基于 PCA 和 LDA 的方言辨识①

何 艳，于凤芹

(江南大学 物联网工程学院，无锡 214122)

摘 要：针对 PCA 没有有效利用样本的类别信息而导致方言识别率低的问题，采用 PCA 和 LDA 组合方法进行特征提取。首先用 PCA 对普通话、上海话、广东话和闽南话四种方言进行降维，然后在降维后的空间中用 LDA 进一步特征提取，最后将该特征向量送入 BP 神经网络进行辨识。仿真实验结果表明，基于 PCA 和 LDA 的方言识别的平均识别率高达 85%。

关键词：方言辨识；主成分分析；线性鉴别分析；BP 神经网络

Dialect Identification Based on PCA and LDA

HE Yan, YU Feng-Qin

(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: In order to solve the low dialect identification rate because PCA doesn't effectively use the sample classification information, a method of feature extraction using PCA and LDA is employed. In this paper, PCA is used to effectively reduce the dimensions of Mandarin, Shanghainese, Cantonese, Minnanese, and then LDA is adopted to extract feature vectors from the dimension-reduced space as the input vectors with BP neural network to recognize. The Simulation results demonstrate that the average dialect identification rate based on PCA and LDA can be up to 85%.

Key words: dialect identification; PCA; LDA; BP neural network

方言识别在公安刑侦工作和语音识别技术的推广和应用中有着重要意义，已越来越受到相关领域研究人员的重视。在方言辨识中，提取语音信号的特征至关重要。主成分分析法（Principal Component Analysis，PCA）是目前常用的特征提取方法，此变换可以达到降维的目的，且降维后能保存样本的主要信息[1]。PCA 广泛应用与语音识别中，文献[2]将语音信号分割为子词后，对各子词单元内各帧语音的特征矢量进行 PCA，且用 DTW 进行语音识别，识别率为 90%左右。

PCA 由于没有有效利用样本的类别信息，所以用 PCA 算法得到的特征并不是最有辨别力的特征。线性鉴别分析（Linear Discriminant Analysis，LDA）也是通过求取一个变换矩阵再做线性转换来达到降维的目的，但与 PCA 不同的是，LDA 使模式样本内的分布凝聚而使样本间的分布疏远[3]。LDA 算法的目的在于从高维特征空间提取出具有辨别力的低维特征。文献[4]采用 LDA 特征提取方法在中文大词汇量连续语音识别系统中音节识别率达到 82.16%。文献[5]证明了 LDA 不仅可以应用在 PCA 降维后的空间中，并且通过 PCA 降维可以使 LDA 散布矩阵的维数进一步减小，从而在一定程度上避免 LDA 的小样本问题，提高 LDA 算法的可用性。

因此，本文提出了 PCA 和 LDA 相结合的特征提取方法，首先对普通话、上海话、广东话和闽南话四种方言的语音信号用 PCA 方法进行有效的降维，然后为了提取具有辨别力的低维特征，在降维后的空间中继续用 LDA 进行特征提取，最后将该特征向量作为 BP 神经网络的输入来进行辨识。仿真实验结果表明，基于 PCA 和 LDA 组合的识别率为 85%，PCA 方法的识别率为 72.5%，LDA 方法的识别率为 75%。

① 基金项目:国家自然科学基金(61075008)
收稿时间:2011-08-18;收到修改稿时间:2011-09-26

1 算法原理

1.1 PCA

PCA 是由 Turk 和 Pentlad 提出来的，以 K-L 变换为基础的一种常用的特征提取技术。PCA 方法是将研究对象的多个属性指标化为少数几个不相关变量的一种多元统计方法[6]。

给定输入数据矩阵 $X_{m\times n}=\{x_i\}_{i=1}^{n}$(通常 $m<n$)，其中 $x_i\in R^n$。首先定义样本集的协方差矩阵：

$$C=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T \tag{1}$$

其中，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}(x_i)$ 是样本集的平均向量，该矩阵表现了样本集向量中各个元素的相关性。

此时对 C 进行特征分解，得到 $\lambda=\{\lambda_1,\cdots,\lambda_n\}$ 为 C 从大到小的特征值，$U=\{u_1,\cdots,u_n\}$ 为对应的特征向量，保留若干个大于零的特征值($\lambda_1\geq\cdots\geq\lambda_d>0$)对应的特征向量组成样本集的特征空间 $P=(u_1,u_2,\cdots,u_d)$。把样本向特征空间上投影，即可得到样本的特征向量。

PCA 的主分量具有如下的特征[7]：

1) 行矢量 $P(i),i=1,\cdots,d$ 互不相关；

2) $P(i),i=1,\cdots,d$ 顺序地具有最大的方差；

3) 用最前面的几个主分量表示原输入，其均方逼近误差最小。

1.2 LDA

LDA 方法是基于 Fisher 准则，寻找一组将高维样本投影到低维空间的最佳的判别投影向量，使所有的投影样本类内离散度最小且类间离散度最大，又被称为 Fisher 线性判别分析(FLD)[8]。

LDA 算法首先定义样本集类内散布矩阵和类间散布矩阵，此二矩阵分别代表样本集每个类别内部向量各元素的相关性以及各个类别均值向量各元素的相关性，算法需要找到一个最有投影向量，并且该向量可以同时对两个散布矩阵进行对角化，并保证对角化后的类间散布矩阵和类内散布矩阵比值最大，此向量即为样本集的 LDA 特征空间。

假设 $X=\{x_1,x_2,\cdots x_n\}$ 是 m 维列向量样本集，有 C 个模式类别 $\{\omega_1,\ \omega_2,\cdots\omega_c\}$，每类语音的个数为 n_j，其中 $x_i\in\omega_j,\ j=\{1,\cdots c\}$，$\omega_j$ 类语音的均值为 $m_j=\frac{1}{n_j}\sum x$，总体语音的均值为 $m=\frac{1}{n}\sum x$，定义样本集类内散布矩阵 S_ω 和类间散布矩阵 S_b 分别为：

$$S_b=\sum_{j=1}^{c}n_j(m_j-m)(m_j-m)^T \tag{2}$$

$$S_\omega=\sum_{j=1}^{C}\sum_{x\in X}(x-m_j)(x-m_j)^T \tag{3}$$

为了使原始数据经过 LDA 投影降维后，在低维空间更加容易分类，我们希望确定投影方向 ξ，使得在 ξ 方向 S_b 和 S_ω 的投影比值最大，也就是满足 Fisher 准则：

$$\arg_\eta\max\frac{\left|\xi^TS_b\xi\right|}{\left|\xi^TS_\omega\xi\right|} \tag{4}$$

由线性代数理论可知，当 S_ω 非奇异时，ξ 是 $S_\omega^{-1}S_b$ 的最大特征值对应的特征向量。通常情况下只有一个特征向量是不够的，设 $\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_d,\cdots\lambda_m)$ 是 $S_\omega^{-1}S_b$ 的特征向量，其中 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d>0$，$\xi=(\xi_1,\xi_2,\cdots,\xi_d)$ 是对应的特征向量，我们取 $Q=(\xi_1,\xi_2,\cdots,\xi_d)$ 作为 LDA 算法的特征空间。

2 算法实现

由上述对 PCA 和 LDA 方法的分析可知，PCA 可以达到降维的目的，但没有考虑到样本的类别信息，而 LDA 充分考虑了样本的类别信息。因此本文采用了一种将 PCA 和 LDA 相结合的方法，对语音信号进行特征提取。

下面是基于 PCA 和 LDA 的方言辨识的具体实现步骤：

1) 读取普通话、上海话、广东话和闽南话四类方言的语音信号的训练集为 X，语音信号的总数为 N。先将训练语音进行分类，把属于同一种方言的语音信号归为一类，记为 X_j，X_j 类中的语音数目为 N_j，其中 $j=1,2,3,4$，则有 $\sum_{j=1}^{4}N_j=N$，$x_i\in X_j$（$i=1,\cdots,N$），$X_j\in X$。

2) 计算所有训练语音信号的总体向量均值为 $m=\frac{1}{N}\sum x_i$，每类语音信号的均值为 $m_j=\frac{1}{N_j}\sum x_i$，并将每个语音信号减去总体语音的向量均值，则有 $\overline{x_i}=x_i-m$。

3) 计算样本集类内散布矩阵 S_ω 和类间散布矩阵 S_b 分别为 $S_b=\sum_{j=1}^{c}n_j(m_j-m)(m_j-m)^T$ 和 $S_\omega=\sum_{j=1}^{C}\sum_{x\in X}(x-m_j)(x-m_j)^T$。

4) 计算协方差矩阵 XX^T 的特征值，并从大到小排列为 $\lambda=\{\lambda_1,\cdots,\lambda_N\}$，$U=\{u_1,\cdots,u_N\}$ 为对应的特征向量，保留若干个大于零的特征值($\lambda_1\geq\cdots\geq\lambda_d>0$)对应的特征向量组成样本集的特征空间 $P=(u_1,u_2,\cdots,u_d)$。

5) 将语音信号类内均值和总体均值投影到 PCA 子空间，则有 $m_j'=Pm_j$，$m'=Pm$，$x_i'=Px_i$。

6) 计算第 j 类的类内散布矩阵 $S_j=\sum_{j=1}^{4}\sum_{x\in X}(x'-m_j')(x'-m_j')^T$，总的类内散布矩阵 $S_\omega=\sum_{j=1}^{4}S_j$，以及类间散布矩阵 $S_b=\sum_{j=1}^{c}n_j(m_j'-m')(m_j'-m')^T$。

7) 求解类内散布矩阵 S_ω 和类间散布矩阵 S_b 的广义特征值，并按从大到小的顺序排列，则相应的特征向量为 Q。

8) 组合 PCA 和 LDA 的子空间 $W=P^TQ^T$，则语音信号 x 在该空间的投影下可以得到 $z=P^TQ^T(x-m)$。

9) 把向量 $z=P^TQ^T(x-m)$ 作为语音信号的特征向量，送入 BP 神经网络进行方言辨识。

3 仿真实验

实验将普通话、上海话、广东话和闽南话四种方言作为研究对象，用 11025Hz 采样频率，16 位采样精度对语音信号进行采样。语音信号来自网页，每种方言有 60 个语音，男生、女生各一名同学朗读，其中 40 个语音为训练语音，20 个语音为测试语音。

该实验首先对普通话、上海话、广东话和闽南话四种方言的语音信号用 PCA 方法进行有效的降维，然后考虑到提取最具辨别力的低维特征，在降维后的空间中继续用 LDA 进行特征提取，最后将该特征向量作为 BP 神经网络的输入来进行辨识。为了对比，分别用 PCA 方法和 LDA 方法对语音信号降维后的特征向量作输入向量，送入 BP 神经网络进行辨识，比较各自的识别性能。表 1 为用 PCA 提取语音特征时各个方言的识别率，表 2 为用 LDA 提取语音特征时各个方言的识别率，表 3 为用 PCA+LDA 提取特征时各个方言的识别率。

表 1 PCA 的方言识别率

识别率	普通话	上海话	广东话	闽南话
普通话	80%	20%	0%	0%
上海话	10%	75%	5%	10%
广东话	0%	10%	75%	15%
闽南话	10%	0%	30%	60%

表 2 LDA 的方言识别率

识别率	普通话	上海话	广东话	闽南话
普通话	75%	25%	0%	0%
上海话	25%	65%	10%	0%
广东话	10%	0%	70%	20%
闽南话	0%	5%	%	90%

表 3 PCA+LDA 的方言识别率

识别率	普通话	上海话	广东话	闽南话
普通话	90%	10%	0%	0%
上海话	10%	80%	10%	0%
广东话	0%	0%	100%	0%
闽南话	0%	0%	30%	70%

从表 1、2 和 3 中可以看出，PCA 的平均识别率为 72.5%，LDA 的平均识别率为 75%，PCA+LDA 的平均识别率为 85%，所以 PCA+LDA 方法要优于 PCA 和 LDA 方法。由于 PCA 算法侧重表达原始模式特征，没有充分利用样本的类别间的信息，而 LDA 侧重于反映不同类模式之间的差别，所以 PCA+LDA 的识别率明显增加。

4 结论

PCA 由于算法本身的原因，并没有针对类别间的差别进行分析，通常并不能找到最完整和最优的特征空间。而 LDA 利用了样本的类别信息在一定条件下可以找到最优的特征空间。因此，本文采用了 PCA 和 LDA 组合的特征提取方法，首先用 PCA 对普通话、上海话、广东话和闽南话四种方言进行降维，然后用 LDA 在降维后的空间中进一步进行特征提取，最后用 BP 神经网络进行辨识。仿真实验结果表明，PCA 和

(下转第 179 页)

session 中保存一个已处理的结果集条数，前台根据该条数计算出处理数据的进度，显示在页面的就是一个随着时间滚动的进度条，提示用户消耗的时间，提高了系统界面的友好性。

系统通过 Ajax 技术与后台保持联系，由线程去解决大量数据的查询问题，前台页面只需要定时访问 session 中的查询结果即可，极大的改善了同步调用产生的页面失效的问题。

在改进后的系统中查询一年的报文数量统计时，不再出现页面无法显示的情况。显示效果如图 4 所示。

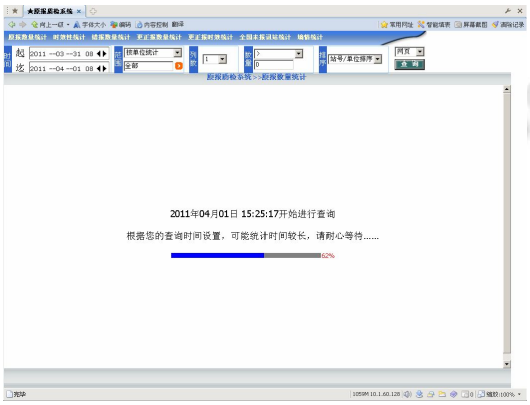

图 4 系统页面

5 结语

通常情况下，BS 架构的应用在做实时大数据量统计分析的时候，当查询结果在 1 个小时内没有返回时，虽然服务器端仍旧在进行运算，但是客户端浏览器页面会产生无法显示的问题。本文提出采用 Ajax 和线程相结合的技术来解决上述问题，并在原报质检系统的报表运算分析上得到了应用。在服务器进行运算的同时通过 ajax 调用运行状态来获取查询进度，并通过进度条进行展示，一方面保证页面的时效性，一方面便于用户了解查询操作的进展，希望对其他系统在遇到相同问题时起到一定的借鉴作用。

参考文献

1 徐驰.Ajax 模式在异步交互 Web 环境中的应用.计算机技术与发展,2006,16(11):228−233.
2 刘晓华,张健,周慧贞.JSP 应用开发详解.第 3 版.北京:电子工业出版社,2007.445−449.
3 余翔宇.AJAX 技术及其框架实现.软件导刊,2006,(9):28−30.
4 骆斌,费翔林.多线程技术的研究与应用.计算机研究与发展,2000,37(4):407−412.

(上接第 171 页)

LDA 组合的识别率最高，LDA 的识别率次之，PCA 的识别率最低。由于 PCA+LDA 的计算复杂度增加，系统的运行时间相应延长，因此以后的研究重点是简化算法，节省系统的运行时间。

参考文献

1 夏鹏,张浩然,徐展敏.一种增量 PCA 算法及其在人脸识别中的应用.计算机工程与应用,2008,44(6):228−230.
2 史笑兴,王太君,何振亚.基于主元分析的语音特征提取.第九届全国信号处理学术年会,1999:258−261.
3 王海珍.基于 LDA 的人脸识别技术研究.西安:西安电子科技大学,2010.
4 王安娜,王勤万,刘俊芳,袁文静.改进的语音特征提取方法及其应用.计算机工程,2008,34(5):196−200.
5 Yang J, Yang JY. Why can LDA be performed in PCA transformed space. Pattern Recognition, 2003,36:563−566.
6 Myoung SP, Jin HN, Jin YC. PCA-based feature extraction using class information. Proc. of 2005 IEEE International Conference on Systems, Man and Cybernetics. 2005: 341−345.
7 Dagher I. Incremental PCA-LDA algorithm. Proc. of 2010 IEEE International Conference on Computational Intelligence Measurement Systems and Applications. 2010: 97−101.
8 庄哲民,张阿妞,李芬兰.基于优化的 LDA 算法人脸识别研究.电子与信息学报,2007,29(9):2047−2049.