

融合 LLE 和 ISOMAP 的非线性降维方法^{*}

张少龙¹, 巩知乐¹, 廖海斌^{2†}

(1. 平顶山工业职业技术学院 计算机与软件工程学院, 河南 平顶山 467000; 2. 湖北科技学院 计算机科技与
技术学院, 湖北 咸宁 437100)

摘要: 局部线性嵌入(LLE)和等距映射(ISOMAP)在降维过程中都只单一地保留数据集的某一种特性结构, 从而使降维后的数据集往往存在顾此失彼的情况。针对这种情况, 借助流形学习的核框架, 提出融合 LLE 和 ISOMAP 的非线性降维方法。新的融合方法使降维后的数据集既保持着数据点间的局部邻域关系, 也保持着数据点间的全局距离关系。在仿真数据集和实际数据集上的实验结果证实了该方法的优越性。

关键词: 人脸识别; 流形学习; 数据降维; 全局距离保持; 局部结构保持

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1001-3695(2014)01-0277-04

doi:10.3969/j.issn.1001-3695.2014.01.065

Nonlinear dimensionality reduction method by fusing LLE and ISOMAP

ZHANG Shao-long¹, GONG Zhi-le¹, LIAO Hai-bin^{2†}

(1. School of Computer & Software Engineering, Pingdingshan Institute of Industry Technology, Pingdingshan Henan 467000, China;
2. School of Computer Science & Technology, Hubei University of Science Technology, Xianning Hubei 437100, China)

Abstract: LLE(local linear embedding) and ISOMAP(Isometric map) only preserved one specific feature of the data sets during the dimensionality reduction process, and ignored other meaningful features. So the features of the original data sets could not be preserved as well as possible after dimensionality reduction. This paper proposed a new method that could better solve this problem. This method could preserve both the neighborhood relationships and the global pairwise distances of the high-dimensional data sets. Experiments on both artificial and real data sets prove effectiveness of the proposed method.

Key words: face recognition; manifold learning; data dimensionality reduction; global distances preservation; local structures preservation

0 引言

寻求高维数据的低维表示一直是模式识别领域的一项重要任务。国内外学者们在近几十年的研究中提出了许多方法, 其中, 最为经典且成功应用在模式识别领域的方法是各类子空间分析法(subspace analysis method, SAM)^[1], 如主成分分析(PCA)^[2]和 Fisher 判别分析等^[3]。但是, 这类方法主要用于处理线性数据集, 当面对非线性数据集时显得无能为力。2000 年之后, 一批能够处理非线性数据集的流形学习算法被相继提出。其中, 代表性算法包括局部线性嵌入(LLE)^[4,5]和等距映射(ISOMAP)^[6]等。与线性降维方法不同, 这类方法不依赖于对原始数据的线性假设, 而是致力于保持原数据集某一方面的特性在降维之后保持不变。比如, LLE 方法就能保持原数据集中各点的邻域关系不变, 而 ISOMAP 方法则能最大程度地保持原数据集中各点间的测地距离。LLE 和 ISOMAP 有许多共同的优点, 如算法思路清晰、步骤简洁、能收敛于全局最优解等。但是, 它们也都有各自的局限, 其中一个比较明显的不足就是它们都只能针对数据集的某一种特性进行保持, 而无法兼顾其

他同样有价值的特性。图 1 给出了一个具体的例子, 从图中可以看出, 对 Swiss roll 数据集(图 1(a)), LLE 的低维输出很好地保持了数据的邻域结构(即高维数据中相邻点的相互关系在低维表示中仍然得到了保持), 但是其全局各点间的相互距离并没有得到很好的保持, 如图 1(b)所示。ISOMAP 算法的低维输出结果如图 1(c)所示, 其中低维表示中各点的欧氏距离很好地逼近了高维空间中各点的测地距离, 但是它的局部邻域结构显然又与原始数据相差甚远。如果算法的目标就是单一地保持数据集某一个性质, 那么这样的输出结果当然是令人满意的。但是, 如果对于要保持哪一种性质没有先验的认识, 那这样的结果就不足以反映高维数据集的全貌。

那么, 有没有一种方法使数据降维后既能保持数据间的局部邻域关系, 同时又能保持其全局距离关系呢? 2006 年 Varini 等人^[7]曾提出了一种名为 ISOLLE 的方法, 试图达成这一目标。在此方法中, 作者针对 LLE 算法在邻域选择的步骤中有可能产生短路问题这一缺陷, 提出可以使用测地距离替代原来的欧氏距离来进行邻域选择的步骤, 从而有效解决短路问题, 并使得选择出的邻域点能更好地反映高维数据集的局部结构信息, 进而提升 LLE 算法的性能。此方法确实一定程度

收稿日期: 2013-05-04; **修回日期:** 2013-06-14 **基金项目:** 中央高校基本科研业务费专项资金资助项目(20102120103000004); 河南省重大科技攻关项目(072SGZS38042)

作者简介: 张少龙(1972-), 男, 吉林梅河口人, 高级程序员, 讲师, 硕士, 主要研究方向为软件开发、图像处理与模式识别; 巩知乐(1981-), 女, 河南平顶山人, 讲师, 硕士, 主要研究方向为人工智能等; 廖海斌(1982-), 男(通信作者), 湖北咸宁人, 讲师, 博士, 主要研究方向为图像处理与模式识别(liao_haibing@163.com)。

上整合了 LLE 和 ISOMAP 两种算法的优点,在降维过程中同时利用了局部结构信息和距离信息。但是,此方法只是对算法的一个中间步骤进行了优化,意图在原有算法的基础上提升性能,并没有对算法的核心目标进行改进,所以该方法并不能在最终的降维结果中体现对邻域关系和全局距离的同时保持。

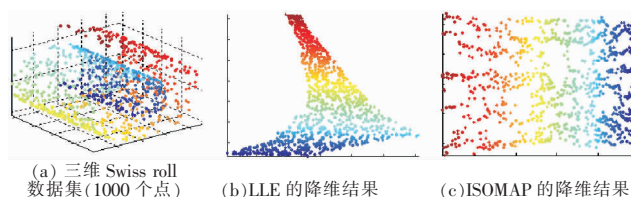


图 1 使用 LLE 和 ISOMAP 方法对同一个数据集进行降维的结果

目前的流形学习算法大致可分为邻域保持 (LLE)、拉普拉斯特征映射 (LEM)^[8]、全局距离保持 (ISOMAP)、扩散距离保持 (扩散映射 (DFM)^[9]、切空间保持 (局部切空间保持 (LTSA)^[10]) 等。这些不同的算法都能很好地保持数据集某一方面的性质。但是,由于高维数据的复杂性,在降维过程中仅从某个单一的角度来保持高维数据的特征难免会损失其他信息。这时,如果能够将各种对数据集的不同考量有机地融合到同一个方法中,使得新方法能尽可能多地保持原数据集的不同性质,就会使降维结果与原始高维数据尽可能地贴近,从而获得令人满意的降维效果。本文的工作是上述思路的一种有效尝试。

由于 LLE 和 ISOMAP 方法都能统一在核方法的框架下,即它们的区别可唯一地归结为使用不同的核。因此,将这两种方法的核进行合理的融合,形成的一种新方法就能达成同时保持邻域关系和全局距离的降维目标。这也正是本文所做的工作,其意义在于为实现多特征保持的降维算法提供一种新思路。

1 LLE 和 ISOMAP 算法介绍

1.1 局部线性嵌入算法 (LLE)

2000 年 Saul 等人第一次提出了局部线性嵌入算法 (LLE)。它是一种非监督的学习算法,能够计算高维数据的低维嵌入坐标。这种嵌入是通过高维数据中的局部邻域进行线性重构而获得的,它能够保持高维数据的邻域性质。嵌入结果的计算最终归结于一个求解稀疏矩阵特征值的问题。算法的具体过程可描述如下:

a) 为每个数据点确定好邻域。确定邻域的方法可以用 k -邻域法或者 ε -半径法。

b) 确定每个点由其邻域进行重构的线性系数。这是通过求解使得代价函数 $\varepsilon = \|X_i - \sum_j w_{ij} X_j\|^2$ 取得最小值的 w_{ij} 来实现的。这里的 w_{ij} 表示由第 j 个点重构第 i 个点的系数,所有的 w_{ij} 就构成了权重矩阵 W 。

c) 寻找能最优地保持权重矩阵 W 的低维嵌入坐标 Y 。代价函数为

$$\Phi(Y) = \sum_i \|Y_i - \sum_j w_{ij} Y_j\|^2 \quad (1)$$

使得此代价函数取得最小值的 Y 就是要寻找的低维嵌入坐标。式 (1) 可改写为: $\Phi(Y) = \sum_{ij} M_{ij} (Y_i Y_j)$ 。其中,矩阵 M 定义如下: $M = (I - W)^T (I - W)$ 。 M 是一个对称的、半正定的矩阵。又为了使最终结果唯一化,还需要去除 Y 中的平移、旋转

及缩放自由度。这是通过添加两项约束来实现的:

$$\sum_i Y_i = 0 \quad (2)$$

$$\frac{1}{N} \sum_i Y_i \otimes Y_i = I \quad (3)$$

满足这两个约束的低维坐标 Y 就是中心化的,且具有单位协方差的降维结果。

最终,对 Y 的求解是通过矩阵 M 进行特征分解来获得的。 M 的第 2 到第 $(d+1)$ 小的特征值所对应的特征向量就构成了 d 维嵌入坐标 Y 。 Y 既能使得式 (1) 的代价函数取得最小值,同时又满足上述两个约束条件,从而形成 LLE 算法的最终输出。

1.2 等距映射算法 (ISOMAP)

与 LLE 齐名的流形学习代表性算法还有 Tenenbaum 等人提出的等距映射算法。与 LLE 算法致力于保持数据集邻域特性不同,ISOMAP 算法的目标是保持所有数据点间的相互测地距离。其算法的主要过程如下:

a) 确定邻域。第一步与 LLE 算法相似,都是为每个点确定一个邻域,不同的是在 ISOMAP 中不仅要确定邻域,还要计算每个点和它所有邻域点间的距离。

b) 计算图距离来逼近测地距离。有了第一步得到的邻域距离之后,再计算所有点间的图距离 (即两点间的最短路径),将所得结果写成矩阵的形式,记为 D 。可以证明,当数据点的个数趋向于无穷的时候,各点间的图距离也将趋向于真实的测地距离。所以这里就用计算得到的最短路径作为测地距离的替代。

c) 利用古典 MDS 方法求得数据的低维表示。这主要是通过使如下代价函数取得最小值来实现的:

$$\varepsilon = \sqrt{\sum_{ij} (\tau(D) - Y_i \cdot Y_j)^2} \quad (4)$$

其中: $\tau(D) = -HS/2$, S 是将 D 中各元素平方后得到的矩阵, $H = I - ee^T/n$, $e = [1, 1, \dots, 1]^T$ 。其约束条件为

$$Y^T Y = \Lambda \quad (5)$$

其中: Λ 为对角线元为常数的对角矩阵。

最终,对 Y 的求解是通过矩阵 $\tau(D)$ 作特征分解来实现的。记矩阵 $\tau(D)$ 的特征值为 $s_1, s_2, \dots$ (按降序排列), $v_1, v_2, \dots$ 为与之相对应的特征向量 (v_1, v_2 为列向量)。那么矩阵 $Y = [\sqrt{s_1} v_1, \sqrt{s_2} v_2, \dots, \sqrt{s_d} v_d]$ 就构成了作为算法输出的 d 维嵌入坐标。

2 融合 LLE 和 ISOMAP 的降维方法

由第 1 章的介绍可以看出,LLE 和 ISOMAP 方法虽然目标不同,但是算法的过程却有很多相似之处。比如都要对数据点的邻域进行分析、都将从数据集中提取出的某一类信息以矩阵形式存储起来、算法输出的低维坐标都是通过对之前得到的矩阵进行特征分解而获得等。其中最关键的一点就是这两种算法都归结为针对某一特定矩阵求解特征向量的问题,而这些矩阵就代表了算法想要保持的数据集特征,而且它们往往都是正定的。一个最直观的融合想法就是能不能将 LLE 和 ISOMAP 的特定矩阵相加,这样便能融合两种方法的特性。

2.1 融合可行性分析

Ham 等人^[11]曾对这一问题进行过专门的阐述。他们认为

对于 LLE 和 ISOMAP 这两种方法,最后都归结为对一个矩阵进行特征分解,而这个矩阵就是核方法中的核矩阵。比如对于 LLE 方法,它的核矩阵就可以写做 $K = (\lambda_{\max} I - M)$ 。其中 M 如 1.1 节中定义, $\lambda_{\max}$ 是 M 的最大特征值。原来的代价函数式 (1) 就可以改写成:

$$\Phi(Y) = \text{trace}(Y^T KY) \quad (6)$$

这样,原来 LLE 方法的目标是要寻找 M 的最小特征值对应的特征向量。这时就变为了寻找 K 的最大特征值对应的特征向量。由于 M 是正定的,而 $\lambda_{\max}$ 是其最大特征值,可以证明, K 也是正定的,因而 K 满足核矩阵的条件。

再对于 ISOMAP 方法,它的核矩阵可以写做 $T = -HSH/2$ 。这里的 H, S 如 1.2 节中定义。算法的目标就变成使如下的代价函数取得最大值:

$$\Phi(Y) = \text{trace}(Y^T TY) \quad (7)$$

在文献[11]中作者提到,如果 T 中元素是各点欧氏距离的平方,那 T 将是正定矩阵。由于 T 实际上是测地距离的平方,所以 T 并不能保证是正定的。但是,在满足连续性条件的光滑流形上,数据点间的测地距离是与它们在低维空间中的欧氏距离成比例的。所以,在连续性条件得到满足时, T 也是正定的,因而也满足核矩阵的条件。

可见, LLE 和 ISOMAP 都可以被统一在核方法的框架下,它们的区别主要体现为核矩阵的区别,而其他步骤则基本是相似的。因而一个能真实反映数据集某一方面性质的核矩阵是一个算法能够取得成功的基础。根据核矩阵的性质显然有:若 K_1, K_2 是核矩阵,则 $K_1 + K_2$ 是核矩阵, $\alpha K_1 (\alpha \geq 0)$ 也是核矩阵。

因此,一种最简单的能同时保持数据集局部和全局特性的特征矩阵就可写做 $K + T$ 。

2.2 融合算法

由前面的分析可知,在核方法的框架下,一个算法的核矩阵就集中体现了该算法想要保持的数据集的性质。因此,本文对核矩阵进行重新设计,使其同时保持数据集的多种性质。最终算法的目标就是要使得如下的代价函数取得最大值:

$$\Phi(Y) = \text{trace}(Y^T (K + T) Y) \quad (8)$$

即,使得该代价函数取得最大值的低维坐标 Y 就是能同时保持原数据集全局距离及邻域结构的低维表示。

为了去除低维输出 Y 中的平移、旋转及缩放自由度,这里对于 Y 仍然有式 (2) 和 (5) 描述的两个限制条件: $\sum_i Y_i = 0, Y^T Y = \Lambda$ 。之所以采用式 (5) 而不用式 (3) 是因为前者是一个更为一般的限制条件。

这样,对新的核矩阵进行特征分解,得到特征值 $s_1, s_2, \dots, s_n$ (按降序排列) 及对应的特征向量 $v_1, v_2, \dots, v_n$, 那么矩阵 $Y = [\sqrt{s_1}v_1, \sqrt{s_2}v_2, \dots, \sqrt{s_d}v_d]$ 就是使得式 (8) 取得最大值的 Y , 也是整个算法的输出结果。这里,之所以要让每个特征向量乘以对应特征值的 $1/2$ 次方是为了满足第二个限制条件。由于 K 和 T 都是正定的,可知 $K + T$ 也是正定的,所以它必有一个 0 特征值(对应的特征向量所有元素相等),所以第一个限制条件也得到了满足。可以看出,这里得到的数据集低维嵌入坐标 Y 就是能同时反映原始高维数据邻域结构和全局距离的最优低维表示。

由于 K 和 T 分别为两个不同参考系下的量,在数值上存

在很大的差距。因此,需要在这个核中引入一个平衡参数,以灵活调整这两种考量标准在整个算法中的平衡关系。令引入的平衡参数为 α , 则新的核矩阵可表示为 $(1 - \alpha)K + \alpha T$, 相应的代价函数即为

$$\Phi(Y) = \text{trace}(Y^T ((1 - \alpha)K + \alpha T) Y) \quad (9)$$

相应的算法输出即是对新核矩阵 $(1 - \alpha)K + \alpha T$ 进行特征分解后得到的结果。融合算法具体步骤如下:

- 确定邻域。用 k -近邻法或者 ε -半径法确定每个数据点的邻域。
- 获得保持近邻关系核 K 。求解使代价函数 $\varepsilon = \|X_i - \sum_j w_{ij} X_j\|^2$ 取得最小值的邻域矩阵 W , 再由 $K = (\lambda_{\max} I - M)$ 计算得到算法中用于保持近邻关系的核 K 。其中 $M = (I - W)^T (I - W)$, $\lambda_{\max}$ 是 M 的最大特征值。
- 获得保持全局距离的核 T 。计算所有点间的最短路径距离矩阵 D , 由 $T = -HSH/2$ 计算得到保持全局距离的核 T 。其中, H 为中心化矩阵, S 是取 D 中元素平方的矩阵。
- 确定低维坐标 Y 。对新核 $(1 - \alpha)K + \alpha T$ 进行特征分解, 得到特征值 $s_1, s_2, \dots, s_n$ (按降序排列) 及对应的特征向量 $v_1, v_2, \dots, v_n$, 矩阵 $Y = [\sqrt{s_1}v_1, \sqrt{s_2}v_2, \dots, \sqrt{s_d}v_d]$ 即为算法输出。

融合算法具有如下特性:

- 算法有确定的全局最优解。由于算法的最终输出是由计算核矩阵的特征向量得到的, 因此对于一个给定的数据集, 算法总能给出一个全局最优的低维输出结果。这样就避免了算法陷于局部最小值的危险。
- 算法有确定的时间和空间开销。算法中最消耗时间的步骤是计算所有点对间最短路径的步骤, 其时间复杂度为 $O(N^3)$, 因而整个算法的时间复杂度也为 $O(N^3)$ 。算法的空间开销主要用于存储核矩阵中的元素, 而核矩阵的大小只和元素个数有关, 因而算法的空间复杂度为 $S(N^2)$ 。与一些需要不断迭代运算来计算最优解的算法相比较, 本算法能保证在一定的时间和空间限制下获得最终结果。
- 算法可根据实际需要灵活调整两种保持标准的权衡关系。本算法由一个平衡因子来平衡算法在两种保持间的偏重, 可以根据实际需要灵活决定配比, 可以在不断实验的过程中发现对于特定数据集最适用的 α , 从而使降维效果最好。

3 实验结果

3.1 数据降维结果

首先采用流形学习中非常经典的 Swiss roll 数据集作为实验数据集。将此数据集分别用 LLE、ISOMAP 和 GLPP 降至二维, 得到的结果如图 2 所示。可以看出, 相较于 LLE 算法仅仅保持了近邻点的相互位置关系, 本文算法能更好地保持全局各点的相互位置; 再与 ISOMAP 算法只保持了数据集全局性质相比较, 本文的算法又能兼顾近邻点间的相互位置关系。

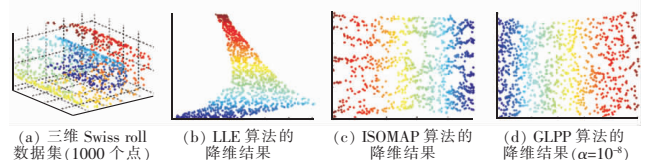


图 2 使用三种方法对 Swiss roll 数据集进行降维的结果

进一步, 本文将 Swiss roll 数据集进行了适当扩展, 使其在原有基础上进行更多次的卷绕, 如图 3(a) 所示。对于这样一个更为复杂的数据集, 如果从单一的角度仅仅保持近邻点间的

相互关系,或者仅保持数据点的全局性质,都无法很好地将数据集的本质特征在二维空间上刻画出来。而本文算法由于在降维的过程中同时考虑了数据集这两方面的性质,因而能在此数据集上获得相对更为优越的降维效果。

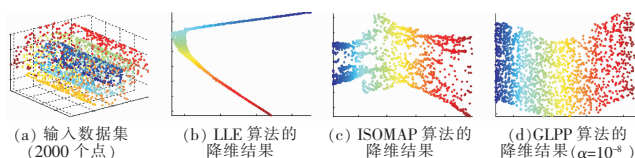


图3 对更复杂的 Swiss roll 数据集进行降维的结果

3.2 在真实数据集上进行降维后的分类结果

前面展示了本文算法在仿真高维数据集上的降维效果,本节将展示算法在两个实际的高维数据集上的分类效果。本节的实验首先使用 ISOMAP、LLE、文献[7]的融合方法以及本文的融合方法四种不同的降维算法对两个实际的高维数据集进行降维;然后对得到的低维结果进行分类识别;最后对各自的识别结果进行比较分析。

第一个数据集由笔者实验室创建,包括 100 个志愿者的 1 200 幅彩色人脸图像。每个人包括 7 幅不同姿态变化的人脸,2 幅不同表情变化的人脸和 3 幅有装饰的人脸,其部分尺度归一化后的人脸图像如图 4 所示。从图中可以看出,此人脸库包括巨大的姿态、装饰变化,导致人脸图像之间呈现出非线性变化。每个人取其中 7 幅作为训练库,剩下的 5 幅作为测试图像。图 5 显示了不同维数特征下的人脸识别率情况。可以看出,本文算法在将数据降至不同的维度时所获得的平均识别率要优于其他几种算法。但是随着目标维数 d 的不断增大,本文方法的识别率会低于 ISOMAP,这是因为 ISOMAP 是一种以严格保持原始数据点间相互距离为目标的降维算法,因而随着 d 的增大,原始数据点间的相互距离能够在低维空间中得到更好的保留,而实验中采用的分类器又是基于距离的,因而由 ISOMAP 的识别率会随 d 的增大而提高。当数据降至二维和三维这两个具有可视化效果的维度再进行分类时,本文算法还是体现出了较为明显的优势。

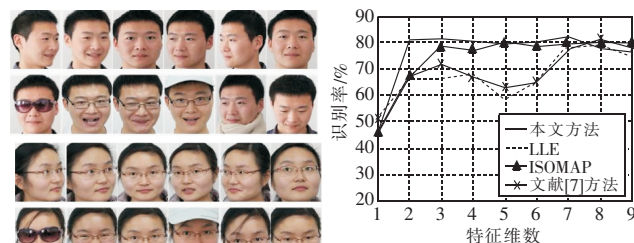


图4 人脸图像数据集中部分图像

图5 将人脸数据集降至不同维度的识别率

第二个数据集是 Yale 人脸库包括 15 个人的 165 幅灰度人脸图像^[2]。每个人由 11 幅照片构成,每人的每幅图像均是从不同视角获得的,有较大的表情变化、光照变化以及部分缺损。部分图像如图 6 所示。采用不同方法将此数据集都降至三维,再对其进行分类识别,不同方法随着训练样本数变化的识别结果如图 7 所示。从图中可以得出,本文算法的分类效果较之于其他算法都具有更好的表现。

4 结束语

基于流形学习算法的核化理论,本文将两种已有流形学习

算法 LLE 和 ISOMAP 的核以一定的方式结合在一起,形成了一种能有效保持数据集两方面性质的新算法。实验结果证实了该方法能成功地保持数据集的邻域结构和全局相互距离。由于本文的工作主要是将已有算法的核加以整合,得到的新核的效果比两者都要好,这也给了一个启示:如果能在同一个算法中同时保持高维数据集的多种不同方面的性质,那么所得到的低维嵌入结果将能更好、更忠实地反映原始数据集的真实性质。而新近提出的核方法框架刚好为这种想法提供了很好的平台,即只要设计出新的能满足需要的核即可。如何设计出或者合理利用现有的核,使其能更好、更多地保持高维数据集各种不同类型的性质,使得高维数据的低维表示尽可能与原始数据贴近,将是下一步要开展的工作。



图6 Yale 人脸库

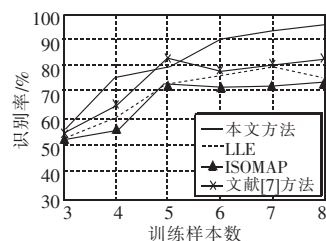


图7 在特征维数为3时不同方法的比较结果

参考文献:

- [1] YANG Jian-chao, YAN Shui-cheng, HUANG T S. Ubiquitously supervised subspace learning [J]. *IEEE Trans on Image Processing*, 2009, 18(2): 241-249.
- [2] BELHUMEUR P N, HESPANHA J P, KRIENGMAN D J. Eigenfaces versus fisherfaces: recognition using class specific linear projection [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 1997, 19(7): 711-720.
- [3] LU Gui-fu, ZOU Jian, WANG Yong. Incremental learning of complete linear discriminant analysis for face recognition [J]. *Knowledge-Based Systems*, 2012, 31: 19-27.
- [4] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding [J]. *Science*, 2000, 290(5500): 2323-2326.
- [5] SAUL L K, ROWEIS S T. Think globally, fit locally: unsupervised learning of low dimensional manifolds [J]. *Journal of Machine Learning Research*, 2003, 4(12): 119-155.
- [6] TENENBAUM J B, De SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. *Science*, 2000, 290(5500): 2319-2323.
- [7] VARINI C, DEGENHARD A, NATTKEMPER T W. ISOLLE: LLE with geodesic distance [J]. *Neurocomputing*, 2006, 69(13-15): 1768-1771.
- [8] JULIANA V A, ANDRES A M, GENARO D S. Multiple manifold learning by nonlinear dimensionality reduction [C]//Proc of the 16th Ibero-American Congress on Pattern Recognition. 2011: 15-18.
- [9] ZHENG Ya-li, FANG Bin, TANG Yuan-yan. Learning orthogonal projections for Isomap [J]. *Neurocomputing*, 2012, 103: 149-154.
- [10] LEI Ying-ke, XU Yang-ming, YANG Jun-an. Feature extraction using orthogonal discriminant local tangent space alignment [J]. *Pattern Analysis and Applications*, 2012, 15(3): 249-259.
- [11] HAM J, LEE D, MIKA D S, et al. A kernel view of the dimensionality reduction of manifolds [C]//Proc of the 21st International Conference on Machine Learning. 2004.