计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(12):2072~2077, 2007

一种新的有监督流形学习方法

孟德宇[1,2] 徐宗本[1] 戴明伟[2]
[1](西安交通大学信息与系统科学研究所 西安 710049)
[2](西安交通大学电子与信息工程学院 西安 710049)
(dymeng@mail.xjtu.edu.cn)

A New Supervised Manifold Learning Method

Meng Deyu[1,2], Xu Zongben[1], and Dai Mingwei[2]
[1](*Institute for Information and System Science, Xi'an Jiaotong University, Xi'an* 710049)
[2](*School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049)

Abstract A new supervised manifold learning method is proposed in this paper, in order to present a new strategy to efficiently apply manifold learning and nonlinear dimensionality reduction methods to supervised learning problems. The new method realizes efficient supervised learning mainly based on integrating the topology preserving property of the manifold learning methods (Isomap and LLE) and some prominent properties of support vector machine such as efficiency on middle and small sized data sets and essential capability of support vectors calculated from support vector machine. The method is realized via the following steps: first to apply Isomap or LLE to get the embeddings of the original data set in the low dimensional space; then to obtain support vectors, which are the most significant and intrinsic data for the final classification result, by using support vector machine on these low dimensional embedding data; subsequently to get support vectors in the original high dimensional space based on the corresponding labels of the obtained low dimensional support vectors; finally to apply support vector machine again on these high dimensional support vectors to gain the final classification discriminant function. The good performance of the new method on a series of synthetic and real world data sets verifies the feasibility and efficiency of the method.

Key words manifold learning; support vector machine; Isomap; LLE; classification

摘 要 提出了一种新的有监督流形学习方法,目的是提供将流形学习降维方法高效应用于有监督学习问题的全新策略.算法的核心思想是集成流形学习方法对高维流形结构数据的降维有效性与支撑向量机(SVM)在中小规模分类数据集上的优良特性实现高效有监督流形学习.算法具体实现步骤为:首先利用SVM在流形学习降维数据中选出对分类决策最重要的数据集,即支撑向量集;按标号返回可得到原空间的支撑向量集;在这个集合上再次使用SVM即可得到原空间的分类决策,从而完成有监督流形学习.在一系列人工与实际数据集上的实验验证了方法的有效性.

关键词 流形学习方法;支撑向量机;等距特征映射;局部线性嵌入;分类

中图法分类号 TP391.4

数据挖掘是数据库知识发现中最重要的步骤之一,其目标是从获取的数据中高效准确地挖掘出我们所需要的信息.在实际应用中,数据往往呈现海量、高维、非线性等特性,这些特性给数据挖掘带来

收稿日期:2006-10-10;修回日期:2007-06-11
基金项目:国家自然科学基金项目(70531030,60575045)

了很多问题,例如海量特性导致的计算效率低下问题、高维特性带来的维数灾难问题和非线性特性引起的线性模型失效问题等. 幸运的是,实际中高维数据的属性之间往往存在一定的规律性和相关性,即实际数据经常存在着外在与内在两个维数. 在这样的情况下,理论上只需得到对高维数据的本质低维表示便可以从中挖掘出我们所需要的信息,即存在把高维数据降维从而避开维数灾难的可能性.

传统的线性降维方法主要包括广为人知的主成分分析(PCA[1])与高维尺度分析(MDS[2])等方法. PCA 主要思想是通过估计数据二阶统计性质来发现数据集的本质线性维数,通过计算其本质维上的坐标表示实现降维;MDS 主要思想是通过保持降维前后的数据间距离来实现降维. 线性降维方法能够对具备线性结构的数据进行降维处理,但从本质上说,不能有效地对高度非线性分布的数据集进行降维.

流形学习降维是近年来兴起的专门针对非线性分布数据降维的方法,主要包括等距特征映射(Isomap[3])与局部线性嵌入(LLE[4])等方法. 与线性降维方法相比,流形学习降维方法具有很多优势:对非线性流形结构数据具有自适应性;只涉及到较少的参数选择问题(只需确定少量邻域参数);基于非常易于理解的模型构造方式,降维后的数据特征具有很好的可解释性. 在人脸识别、手写数字辩识、文本分类等方面的成功应用[3-4]也证明了流形学习降维方法的有效性.

然而,此类方法目前的有效性主要体现在无监督学习的应用中,对于有监督学习,流形学习降维方法还不能被很好的应用. 主要原因是其均只建立了对已知高维数据 $\{x_i\}_{i=1}^{l}\subset\mathbb{R}^n$(即训练数据)的降维表示 $\{y_i\}_{i=1}^{l}\subset\mathbb{R}^m$,$m<n$,却不能直接得到一个新的高维数据 x'(即测试数据)的严格流形降维表示 y'. 这影响到需要预测功能的模式分类方法应用的效率甚至可行性. 为了解决这个问题,本文提出了一种有监督流形学习方法,此方法集成了一种高效模式分类算法,支撑向量机(support vector machine,SVM)[5]并因而兼有了 SVM 的两个优良特性:能够从训练数据中选出本质影响分类结果的少量数据(即支撑向量);能够高效处理中小规模的模式分类问题. 更具体些说,新方法的主要思想是利用 SVM 从流形降维后的低维数据集中筛选对分类问题起本质作用的少量数据,即支撑向量,并在原空间中对应的少量数据集上应用 SVM 对其进行训练得到原空间上的分类决策函数. 在一系列人工与实际数据上的模拟仿真实验证实了新方法的可行性与有效性.

1 相关背景知识

在本节我们将对相关背景知识进行简要回顾.

1.1 Isomap 与 LLE

Isomap 的主要思想是利用局部邻域距离近似计算数据点间的流形测地线距离,通过建立原数据的测地线距离与降维数据间的空间距离的对等关系完成数据降维. 其实现分为 3 个步骤:

首先 Isomap 需要基于原高维空间数据间($\{x_i\}_{i=1}^{l}\subset\mathbb{R}^n$)的距离判断每个数据的邻域($K$ 邻域或 ε 邻域)数据,将各个数据与其邻域连接从而构成高维空间上的图,图中每条边的权值即为其邻域距离 $d_X(i,j)=\|x_i-x_j\|$;然后计算图中任意两数据间的最短路径 $d_G(i,j)$,即近似得到流形上两数据间的测地线距离;最后运用经典的 MDS 算法建立测地线距离矩阵 $\boldsymbol{D}_G=\{d_G(i,j)\}_{l\times l}$ 与降维数据间($\{y_i\}_{i=1}^{l}\subset\mathbb{R}^m$,$m<n$)的空间距离矩阵 $\boldsymbol{D}_Y=\{d_Y(i,j)\}_{l\times l}$($d_Y(i,j)=\|y_i-y_j\|$)的对等关系,即最小化测度函数 $E=\|\tau(\boldsymbol{D}_G)-\tau(\boldsymbol{D}_Y)\|_{L^2}$(其中 L^2 矩阵距离 $\|\boldsymbol{A}\|_{L^2}=\sqrt{\sum_{i,j}A_{ij}^2}$,$\tau(D)=-HSH/2$,$S_{ij}=D_{ij}^2$,$H_{ij}=\delta_{ij}-1/N$),从而实现降维.

LLE 的主要思想则是建立原高维空间数据的邻近数据局部线性表示,通过在降维空间中尽可能保持其局部线性表示特征来实现降维,也包括 3 个步骤:

首先确定每个数据的邻域数据;然后通过解下列约束优化问题得到约束矩阵 $\boldsymbol{W}$:

$$\min\ \varepsilon(\boldsymbol{W})=\sum_i\left|X_i-\sum_j W_{ij}X_j\right|^2,$$
$$\text{s.t. } W_{ij}=0,\ \text{若 } X_j \text{ 不位于 } X_i \text{ 邻域},$$
$$\sum_j W_{ij}=1;$$

最后计算如下无约束优化问题来得到降维结果 $Y=\{y_i\}_{i=1}^{l}$:

$$\min\ \Phi(Y)=\sum_i\left|y_i-\sum_j W_{ij}y_j\right|^2.$$

1.2 SVM

支撑向量机是以统计学习理论为基础,以结构风险最小化为学习原则构造的学习算法,是目前公

认较为有效的一种模式分类方法. SVM 主要通过计算异类数据的最大间隔超平面从而得到分类决策函数的分类面,其线性模型如下:

$$\min_{w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i,$$
$$\text{s.t. } y_i[w\cdot x_i-b]\geqslant 1-\xi_i,\ i=1,\cdots,l,$$
$$\xi_i\geqslant 0.$$

其中 $\xi_i\geqslant 1$ 为松弛变量,$C>0$ 为惩罚因子. 可利用核函数 $K(x_i,x_j)=(\Phi(x_i),\Phi(x_j))$ 将模型推广至非线性可分情形,并可利用拉格朗日乘子法来解决此优化问题从而得到分类结果.

SVM 有一些比较突出的优势:其计算结果完全由包含在数据中的少量支撑向量决定;其计算模型是一个标准的二次规划模型,存在诸多标准解法;仅存在惟一极值,不存在局部寻优问题;拥有完整的理论体系,模型推广能力可从本质上得以保证等. 但迄今为止,SVM 大部分的成功应用还是在中小规模(较低维或样本数目较小)数据上,对于较大规模的数据,其计算效率往往不能得以保证,其中一个很大原因是由于尽管结果只与少量支撑向量有关,但一般缺少先验信息来确定这些支撑向量,只能通过在全数据上的计算才能得到最终结果.

1.3 现有有监督流形学习方法

流形学习方法已经比较有效地应用到无监督学习(聚类)与可视化的问题中,然而对于有监督学习问题,流形学习方法的应用较少. 现有的此方面的应用均以如下几个步骤实现:

首先定义数据间的相似度度量,即数据间距离,现有的定义方法大多试图同时保持同类间距离与拉大异类间距离,如以下两种定义方式[6](假定分类数据为 $\{x_i,z_i\}_{i=1}^{l}\subset\mathbb{R}^n\times\{0,1\}$):

$$D(x_i,x_j)=\begin{cases}\sqrt{1-\mathrm{e}^{\frac{-d^2(x_i,x_j)}{\beta}}}, & z_i=z_j,\\ \sqrt{\mathrm{e}^{\frac{d^2(x_i,x_j)}{\beta}}}-\alpha, & z_i\neq z_j.\end{cases}\tag{1}$$

$$D(x_i,x_j)=d(x_i,x_j)+\alpha\max_{k,m}(d(x_k,x_m))\Delta_{i,j},\tag{2}$$

其中 α,β 均为指定常数,$\Delta_{i,j}=|z_i-z_j|$;在定义了距离之后使用 Isomap 或 LLE 对数据进行降维,得到低维数据 $\{y_i\}_{i=1}^{l}\subset\mathbb{R}^m$;将降维分类数据($\{y_i,z_i\}_{i=1}^{l}$)作为训练数据,采用某种分类方法得到降维空间上的分类决策函数 $g:\mathbb{R}^m\to\{0,1\}$;利用某种回归方法(如神经网络回归[6])将 $\{x_i,y_i\}_{i=1}^{l}$ 作为训练数据建立从 $\mathbb{R}^n$ 到 $\mathbb{R}^m$ 空间的映射函数 $f:\mathbb{R}^n\to\mathbb{R}^m$;对任意新输入数据 $x'\in\mathbb{R}^n$,将其映射至低维空间得到其低维表示 $y'=f(x')$,则可求得其对应的分类决策为 $g(y')=g(f(x'))$.

尽管这些有监督流形学习方法保证了流形学习降维后有监督模式分类的可行性,但这些方法依然存在一些问题:首先,尽管已有方法尽可能保持了同类数据间的距离,但其对异类数据间距离的放大可能直接导致数据内部拓扑的破坏,从而可能造成降维方法的失效;其次,已有方法必须运用回归建立从原高维空间向降维空间的映射,这会带来两个问题:一是增加了算法的复杂度(由于 $\{x_i\}_{i=1}^{l}$ 往往是高维且数量庞大的数据),降低了算法效率;二是神经网络等回归方法往往涉及许多参数的选择问题,从而增加了算法模型选择的难度.

下一节中我们将提出一种新的有监督流形学习方法,旨在从一定程度上解决这些问题.

2 新的有监督流形学习算法

流形学习算法能够近似得到高维空间中呈现流形结构数据的低维同构映射[7]. 对于无监督学习问题,流形学习降维算法有效的本质是其保存了原流形结构数据的聚类结构,从而在降维数据上聚类等价于在原数据上的聚类(在标号对应的意义上);而对于有监督学习问题,流形学习算法尽管也保持了有监督数据的流形结构,但问题是在对降维数据学习得到降维空间上的分类决策信息之后,如何高效地得到一个原高维空间新输入数据的分类决策.

注意到,在流形降维空间中利用 SVM 能够得到一类关键的分类决策信息——支撑向量集合,即对于分类决策函数起本质作用的分类数据. 支撑向量是距离分类决策面最近的数据,而由于降维数据保持了原数据的流形拓扑结构,因此这些低维支撑向量按标号对应的原高维空间数据集也近似为距离原空间分类决策面最近的数据,即原高维分类数据的支撑向量. 那么我们只需在这些数据上运行 SVM 即可得到高维空间的分类决策函数. 这就是我们的算法构造思想,具体实现步骤如下:

Step1. 应用 Isomap 或 LLE 等流形降维方法将高维数据 $\{x_i\}_{i=1}^{l}\subset\mathbb{R}^n$ 降维至低维数据 $\{y_i\}_{i=1}^{l}\subset\mathbb{R}^m$;

Step2. 对分类训练数据 $\{y_i,z_i\}_{i=1}^{l}$ 进行训练,得到低维空间上的支撑向量集合 $D_{SV}=\{y_i'\}_{i=1}^{l_{SV}}\subset\mathbb{R}^m$;

Step3. 将低维支撑向量集合 D_{SV} 按标号返回原空间,得到对应的高维数据集合 $D'_{SV}=\{x'_i\}_{i=1}^{l_{SV}}\subset\mathbb{R}^n$;

Step4. 将集合 $\{x'_i,z'_i\}_{i=1}^{l_{SV}}$ 作为训练数据再一次运行 SVM,得到最终的分类决策函数 $f:\mathbb{R}^n\to\{0,1\}$.

算法中,$z'_i\in\{0,1\}$ 为对应 x'_i 的标号. 方便起见,我们将新方法简称为 ASMLM(advanced supervised manifold learning methods).

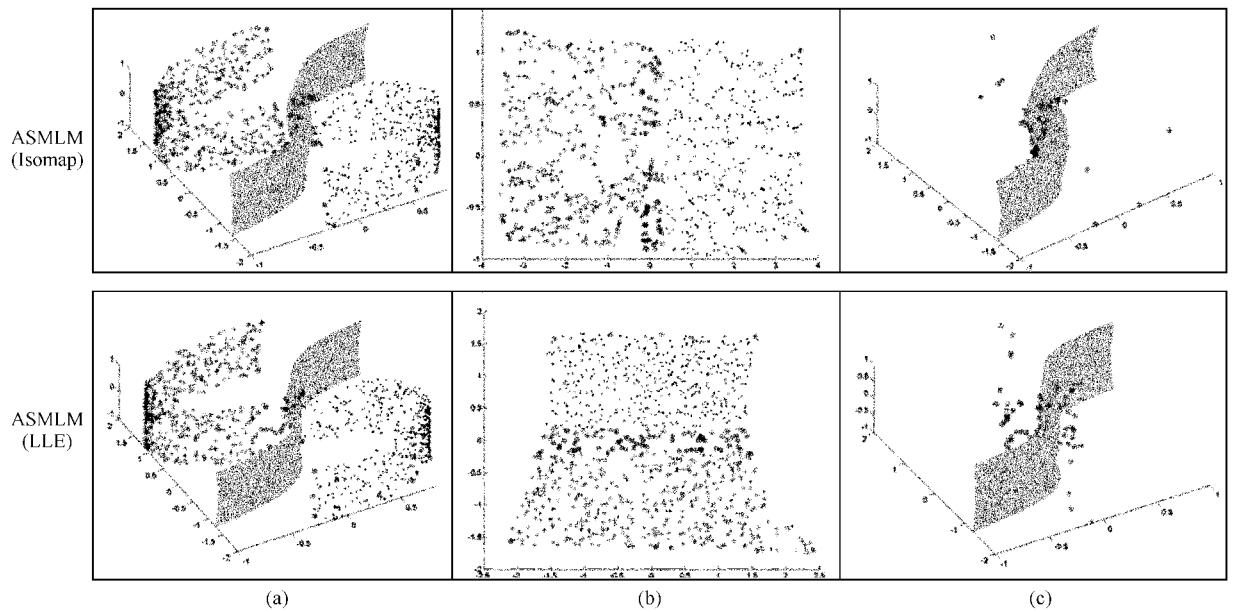
ASMLM 方法有如下特点:在得到降维数据分类决策信息(支撑向量集合)时不改变原数据间的距离,不破坏其内部拓扑结构;分别在低维数据集,少量的支撑向量集上两次运用 SVM,利用 SVM 在中小规模数据上的高效性保证算法效率,不涉及对高维且数据庞大数据集上的学习问题;直接得到原空间而非降维空间上的分类决策函数,在判断新数据类别时不需要使用神经网络等回归工具首先将其近似降维,降低了问题的复杂性和不确定性.

3 数值实验

本节将介绍 ASMLM 算法在人工与实际分类数据集上的模拟实验效果,从而验证新算法的可行性与有效性. 所有的程序均用 Matlab 语言编制,在 Matlab 7.0 平台上运行,计算机配置为:Pentium Ⅳ 1.7GHz 中央处理器,1GB 内存与 Windows XP 操作系统.

3.1 S-curve 型人工数据实验效果

第 1 组实验为 S-curve 型分类数据集上的模拟实验,数据分布如图 1(a)所示. 分别采用 SVM, Isomap + ASMLM 与 LLE + ASMLM 方法对数据进行分类训练. 其中,SVM 参数选择方法采用 5 倍交叉法[8],Isomap 与 LLE 均选取 7-邻域. 计算效果如图 1 所示.



'+' and '.' denote positive and negative points; squared and circled point denote positive and negative support vectors and the 2-D surface denotes the obtained classification hyperplane.

Fig. 1 Comparisons of the classification results. (a) By SVM; (b) Showing 2-D embedding result calculated by Isomap and LLE; and (c) By ASMLM.

图 1 分类结果比较. (a) SVM 结果;(b) 分别为运用 Isomap 与 LLE 在降维空间中提取支撑向量的效果;(c) ASMLM 结果.

容易观察到,运用 Isomap 或 LLE 降维后,可利用 SVM 找到降维空间中处于分类面附近的支撑向量,从图 1(b)可看出,这些支撑向量只是原数据的一小部分. 利用标号将这些支撑向量返回后,得到对应的高维数据集仍位于分类面附近,从而也是高维空间分类决策的支撑向量,也就是说,对这些少量的支撑向量进行分类训练基本等价于对原数据集的分类训练. 从图 1(c)可以观察到,在这些少量数据上运用 SVM 得到的分类决策面基本与在整体数据上得到的分类决策面相同. 这直观地展示了 ASMLM 的构造机理及验证了新算法的有效性.

3.2 手写数字数据实验效果

第 2 组实验为 NSDL 标准手写数字数据库中'0'-'1'分类数据集上的模式分类实验. 在此数据集

中,共包含 5923 个‘0’与 6742 个‘1’手写数字,均由 $28\times28=784$ 维像素构成. 我们分别抽取前 1000 个‘0’与‘1’构成训练数据集,将剩余的 10665 个数据组合构成测试数据集. 分别采用 SVM,Isomap + ASMLM,LLE + ASMLM 与第 1.3 节中所介绍的方法(记为 M 方法)对训练数据集进行分类学习. 其中,SVM 参数选择方法采用 5 倍交叉法[8],Isomap 与 LLE 均选取 6-邻域,降维空间维数设置为 3,M 方法中采用 3 层前向神经网络进行回归学习,网络隐层单元数设为 100,输出层单元数设置为 3,权值采用 BP 算法进行求解. 最终在测试数据集上进行测试可得分类正确率. 算法效果如表 1 所示:

Table 1 Comparisons of Five Methods

表 1 5 种方法分类效果比较

Method	Training Time (s)	Classification Rate (%)
SVM	3248.9	99.91
Isomap + M	31325.0	94.35
LLE + M	3278.6	98.50
Isomap + ASMLM	828.2	99.91
LLE + ASMLM	381.0	99.64

从表 1 可看出,与 Isomap 结合的 ASMLM 方法不仅训练时间相对 SVM 与以往有监督流形学习方法有显著减少,并且可达到与 SVM 几乎相同的精度,几乎将测试集中全部数据分类正确;而与 LLE 结合的 ASMLM 方法尽管分类测试精度略差,但也远强于 M 方法的测试精度,而且其算法训练时间相对其他方法大副减小,计算效率大大提高. 这说明了 ASMLM 算法的高效性与应用潜力.

为了进一步展示 ASMLM 的运行原理与计算效果,我们在图 2 与图 3 中分别展示了利用 SVM 方法在 Isomap 与 LLE 降维数据中所求得的支撑向量. 可观察到,这些位于低维分类决策平面附近的‘0’‘1’支撑向量具有一定的相似性与模糊性,说明它们的确是位于分类边缘的数据点. 这进一步说明了 ASMLM 计算的有效性.

Fig. 2 Demonstrations of support vectors found by Isomap.

图 2 利用 Isomap 所找到的支撑向量集合图示

Fig. 3 Demonstrations of support vectors found by LLE.

图 3 利用 LLE 所找到的支撑向量集合图示

4 结 论

本文提出了一种新的有监督流形分类学习方法,旨在从一定程度上解决流形学习方法应用于有监督模式分类中存在的问题. 新方法的构造主要基于现有流形学习算法对高维流形结构数据降维的保拓性和 SVM 对中小规模数据的高效性及其提取出的支撑向量集对分类决策的本质性. 其主要步骤是首先利用 SVM 在使用流形学习算法降维后的数据中选取出支撑向量集;按标号返回得到原空间的支撑向量集;在此集合上再次使用 SVM 进而求得原空间的分类决策函数. 新方法与之前的有监督流形学习方法的主要区别是:在得到降维数据分类决策信息(支撑向量集合)时不破坏其内部拓扑结构;分别在降维数据集(低维数据),高维支撑向量集(小量数据)上两次运用 SVM,不涉及对大规模(即高维且数据庞大)数据集的学习问题;直接得到了原空间而非降维空间上的分类决策函数,在判断新数据类别时不需使用神经网络等回归工具将其首先近似降维. 人工与实际数据实验验证了新方法的有效性.

参 考 文 献

[1] I Jolliffe. Principal Component Analysis [M]. New York: Springer-Verlag, 1989

[2] T Cox, M Cox. Multidimensional Scaling [M]. London: Chapman & Hall, 1994

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

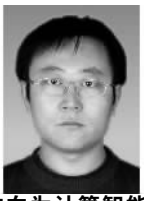
[4] S T Roweis, L K Saul. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290: 2323-2326

[5] V N Vapnik. The nature of statistical learning theory [M]. New York: Spinger-Verlag, 1995

[6] Xin Geng, Zhan Dechuan, Zhou Zhihua. Supervised nonlinear dimensionality reduction for visualization and classification [J]. IEEE Trans on SMC B, 2005, 35(6): 1098-1107

[7] Luo Siwei, Zhao Lianwei. Manifold learning algorithms based on spectral graph theory [J]. Journal of Computer Reaserch and Development, 2006, 43(7): 1173-1179 (in Chinese)
(罗四维, 赵连伟. 基于谱图理论的流形学习算法[J]. 计算机研究与发展, 2006, 43(7): 1173-1179)

[8] Y Bengio, Y Grandvalet. No unbiased estimator of the variance of K-fold crossvalidation [J]. Journal of Machine Learning Research, 2004, 5: 1089-1105



Meng Deyu, born in 1978. Ph. D. candidate in Xi'an Jiaotong University. His current research interests include computational intelligence and data mining.
孟德宇,1978 年生,博士研究生,主要研究方向为计算智能与数据挖掘.

Xu Zongben, born in 1955. Professor and Ph. D. supervisor in mathematics and computer science. His current research interests include nonlinear functional analysis, mathematical foundation of information technology and data mining theory and applications.
徐宗本,1955 年生,教授,博士生导师,主要研究方向为非线性泛函分析、计算智能的数学基础、数据挖掘及应用.

Dai Mingwei, born in 1980. Received his master degree from Xi'an Jiaotong University in 2007. His current research interests include computational intelligence.
戴明伟,1980 年生,硕士,主要研究方向为信息与计算智能算法.

Research Background

Manifold learning methods for nonlinear dimensionality reduction have attracted more and more attentions in the recent decade due to their excellent performance especially on unsupervised learning and data visualization. However, these methods still can't be applied to supervised learning problem very efficiently. In this paper, a new supervised manifold learning method is proposed by integrating SVM and manifold learning methods. Because of the prominent properties of both adopted methods, the new method has excellent ability to deal with the supervised learning problem. Our work is supported by the projects of the National Natural Science Foundation of China (70531030 and 60575045).