文章编号: 1001-0920 (2011) 06-0826-05

一种基于ISOMAP的分类算法

程起才[1], 王洪元[1], 吴小俊[2], 刘锁兰[1]

(1. 常州大学 信息学院，江苏 常州 213164；2. 江南大学 物联网学院，江苏 无锡 214122)

摘 要: 提出一种解决分类任务的等测距映射算法, 该算法利用类标签信息指导高维数据的降维. 首先根据类标签在属于某个类的数据集上构造类内邻域图; 然后寻找类间最短距离相邻边, 并将其乘以大于1的尺度变化因子, 使得降维后的类内数据更加紧凑、类间数据更加分开; 最后利用BP神经网络构建一个近似的从原始高维数据集到低维数据集之间的映射函数, 通过遗传算法对BP神经网络的初始权值和阈值进行优化, 以避免使用剃度下降算法所带来的局部最优问题. 实验结果表明, 分类性能有较大提高, 并对噪声有一定的鲁棒性.

关键词: 分类；流形学习；等测距映射；类内邻域图；遗传算法

中图分类号: TP311 **文献标识码:** A

An algorithm for classification based on ISOMAP

CHENG Qi-cai[1], *WANG Hong-yuan*[1], *WU Xiao-jun*[2], *LIU Suo-lan*[1]

(1. School of Information，Changzhou University，Changzhou 213164，China；2. School of Internet of Things Engineering，Jiangnan University，Wuxi 21412，China．Correspondent：WANG Hong-yuan，E-mail：hywang@cczu.edu.cn)

Abstract: An improved isometric feature mapping(ISOMAP)algorithm for classification task, called ISOMAP-C, is proposed, which employs label information to guide the dimensionality reduction for high dimensional datasets. Firstly, within-class neighborhood graphs are constructed over each sub dataset belonging to the same class according to label information. Secondly, the between-class neighborhood edges with the shortest distance are searched for, which is multiplied by scaling factor greater than one so that low dimensional dataset after mapping become more compact within class and more separate between classes. Finally, the mapping function from original high dimensional space to low dimensional space can be approximately modeled by using Back-Propagation neural network, whose initial weights and thresholds are optimized by using genetic algorithms to avoid local minimum using gradient decent techniques. The experimental results show that the performance of classification is greatly enhanced and the alsorithm has robust for noisy data.

Key words: classification；manifold learning；ISOMAP；within-class neighborhood graphs；genetic algorithms

1 引 言

降维方法通常分为线性降维和非线性降维两类. 经典的线性降维方法主要有主分量分析法[1]、独立分量分析法[2]、多尺度变换算法[3]、线性判别分析法[4]等. 但现实中, 很多数据呈非线性结构, 为了能够处理这类数据, 提出了多种非线性降维算法, 如自组织映射算法[5]、主曲线方法[6]、生成拓扑映射方法[7]和基于核主分量分析的方法[8]等. 但这些算法存在训练参数过多、代价函数设计困难、没有考虑数据可能位于内在的低维流形结构特征等缺陷. Seung等人[9]提出了人脑感知可能以流形方式存在, 在此理论基础上, 基于流形假设的非线性降维方法成为机器学习和模式识别等领域的研究热门. 等测距映射(ISOMAP)算法[10]是其中一个经典的流形学习算法, 它以测地线距离代替欧氏距离, 然后用多尺度变换求出低维映射结果. 该算法在数据均匀抽样于一个内在扁平的单个流形时, 低维嵌入结果效果较好, 但缺点是它们均以无监督方式工作, 即不能直接用于分类目的, 且当数据采样于多个分离的类簇时, 在构造每个数据点的邻域时会产生“短路边”, 导致低维嵌入不能正确反映原

收稿日期: 2010-03-08；**修回日期:** 2010-04-27.

基金项目: 国家自然科学基金项目(60973094, 61070121)；江苏省自然科学基金项目(BK2009538)；江苏省高校自然科学基金项目(09KJB520004)；江苏工业学院青年创新基金项目(JQ200806).

作者简介: 程起才(1981−), 男, 讲师, 硕士, 从事机器学习与数据挖掘的研究；王洪元(1960−), 男, 教授, 从事机器学习与模式识别等研究.

始高维空间中数据间的拓扑结构.

本文基于多类多流形算法 (MCMM-ISOMAP)[11] 对传统的 ISOMAP 方法进行改进, 提出一种新的能够完成分类目的的 ISOMAP-C(ISOMAP for classification) 算法. 该算法可以解决经典 ISOMAP 算法不适合数据分布于多个分散类簇的难题, 同时也解决了 KNN 算法在高维空间直接进行分类所带来的维数灾问题. 首先在使用 ISOMAP 算法进行降维时, 利用有标签信息使得类间间隔在低维空间中可以任意放大, 类内间隔可以任意缩小; 然后通过 BP 神经网络显式地近似构造原始高维空间到低位空间的映射函数, 其中 BP 神经网络的初始化权值和阈值通过遗传算法来选取, 以避免初始权值选取不当引起网络振荡而不能收敛, 或陷入局部极小值而影响网络的泛化能力; 最后在低维空间中通过 KNN 分类器对需要预测的数据进行分类.

2　经典 ISOMAP 概述

ISOMAP 算法是一种非线性降维, 它在降维过程中通过计算点对点之间的测地距离, 并采用多尺度变换算法来获取全局最优的几何结构, 获得了较好的效果. 但该算法也存在较多缺陷:

1) ISOMAP 算法以无监督方式工作, 不能直接用于分类目的. 针对该问题, 国内外学者做了相关研究, Yang[12]提出了基于 LDA 的 ISOMAP 算法用于人脸识别并取得了一定效果, 但该方法以到其他点的近似测地线距离作为该点的特征向量, 这样对于一个要分类的数据点, 需要重新计算该距离, 耗时较多; Geng 等人[13]提出了有监督 ISOMAP 算法, 在数据可视化和分类上取得了较好的效果, 但该方法不适合于分散的类簇情况. 本文在 MCMM-ISOMAP 方法的基础上提出适合多个分散类簇情形的数据.

2) ISOMAP 算法只能对一个流形产生较好的低维嵌入结果, 若数据集采样于多个流形, 则该算法可能失效; 如 Wu 等人[14]提出的多类 ISOMAP 算法仅适合于数据可视化, 不能完成分类任务.

3) ISOMAP 算法不适于学习有较大内在曲率的流形, 因为在寻找每个数据点的邻域点时, 若某个点所在的流形位置处曲率较大, 则附近的点之间可能会产生大量“短路”边, 破坏了数据的拓扑结构.

4) ISOMAP 算法不能产生从高维空间到低维特征空间的映射函数.

本文针对上述 4 个问题对经典 ISOMAP 算法进行改进, 提出了一种实现分类目的的算法 ISOMAP-C.

3　ISOMAP-C 算法

就分类器设计而言, 所给的数据集类间间隔越大越好, 所以若存在一个映射函数 $f: R^n \to R^m$ (通常 $m \ll n$), 使得所映射低维特征空间的数据类间间隔可以任意变大, 类内距离可以任意变小, 则在该低维特征空间中进行分类时, 准确率会得到较大提高. 假设一组具有类别标签的高维样本集 $X = \{(x_1, l_1), (x_2, l_2), \cdots, (x_N, l_N)\}$, $x_i \in R^n$, $l_i \in \{1, 2, \cdots, c\}$, 即输入样本属于 c 个类. 为了叙述方便, 引入如下定义.

定义 1　对于给定的有标签数据集, 类标签为 l_i 位于数据集内, 在特定的邻域参数 K 或 ε 下, 构造一个邻域图 NG_i, 称该邻域图为 l_i 类内邻域图.

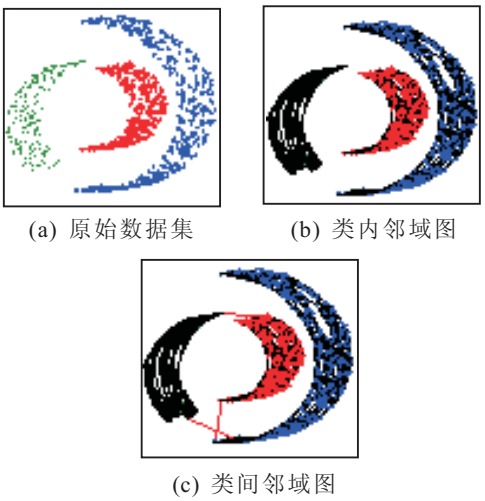
注 1　从定义 1 可以看出, 对于给定的样本集 X, 会产生 c 个分离的邻域图.

注 2　在构造类内邻域图时, 可以选择一个合适的邻域参数, 使每个类内邻域图不产生短路边, 即该邻域图能够近似反映类内数据间的邻域拓扑关系, 因此, ISOMAP-C 算法能够保持类内的内在拓扑结构不变.

定义 2　在第 i 类与第 j 类数据间的欧氏距离中, 将最短的欧氏距离 d_{ij} 所对应的两个点称为第 i 类与第 j 类数据之间的最短距离点, 且其边 e_{ij} 称为类间最短距离相邻边.

定义 3　在 c 个类内邻域图 $\mathrm{NG}_i (i = 1, 2, \cdots, c)$ 中, 将类间最短欧氏距离的两个点用一条边相连, 使得 c 个类内邻域图变成 1 个连图的邻域图 G, 称 G 为类间邻域图.

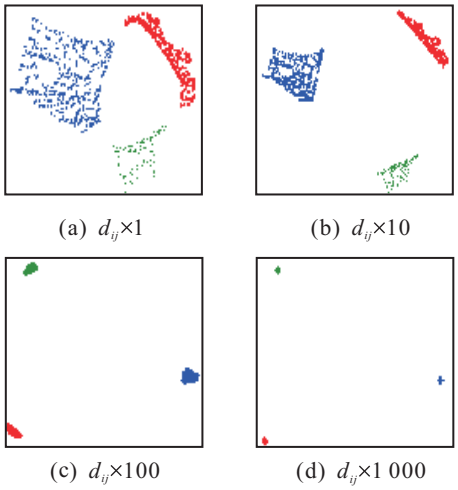
图 1(a) 为原始 Swiss roll 数据集, 图 1(b) 给出了在邻域参数 $K = 8$ 时得到的类内邻域图, 图 1(c) 给出 3 个类间最短距离相邻边连接起来的类间邻域图. 在图 1(c) 的基础上, 求出点对点之间的最短路径, 用该路径的长度代替测地线距离, 然后用多尺度变换算法将数据集映射到二维欧氏空间中, 结果如图 2 所示.



(a) 原始数据集　(b) 类内邻域图

(c) 类间邻域图

图 1　原始数据集及相应邻域图



(a) d_{ij}×1 (b) d_{ij}×10

(c) d_{ij}×100 (d) d_{ij}×1 000

图 2 最短欧氏距离乘以不同系数的低维映射

图 2(a)~图 2(d) 显示了在类间最短欧氏距离乘以不同系数下时所得到的类间分隔程度, 很明显, 系数越大类间间隔越大, 最终使得每个类内的数据缩成一个点.

由于通过 ISOMAP 算法求得的从高维空间到低维空间的映射函数 f 是未知的, 本文利用 BP 神经网络和遗传算法相结合的方法来近似构造映射函数 f. 神经网络以原始数据集 X 作为输入, 以相应的低维嵌入作为输出来构造近似的映射函数 $\bar{f}$, 通过 $\bar{f}$ 将需要分类的未标签样本 x_0 映射到低维空间.

具体的算法步骤如下: 假设原始高维有标签数据集为 $X = \{(x_1, l_1), (x_2, l_2), \cdots, (x_N, l_N)\}$, 按照类别标签将输入样本集进行划分, 得到 c 个子集 $X = X_1 \bigcup X_2 \bigcup \cdots \bigcup X_c$, 很明显 $X_i \bigcap X_j = \varnothing (1 \leqslant i, j \leqslant c, i \neq j)$. 假设第 i 类样本子集为 $X_i = \{x_i^1, x_i^2, \cdots, x_i^{N_i}\}$, 显然有 $\sum_{i=1}^{c} N_i = N$, 则 $x_i^p (1 \leqslant p \leqslant N_i)$ 表示属于第 i 类的第 p 个样本. 相应的低维嵌入输出数据集 $Y = Y_1 \bigcup Y_2 \bigcup \cdots \bigcup Y_c$, 其中 $y_i^p \in Y_i \subset R^m$ 表示 n 维欧氏空间中的样本点 x_i^p 在 m 维欧式空间中的嵌入.

Step 1: 构造邻域图. 选取合适的邻域参数值 K 或 ε, 该值要保证在每个类所对应的流形内部不产生短路边, 且能够求出任意样本点 x_i^p 所对应的邻域点集合 $\mathrm{NE}(x_i^p)$.

Step 2: 求出各类数据集间最短的欧氏距离所对应的数据点, 即下式的解:

$$\mathrm{WM}_i, \mathrm{WM}_j = \arg \min_{w_i \in X_i, w_j \in X_j} \min_{1 \leqslant i, j \leqslant c, i < j} (\|X_i - X_j\|). \tag{1}$$

其中: $\mathrm{WM}_i \in X_i$, $\mathrm{WM}_j \in X_j$.

Step 3: 求出 Step 2 中得到的各对类间最短距离点的欧氏距离 $d_{ij} = \|\mathrm{WM}_i - \mathrm{WM}_j\|$, 且将 d_{ij} 更新为 $d_{ij} = d_{ij} * \gamma$, $\gamma > 1$.

Step 4: 对 Step 2 中得到的样本点 WM_i, WM_j 所对应的邻域集合 $\mathrm{NE}(\mathrm{WM}_i)$, $\mathrm{NE}(\mathrm{WM}_j)$ 进行修正: 将 WM_i 设置为 WM_j 的邻域点, 将 WM_j 设置为 WM_i 的邻域点, 即使得

$$\mathrm{WM}_i \in \mathrm{NE}(\mathrm{WM}_j), \ \mathrm{WM}_j \in \mathrm{NE}(\mathrm{WM}_i).$$

Step 5: 求出所有数据点间的最短路径. Step 4 结束后, 在整个数据集上形成了唯一邻域图, 在一个以欧氏距离加权的无向邻域图中, 数据点间的最短路径可以用经典 Floyd 算法或 Dijkstra 算法求得.

Step 6: 构建 m 维欧氏空间中数据的低维嵌入. 以 Step 5 所得到的点对点之间的最短路径作为经典多尺度变换的输入, 得到 m 维的低维嵌入 Y.

Step 7: 以高维样本集 X 作为输入 BP 神经网络的输入, 以低维嵌入 Y 作为其输出, 构造一个 BP 神经网络结构逼近该映射函数, 并用遗传算法优化该神经网络的初始权值和阈值.

Step 8: 将给定的需要预测的样本点 $x_0 \in R^n$ 作为 BP 神经网络的输入, 得到输出 $y_0 \in R^m$.

Step 9: 以 Y 作为 KNN 算法的训练样本集, 预测出 y_0 的类别标签, 该标签即为 x_0 所属的类别.

4 实验过程

4.1 实验准备

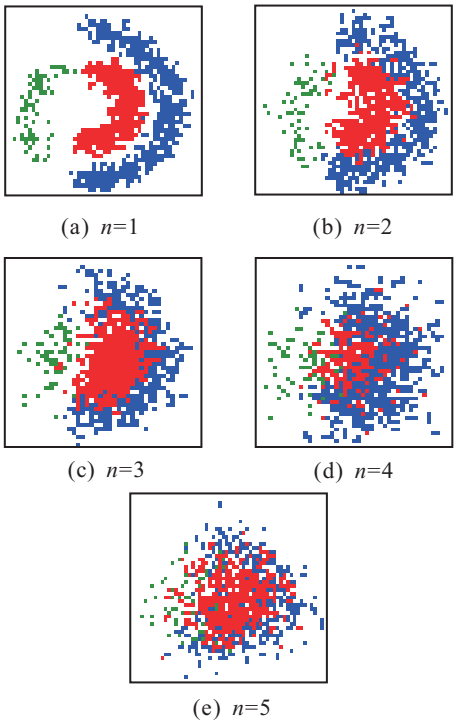
实验采用的数据集包括人工数据集和真实数据集, 这些数据集的属性均为数值属性. 人工数据集是如图 1(a) 所示的数据, 真实数据集来源于 UCI 机器学习数据库中的 5 个数据集[15], 数据集去掉了带有缺失属性值的样本. 经过整理后的数据集大小及其属性个数如表 1 所示.

表 1 实验所用数据集

数据集	数据集简写	大小	属性总数	类别数
balance-scale	bal	625	4	3
breast-w	brew	683	9	2
diabetes	dia	768	8	2
glass	gla	214	9	7
sonar	son	208	60	2
swiss roll	swr	1036	3	3

将 ISOMAP-C 方法与 BP 神经网络[16], RBF 神经网络[17], K-近邻分类器, C4.5 决策树[18], SVM[19-20], ISOMAP 在 6 个数据集上进行分类性能比较. 其中: BP, RBF, K-NN, C4.5 在基于 WEKA 平台下实现[21], 参数为 WEKA 中的默认参数. SVM 采用 LIBSVM 工具箱[22], 并将其导入 WEKA 中实现. 这里用 ISOMAP 算法完成分类任务, 实质上是将传统的 ISOMAP 算法用作分类器的一种最简单的扩充, 在原理上, 与 C-ISOMAP 方法不同的是: ISOMAP 算法在不利用任

何类标签信息的情况下, 在训练数据集上构造唯一的邻域图, 余下的步骤与 C-ISOMAP 方法一致. BP 神经网络采用 3 层结构, 输入层神经元个数为 X 的维度 n, 隐层神经元个数统一为 20, 输出层神经元个数为 Y 的维度 m. 本次实验取 $m=2$, 即将高维数据降到 2 维特征空间. 遗传算法在基于 GAOT 工具箱平台上采用实数编码, 编码长度为 $n\times 20+20\times 2+20+2$, 初始种群为 50, 遗传代数为 100. 为了验证 ISOMAP-C 算法在有噪声数据上的分类性能, 在图 1(a) 所示的数据集上分别加上 1~5 倍的均值为 0, 标准差为 1 的高斯噪声, 如图 3(a)~图 3(e) 所示.



(a) n=1　(b) n=2　(c) n=3　(d) n=4　(e) n=5

图 3　n 倍均值为 0 标准差为 1 的高斯噪声 swiss roll 数据集

4.2　性能评估

根据有限数据的实验结果进行分类预测性能评估是机器学习领域内一个存在较多争议的研究领域, 这是因为在分类模型预测性能评估体系中存在很多模型评估准则, 且有多种不同的模型评估方法. 实验中, 在 10 次 10 折交叉验证 (ten times ten-fold cross validation) 的基础上计算各分类器的准确率. 为了判定两个分类器的平均准确率是否存在显著差异, 采取纠正重复取样 t 测试[23], 而不是标准的配对 t 测试, 因为在交叉验证过程中, 该方法能够有效解决因训练样本重叠导致所得交叉验证估计不独立的问题.

4.3　结果与分析

将上述 7 种算法在 6 个数据集上进行实验, 表 2 列出了 7 种算法在每个数据集上交叉验证的平均正确率和标准差. 表 3 是在显著性水平为 0.05 的情况下, ISOMAP-C 与其他算法比较的双侧成对 t 测试结果. 表 3 中, “win”代表 ISOMAP-C 算法优于其他算法, “tie”代表无显著差别, “loss”代表 ISOMAP-C 算法逊于其他算法. 表 3 的最后 1 行统计了 ISOMAP-C 算法与其他算法在所有数据集上比较的 win/tie/loss 个数, 最后 1 列统计了在某个特定数据集上, ISOMAP-C 算法与其他所有算法分别比较的 win/tie/loss 个数.

由表 2 可见, ISOMAP-C 在 6 个数据集中的 4 个数据集上取得了最好的分类精度, 而在 sonar 数据集上的效果反而比直接使用 BP 差. 原因是该数据集的属性个数过多, 而样本个数相对较少, 导致在降维过程中有过多的信息丢失, 从而在使用 BP 近似构建高维数据到低维数据的映射函数时出现偏差. 从表 3 的最后 1 行可以推断出, ISOMAP-C 算法的分类性能优于其他算法, 尤其是对于 ISOMAP 算法, 前者在 5 个数

表 2　不同算法的分类准确率　%

dataset	ISOMAP-C	ISOMAP	BP	RBF	K-NN	C 4.5	SVM
bal	**96.49(1.57)**	73.62(4.03)	90.69(3.04)	86.34(3.41)	90.26(1.95)	77.82(3.42)	89.87(1.67)
bre	95.63(1.00)	73.48(3.43)	96.10(2.18)	96.32(2.18)	**96.92(2.08)**	95.44(2.64)	96.33(2.04)
dia	**74.60(2.41)**	69.28(5.17)	74.55(4.90)	74.04(4.91)	72.94(4.26)	74.49(5.27)	65.11(0.34)
gla	**67.80(8.55)**	61.08(9.69)	67.32(8.64)	64.92(9.90)	63.26(8.51)	67.63(9.31)	68.34(8.25)
son	76.81(4.35)	64.35(10.21)	**81.61(8.66)**	72.62(9.91)	75.25(9.91)	73.61(9.34)	64.99(7.66)
swr	**100(0.00)**	99.78(0.17)	**100(0.00)**	99.85(0.62)	**100.00(0.00)**	99.79(0.47)	99.84(0.39)

表 3　ISOMAP-C 与其他算法比较的 win/tie/loss 表

dataset	ISOMAP	BP	RBF	K-NN	C 4.5	SVM	win/tie/loss
bal	win	win	win	win	win	win	6/0/0
bre	win	tie	tie	tie	tie	tie	1/5/0
dia	win	tie	tie	tie	tie	win	2/4/0
gla	win	tie	tie	tie	tie	tie	1/5/0
son	win	tie	tie	tie	tie	win	2/4/0
swr	tie	tie	tie	tie	tie	tie	0/6/0
win/tie/loss	5/1/0	1/5/0	1/5/0	1/5/0	1/5/0	3/3/0	

表 4 不同算法在不同噪声的 swiss roll 数据集上的分类准确率 %

噪声程度/倍	ISOMAP-C	BP	RBF	K-NN	C4.5	SVM
1	99.51(0.53)	99.52(0.68)	98.74(0.79)	**99.71(0.65)**	98.84(1.19)	97.88(2.02)
2	**94.57(1.83)**	93.72(2.86)	92.47(2.08)	93.05(2.02)	92.85(2.64)	86.78(5.19)
3	**84.76(4.43)**	83.49(3.79)	83.01(4.08)	83.78(4.43)	81.95(3.61)	73.65(4.49)
4	75.24(3.79)	**76.13(4.29)**	74.99(4.47)	75.05(4.05)	73.61(3.80)	69.62(4.07)
5	67.30(1.86)	**68.91(5.13)**	67.26(4.44)	66.88(4.49)	67.47(4.94)	51.93(5.21)

据集上显著优于后者. 从表3最后1列可以看出, 在balance-scale数据集上, ISOMAP-C算法显著优于其他所有分类算法, 在其他数据集上, ISOMAP-C算法也取得了较好的效果.

将不同算法在图3所示的5种不同程度的高斯噪声数据集上进行分类, 分类结果如表4所示. 由表4可见, ISOMAP-C对于噪声具有较好的鲁棒性, 反而SVM受噪声影响较大, 其原因在于: 在构造子邻域图时利用类标签, 无论噪声多大, 在属于每个类的子数据集上, 仅产生一个子邻域图, 属于不同类的数据不会同时出现在一个子邻域图中, 从而在降维后能够保证类间距离足够大.

5 结 论

本文对传统的以无监督方式工作的ISOMAP算法进行改进, 提出了一种以完成分类任务为目的的ISOMAP算法, 即ISOMAP-C. 该算法在降维过程中充分利用有用的类标签信息指导高维数据降维, 使得降维后类内数据更加紧凑, 类间数据更加分开. 采用BP神经网络和遗传算法近似构建从高维数据到低维数据的映射函数, 并在低维空间中完成分类任务. 将该算法与其他常用的分类器进行性能比较, 结果显示, 该算法的分类精度在总体上优于其他分类器, 且在某些数据集上显著超越其他分类器的性能.

该算法的缺点是需要计算各样本间的最短路径, 时间复杂度较高, 且目前只能处理数值属性的数据集. 对于含有分类属性的数据集以及高维的稀疏数据集, 如何提高分类速度和分类精度是下一步的研究工作.

参考文献(References)

[1] Turk M, Pentland A. Eigenfaces for recognition[J]. J of Cognitive Neuroscience, 1991, 3(1): 71-86.

[2] Amari S I, Cichocki A, Yang H. A new learning algorithm for blind source separation[C]. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 1996: 757-763.

[3] Cox T, Cox M. Multidimensional scaling[M]. London: Chapman and Hall, 1994.

[4] Duda R O, Hart P E, Stork D G. Pattern classification[M]. 2nd ed. Beijing: China Machine Press, 2004.

[5] Kohonen T. Self-organizing maps[M]. 3rd ed. Berlin: Springer-Verlag, 2001.

[6] Smola A J, Mika S, SchÖlkopf B, et al. Regularized principal manifolds[J]. J of Machine Learning Research, 2001, 1(3): 179-209.

[7] Bishop C M, Svensen M, Williams C K I. GTM: The generative topographic mapping[J]. Neural Computation, 1998, 10(1): 215-234.

[8] Yang J, Frangi A F, Yang J Y, et al. KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition[J]. IEEE TPAMI, 2005, 27(2): 230-244.

[9] Seung H S, Lee D D. Manifold ways of perception[J]. Science, 2000, 290(5500): 2268-2269.

[10] Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[11] Cheng Q C, Wang H Y, Feng Y, et al. A multi-class multi-manifold learning algorithm based on ISOMAP[C]. Proc of the 1st CJK Joint Workshop on Pattern Recognition. Nanjing, 2009, 2: 813-817.

[12] Yang M H. Face recognition using extended Isomap[C]. Proc of IEEE Int Conf of Image Processing. New York: 2002, 2: 117-120.

[13] Geng X, Zhan D C, Zhou Z H. Supervised nonlinear dimensionality reduction for visualization and classification[J]. IEEE Trans on Systems, Man and Cybernetics, 2005, 35(6): 1098-1107.

[14] Wu Y M, Chan K L. An extended Isomap algorithm for learning multi-class manifold[C]. Proc of ICMLC. Shanghai, 2004, 6: 3429-3433.

[15] Blake C L, Merz C J. UCI repository of machine learning databases[D]. Irvine: Department of Information and Computer Science, University of California, 1998.

[16] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back propagating errors[J]. Nature, 1986, 323(9): 318-362.

[17] Chen S, Cowan C F, Grant P M. Orthogonal least squares learning algorithm for radial basis function networks[J]. IEEE Trans on Neural Networks, 1991, 2(2): 302-309.

[18] Quinlan J R. C4.5: Programs for machine learning[M]. San Francisco: Morgan Kaufmann, 1993.

(下转第836页)