这些领域包括：针对说话人，可分为特定说话人语音识别和非特定说话人语音识别；针对词汇量，可划分为小词汇量、中词汇量和大词汇量的识别；按说话方式，可分为孤立词识别和连续语音识别等。最简单的研究领域是特定说话人、小词汇量、孤立词的识别，而最难的研究领域是非特定说话人、大词汇量、连续语音的

① 基金项目:浙江省教育厅项目(Y200803141)
收稿时间:2010-06-08;收到修改稿时间:2010-07-09

识别[1]。

从 20 世纪 70 年代起人工智能技术开始引入到语音识别中来。80 年代开始，语音识别的一个重要进展就是识别算法从模式匹配技术转向基于统计模型的技术，更多地追求从整体统计的角度来建立最佳的语音识别系统。80 年代后期和 90 年代初期开始，人工神经网络的研究异常活跃，并且也被应用到语音识别的研究中[2]。也有人使用遗传编程的方法进行研究[3]。近年来国内有人用支持向量机、高斯混合模型等方法进行研究[4-7]。

尽管语音识别技术研究已经取得了很大的成绩，但很多因素影响着语音识别系统的性能，例如实际环境中的背景噪声、传输通道的频率特性、说话人生理或心理情况的变化以及应用领域的变化等都会导致语音识别系统性能的下降，甚至使系统不能工作。语音识别通常是指能识别出相应的语音内容。在进行语音信号数字处理是，最先接触最直观的是它的时域波形。通常是将语音信号用话筒转换成电信号，再用模拟-数字信号转换器将其转换成离散的数字采样信号后存入计算机中。时域波形虽然简单直观，但对于语音这样复杂的信号而言，一些特性要在频域中才能体现出来，并且无论是从发音器官的共振角度还是从听觉器官的频率响应角度来看，频谱都是表征语音特性的基本参数。其中共振峰就是一个典型的频域参数，它可以决定信号频谱的总体轮廓或谱包络，对于声道而言，它的共振频率不止一个，一般元音可以有 3-5 个共振峰。频谱分析只能反映出信号的频率变化，而不能表示信号的时间变化特性，由于语音信号是一种短时平稳信号，可以在每个时刻用其附近的短时段语音信号分析得到一种频谱，将语音信号连续地进行这种频谱分析，可以得到一种二维图谱，它的横坐标表示时间，纵坐标表示频率，每个像素的灰度值大小反映相应时刻和相应频率的能量。这种时频图称为语谱图[1]。

2 实验数据的说明

前面说过语音识别就说话人分可以分为特定说话人和非特定说话人两种，对于后者，机器能识别任意人的发音，由于语音信号的可变性很大，这种系统要能从大量的不同人的发音样本中学习到非特定人的发音速度、语音强度、发音方式等基本特征，并归纳出其相似性作为识别的标准。使用者无论是否参加过训练都可以共用一套参考模板进行语音识别。

在网站 www-stat.stanford.edu/ElemStatLearn 上下载的元音数据分为训练和测试两部分。数据维数为 10 维，类别有 11 类。其中训练数据 528 条，测试数据 462 条，为不同人所发的元音数据。这些 11 类不同的元音是用 11 个不同的单词来发出的，参看表 1。

表 1 元音-单词对照

vowel	word	word	vowel
i	heed	O	hod
I	hid	C:	hoard
E	head	U	hood
A	had	u:	who'd
a:	hard	3:	heard
Y	hud		

3 模型的理论基础

利用这些元音训练数据，本文分别用降秩线性判别分析(RRLDA)、降秩二次判别分析(RRQDA)和主成分分析+线性判别分析(PCA+LDA)三种模型对数据进行了分析[8]。

3.1 降秩线性判别分析(RRLDA)

由分类的判定理论可知，为了最优分类，需要知道后验概率 Pr(G|X)。由贝叶斯定理知道了 $fk(x)$(类的类条件密度)就可以知道后验概率。假定用多元高斯分布对每个类 $fk(x)=\frac{1}{(2\pi)^{p/2}|\sum^{K}|^{1/2}}e^{-\frac{1}{2}(x-\mu k)^{T}\sum k^{-1}x-\mu k)}$ 线性判别分析(LDA)在当假定类具有共同的协方差矩阵 $\sum_{k}=\sum\forall k$ 时出现。线性判别分析(LDA)使用了线性判别函数

$$\delta k(x)=x^{T}\sum^{-1}\mu k-\frac{1}{2}\mu_{k}^{T}\sum^{-1}\mu k+\log\pi_{k}$$

其中 $G(x)=\arg\max_{k}\delta_{k}(x)$

在实践中，不知道高斯分布的参数，而需要用训练数据估计它们：

$\hat{\pi}k=N_{k}/N, whereNkist\ henumberof\ class-kobservati\ ons;$

$$\mu_{k}=\sum_{gi}=k^{Ti}/Nk;$$

$$\sum=\sum_{K}^{K}=1\sum_{gi=k}(xi-\mu k)(xi-\mu k)^{T}/(N-K)$$

对于两个类，线性判别分析与用最小二乘方分类之间存在一个简单的对应。如果，

$$x^T\Sigma^{-1}(\mu_2-\mu_1)>\frac{1}{2}\mu_2^T\Sigma\mu_2-\frac{1}{2}\mu_1\Sigma\mu_1+\log(N_1/N)-\log(N_2/N)$$

LDA 规则将 x 分到类 2，否则将 x 分到类 1。当类多于两个时，LDA 与类指示矩阵的线性回归不同，并且它避免了与那种方法有关的屏蔽问题。

P 维输入空间的 K 个形心在一个小于或等于 K-1 维的仿射子空间中，并且如果 p 比 K 大得多，则可以相当可观地降低维数。此外，为了确定最近的形心，可以忽略正交于该子空间的距离，因为它们对每个类所起的作用相等。这样，也可以将 X*投影到这个形心生成的子空间 HK-1，并在那里做距离比较。这样，LDA 中存在一个基本的维归约，即最多需要在 K-1 维子空间上考虑数据。例如，K=3 的话，可以在一个 2 维图上观察数据，用颜色对每个类编码。这样就不必放弃 LDA 分类所需要的任何信息。如果 K>3，可以寻找某个 L<K-1 子空间 $H_L \subseteq H_{k-1}$，在某种意义下对 LDA 是最佳的。这实际上等价于找形心本身的主成分子空间。

寻找 LDA 最佳子空间

概括地说，寻找 LDA 的最佳子空间序列设计如下步骤：

① 计算 K×p 的类形心矩阵 M 和公共协方差矩阵 W(关于类内协方差)；

② 使用 W 的本征分解计算 M*=MW-1/2；

③ 计算 M*的协方差矩阵 B*(B 表示类间协方差)和它的本征分解 B*=V*DBV*T。V*的列 v_l^* 从第一个到最后一个依次定义最佳子空间坐标。

将所有这些操作结合在一起，第 l 个判别变量由

$Z_l=v_l^TX$ 给出，这里 $v_l=W^{-1/2}v_l^*$

3.2 降秩二次判别分析(RRQDA)

在一般的判别问题中，如果不假定 $\sum K$ 相等，则可以得到二次判别函数(QDA)

$$\delta_K(x)=-\frac{1}{2}\log|\sum_K|-\frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k)+\log\pi_k$$

先降秩再进行二次判别，可优化结果。

3.3 主成分分析+线性判别分析(PCA+LDA)

在许多情况下，有大量的输入，它们常常是很相关的。主成分回归对某 $M\le p$，形成导出的输入列 $z_m=X_{vm}$，然后在 z_1，z_2，…，z_m 上对 y 回归。由于这些 z_m 是正交的，所以该回归就是一元回归的和：

$y^{per}=y+\sum_{m=1}^{M}\theta_mZ_m$，其中 $\theta_m=<Z_m,y>/<Z_m,Z_m>$。

由于每个 z_m 都是原来的 x_j 的线性组合，可以用 x_j 的系数表示解：

$$\beta^{per}(M)=\sum_{m=1}^{M}\theta_mU_m$$

与岭回归一样，主成分依赖于输入的定标。因此，通常首先要对输入标准化。注意，如果 M=p，回到通常的最小二乘方估计。

IR^p 中数据集合的主成分提供了对所有秩 $q\le p$ 的数据的一系列最佳线性逼近。记观测为 x_1，x_2，…，x_n，考虑表示它们的秩 q 线性模型：

$$f(\lambda)=\mu+V_q\lambda$$

其中 μ 是 IR^p 中的定位向量，V_q 是其 q 个列为正交单位向量的 p×q 矩阵，λ 是参数的 q 向量。这是秩 q 的一个仿射超平面的参数表示。通过最小二乘方对数据拟合这样一个模型相当于极小化重构误差

$$\min_{\mu_i,(\lambda_i)_{Vq}}\sum_{i=1}^{N}\|x_i-\mu-V_q\lambda_i\|^2$$

可以对 μ 和 λi 分别进行优化，得到

$\hat{\mu}=\bar{x}_1$　$\hat{\lambda}_i=V_q^T(x_i-\bar{x})$

接下来求解正交矩阵 V_q

$$\min_{V_q}\sum_{i=1}^{N}\|(x_i-\bar{x})-V_qV_q^T(x_i-\bar{x})\|^2$$

将中心化的观测放入一个 N×p 矩阵 X 的行中，构造 X 的奇异值分解 $X=UDV^T$

UD 的列称做 X 的主成分。

式子中 N 个最优的值由前 q 个主成分给出，也就是 N×q 矩阵 U_qD_q 的 N 个行。

4　实验结果

利用元音训练数据，本文分别用降秩线性判别分析(RRLDA)、降秩二次判别分析（RRQDA）和主成分分析+线性判别分析(PCA+LDA)三种模型对数据进行了分析，并在测试数据集上进行了测试，分别画出了这三种模型的误分类率曲线，画出了 RRLDA 和 PCA+LDA 分别降至二维后的最优分类面,如图 1~图 5 所示。

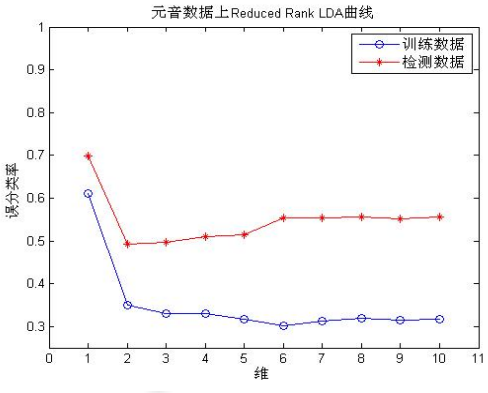

图 1　元音数据 Reduced Rank LDA 曲线

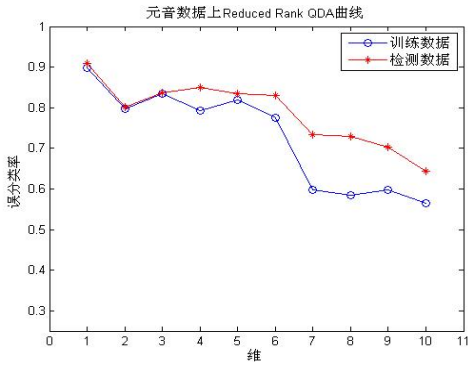

图 2　元音数据 Reduced Rank QDA 曲线

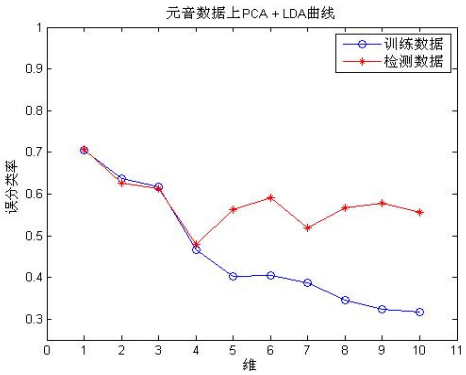

图 3　元音数据上 PCA+LDA 曲线

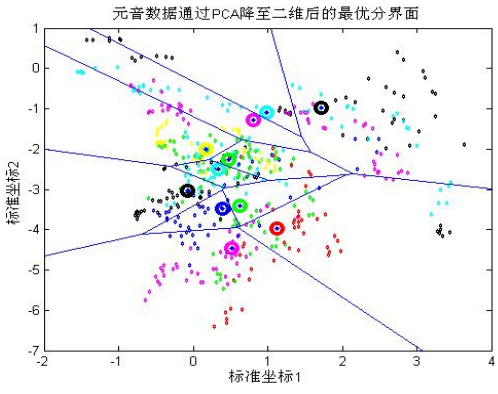

图 4　元音数据通过 PCA 降到二维后的最优分界面

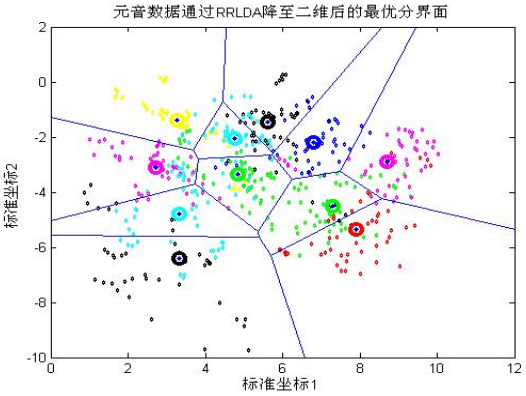

图 5　元音数据通过 RRLDA 降到二维后的最优分界面

5　结果讨论

在元音数据的这种最佳二维子空间的图中，显示了在 10 维输入空间上有 11 个类，每个类是一个不同的元音。在此情况下，形心需要整个空间，因为 K-1=p，但显示了一个最佳 2 维子空间。维是有序的，因此可以顺序计算附加的维。

作为一个受益于降秩限制的例子，元音数据共有 11 个类，10 个变量，因此分类法有 10 个可能的维，可以在每个层次子空间计算训练误差和检验误差，并画出基于二维 PCA 和 RRLDA 解的分类法的判定边界。

用 LDA 方法对元音数据进行分类时，将数据分类到变换后空间中的最近的类形心。这样就需要计算空间中两个数据点的距离，而在高维空间中，这种距离计算不仅耗时，而且容易误差积累最终导致模型失败。虽然这里的元音数据只有 10 维，但是本

实验中还是先对对数据进行降维，然后再进行分类。本实验中降维的模型有三种，RRLDA/RRQDA 和 PCA。

首先对于 RRLDA 和 RRQDA，这两种模型子空间坐标的选取是通过球形化数据以后，求类间协方差矩阵的正交本证特征向量得到的。它保证了数据在子空间上的投影具有同类数据尽量紧凑，不同类数据尽量分开的特性。而对于 PCA，这一模型子空间的坐标的选取是通过对数据直接求协方差矩阵的正交特征向量得到的。它保证了数据在子空间上的投影具有最大的方差，也就是数据在子空间上尽量的分散。

这几种模型的运用，都需要知道子空间的维数，子空间的维数作为模型的可变参数具有重要意义。

从实验结果中可以发现，RRLDA 模型的实验结果优于PCA+LDA 模型，而 RRQDA 的误分类率相当的高。理由很简单，因为 PCA 在降维过程中仅仅要求数据分散，而忽略了数据的类内和类间的信息。同时，曲线提示 RRLDA 在子空间的维数取 2 时具有最好的泛化能力，PCA+LDA 在子空间的维数取 4 时具有最好的泛化能力，RRQDA 在第 10 维才有最好的检验误差率。

除了必须对每个类分别估计协方差矩阵外，QDA 估计与 LDA 估计类似，但误分类率为什么会相差这么大呢？可能是因为当 p 较大时，这可能意味参数急剧增加，由于判定边界是密度参数的函数，计算参数的数目必须小心。对于 LDA 看来有(K-1)×(p+1)个参数，这是因为只需要判别函数的差 δk(x)- δK(x)，其中 K 是某个预先选定的类，而每个差需要 p+1 个参数，对于 QDA 有(K-1)×{p(p+3)/2+1}个参数。

参考文献

1 韩纪庆,张磊,郑铁然.语音信号处理.北京:清华大学出版社,2004.1－328.

2 Iulian B. Ciocoiu. Hybrid Feedforward Neural Networks for Solving Classification Problems. Neural Processing Letters, 2002,16(1):81－91.

3 Markus Conrads, Peter Nordin, Wolfgang Banzhaf. Speech sound discrimination with genetic programming. Heidelberg: Springer Berlin, 2006. 113－129.

4 赵培.中文语音识别结果文本分类的研究与实现[硕士学位论文].大连:大连理工大学,2008.

5 黄锋,尹俊勋.一种改进的基于 GMM 模型的语音序列评分和分类方法.湖南大学学报(自然科学版),2008,35(11):79－82.

6 侯雪梅.一种 SVM 多类分类算法用于抗噪语音识别.西安邮电学院学报,2009,14(5):100－102,135.

7 赵培,牛纪桢,史金艳.改进的 SVM 在语音识别文本分类中的应用.广西师范大学学报(自然科学版),2009,7(1):137－140.

8 范明,柴玉梅,咎红英,译.统计学习基础——数据挖掘、推理与预测.北京:电子工业出版社,2007.55－77.

(上接第 151 页)

运行速度之快很明显[8]。

因此，协议方案具有良好的安全性、运行速度快、计算负担轻、存储需求低以及带宽需求少等特点。

5 结论

在分析已有双向认证协议的基础上，指出利用 ECC 技术的迫切需求。以 Helsinki 认证协议草案为基础，结合 Mitchell-Yeun 改进协议的思想，设计了一个基于 ECC 的双向认证协议。与以往的双向认证协议相比，本文协议具有较大的优越性，可以广泛应用在数据通信网络中，特别是无线数据通信中。

参考文献

1 王亚弟,束妮娜,等.密码协议形式化分析.北京:机械工业出版社,2006.

2 Boyd C, Mathuria A. Protocols for Authentication and key Establishment. Berlin: Springer-Vergles, 2003. 300－370.

3 Aydos M, Sunar B, Koc CK. An Elliptic Curve Cryptography based Authentication and Key Agreement Protocol for Wireless Communication. Workshop Discrete Algorithms and Methods for Mobility (DIAL'98). Dallas. TX, Oct. 1998.

4 Horn G, Preneel B. Authentication and Payment in future mobile systems. In: Quisquater JJ, et a1. eds. ESORICS98. LNCS 1485, Springer-Verlag, 1998. 175－191.

5 Sun HM, Hsieh BT, Tseng SM. Cryptoanalysis of Aydos, et al. ECC-based wireless authentication protocol. Proc. IEEE International Conference on e-Technology, e-Commerce and e-Service. Los Alamitos: IEEE Computer Society, 2004. 563－566.

6 Kapoor V, Abraham VS. Elliptic Curve Cryptography. ACM Ubiquity, 2008,9(20).