一种二次投影识别蛋白质谱数据的新方法*

蒋胜利[1,2]，张军英[1]，许 进[3]
（1. 西安电子科技大学计算机学院，陕西 西安 710071；
2. 洛阳师范学院信息技术学院，河南 洛阳 471022；
3. 华中科技大学控制科学与工程系，湖北 武汉 430074）

摘 要：作为一种蛋白组学工具，质谱法的使用对疾病的早期诊断和治疗带来了革命性的变化。然而，由于面临“维数灾难”问题，大部分机器学习方法不能直接用于识别蛋白质谱数据，同时这些方法也面临着识别性能较低的问题。借鉴主分量分析（PCA）与局部线性判别嵌入（LLDE）在人脸识别方面取得的较好效果，提出了用于蛋白质谱数据识别的二次投影法（DTP）及改进的二次投影法（MDTP）。该方法先对数据去噪并用T检验降维，再提取均方误最小的第一次投影特征向量与可分性最好的第二次投影特征向量，最后将预处理过的数据先后在二次特征向量空间投影并分类。在卵巢癌蛋白质谱数据上的实验表明，二次投影及其改进方法识别性能较好，优于现有各方法。

关键词：蛋白质谱；主分量分析；局部线性嵌入；最大边界准则；模式识别
中图分类号：TP391.41 **文献标识码**：A **文章编号**：0529-6579（2009）06-0027-07

A New Method for Recognition Proteomic Mass Spectrometry Data Using Double-Time Projections

JIANG Shengli[1,2], *ZHANG Junying*[1], *XU Jin*[3]
（1. School of Computer Science and Engineering, Xidian University, Xi'an 710071, China;
2. Academy of Information Technology, Luoyang Normal University, Luoyang Henan 471022, China;
3. Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China）

Abstract: The use of mass spectrometry as a proteomics tool is poised to revolutionize early disease diagnosis and treat. Unfortunately, due to existing "the curse of dimensionality", most standard machine learning techniques cannot be directly applied to recognition proteomic mass spectrometry, and these methods also faced with the problem of poor recognition performance. For better efficiency to use principal component analysis (PCA) and local linear discriminant embedding (LLDE) for face recognition, a Double-Time Projections method and a Modified Double-Time Projections method are proposed for recognition proteomic mass spectrometry. The proposed methods first do de-noising and dimension reduction by T-test, and then obtain first projection feature vectors with minimum mean square error and get second projection feature vectors with maximum separability. Finally, preprocessed data are projected in the sub space based on two feature vectors and, are classified. The experimental result from the dataset of ovarian cancer proteomic mass spectrometry indicates that the proposed methods have a better accuracy than available methods.

Key words: proteomic mass spectrometry; principal component analysis; locally linear embedding; maximum margin criterion; pattern recognition

* 收稿日期：2008-11-28
基金项目：国家自然科学基金资助项目（60533010）
作者简介：蒋胜利（1968年生），男，博士生，讲师；E-mail：jiangshl@163.com

卵巢癌是女性生殖器官常见的肿瘤之一，提高该病的早期诊断率一直是医学界关注的一个热点[1-2]。质谱数据技术的出现使一次获得大量生物信息成为现实，为研究隐藏的规则提供了新方法。目前，用机器学习的方法将蛋白质谱数据用于癌症早期诊断已成为模式识别研究的热门领域[3-4]。Yu 等提出了使用 K_ S 检验与小波分析等四步策略可以使检验的特异性达到93%[1]。Wang 等[2]提出了简约独立门限特征选择方法（Parsimonious Threshold-Independent Feature Selection，PTIFS），在获得较少特征的情况下取得了较好的分类效果。Robert 等提出了使用峰度概率对比的技术对蛋白质谱数据分类取得了较好的效果[5]。Tang 等提出了使用决策树、支持向量机与神经网络相结合的方法从病人与健康人的二种蛋白质谱数据中提取特征，从而区分出两类人员[6]。Alexe 等[7]提出了一个基于组合优化的逻辑数据分析（Logical Analysis of Data，LAD）方法来构建卵巢癌诊断模型，他们所构建的模型仅由 7 ~ 9 个质荷比值构成，特异性几乎可以达到 100%。Levner 等[8]在癌细胞蛋白质谱数据中进行试验，测试了最临近重心分类算法（the Nearest Centroied Classifier，NCC）与 K-S 检验（Kolmogorov-Smirnov test）、T 检验（T-test）和 P 检验（P-test）这三种统计方法分别结合进行识别的效果。然后，他采用 PCA，PCA 结合 LDA，和他提出的一种基于 boosting 特征选择的改进方法对相同的数据集进行试验，并指出这种基 boosting 特征选择的改进方法在他的试验中取得了较好的效果。Kirby 等[9]先用 PCA 方法对卵巢癌质谱数据进行特征提取，并采用 LDA 结合最临近重心分类算法的结合方法进行分类，并提出此方法的效果要好于单独用 LDA 和最邻近重心分类算法的效果。

根据是否考虑类信息，这些方法可以分为两类：监督的与非监督的。也可以简单地分为线性与非线性两类。线性特征提取方法是把高维的输入数据通过线性转换投影到有意义的低维子空间。这些方法中最著名的是主分量分析（Principal Components Analysis，PCA）与线性判别分析（Linear Discriminant Analysis，LDA）。然而，PCA 不考虑类信息，是纯粹的非监督方法，可能会丢失有用的分类信息。与 PCA 不同，LDA 使用了类信息，从而提高了识别能力。然而 LDA 也有一些缺陷，当样本点的维数比样本数大时可能会产生小样本（Small Sample Size，SSS）问题。当然，也可以通过核方法解决这些问题，但是核方法也面临着计算代价过大与核参数如何设置的问题。与核方法不同，流形学习也被广泛应用在模式识别方面，发现隐藏在高维数据中的本质非线性结构。局部线性嵌入（Locally Linear Embedding，LLE）[10]是流形学习方法的代表。然而，LLE 在模式识别中的一个局限就是未知类别的新样本处理过程麻烦，另一个局限就是 LLE 忽略了类信息，这可能影响识别的精确性。为此，Li[11]提出融合了 LLE 与最大边界准则（Maximizing Margin Criterion，MMC）两种方法提出一种监督版的 LLE，即局部线性判别嵌入（Locally Linear Discriminant Embedding，LLDE），用于人脸识别取得了较好的效果。

然而，LLDE 应用在蛋白质谱数据上也存在许多的限制。首先是蛋白质谱数据的维数高达上万维，不可能直接应用 LLDE 方法。再者，高维的数据在构建重建权重过程中计算代价巨大，从而失去计算的可能性。其次，MMC 仅是基于最大化类间散度与类内散度的差，没有考虑到最小均方误。最后，LLDE 中重建权重矩阵没有考虑到训练数据的先验类信息。在解决 LLDE 诸多限制的基础上，我们提出了一种与上述诸特征提取方法不同的蛋白质谱数据识别新方法即二次投影（Double-Time Projections，DTP）方法。这种新方法对预处理过的测试样本数据，先后两次投影到低维子空间，再用分类器分类。与 LLDE 方法相比，我们提出的 DTP 方法在考虑了数据最小均方误的前提下，有效地降低了蛋白质谱数据的维数，极大地降低了计算代价。并且，为充分利用已有的类信息，提出了改进的二次投影方法（Modified Double-Time Projections，MDTP）。作为蛋白质谱数据识别的一般框架，与现有的蛋白质谱特征提取方法相比，在卵巢癌蛋白质谱数据上实现表明，DTP 及 MDTP 两种方法识别性能接近 100%，取得了较好的效果，优于现有各方法。

1 LLDE

1.1 LLE

LLE 算法是 Roweis 等[10]针对非线性数据提出的一种无监督的降维方法，能够使降维的数据保持原有的拓扑结构。设 $X = [X_1, X_2, \cdots, X_N] \in \mathbb{R}^{D\times N}$ 表示 D 维空间中的 N 个样本点，LLE 把高维空间数据点 X_i 按维数映射到低维嵌入空间为 Y_i，即 φ：$X_i \rightarrow Y_i$，步骤如下：

第 1 步：计算每一个点 X_i 的近邻点，一般采用 K 近邻或者 ε 邻域。

第 2 步：计算重构权值 W_{ij}。使得 X_i 用它的 K 个近邻点线性表示的误差最小，即在式（2）的约束条件下，根据式（1），求出 W_{ij}。

$$\varepsilon_i(W) = \operatorname{argmin}\left\| X_i - \sum_{j=1}^{K} W_{ij}X_j \right\|^2 \tag{1}$$

$$\begin{cases} \sum_{j=1}^{k} W_{ij} = 1 & if \quad X_j \in \mathbb{N}(X_i) \\ W_{ij} = 0 & if \quad X_j \notin N(X_i) \end{cases} \tag{2}$$

第 3 步：根据权值 W_{ij}，求 X_i 在低维空间的象 Y_i，使得低维重构误差最小，即满足式（3），基于权重矩阵 W 可以定义一个稀疏对称半正定的矩阵 $M = (I - W)^T(I - W)$，则式（3）可以表示成式（4）。由 Rayleigh-Ritz 原理，使式（4）最小化可以通过求矩阵 M 的最小非零特征值相对应的特征向量来完成。

$$\varepsilon(Y) = \operatorname{argmin} \sum_i \left\| Y_i - \sum_{j=1}^{K} W_{ij}Y_j \right\|^2 =$$

$$\operatorname{argmin}\operatorname{tr}\left\{ \sum_{ij} Y_j(\delta_{ij} - W_{ij})(\delta_{ij} - W_{ij})^{\mathrm{T}}Y_i^{\mathrm{T}} \right\} \tag{3}$$

$$\varepsilon(Y) = \operatorname{tr}\left\{ \sum_{ij} M_{ij}Y_i^{\mathrm{T}}Y_j \right\} = \operatorname{tr}\{YMY^{\mathrm{T}}\} \tag{4}$$

1.2　LLDE

Li 等[12]将 LDA 准则中类间离散度矩阵与类内离散度矩阵的比值关系改为相减关系，提出基于 MMC 的特征提取方法。Li 等[11]提出了融合 LLE 与 MMC 的 LLDE 方法。LLDE 方法有如下特点：构建一个可以增强 LLE 识别能力的平移与距离缩放模型向量，同时 LLE 具有两个属性，一个是嵌入低维的平移与缩放代价函数不变，另一个是引入了修改的 MMC。基于第一个属性，低维嵌入在构建误不变的情况下，数据可被平移到任何低维子空间，基于第二个属性，由输入点的类间离散度与类内离散度线性组成的向量，在最大化边界准则约束下使分类能力最大。LLDE 步骤如下：

为处理新样本，引入一个线性变换 $Y = V^{\mathrm{T}}X$，这样源 LLE 的目标函数式（4）可以变换为式（5）。修改的 MMC 的目标函数如式（6）。

$$J_1(V) = \min \operatorname{tr}\{YMY^{\mathrm{T}}\} = \min \operatorname{tr}\{V^{\mathrm{T}}XMX^{\mathrm{T}}V\} \tag{5}$$

$$J_2 = \max \operatorname{tr}(S_b - \mu S_w) \tag{6}$$

其中 S_b 为类间离散度矩阵，S_w 为类内离散度矩阵，μ 为调节参数。如果一个线性变换可以使 J_2 最大化，则一个最优的用于分类的子空间就可以确定，这是因为这个线性变换的目标是同类投影后距离较近，异类较远。或者说寻找一个用于分类的较优线性子空间意味着使式（7）的优化函数最大化。

$$J_3(V) = \max \operatorname{tr}\{V^{\mathrm{T}}(S_b - \mu S_w)V\} \tag{7}$$

由于受到 LLE 中式（2）约束，使得 $V^{\mathrm{T}}XX^{\mathrm{T}}V = nI$。解决 J_1 与 J_3 这个多目标优化问题可使用 Lagrangian 乘子法。于是，得到式（8）

$$(XMX^{\mathrm{T}} - (S_b - \mu S_W))V = \lambda XX^{\mathrm{T}}V \tag{8}$$

其中，λ 是 $(XMX^{\mathrm{T}} - (S_b - \mu S_W))$ 和 XX^{T} 的广义特征值，V 是相应的特征向量。因此，当取广义特征分解的前 d 个最小特征值相应的特征向量，多目标优化函数最小化，求出的 V 即为低维子空间的投影向量。

2　二次投影（DTP）方法

2.1　数据预处理

一个蛋白质谱数据样本集合可表示为 $X = [X_1, X_2, \cdots, X_N] \in \mathbb{R}^{D\times N}$，其中 D = 15 154 表示样本数据的维数，N = 253 表示样本的数量。样本数据的每一维为一定质荷比（M/Z）对应的强度（蛋白质相对含量）。高维的蛋白质谱数据中有较多冗余信息，根据 Alexe 等[7]提出的理论，可把质荷比（M/Z）值小于 500 的点作为噪声去掉。这样，每个样本剩下了 12 757 维的数据。

可以看出，去噪后的数据集合 X 是一个 12 757 行 253 列的矩阵，维数依然偏高。根据 T 检验降维有利于分类的理论[1]，对矩阵 X 做双边 T 检验，统计量 t 计算方法如式（9）。

$$t = (\bar{x} - \bar{y})/s\sqrt{\frac{1}{n} + \frac{1}{m}} \tag{9}$$

其中，$\bar{x}$ 和 $\bar{y}$ 分别代表所有癌症与非癌症样本在每一质荷比所对应强度值的平均值，S 是两类样本共同组成的每个质荷比数据所对应强度值的标准差，n 和 m 分别代表癌症患者的数目与健康人的数目。计算每一个质荷比对应的统计量 t 值，根据 t 值生成所有质荷比对应的 P 值，P 值越小，表示对应的质荷比位置越能体现两类样本之间的差异，把 t 检验得到的最小 n 个 P 值所对应的强度值取出作为预处理后的数据。

2.2　PCA

主成分分析也称主分量分析、主元分析，它的思想来源于 K－L 变换，目的是通过线性变换找一组最优的单位正交向量基，用它们的线性组合来重建原样本，并使重建后的样本和原样本的误差最小。经过预处理后的蛋白质谱数据是一种小样本问题，后续处理中类内散布矩阵可能存在奇异或秩亏，因而先用主分量分析降维，既可以放宽小样本

问题的局限，又可以降低后续算法的计算代价。寻找主分量的公式如下：

$$C = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}} \quad (10)$$

$$CV = \lambda V \quad (11)$$

其中 N 表示样本的总个数，$\bar{X}$ 是所有样本的均值。对式（11）求特征值，把特征值按降序排列 $\lambda_i \geqslant \lambda_{i+1}$，选择对应前 M（通常 $M << N$）个非零特征值的特征向量作为主分量。

2.3 二次投影（DTP）方法

二次投影方法的步骤如下：

第1步：预处理所有样本数据，得到样本矩阵 X。

第2步：使用PCA寻找第一次投影方向 $V_1 \in \mathbb{R}^{d\times p_1}$，即均方误最小约束下的投影方向。将预处理过的 M 个训练样本投影，得到低维子空间矩阵 $Y_1 = V_1^{\mathrm{T}} \times X, Y_1 \in \mathbb{R}^{p_1\times M}$。

第3步：使用LLDE寻找第二次投影方向 $V_2 \in \mathbb{R}^{p_1\times p_2}$。将第2步投影到新空间的数据 Y_1 投影到新的低维空间，形成新的可分性好的特征数据 $Y_2 = V_2^{\mathrm{T}} \times Y_1$，$Y_2 \in \mathbb{R}^{p_2\times M}$。

第4步：对预处理过的未知类别的 N 个测试数据矩阵 T 第一次投影到低维子空间，得到测试数据 $T_1 = V_1^{\mathrm{T}} \times T, T_1 \in \mathbb{R}^{p_1\times N}$。

第5步：对经过第一次投影的数据 T_1 二次投影，映射到新的低维子空间，得到投影后的测试数据，$T_2 = V_2^{\mathrm{T}} \times T_1, T_2 \in \mathbb{R}^{p_2\times N}$。

第6步：用分类器测试 T_2 在训练样本 Y_2 中的类别。

2.4 改进的二次投影（MDTP）方法

LLDE算法在构建重建权值时，在最小构建误的约束下，仅考虑了把重建点到邻居的距离作为构建重建权值的依据，没有利用重建点邻居的已知类信息。一个点的邻居距离这个点越近，这个邻居的重建权值越大，对构建个这点所起的作用也越大，一个点的邻居越远，则相反。在已知邻居类信息的情况下，如果增大这个点的同类邻居的权值，相当于拉近同类邻居点的距离，相反，减少这个点的异类点的重建权值，相当于把异类点推向更远的距离。对于分类问题，在重建空间中，不同类别更容易区分。鉴于此，本文提出了一个利用样本已有类信息改变重建权值的简单算法，通过设置式（12）两个参数 θ_1 与 θ_2 分别用来改变同类与异类邻居的权重。某个点 X_i 有 K 个邻居，它的权值有 K 个分量即 $W_{ij}, j = 1 \sim K$，每个分量与一个邻居对应。

$$\theta_2 = 1 - 0.1 \times \theta_1, 1 \leqslant \theta_1 < 10 \quad (12)$$

每个样本点的同类邻居权重用参数 θ_1 调整，异类用参数 θ_2 调整。改变后的权重再重新计算每个权值的比例，满足式（2）的约束条件。算法步骤如下：

第1步：取出一个样本点 X_i 的类别 L_i 及其所有邻居的类别信息 L。

第2步：判断 X_i 的类别 L_i 与一个邻居类别 L_j 是否相同。如果相同，执行 $W_{ij} = W_{ij} \times \theta_1$，否则执行 $W_{ij} = W_{ij} \times \theta_2$。

第3步：对 X_i 的所有邻居权重执行第2步。

第4步：计算 X_i 的所有权重之和，$S_i = \sum_{j=1}^{K} W_j$。

第5步：重新计算这个样本点的所有邻居权重，$W_{ij} = W_{ij}/S_i$。

第6步：对所有样本点重复1-5步。

3 实验结果及分析

3.1 数据及性能指标说明

本文实验数据采用了卵巢癌质谱数据（数据来源：National Ovarian Cancer Early Detection Program clinic at Northwestern University Hospital, www.home.ccr.cancer.gov），数据分为两类，健康人的卵巢蛋白质谱数据与患卵巢癌病人的蛋白质谱数据。本试验有253个样本，其中162个为正常的样本，91个为癌变组织的样本。

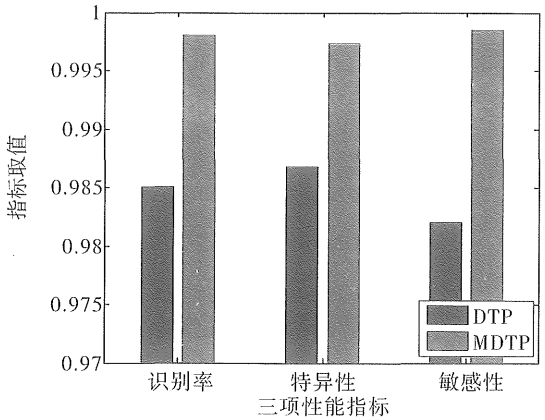

图1 DTP与MDTP性能对比

Fig. 1 Accuracy comparison of recognition rates using DTP and MDTP

本文采用了识别率（Accuracy）、灵敏性（Sensitivity）与特异性（Specificity）三个指标来衡量算法的性能，灵敏性定义为TP/（TP+FN），特异性定义为（TN/TN+FP），识别率定义为（TP

+TN）/（TN + FP + TP + FN），其中 TP、TN、FP 与 FN 分别代表正确识别出癌变组织样本的数量、正确识别出正常组织样本的数量、没有正确识别出癌变组织样本的数量与没有正确识别出正常组织样本的数量[8]。

3.2　参数设置

本文实验预处理后的数据降至200 维，PCA 找出的第一次投影的子空间的维度为 50，第二次投影子空间的维度为 1，构建重建权值时邻居个数为 5，缩放系数 θ_1 为 4。所有数据包括测试数据都进行了预处理。对预处理后的 253 个样本，采用 5 - fold 交叉验证方法，提取出 20% 数据作测试，80% 数据作训练，循环 5 次，一共做 30 组，各项性能指标取的是 30 组的平均值，最后选用 matlab7. 4 提供的 K 近邻分类器（Knnclassify）分类，近邻个数为 1。本实验软件平台是 MatLab 7. 4。

3.3　各算法性能指标对比实验

表 1 所列是各种方法测试蛋白质谱数据的识别性能。其中，1 -7 行是 Levner[8] 在相同卵巢癌数据集相同参数设置下的试验结果，8、9 行是 DTP 方法与 MDTP 方法的识别性能。从表 1 中可以看出，DTP 与 MDTP 两种二次投影方法各项性能指标要明显优于其它方法的性能。

表 1　各种方法识别性能

Table 1　Recognition performance using several methods

序号	算法[1)]	识别率（Accuracy）	特异性（Specificity）	灵敏性（Sensitivity）
1	No FE	0. 773	0. 828	0. 717
2	PCA	0. 773	0. 687	0. 677
3	PCA + LDA	0. 899	0. 889	0. 909
4	SFS	0. 949	0. 980	0. 919
5	SBS	0. 854	0. 929	0. 778
6	P-test	0. 944	0. 970	0. 919
7	Boosted	0. 965	1. 00	0. 929
8	DTP	0. 985 1	0. 986 9	0. 982 1
9	MDTP	0. 998 1	0. 998 5	0. 997 4

1）No FE：非特征提取，PCA：主分量分析，PCA/LDA：主分量分析与线性辨别分析相结合的方法，SBS：顺序后退法，SFS：顺序前进法，P-test：P 检验法，Boosted：基于 Boosted 决策树，DTP：二次投影法，MDTP：改进的二次投影法。

另外，在参数设置相同的情况下，DTP 与 MDTP 方法进行了对比试验，MDTP 的缩放系数 $\theta_1 = 4$。从图 1 中可以看出，MDTP 方法的各项性能要明显高于 DTP 方法的性能。

3.4　主分量个数对比实验

PCA 是在均方误最小约束下寻找最优投影子空间的一种方法，它的误差主要取决于选取协方差矩阵的特征值个数，也就是投影到新的子空间的维度。从理论上讲，投影后子空间的维度越高，保存原数据的能量也就越多。对于蛋白质谱数据，从上万维降低到几百维，数据的能量保存是否合理，判断的依据应该是数据的识别率高低。由于无法自动生成主分量个数，在本算法后续步骤对数据约束条件下，我们进行了对比试验。后续步骤中要求我们提供的数据维度不小于用于重建某点数据的邻居个数 K（本实验取 5），因此，我们选取子空间维度最小不低于 K。而后续步骤中为求邻居点要计算每个数据点的之间的距离，如果维度过高，计算代价过大，同时，为了求第二次投影后子空间的维度，需要对权重矩阵求逆，矩阵过大时同样计算代价价过大。所以根据这些限制，我们对原数据第一次投影后子空间的维度确定为不大于 130。主分量取不同个数相应识别率的对比见图 2，横坐标是保留的主分量个数，从 5 到 130（投影后子空间的维度），纵坐标是数据的识别率。从图中可以看出，主分量个数在 5 时数据的识别率最低，为 0. 848，随着主分量个数的增加，识别率急剧升高。主分量个数 40 到 90 之间达到最高值，识别率有 0. 001 左右的波动。这充分说明了，主分量个数太少时，数据的能量保留较少，识别率较低，而主分量个数太多时，对于识别率的提高没有太大的作用，反而会增加计算代价。结论表明，我们实验中主分量个数取 50 个左右可以取得较优的实验结果。

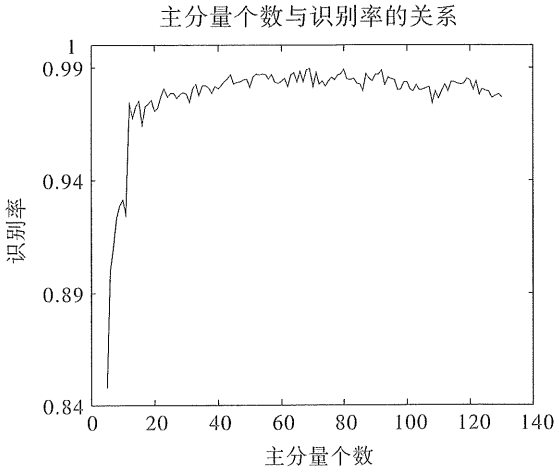

图 2　主分量个数对比

Fig. 2　Recognition accuracy vs. the number of principal component

3.5 重建权值邻居个数的对比实验

寻找第二次投影向量时，在最小构建误的约束下，需要构建每个点的重建权值，而权值是由某点与近邻点的距离决定的，近邻个数作为一个参数如何设置，我们做了对比试验，如图3所示。从图中可以看出，随着近邻个数的增加，数据的识别率从最初的0.977 9，缓慢上升，当近邻点的个数增加到12时，数据的识别率达到最高值1，之后随着近邻点数的增加，识别率的值稳定在1。这说明增加近邻点的个数能有效减少构建误，使构建误达到最小值，从而使测试点在低维子空间中能正确地分类。但邻居点个数过多会增加计算代价，同时，为了对蛋白质谱数据识别效果的有所保留，本实验中采用的近邻个数为5。

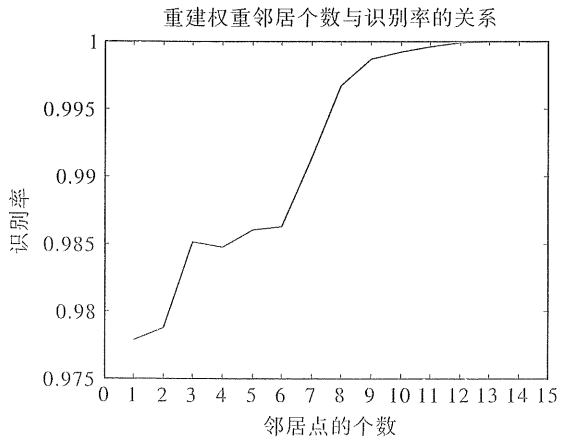

图3 邻居个数性能对比

Fig. 3 Recognition accuracy vs. number of nearest neighbour point

3.6 第二次投影子空间维数对比实验

第二次投影子空间的维数与蛋白质谱数据识别率的关系示于图4中。从图中可以看出随着投影子空间维数的增加，数据的识别率成下降趋势。将数据投影到1维子空间时，识别率最高。这也充分说明了，第二次投影与第一次投影的不同，第二次投影是寻找可分性最好的投影子空间，随着子空间维度的增加，数据的可分性变得越来起差。

3.7 改进的二次投影缩放系数设置实验

为了进一步提高二次投影的效果，我们设置了一组缩放系数，图5中的缩放系数 θ_1 用于同类数据样本的权值修改，异类样本的 θ_2 可以根据式（12）计算。从图中可以看出，随着缩放系数增大，识别率明显有所提高，特别是缩放系数在1到1.5时最为明显。但当缩放系数达到5以后时，识别率的增加趋于平缓。这充分说明了在缩放系统设置合理时，本文提出的MDTP方法的识别性能要明显高于DTP方法。

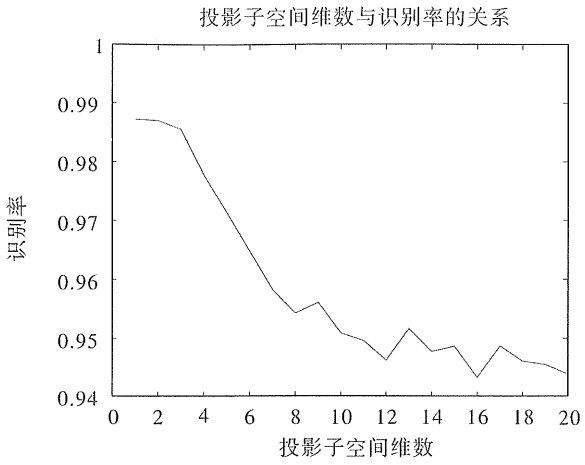

图4 投影维数性能对比

Fig. 4 Recognition accuracy vs. dimensionality reduction

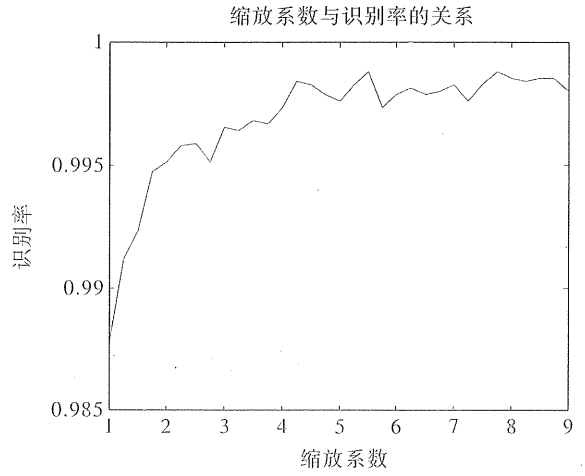

图5 不同缩放系数 性能对比

Fig. 5 Recognition accuracy vs. scaling coefficient θ_1

4 总 结

根据PCA与LLDE在人脸识别方面取得的较好效果，本文提出了二次投影法，并将其应用到卵巢癌蛋白质谱数据识别当中，该方法先后提取了两种不同目的的投影向量，将测试数据对其先后投影到较低的子空间再分类。文中还对PCA投影后保留的主分量个数、低维嵌入时近邻的个数、低维嵌入子空间的维度以及利用类信息的权值缩放系数等参数做了对比实验，得出了最好识别性能的各参数值。实验结果表明，本文提出的DTP及MDTP方法在识别性能方面取得了较好的效果。然而，这种方法应用在其它蛋白质谱数据或者基因数据上是否能取得较好的效果，值得我们做进一步研究。

（下转第37页）

2007, 57: 1091 - 1096.

[7] REIMANN M S , MANNINEN M . Electronic structure of quantum dots[J]. Rev. Mod. Phys. , 2002, 74 :1283 - 1342.

[8] REIMANN M S, KOSKINEN M, KOLEHMAINEN J, et al. Electronic and magnetic structure of artificial atoms [J]. Eur. Phys. J. D, 1999 ,9:105 - 110.

[9] BAO C G. The symmetry backgyound underlying the ring structures fo quantum dots and a classification scheme [J]. J. Phys. cond. metter, 2002,14: 8549 - 8561.

[10] ZHANG X W, XIA J B. Effects of magnetic field on the electronic structure of wurtzite quantum dots: Calculations using effective-mass envelope function theory[J]. Phys. Rev. B, 2005,72: 075363.

[11] CHUTIA S, BHATTACHARJEE A K. Electronic structure of Mn-doped III-V semiconductor quantum dots [J]. Phys. Rev. B, 2008 ,78: 195311.

[12] HUANG G M ,LIU Y M , BAO C G. Electronic structures of donor states under a strong magnetic field [J]. Phys. Rev. B, 2005,71:075302.

[13] HUANG G M ,LIU Y M , BAO C G. Symmetry constraints and the electronic structures of a quantum dot with thirteen electrons[J]. Phys. Rev. B, 2003,68: 165334.

(上接第 32 页)

参考文献：

[1] YU J S, ONAGDLO S, FIEDLER R. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data [J]. Bioinformatics, 2005, 21(10):2200.

[2] WANG Z F, CHANG Y C. A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve [J]. Bioinformatics,2007, 23 (20): 2788 - 2794.

[3] BOCKER S, MAKINEN V. Combinatorial approaches for mass spectra recalibration [C]. IEEE/ACM Transactions on computational Biology and Bioinformatics, Jan. -Mar. , 2008.

[4] LI X, SHU L. Kernel-based non-linear dimensionality reduction and classification for genomic microarray [J]. Expert Systems with Applications, doi:10.1016,2008.9: 70.

[5] ROBERT T, TREVOR H, BALASUBRAMANIAN N, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts' [J]. Bioinformatics, 2004, 20(17):3034 - 3044.

[6] TANG H, MUKOMEL Y, FINK E. Diagnosis of ovarian cancer based on mass spectra of blood samples [C]// 2004 IEEE International Conference on Systems, Man and Cybernetics,2004.

[7] ALEXE G, ALEXE S, LIOTTA L A, et al. Ovarian cancer detection by logical analysis of proteomic data [J]. Proteomics, 2004, 4(3):766.

[8] LEVNER I. Feature selection and nearest centroid classification for protein mass spectrometry [J]. BMC Bioinformatics, 2005, 6:68.

[9] KIRBY M. Geometric data analysis: an empirical approach to dimensionality reduction and the study of patterns [P]. John Wiley & Sons, New York, 2001.

[10] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000,290: 2323 - 2326.

[11] LI B, ZHENG C H, HUANG D S. Locally linear discriminant embedding: An efficient method for face recognition [J]. Pattern Recognition,2008,41: 3813 - 3821.

[12] LI H, JIANG T, ZHANG K. Efficient and robust feature extraction by maximum margin criterion [J]. IEEE Trans Neural Networks, 2006, 17 (1): 157 - 165.