软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(4):748−764 [doi: 10.3724/SP.J.1001.2012.04039]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

监督式谱空间分类器*

何 萍[1], 徐晓华[2], 陈 崚[1,2+]

[1](南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

[2](扬州大学 信息工程学院 计算机系,江苏 扬州 225009)

Supervised Spectral Space Classifier

HE Ping[1], XU Xiao-Hua[2], CHEN Ling[1,2+]

[1](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

[2](Department of Computer Science, College of Information Engineering, Yangzhou University, Yangzhou 225009, China)

+ Corresponding author: E-mail: yzulchen@gmail.com

He P, Xu XH, Chen L. Supervised spectral space classifier. *Journal of Software*, 2012,23(4):748−764.
http://www.jos.org.cn/1000-9825/4039.htm

Abstract: This paper proposes a nonlinear classification algorithm S^3C (supervised spectral space classifier), short for supervised spectral space classifier. S^3C integrates the discriminative information into the construction of the low-dimensional supervised spectral space. The input training data is mapped into the supervised spectral space, followed by the optimization of the partitioning hyperplane with maximum margin. The test data is also transformed into the same feature space via an intermediate "bridge" between the original feature space and the target feature space. The classification result of S^3C is obtained by applying the optimal partitioning hyperplane to the transformed test data, directly. S^3C enables researchers to examine the transformed data in the supervised spectral space, which is beneficial to both algorithm evaluation and parameter selection. Moreover, the study presents a supervised spectral space transformation algorithm (S^3T) on the basis of S^3C. S^3T (supervised spectral space transformation) estimates the class indicating matrix by projecting the data from the supervised spectral space to the class indicating space. S^3T can directly deal with multi-class classification problems, and it is more robust on the data sets containing noise. Experimental results on both synthetic and real-world data sets demonstrate the superiority of S^3C and S^3T algorithms compared with other state-of-the-art classification algorithms.

Key words: classification; spectral method; dimension reduction; manifold mapping; supervised spectral space

摘 要: 提出了一种非线性的监督式谱空间分类器(supervised spectral space classifier,简称 S^3C).S^3C 首先将输入数据映射到融合了训练数据判别信息的低维监督式谱空间中,然后在该监督式谱空间中构造最大化间隔的最优分割超平面,并把测试数据以无监督的方式也映射到与训练数据相同的新特征空间中,最后,直接应用之前构建的分类超平面对映射后的测试数据进行分类.由于 S^3C 使研究者可以直观地观察到变化后的特征空间和映射后的数据,因此有利于对算法的评价和参数的选择.在 S^3C 的基础上,进一步提出了一种监督式谱空间分类器的改进算法(supervised spectral space transformation,简称 S^3T).S^3T 通过采用线性子空间变换和强迫一致的方法,将

* 基金项目: 国家自然科学基金(61003180, 61070047, 61103018); 江苏省自然科学基金(BK2010318, BK2011442); 江苏省教育厅自然科学基金(09KJB200013)

收稿时间: 2010-03-15; 修改时间: 2010-08-13; 定稿时间: 2011-04-25

映射到监督式谱空间内的数据再变换到指定的类别指示空间中去,从而获得关于测试数据的类别指示矩阵,并在此基础上对其进行分类.S^3T 不仅保留了 S^3C 算法的各项优点,而且还可以用于直接处理多分类问题,抗噪声能力更强,性能更加鲁棒.在人工数据集和真实数据集上的大量实验结果显示,S^3C 和 S^3T 与其他多种著名分类器相比,具有更加优越的分类性能.

关键词: 分类;谱方法;维数约减;流形映射;监督式谱空间

中图法分类号: TP181 **文献标识码:** A

分类是机器学习领域的一个核心研究内容,有着非常广泛的实际应用,例如人脸识别[1]、文本分类[2]、蛋白质结构预测[3]和入侵检测[4]等.近年来,由于核方法的提出,非线性分类器的设计得到了长足的发展.一大批具有内积形式的线性分类器借助于核技巧被发展为相应的非线性核分类器,包括核感知机学习(kernel perceptron learning)、核费歇尔判别(kernel Fisher discriminant)、相关向量机(relevance vector machine)、高斯过程(Gaussian processes)、贝叶斯点机(Bayes point machine)和支持向量机(support vector machine,简称 SVM)等[5,6].它们的基本思想是,将线性分类器中的每个点积用一个核函数来代替,把输入数据映射到某个高维的特征空间,然后通过在高维特征空间中构造线性的分类决策面,从而达到对原始数据非线性分类的目的.在这些种类繁多的核分类器中,应用最为广泛的就是由 Boser,Guyon 和 Vapnik 三人所提出的支持向量机(SVM)[7].SVM 结合核化和最大化间隔的思想,通过将原始数据映射到高维或无穷维的特征空间,寻找在新特征空间中最大化异类数据之间几何间隔的最优分割超平面,从而获得在原始空间中泛化性较强的非线性分类决策面.由于 SVM 基于核方法所进行的空间映射是隐式的,并且该隐式特征空间又通常是高维或者无穷维的,因此想要直接观察到升维后的特征空间以及映射后的数据是非常困难的,甚至是不可能的(在无穷维特征空间中).虽然我们已经知道升维的确可以在大多数情况下帮助 SVM 将不同类别的数据在高维空间中线性地分开,但是升维是否是实现数据线性分离的唯一手段,能否通过降维达到同样的效果,如果可以又该选择什么样的低维空间进行映射,这些都是值得我们探讨的重要问题.虽然目前我们有大量的降维算法可供选择,尤其是近年来受到广泛关注的流形学习算法 Isomap[8],LLE[9],Laplacian Eigenmap[10],NPE[11]和 LPP[12]等,但是众所周知,无监督降维的目标是最大化数据方差或维持近邻关系,与有监督分类寻找数据和类别之间映射关系的目标相差甚远,所以一般认为,无监督降维与有监督分类器的简单组合不能产生良好的分类性能.因此,本文寻找的是一种有监督的降维方法.从对监督信息的利用角度来看,SVM 在处理分类问题时只在其优化目标中使用正负号来表示数据间的同类或异类关系,对判别信息的利用较为简单.当使用者选择了不合适的核函数时,SVM 会因为其自身对判别信息利用的有限性,使分类性能受到较大的影响.因此,我们还将讨论是否有可能将分类问题的判别信息直接融入到数据的空间映射中去,从而使分类器变得更加鲁棒,而不再仅仅依赖于对核函数的选择.一种当前最为常用的将数据的判别信息融入到维数约减过程中的监督式降维分类算法称为线性判别分析(linear discriminant analysis,简称 LDA),其核化版本简称为 KDA.LDA 和 KDA 的缺点在于,它们假设每个类的数据都服从一个高斯分布,而且所有类的高斯分布都共享一个相同的协方差矩阵.类似地,无监督降维的流形学习算法也普遍依赖于数据的低维潜在流形假设.所以,本文探讨的是一种独立于数据分布假设的有监督降维方法.

谱方法(spectral method)[13]是近几年来机器学习领域的另一个研究热点.它建立在图论的谱图理论基础之上,通过求解关于图的拉普拉斯矩阵特征值分解来解决图的最优割问题.谱方法作为一种无监督的学习方法,最初被成功地应用于聚类、降维等无监督学习领域[14,15],近年来又逐渐发展到半监督学习领域.其中,Zhu 等人[16]提出了一种基于带约束的凸优化方法来构建半监督谱核,并采用二次约束的二次规划方法(QCQP)来求解该优化问题;Liu 等人[17]结合正则化方法和非参数化核学习提出了一种转导谱核(transductive spectral kernel),通过线性规划的方法来求解最优核矩阵;Johnson 等人[18]提出了一种使用无监督谱核的半监督学习框架,然后将其与以往基于图的半监督学习方法相联系.尽管如此,谱方法在监督学习领域的应用目前还较为少见.Li 等人[19]从核方法的角度重新看待谱方法,并设计了一种谱核,提供了对未知数据的映射方法.但是该方法在核的构建过程中并未加入任何训练数据的类别信息,并且它对测试数据的映射也是基于核方法推导而来的.

本文提出了一种融合了数据判别信息的监督式谱空间分类器(supervised spectral space classifier,简称 S^3C). S^3C 在对原始数据的处理上与 SVM 等核分类器正好相反,它将输入数据从原始空间显式地映射到低维的监督式谱空间.不同于核方法将数据映射到不可见的高维或无限维隐特征空间,S^3C 指定其目标映射空间具有特定的流形结构(单位超球面),因此研究者可以直观地观察到变化后的特征空间和映射后的训练和测试数据.为了保证不同类别的数据能够在新的特征空间中被尽可能清楚地分离开来,我们还在谱空间变换中融入了训练数据的判别信息,并称之为"监督式谱空间".在该监督式谱空间中,S^3C 采用最大化间隔的方法求解映射后训练数据的最优分割超平面;然后通过在原始空间和目标空间(即监督式谱空间)之间构造的一个过渡性的"桥",将测试数据映射到与训练数据相同的目标空间中去;最后,S^3C 根据之前在训练数据上构建的分类超平面对测试数据进行类别预测.考虑到 S^3C 在处理多分类问题时,还需使用额外的组合策略结合多个分割超平面进行投票才能获得最终的分类结果,我们在 S^3C 的基础上进一步提出了一种改进的监督式谱空间变换多分类算法,简称为 S^3T(supervised spectral space transformation).S^3T 采用线性子空间变换和强迫一致的方法替换 S^3C 中最大化间隔的分类超平面构建,通过将监督式谱空间中的映射数据再线性变换到指定的类别指示空间中去,根据所得类别指示矩阵,从而获得对测试数据的类别预测.S^3T 不但可以用于直接处理多分类问题,无需任何组合策略,而且还适用于可能存在类别标号错误的数据集,抗噪声能力更强,分类性能也更加优越.大量基于人工数据集和真实数据集的实验结果显示,S^3C 和 S^3T 的分类性能优于 C4.5 决策树算法、SVM 核分类器、流形学习分类器 KLPP+Linear SVM、有监督降维分类算法 LDA 及其核化版本 KDA,S^3C 和 S^3T 对参数的敏感性也远低于 RBF-SVM,KDA 和 KLPP+Linear SVM.此外,本文还深入讨论了 S^3C(S^3T)的各项参数对其分类性能的影响、交互作用和显著性.在对参数α的讨论过程中,我们还实验验证了采用监督式谱空间变换的 S^3T 算法的分类错误率远低于采用无监督降维方法的谱方法降维分类算法.

S^3C(S^3T)与以往的非线性分类算法相比,主要有以下几个特点:

1) SVM 通过把数据升到高维或无穷维特征空间,从而把不同类别的数据线性地分开,但是这种方法既不可观测也不易理解.S^3C(S^3T)通过将数据映射到低维的监督式谱空间,不但达到了相同甚至更好的数据分离效果,而且更简单、更直观,也更容易理解.

2) 采用无监督降维再分类的算法忽视了监督信息在降维中的指导作用,将降维与分类两个过程割裂开来,导致输入数据可能会被映射到不利于分类的低维空间中去.SVM 没有充分利用训练数据的判别信息,只在其目标函数中用正负号表示了数据间的同类或异类关系,因此它的空间映射在很大程度上依赖于对核函数的选择.S^3C(S^3T)则更加充分地利用了训练数据的判别信息,它将数据间的类别关系直接融入到输入数据的监督式谱空间映射中去,因此可以确保不同类别的数据能够在变换后的监督式谱空间中被尽可能清楚地分离开来.

3) 无监督降维的流形学习算法依赖于数据的潜在低维流形假设,有监督降维的 LDA 和 KDA 算法依赖于每个类别的数据服从共享同一个协方差矩阵的高斯分布假设.S^3C(S^3T)则通过采用不对数据分布作任何估计和假设的谱方法,在不同分类问题上表现出更加鲁棒的分类性能.

4) 传统维数约减算法的目的在于发现潜藏在数据内部的低维流形结构,因此在发生数据坍塌时,会因为丢失必要的结构信息而失效.S^3C(S^3T)算法的目的在于挖掘输入数据与类别之间的非线性关系,因此数据坍塌作为一种数据压缩的表现形式,不但不一定会导致分类性能的降低,反而可能会因为不同类别的训练数据被压缩为几个具有代表性的数据点而起到简化并改进分类决策面的效果.

5) SVM 的分类性能对参数设置非常敏感,参数的取值范围也非常广,使得 SVM 的优化过程也变得较为繁琐.S^3C(S^3T)采用了一种基于连续 k-近邻的局部尺度全连接图构建方式,在优化性能时只需在有意义的小范围内(通常 $k\in(0,10)$)以一定精度搜索最佳的 k 值,因此能够更好地平衡参数鲁棒性和灵活性两者之间的关系.

本文第 1 节简要介绍 SVM 和谱方法的基本思想.第 2 节提出监督式谱空间分类器(S^3C).第 3 节提出监督式谱空间分类器的改进算法(S^3T).第 4 节对 S^3C 和 S^3T 进行全面的分类性能评价.最后,第 5 节对全文给出总结.

1 基本理论

1.1 支持向量机

SVM 是建立在统计学习理论基础上的一种具有直观几何解释和易理解代数表达形式的学习方法.SVM 的核心思想在于最大化异类数据之间的几何间隔,而核化则是其对线性不可分问题的扩展.

通常我们考虑二分类问题,假设训练数据集$\{(x_1,y_1),\ldots,(x_m,y_m)\}$,$x_i\in R^N$,$y_i\in\{-1,1\}$.SVM 的优化目标是寻找一个能够最大化两类数据之间几何间隔的分割超平面;如果不存在这样的超平面将所有数据都分类正确,就采用软边界技术,对错误分类的数据引入松弛变量ξ_i和惩罚参数 C,寻找能够最小化两类数据分类误差的超平面作为代替.SVM 的具体目标函数如下所示:

$$\begin{aligned}&\min_{w,b}\ \frac{1}{2}\|w\|^2+C\sum_{l=1}^{m}\xi_i\\&\text{s.t.}\quad y_i(\langle w,x_i\rangle+b)\geqslant 1-\xi_i,\ i=1,\ldots,m,\\&\qquad \xi_i\geqslant 0,\ i=1,\ldots,m\end{aligned}\tag{1}$$

对公式(1)的对偶优化问题进行变形,可得如下等价问题:

$$\begin{aligned}&\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_iy_j\langle x_i,x_j\rangle\alpha_i\alpha_j-\sum_{j=1}^{m}\alpha_j\\&\text{s.t.}\quad \sum_{i=1}^{m}y_i\alpha_i=0,\\&\qquad 0\leqslant\alpha_i\leqslant C,\ i=1,\ldots,m\end{aligned}\tag{2}$$

公式(2)是一个二次优化问题,可通过 SMO 等算法[20]进行求解.此时,SVM 的分类决策函数为

$$y=\operatorname{sgn}\left(\sum_{i=1}^{m}y_i\alpha_i^*(x_i^Tx)+b^*\right)\tag{3}$$

其中,

$$b^*=y_j-\sum_{i=1}^{m}y_i\alpha_i^*(x_i^Tx_j),\ \forall j\ \text{satisfies}\ \alpha_j^*\in(0,C)\tag{4}$$

当学习样本线性不可分时,SVM 采用核化技术,将公式(2)中的内积$\langle x_i,x_j\rangle$用核函数值 $K(x_i,x_j)$来代替,从而将数据映射到高维特征空间,使之在高维空间中线性可分.

SVM 常用的核函数包括多项式核函数和高斯径向基核函数等,其中,高斯径向基核函数的性能最为稳定,因此在实际中应用得也最为广泛.通常,我们把使用高斯径向基核函数的 SVM 简称为 RBF-SVM.

1.2 谱方法

谱方法是建立在谱图理论[21]基础上的一种用于处理无监督学习问题的有效工具.谱方法的基本思想是,将图的最优割问题转化为拉普拉斯矩阵的特征值分解问题.假设在一个无向加权图 $\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{W})$上,其中,$\mathcal{V}$为 $\mathcal{X}$对应的顶点集, $\mathcal{E}$ 为边集, $\mathcal{W}$则为相似度集.图的分割问题就是指将图 $\mathcal{G}$ 划分为 p 个互不连通的子图$\{\mathcal{G}_1,\ldots,\mathcal{G}_p\}$,使得它们的内部相似度最大化且外部相似度最小化.目前,使用最为广泛的是归一化割(normalized cut)[22],其目标函数如下所示:

$$\min\ \mathcal{J}(\mathcal{V}_1,\ldots,\mathcal{V}_p)=\sum_{i=1}^{p}\frac{cut(\mathcal{V}_i,\overline{\mathcal{V}_i})}{cut(\mathcal{V}_i,\mathcal{V})}\tag{5}$$

$\mathcal{V}_i$是子图 $\mathcal{G}_i$的顶点集,$\overline{\mathcal{V}_i}=\mathcal{V}/\mathcal{V}_i$ 是 $\mathcal{V}_i$的补集.

$$cut(\mathcal{V}_i,\overline{\mathcal{V}_i})=\sum_{i\in\mathcal{V}_i,\,j\in\overline{\mathcal{V}_i}}w_{ij}\tag{6}$$

表示 $\mathcal{V}_i$与 $\overline{\mathcal{V}_i}$之间的相似度连接权重.通过引入谱松弛技术,公式(8)可以简化为

$$\begin{aligned}&\min_{F}\ tr(F^TLF)\\&\text{s.t.}\quad F^TDF=I\end{aligned}\tag{7}$$

其中,$F=(f_{ij})_{n\times p}$是取实数值的分割指示矩阵.

定义 1(拉普拉斯矩阵). 图 $\mathcal{G}$ 的拉普拉斯矩阵为

$$L \overset{\text{def}}{=} D - W \tag{8}$$

其中,$W=(w_{ij})_{n\times n}$,w_{ij}表示顶点 v_i 与 v_j 之间的相似度,

$$D=diag(W\mathbf{1}_n) \tag{9}$$

其第 i 个对角元素为 $d_i=\sum_{j=1}^{n} w_{ij}$.

定义 2(归一化的拉普拉斯矩阵). 图 $\mathcal{G}$ 的归一化拉普拉斯矩阵为

$$\mathcal{L} \overset{\text{def}}{=} D^{-1/2}LD^{-1/2} = I - D^{-1/2}WD^{-1/2} \tag{10}$$

由于 $L=D-W$,我们可以通过直接求解如下公式来获得归一化割的等价松弛解:

$$\begin{aligned} &\max_{F}\ tr(F^TWF) \\ &\text{s.t.}\quad F^TDF = I \end{aligned} \tag{11}$$

目前,有 3 种最为常用的相似度矩阵 W 的构建方法:

(1) ε-连接图:只有距离小于ε的顶点是相互连接的;

(2) k-近邻图:每个顶点只与其前 k 个最近邻相连接;

(3) 全连接图:所有顶点都是相互连接的,每条边的权重通常用高斯径向基函数来定义.

$$w_{ij} = \mathrm{e}^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}} \tag{12}$$

近年来,Manor 等人提出了一种基于 k-近邻的局部尺度全连接图构建方式[15],具有能够处理多尺度数据集的优点.

2 监督式谱空间分类器

考虑一般的多分类问题,假设数据集 $\mathcal{X}=\{x_1,\ldots,x_m,x_{m+1},\ldots,x_n\}$,$x_i\in R^N$,其对应的标签集为 $\mathcal{Y}=\{y_1,\ldots,y_m,y_{m+1},\ldots,y_n\}$,其中,$y_i\in C=\{c_1,\ldots,c_p\}$,$C$ 表示类别集合,p 为类别总数.令 $S=\{1,2,\ldots,m\}$表示训练数据的下标集,$T=\{m+1,\ldots,n\}$表示测试数据的下标集,训练数据 $\mathcal{X}_S=\{x_1,\ldots,x_m\}$对应已知的类别标签集 $\mathcal{Y}_S=\{y_1,\ldots,y_m\}$,测试数据 $\mathcal{X}_T=\{x_{m+1},\ldots,x_n\}$对应未知的类别标签集 $\mathcal{Y}_T=\{y_{m+1},\ldots,y_n\}$.

多分类问题的任务就是要预测 $\mathcal{Y}_T$.我们用矩阵 $X_S=[x_1^T,x_2^T,\ldots,x_m^T]^T$ 和 $X_T=[x_{m+1}^T,x_{m+2}^T,\ldots,x_n^T]^T$ 分别表示训练数据和测试数据的输入空间,它们的每个行向量都对应着一个数据点的坐标.

这里,我们提出了一种融合了数据判别信息的监督式谱空间分类器 S^{3}C 用于处理以上分类问题.S^{3}C 的基本思想是,在无监督的谱空间变换中加入训练数据的判别信息,将训练数据映射到低维的监督式谱空间,然后应用最大化间隔的方法确定在监督式谱空间中的最优分割超平面,并借助于在输入空间和目标空间(即监督式谱空间)之间构造的一个“桥”,把测试数据也映射到与训练数据相同的新特征空间中,最后使用之前构建的分类超平面对映射后的测试数据进行分类.

定义 3(条件相似度矩阵). 基于训练数据 $\mathcal{X}_S$ 的条件相似度矩阵为

$$W_{\mathcal{X}_S} = (w_{\mathcal{X}_S})_{m\times m} \tag{13}$$

其元素

$$w_{\mathcal{X}_S}(x_i,x_j) \overset{\text{def}}{=} \mathrm{e}^{-\frac{\|x_i-x_j\|^2}{\sigma_i\sigma_j}} \tag{14}$$

表示数据点 x_i,x_j 之间的相似度,x_i 的近邻尺度σ_i 定义为

$$\sigma_i \overset{\text{def}}{=} \begin{cases} \| x_i - x_i^{\lfloor k \rfloor} \|, & \text{if } \lfloor k \rfloor = \lceil k \rceil \\ k \cdot \| x_i - x_i^{\lceil k \rceil} \|, & \text{if } \lfloor k \rfloor = 0 \\ (k - \lfloor k \rfloor) \cdot \| x_i - x_i^{\lceil k \rceil} \| + (\lceil k \rceil - k) \cdot \| x_i - x_i^{\lfloor k \rfloor} \|, & \text{otherwise} \end{cases} \tag{15}$$

其中, $x_i^{\lfloor k \rfloor}$ 是 x_i 的第$\lfloor k \rfloor$个近邻,而 $x_i^{\lceil k \rceil}$ 则是 x_i 的第$\lceil k \rceil$个近邻.

注意,这里 W_{X_S} 的构建实际上只使用了一个超参数 $k \in R^+$.当 k 为整数时($\lfloor k \rfloor = \lceil k \rceil$),$\sigma_i$ 就是 x_i 到其第 k 个近邻的距离;当 $0<k<1$ 时($\lfloor k \rfloor=0$),σ_i 就是 k 乘以 x_i 到其最近邻的距离;当以上条件都不满足时,σ_i 就等于 x_i 到其第$\lfloor k \rfloor$个近邻和第$\lceil k \rceil$个近邻距离的线性组合,且 k 值越接近于$\lceil k \rceil$(或$\lfloor k \rfloor$)值,其对应的线性权重($k-\lfloor k \rfloor$)(或$\lceil k \rceil - k$)就越大.这就是我们提出的基于连续 k 最近邻的局部尺度全连接图构建方式.该构建方式不仅保留了基于整数 k 最近邻的全连接图构建方式在不抛弃 k-近邻以外数据信息的前提下处理多尺度数据集的优点,同时又克服了原方法由于参数范围过窄而导致的参数调节能力弱化的缺点,从而在参数的鲁棒性和灵活性两者之间达到了一种较好的平衡.

定义 4(类别相似度矩阵). 基于训练数据类别信息 Υ_S 的类别相似度矩阵为

$$W_{\Upsilon_S} \overset{\text{def}}{=} Y_S Y_S^T \tag{16}$$

其中,$Y_S=(y_{ij})_{m\times p}$,

$$Y_S(i,j)=\delta(y_i,c_j), \forall y_i \in \Upsilon_S, c_j \in C \tag{17}$$

即 $y_{ij}=1$,当且仅当 x_i 属于 c_j 类;否则 $y_{ij}=0$.

因此 $W_{\Upsilon_S}(y_i, y_j)=1$,当且仅当 $y_i=y_j$;否则, $W_{\Upsilon_S}(y_i, y_j)=0$.

定义 5(组合相似度矩阵). 基于条件相似度矩阵和类别相似度矩阵的组合相似度矩阵为

$$W_S \overset{\text{def}}{=} \alpha W_{X_S} + (1-\alpha) W_{\Upsilon_S} \tag{18}$$

其中,权衡因子$\alpha \in [0,1]$用于调节条件相似度矩阵 W_{X_S} 和类别相似度矩阵 W_{Υ_S} 两者的比重.

由公式(14)和定义 4 中的 $W_{\Upsilon_S}(y_i, y_j) \in \{0,1\}$ 可知,当$\alpha<1$ 时,组合相似度矩阵 W_S 中同类数据之间的相似度得到了增强,而异类数据之间的相似度则得到了减弱.由于采用了线性组合的形式,W_S 保留了相似度矩阵的两个基本性质:(1) W_S 的每个元素都在[0,1]范围内;(2) W_S 的各对角线元素(即自相似度)都等于 1.

定义 6(归一化的组合相似度矩阵). 归一化的组合相似度矩阵为

$$\tilde{W}_S \overset{\text{def}}{=} D_S^{-1/2} W_S D_S^{-1/2} \tag{19}$$

其中,

$$D_S = diag(W_S \mathbf{1}_m) \tag{20}$$

这里,$\tilde{W}_S$ 的定义形式对应着归一化拉普拉斯矩阵(定义 2)的定义形式.

定理 1. 用组合相似度矩阵 W_S 及其对应的度数对角阵 D_S 替换公式(11)中的一般相似度矩阵 W 及其对应的度数对角阵 D,可得

$$\begin{aligned} &\max_{F_S} \ tr(F_S^T W_S F_S) \\ &\text{s.t.} \quad F_S^T D_S F_S = I \end{aligned} \tag{21}$$

此时,归一化割的松弛解是

$$F_S = D_S^{-1/2}[v_1,...,v_p] \tag{22}$$

其中,$v_1,\ldots,v_p$ 是 $\tilde{W}_S$ 最大的 p 个特征值对应的特征向量,即 $v_1,\ldots,v_p$ 满足方程

$$\tilde{W}_S v_i = \lambda_i v_i \tag{23}$$

且$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$.

证明:令 $\tilde{F}_S = D_S^{1/2} F_S$,则公式(21)可重写为

$$\max_{F_S} tr(\tilde{F}_S^T \tilde{W}_S \tilde{F}_S)$$
$$\text{s.t.} \quad \tilde{F}_S^T \tilde{F}_S = I \tag{24}$$

可以得到 $\tilde{F}_S$ 解的形式为

$$\tilde{F}_S = [v_1,...,v_p] \tag{25}$$

其中,$v_1,\ldots,v_p$ 为两两正交的单位向量.根据 Courant-Fischer 定理可知:

$$tr(\tilde{F}_S^T \tilde{W}_S \tilde{F}_S) = \sum_{i=1}^{n} v_i^T \tilde{W}_S v_i \leqslant \sum_{i=1}^{n} \lambda_i \tag{26}$$

$\lambda_1,\ldots,\lambda_p$ 是 $\tilde{W}_S$ 最大的 p 个特征值,当等号成立时,v_i 是 λ_i 对应的 $\tilde{W}_S$ 的特征向量.故公式(21)的最优解为

$$F_S = D_S^{-1/2} \tilde{F}_S \tag{27}$$

其中, $\tilde{F}_S = [v_1,...,v_p]$,且 $v_1,\ldots,v_p$ 是 $\tilde{W}_S$ 最大的 p 个特征值对应的特征向量. □

为了方便起见,人们通常在谱聚类算法中简单地令

$$F_S=[v_1,\ldots,v_p] \tag{28}$$

并因此称其为“谱”方法.本文中我们遵循这一习惯,并用 $F_S=(f_{ij})_{m\times p}$ 表示监督式谱空间变换后的训练数据新坐标,则它的每个行向量 $f_i=[f_{i1},\ldots,f_{ip}]$就对应着输入数据 x_i 的映射后坐标.由于公式(18)中 W_S 的同类数据相似度得到了强化,异类数据相似度得到了弱化,根据谱方法的优化目标可知,公式(28)的最优解可以确保属于相同类别的训练数据在新的特征空间中被映射得尽可能彼此靠近,而不同类别的训练数据则在新特征空间被映射得尽可能相互远离.

为了防止监督式谱空间映射产生过于靠近原点的数据坐标不利于分类,S^3C 还对 F_S 进行了如下行归一化处理:

$$Z_S = D_V^{-1/2} F_S \tag{29}$$

其中,

$$D_V = diag(diag(F_S F_S^T)) \tag{30}$$

则 $Z_S = [z_1^T, z_2^T,..., z_m^T]^T$ 就是映射到监督式谱空间中的单位超球面上的训练数据坐标矩阵,它的每个行向量 z_i 就是 f_i 行归一化后的结果,也对应着输入数据 x_i.

定理 2. 假设 z_S 在监督式谱空间中线性可分,将 $z_S=\{z_1,\ldots,z_m\}$代入公式(2),用$\langle z_i,z_j\rangle$替换$\langle x_i,x_j\rangle$,则最优化问题:

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \langle z_i, z_j\rangle \alpha_i \alpha_j - \sum_{j=1}^{m} \alpha_j$$
$$\text{s.t.} \quad \sum_{i=1}^{m} y_i \alpha_i = 0, \tag{31}$$
$$0 \leqslant \alpha_i \leqslant C, i = 1,...,m$$

的解就是在监督式谱空间内最大化异类数据之间几何间隔的最优分割超平面.

证明:考虑二分类问题,S^3C 的优化目标是在监督式谱空间中寻找一个能够最大化两类数据之间几何间隔的线性分割超平面;如果不存在这样的超平面,就采用软边界技术,寻找能够最小化两类数据分类误差的超平面作为代替.由此,我们可以推得类似于公式(1)的优化问题:

$$\min_{w,b} \frac{1}{2}\| w\|^2 + C\sum_{l=1}^{m} \xi_i$$
$$\text{s.t.} \quad y_i(\langle w, z_i\rangle + b) \geqslant 1 - \xi_i,\ i = 1,...,m, \tag{32}$$
$$\xi_i \geqslant 0,\ i = 1,...,m$$

通过对公式(32)的对偶问题进行变形,即可得其等价问题公式(31),具体推导过程可参考文献[7].因此,公式(31)的解就是 S^3C 在监督式谱空间内最大化异类数据之间几何间隔的最优分割超平面. □

在处理多分类问题时,S^3C 采用一对一(one-versus-one)的组合策略,即对任意两类样本之间构造一个最优

分割超平面,p 个类别的样本就对应着 $p(p-1)/2$ 个超平面;当对未知样本进行分类时,只需结合 $p(p-1)/2$ 个超平面进行投票,得票最多的类别即为该未知样本的类别.

对于未知类别的测试数据,S^{3}C 首先把测试数据 X_T 映射到与训练数据 Z_S 相同的特征空间中去,然后再用已构建的最优分割超平面对映射后的测试数据进行分类.图 1 演示了将 X_T 映射到 Z_S 所在的监督式谱空间内的示意图.图中 $\mathcal{M}$ 表示输入数据 X 所在的原始空间,$\mathcal{H}$ 表示 Z_S 所在的目标空间,f 表示 $X_S \mapsto Z_S$ 的映射.由于映射 f 需要使用被映射数据的类别信息进行空间变换,因此不可能直接将测试数据 $X_T \in \mathcal{M}$ 通过 f 变化到目标空间 $\mathcal{H}$ 中去.为此,我们在 $\mathcal{M}$ 和 $\mathcal{H}$ 之间构造了一个起过渡作用的"桥"$\mathcal{N}$:只要保证存在一个从 $\mathcal{M}$ 到 $\mathcal{N}$ 的映射 ζ 和一个从 $\mathcal{H}$ 到 $\mathcal{N}$ 的映射 η 及其逆映射 η^{-1},那么只需确定 $\mathcal{M}$ 和 $\mathcal{H}$ 在 $\mathcal{N}$ 中像之间的关系,就可以用两步走的形式将测试数据通过 ζ 和 η^{-1} 将 X_T 合理地映射到 $\mathcal{H}$ 中去.

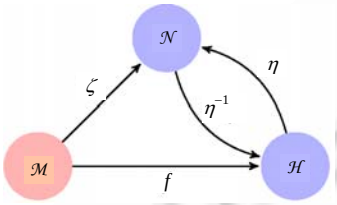


Fig.1 Illustration of the mapping from the original space to the supervised spectral space

图 1 原空间到监督式谱空间的映射示意图

定义 7. 测试数据 X_T 与训练数据 X_S 在原始空间 $\mathcal{M}$ 内的归一化相似度矩阵为

$$\tilde{W}_{X_T} \overset{\text{def}}{=} D_r^{-1/2} W_{X_T} D_c^{-1/2} \tag{33}$$

其中,$W_{X_T} = (w_{X_T})_{(n-m)\times m}$,$w_{X_T}$ 的定义与公式(14)相同:

$$D_r = diag(W_{X_T} \mathbf{1}_m) \tag{34}$$

$$D_c = diag(\mathbf{1}_{n-m}^T W_{X_T}) \tag{35}$$

定义 8($\mathcal{M}$ 到 $\mathcal{N}$ 的映射 ζ). 定义输入空间 $\mathcal{M}$ 到过渡空间 $\mathcal{N}$ 的映射 ζ 为

$$\zeta_{X_S}(X_T) \overset{\text{def}}{=} D_V^{-1/2} \tilde{W}_{X_T} \tag{36}$$

其中,D_V 的定义见公式(30).

定义 9. 任意数据 $Z \in \mathcal{H}$ 与映射后训练数据 Z_S 在目标空间 $\mathcal{H}$ 内的相似度矩阵为

$$W_Z \overset{\text{def}}{=} Z_s \Lambda Z^T \tag{37}$$

其中,$\Lambda=diag(\lambda_1,\lambda_2,\ldots,\lambda_p)$,$\lambda_i$ 是公式(23)中 $\tilde{W}_S$ 的第 i 个最大的特征值.

定义 10($\mathcal{H}$ 到 $\mathcal{N}$ 的映射 η). 定义目标空间 $\mathcal{H}$ 到过渡空间 $\mathcal{N}$ 的映射 η 为

$$\eta_{Z_S}(Z) \overset{\text{def}}{=} D_V^{1/2} W_Z \tag{38}$$

其中,D_V 的定义见公式(30).

由定义 7~定义 10 可知,我们将图 1 中的"桥"$\mathcal{N}$ 定义为 $\mathcal{M}$ 和 $\mathcal{H}$ 上测试数据关于训练数据的局部坐标空间,因此,$X_T \in \mathcal{M}$ 在 $\mathcal{N}$ 上的像应该与其变换后的 $Z_T \in \mathcal{H}$ 在 $\mathcal{N}$ 上的像是一致的.

定理 3. 测试数据 X_T 在监督式谱空间内的最佳映射为 $Z_T = \tilde{W}_{X_T} Z_S \Lambda^{-1}$.

证明:因为 $Z_S = D_V^{-1/2} F_S$ 和 $F_S^T F_S = I_{p\times p}$,由图 1 的示意图可知,测试数据 X_T 在 $\mathcal{H}$ 内的最佳映射应该是能够最小化 $X_T \in \mathcal{M}$ 和 $Z \in \mathcal{H}$ 在 $\mathcal{N}$ 上所成像之间差异的最优 Z 值,即

$$\begin{aligned} Z_T &= \arg\min_{z} \| \zeta_{X_S}(X_T) - \eta_{Z_S}(Z) \|^2 \\ &= \arg\min_{z} \| D_V^{-1/2} D_r^{-1/2} W_{X_T} D_c^{-1/2} - D_V^{1/2} Z_S \Lambda Z^T \|^2 \\ &= (D_c^{-1/2} W_{X_T}^T D_r^{-1/2}) Z_S (Z_S^T D_V Z_S)^+ \Lambda^{-1} \\ &= (D_c^{-1/2} W_{X_T}^T D_r^{-1/2}) Z_S (F_S^T D_V^{-1/2} D_V D_V^{-1/2} F_S)^+ \Lambda^{-1} \\ &= (D_c^{-1/2} W_{X_T}^T D_r^{-1/2}) Z_S (F_S^T F_S)^+ \Lambda^{-1} \\ &= \tilde{W}_{X_T} Z_S \Lambda^{-1} \end{aligned} \tag{39}$$

为了确保映射后的测试数据具有和 Z_S 相同的单位范数,Z_T 也需要进行如下行归一化处理:

$$Z_T = D_Z^{-1/2} Z_T \tag{40}$$

其中,

$$D_Z = diag(diag(Z_T Z_T^T)) \tag{41}$$

最后,我们只需应用由定理 2 构建的最优分割超平面对 Z_T 直接进行分类即可. □

图 2 给出了算法 S^{3}C 的具体描述.其中,第 1 步~第 3 步用于构建关于训练数据的组合相似度矩阵,第 4 步、第 5 步将训练数据映射到监督式谱空间中,第 6 步对变化后的训练数据进行单位超球面投影,第 7 步、第 8 步寻找最大化几何间隔的最优分割超平面,第 9 步~第 11 步将测试数据映射到监督式谱空间,第 12 步对映射后的测试数据进行单位超球面投影,第 13 步、第 14 步使用由第 7 步、第 8 步构建的分类超平面对测试数据进行分类,最后,第 15 步返回测试数据分类结果.

ALGORITHM. S^{3}C($\mathcal{X}_S$,$\mathcal{Y}_S$,$\mathcal{X}_T$).
Input:训练数据集 $\mathcal{X}_S$,训练数据标签集 $\mathcal{Y}_S$,测试数据集 $\mathcal{X}_T$;
Output:测试数据标签集 $\mathcal{Y}_T$.

1 Compute W_{X_S} with $w_{X_S}(x_i,x_j)=\exp(-\| x_i-x_j \|^2/(\sigma_i\sigma_j))$, σ_i, σ_j defined in Eq.(15)
2 $W_{Y_S} \leftarrow Y_S Y_S^T$, where $Y_S(i,j)=\delta(y_i,c_j), \forall y_i \in \mathcal{Y}_S, c_j \in C$
3 $W_S \leftarrow \alpha W_{X_S} + (1-\alpha) W_{Y_S}$
4 Solve $D_S^{-1/2} W_S D_S^{-1/2} = \lambda_i v_i$, where $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_p$, D_S=$diag(W_S \mathbf{1}_m)$
5 $F_S \leftarrow [v_1, v_2, \ldots, v_p]$, $\Lambda \leftarrow diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$
6 $Z_S \leftarrow diag(diag(F_S F_S^T))^{-1/2} F_S$
7 $\alpha^* \leftarrow \arg\min_{\alpha} \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m y_i y_j (z_i^T z_j)\alpha_i\alpha_j - \sum_{j=1}^m \alpha_j$, s.t. $\sum_{i=1}^m y_i\alpha_i = 0$, $0 \leqslant \alpha_i \leqslant C$, $i=1,...,m$
8 $b^* \leftarrow y_j - \sum_{i=1}^m y_i \alpha_i^* (z_i^T z_j)$, $\forall j$ satisfies $\alpha_j^* \in (0,C)$
9 Compute W_{X_T} with $w_{X_T}(x_i,x_j)=\exp(-\| x_i-x_j \|^2/(\sigma_i\sigma_j))$, $\forall x_i \in \mathcal{X}_T$, $x_j \in \mathcal{X}_S$, σ_i, σ_j defined in Eq.(15)
10 $\tilde{W}_{X_T} \leftarrow D_r^{-1/2} W_{X_T} D_c^{-1/2}$, whese $D_r = diag(W_{X_T}\mathbf{1}_m)$, $D_c = diag(\mathbf{1}_{n-m}^T W_{X_T})$
11 $Z_T \leftarrow \tilde{W}_{X_T}^T Z_S \Lambda^{-1}$
12 $Z_T \leftarrow diag(diag(Z_T Z_T^T))^{-1/2} Z_T$
13 $y_j \leftarrow \text{sgn}(\sum_{i=1}^m y_i \alpha_i^* (z_i^T z_j) + b^*)$, $\forall j \in \{m+1, m+2, ..., n\}$
14 $\mathcal{Y}_T \leftarrow \{y_{m+1}, y_{m+2}, \ldots, y_n\}$
15 return $\mathcal{Y}_T$

Fig.2 Description of the supervised spectral space classifier

图 2 监督式谱空间分类器的描述

3 改进的监督式谱空间分类器

S^{3}C 算法在处理多分类问题时需要结合多个分割超平面进行投票才能得到最终的分类结果.为此,我们提出了一种改进的监督式谱空间变换多分类算法 S^{3}T.S^{3}T 采用线性子空间变换和强迫一致的方法,通过求解从监督式谱空间到类别指示空间的线性变换矩阵,将监督式谱空间内的训练和测试数据(Z_S,Z_R)进一步线性变换到指定的类别指示空间中去,从而获得关于测试数据的类别指示矩阵,并在此基础上进行类别预测.S^{3}T 不但可以

直接应用于多分类问题而无需任何组合策略,而且由于其线性子空间变换是基于整体训练数据集的,因此还适用于可能存在类别标号错误的分类问题,性能更加鲁棒.

S^3T 在保留了与 S^3C 相同的监督式谱空间映射的基础上,将定义 4 中的 Y_S 看作一个类别指示空间,通过求解从 Z_S 到 Y_S 的线性转换矩阵 R,使 Z_SR 与 Y_S 尽可能地保持一致,即

$$\min_R \| Z_S R - Y_S \| \tag{42}$$

通常,R 也可以看作是从监督式谱空间到类别指示空间的线性变换矩阵.通过求解公式(42),我们可得

$$R = (Z_S^T Z_S)^+ Z_S^T Y_S \tag{43}$$

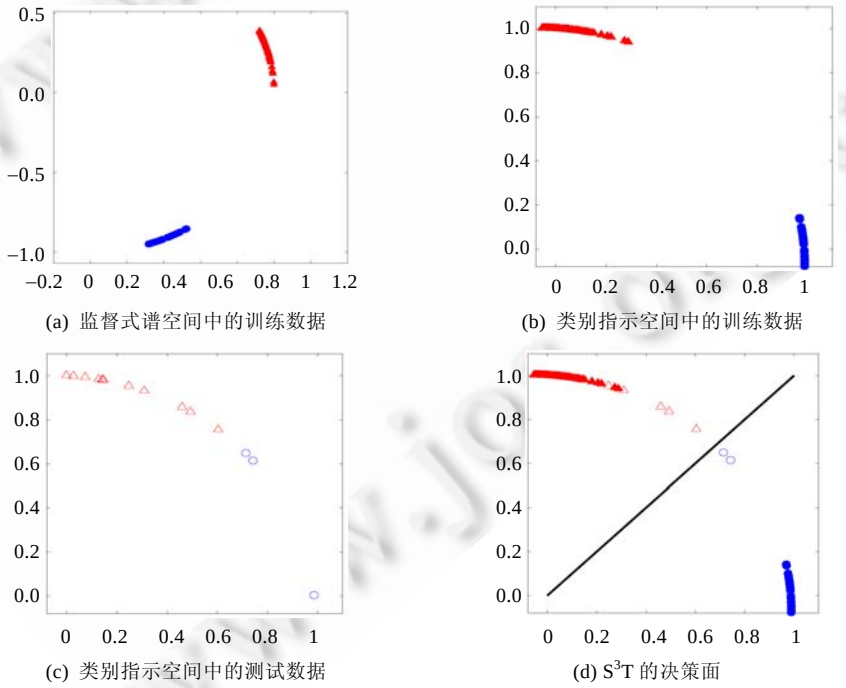
此时,令映射到监督式谱空间的测试数据 Z_T 作同样的线性变换,右乘 R 得

$$F_T = Z_T R = Z_T (Z_S^T Z_S)^+ Z_S^T Y_S \tag{44}$$

从而把测试数据也变换到与 Y_S 相同的类别指示空间中去,则 $F_T=(f_{ij})_{(n-m)\times p}$ 就是对应测试数据 X_T 的类别指示矩阵,其行向量 f_i 就是数据点 x_i 的类别指示向量.最后,S^3T 通过如下公式预测 $x_i \in X_T$ 的类别:

$$y_i = \arg\max_j f_{ij} \tag{45}$$

图 3 演示了 S^3T 算法在 UCI 数据库中的 parkinsons 数据集上所进行的空间映射,以及在变化后的特征空间内所做的分类决策面.在图 3(a)中,训练数据已经被 S^3C 映射到监督式谱空间中的一个单位圆上.由于在训练数据的空间映射中我们已经加入了数据的判别信息,不同类别的训练数据在监督式谱空间中被映射得相互远离.S^3T 算法进一步将监督式谱空间中的训练数据和测试数据变换到与 Y_S 相同的类别指示空间中去(如图 3(b)、图 3(c)所示).由于测试数据的空间映射是建立在与训练数据之间的相似度关系上的,因此大多数测试数据都被较好地映射到与其真实类别相同的训练数据附近,而另一小部分可能会由于数据集本身的噪声或者是测试数据在原始空间中过于接近决策边界的位置而产生了混淆(如图 3(c)所示).根据公式(45)所给出的类别预测公式,S^3T 在类别指示空间中做了一条始于原点且倾斜角为 45°的分类决策面(如图 3(d)所示),对测试数据给出了准确的分类结果.



(a) 监督式谱空间中的训练数据 (b) 类别指示空间中的训练数据

(c) 类别指示空间中的测试数据 (d) S^3T 的决策面

Fig.3 Space transformation and classification result of S^3T on parkinsons data set

图 3 S^3T 对 parkinsons 数据集的空间变换和分类

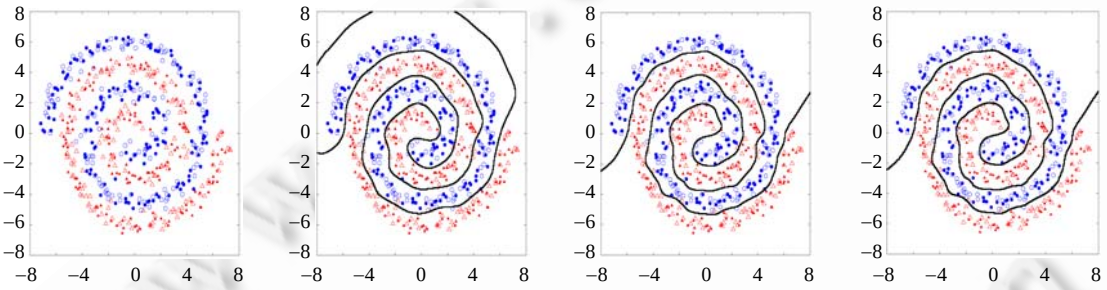
S^3C(S^3T)不同于 SVM 将数据变换到不可观测的高维或无穷维特征空间,而是通过将输入数据映射到低维的监督式谱空间(类别指示空间),使研究者可以直观地观察到变化后的新特征空间、映射后的数据新坐标以及在新特征空间中所做的分类决策面,因此更有利于我们对算法的评价和对参数的选择.

4 实 验

在本节中,我们全面评价了 S^3C 和 S^3T 两种算法的分类性能,并将其与 C4.5 决策树算法、支持向量机 SVM、流形学习分类器 KLPP+Linear SVM、监督降维分类算法 LDA 及其核化版本 KDA 进行了比较.首先,我们演示了 RBF-SVM,S^3C 和 S^3T 在人工数据集上的分类决策面和分类测试错误率.然后,我们使用了 8 个 UCI 真实数据集和 1 个钢笔手写数字识别问题比较了 C4.5,linear SVM,RBF-SVM,KLPP+Linear SVM,LDA, KDA,S^3C 以及 S^3T 的平均分类错误率.最后,我们还讨论了 S^3C 和 S^3T 的两个重要参数——近邻参数 k 和权衡因子 α 对其分类性能的影响.

4.1 人工数据集

图 4 演示了 RBF-SVM,S^3C 和 S^3T 这 3 种算法在 Spiral 人工数据集上的分类决策面.图中实心的符号表示训练数据,空心的符号表示测试数据,圆圈和三角分别表示两种不同类别的数据.表 1 给出了 RBF-SVM,S^3C 和 S^3T 在 Spiral 数据集上的分类测试错误率.



(a) 原始数据集 (b) SVM 的决策边界 (c) S^3C 的决策边界 (d) S^3T 的决策边界

Fig.4 Decision boundaries of SVM,S^3C and S^3T on Spiral data set

图 4 SVM,S^3C 和 S^3T 在 Spiral 数据集上的决策边界

Table 1 Comparison of the test error ratios of RBF-SVM, S^3C and S^3T on Spiral data set

表 1 RBF-SVM,S^3C 和 S^3T 在 Spiral 数据集上的分类错误率比较

数据集	RBF-SVM	S^3C	S^3T
Spiral	6.00	5.00	5.33

从图 4 可以看出,RBF-SVM 倾向于用一条闭合的曲线将一类数据从另一类数据分开(如图 4(b)所示),而 S^3C 和 S^3T 的分类决策面则都用一条开放的曲线将两类数据分开(如图 4(c)、图 4(d)所示),这是由 3 种算法选择不同的映射空间所造成的.尽管从图上来看,RBF-SVM,S^3C 和 S^3T 对图中测试数据的分类结果较为相似,但实际上,如果存在任意一个测试数据点落在 RBF-SVM 所画的闭合曲线之外,那么 S^3C 和 S^3T 对该数据点的分类结果就会与 SVM 截然不同.按照人们通常所接受的近邻数据类别相似的假设,我们没有理由将处于同类训练数据一侧的相邻测试数据划分为不同的类别,而且一般来说,开放的决策边界也比闭合的决策边界具有更好的泛化性.因此,S^3C 和 S^3T 在 Spiral 数据集上所作的开放式决策边界要比 RBF-SVM 所作的闭合式决策边界更加合理.

4.2 真实数据集

为了进一步考察 S^3C 和 S^3T 算法的分类性能,我们选择了 8 个 UCI 真实数据集(包括 4 个二分类问题和 4 个多分类问题)和 1 个钢笔手写数字识别问题,用于比较 C4.5 决策树算法、Linear SVM、RBF-SVM、KLPP+

Linear SVM、LDA、KDA、S^3C 以及 S^3T 算法在这些测试数据集上的分类错误率.其中,KLPP 是近几年来较受关注的流形学习算法 LPP 的核化版本,我们将其与 Linear SVM 分类器相结合作为流形学习分类算法的代表,与本文的监督式谱空间算法进行比较.

在算法实现方面,我们采用了 Fast C4.5[23]作为 C4.5 算法的实现,LibSVM[24]作为线性 SVM 和 RBF-SVM 的实现,LDA,KDA 及 KLPP 采用 Cai 等人的实现(http://www.zjucadcg.cn/dengcai/Data/data.html),并用 Matlab 实现了我们的 S^3C 和 S^3T 算法.在参数设置方面,线性 SVM 的参数 C、RBF-SVM 的参数($\gamma=1/2\sigma^2$,C)、KDA 的参数 γ 和 KLPP+Linear SVM 的参数($\gamma=1/2\sigma^2$,C)均在$\{2^{-15},2^{-13},2^{-11},\ldots,2^{13},2^{15}\}$范围内进行优化,$S^3C$ 的最大化间隔参数 C 被设定为 C=2,S^3C 和 S^3T 的权衡因子被设定为 α=0.9.S^3C 和 S^3T 唯一优化的近邻参数 k 在(0,10)范围内以 0.1 的精度进行搜索.本文中所有的参数选择都是通过 10 折交叉验证来实现的,即将训练数据随机均分为 10 份,其中任意 9 份用于对不同参数进行分类器构建,剩余 1 份用于对分类器性能进行评价.最后对 10 折交叉验证分类错误率进行平均后,我们选取其中平均错误率最低的一组参数作为优化后的参数以进行最终的分类实验.

4.2.1 UCI 数据集

表 2 给出了 8 个 UCI 测试数据集的基本信息,表 3 则汇总了 C4.5,Linear SVM,RBF-SVM,KLPP+Lin SVM,LDA,KDA,S^3C 和 S^3T 这 8 种算法在 UCI 测试数据集上的 10 折交叉验证平均分类错误率及其标准差.

Table 2 Summary of test data sets

表 2 测试数据集汇总

数据集	数据个数	维数	类别数
Parkinsons	195	22	2
Sonar	208	60	2
Glass	214	9	7
Ionosphere	351	34	2
Breast	683	9	2
Vowel	990	10	11
Wine-Red	1 599	11	11
Uspst	2 007	256	10

Table 3 Comparison of the averaged test error ratios of the 8 classification algorithms on the UCI data sets

表 3 8 种分类算法在 UCI 数据集上的平均分类错误率比较

数据集	C4.5	Linear SVM	RBF-SVM	KLPP+Linear SVM
Parkinsons	16.6±7.83	15.39±5.51	11.71±4.61	17.42±3.51
Sonar	26.1±7.95	23.14±9.71	11.57±6.54	24.00±6.24
Glass	31.6±9.31	36.90±7.38	25.28±10.69	47.64±6.59
Ionosphere	10.2±5.48	13.97±6.25	4.83±3.00	35.03±3.68
Breast	4.8±2.15	3.66±1.86	2.78±1.62	3.81±1.72
Vowel	21.4±3.53	23.03±3.62	0.40±0.52	45.76±5.81
Wine-Red	37.6±3.65	41.59±4.42	38.52±2.17	51.09±4.15
Uspst	19.2±2.01	19.57±1.97	5.58±1.32	10.71±2.32
数据集	LDA	KDA	S^3C	S^3T
Parkinsons	16.87±8.94	15.39±5.45	10.76±2.91	**10.29±6.45**
Sonar	25.55±9.24	13.45±7.43	10.10±5.28	**9.74±8.52**
Glass	52.75±11.53	47.19±9.36	**20.41±9.69**	21.00±7.58
Ionosphere	23.99±14.78	5.42±2.85	**2.86±2.33**	**2.86±3.81**
Breast	3.96±2.61	7.46±2.95	**2.20±2.00**	**2.20±1.73**
Vowel	39.9±4.5	**0.20±0.64**	0.51±1.28	0.30±0.49
Wine-Red	64.29±8.98	39.90±3.49	35.58±3.81	**35.27±3.67**
Uspst	13.56±2.1	18.69±1.22	**5.38±1.32**	**5.38±0.66**

在表 3 中,S^3C 和 S^3T 算法的平均分类错误率在 8 个测试数据集上均低于 C4.5 决策树算法、线性 SVM、RBF-SVM、有监督降维分类算法 LDA 和流形学习分类器 KLPP+Linear SVM,并且在 7 个数据集上低于 KDA.LDA 和 KLPP+Linear SVM 的分类性能是最不稳定的,它们在有的数据集上表现良好,但同时又在其他一些数据集上差得离谱.前者是因为 LDA 是线性分类算法,而且 LDA 基于的高斯分布假设并不总是被测试数据所满足,当假设与实际情况存在偏差时,就会出现分类错误;后者则是因为 KLPP 是无监督的降维算法,它的降维结果是

否有利于后续的分类算法完全是无法预测的,没有任何保证.C4.5 的分类性能略优于 LDA 和 KLPP+Linear SVM,也更加稳定一些.KDA 由于采用了核技术,将数据映射到高维隐空间,对假设的依赖性有一定的减弱,分类错误率高于 LDA.相比之下,S^3C 和 S^3T 的分类错误率是 8 种算法中最低的,其中,S^3C 的分类错误率远低于不进行空间变化的 Linear SVM,这说明 S^3C 进行的监督式谱空间变换有利于改进分类器的分类错误率.另一方面,S^3C 的分类错误率也在绝大多数数据集上低于 RBF-SVM,这说明,尽管 S^3C 将输入数据映射到低维的监督式谱空间,但仍可达到比 RBF-SVM 将数据映射到无限维特征空间中更好的数据分离效果.在参数敏感性方面,我们发现 RBF-SVM 对它的两个参数$\gamma=1/2\sigma^2$ 和 C 都非常敏感,KDA 和 KLPP+Linear SVM 也对其参数γ较敏感,而 S^3C 由于在空间映射中加入了监督信息,总能保证不同类别的训练数据在监督式谱空间中被清楚地分离开来,因此其软边界参数 C 实际上并未起到太大的作用,这也是我们一开始就在实验设置中将 S^3C 的参数 C 设定为默认值的原因.此外,改进后的 S^3T 算法在大多数数据集上的分类错误率都明显低于其他 7 种算法,包括 S^3C 在内.

4.2.2 钢笔手写数字的识别

对于大数据集,本文测试了 S^3C 和 S^3T 算法在钢笔手写数字识别(pen-based recognition of handwritten digits)问题上的分类性能.Pendigit 数据集(http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits)是通过收集 44 个人在对压力敏感的平板上每人写 250 个数字所获得的,它总共包含 10 992 个数据、16 个特征属性和 1 个类别属性(数字 0~9).数据集提供者将数据预先划分成训练数据和测试数据,其中,训练数据有 7 494 个样本,是 44 个人中的 30 个人所写的数字样本;测试数据则包含有 3 498 个样本,是另外 14 个人所写的数字样本.这种划分有利于评价分类器独立于写者的分类性能.

表 4 列出了 C4.5,Linear SVM,RBF-SVM,KLPP+Linear SVM,LDA,KDA,S^3C 和 S^3T 在 pendigits 数据集上的分类错误率比较,其中,KDA 和 KLPP+Linear SVM 在运行时出现内存不足现象,无法得到最终结果.从表中易见,S^3T 相对于其他算法在 pendigits 数据集上表现出更低的分类错误率和更好的分类性能.

Table 4 Comparison of the test error ratios of the 8 classification algorithms on pendigits data set

表 4 8 种分类算法在 pendigits 数据集上的分类错误率比较

C4.5	Linear SVM	RBF-SVM	KLPP+Lin-SVM	LDA	KDA	S^3C	S^3T
7.78	4.66	1.8	—	18.50	—	2.77	**0.09**

4.3 参数讨论

S^3C 和 S^3T 算法都涉及两个重要的参数,分别是近邻参数 k 和权衡因子 α.下面我们分别就这两个参数对 S^3T 分类性能的影响进行讨论.

4.3.1 近邻参数 k

图 5 演示了 8 个 UCI 数据集上关于近邻参数 k 的分类测试错误率曲线(α=0.9),图中的圆圈标记了各个数据集上最优实数 k(精度为 0.1)所对应的分类错误率,三角形则标记了最优整数 k 所对应的分类错误率.

在图 5 中,相比于基于整数 k 的全连通图构建方式,我们提出的基于连续 k 近邻的全连通图构建方式使 S^3T 的分类错误率得到了 2%~7%的改进.有趣的是,大多数最优的 k 值都小于或接近于 1.由公式(14)、公式(15)可知,当 k<1 时,每个训练数据与其他数据之间的条件相似度就会变得很低.因此,在我们将条件相似度矩阵 W_{X_S} 和类别相似度矩阵 W_{Y_S} 进行线性组合之后,所得到的组合相似度矩阵 W_S 就会变成一个近似块对角阵.不难想象,对这样的 W_S 进行归一化后特征值分解,所得到的映射后训练数据新坐标将会是坍塌的.也就是说,如果有 p 类训练数据,那它们就会在低维的监督式谱空间中被映射为 p 个数据点.从中我们可以得出以下两个结论:

(1) 分类与传统的维数约减不同,维数约减的目的是尽可能地维持原始数据潜在的流形结构,而分类的目的则是尽可能地挖掘出输入数据与类别标签之间的关系.因此,虽然数据坍塌在传统的维数约减中是一种需要极力避免的不利情况,但它并不一定会对分类产生负面影响;相反,我们的 S^3T 算法事实上还因为数据坍塌所造成的数据压缩而使得分类过程更加简单,分类结果更加优化;

(2) 当k<1时,组合相似度矩阵W_S主要是由类别相似度矩阵W_{Y_S}所构成,这说明在大多数数据集取最优的k≤1参数值时,监督信息对于提高 S^{3}T 算法的分类性能起着重要的作用.此外,虽然在大多数数据集上 k 的最优参数值小于等于 1,但当 k 取其最小值 k=0.1 时,总会导致很高的分类错误率.这可能是因为当 k 取值很小时,S^{3}T 基于定义 7 计算得到的测试数据与训练数据之间的相似度w_{X_T}就会变得很小,无法为测试数据从原始空间到监督式谱空间的映射提供足够的信息,从而影响了 S^{3}T 的分类错误率.

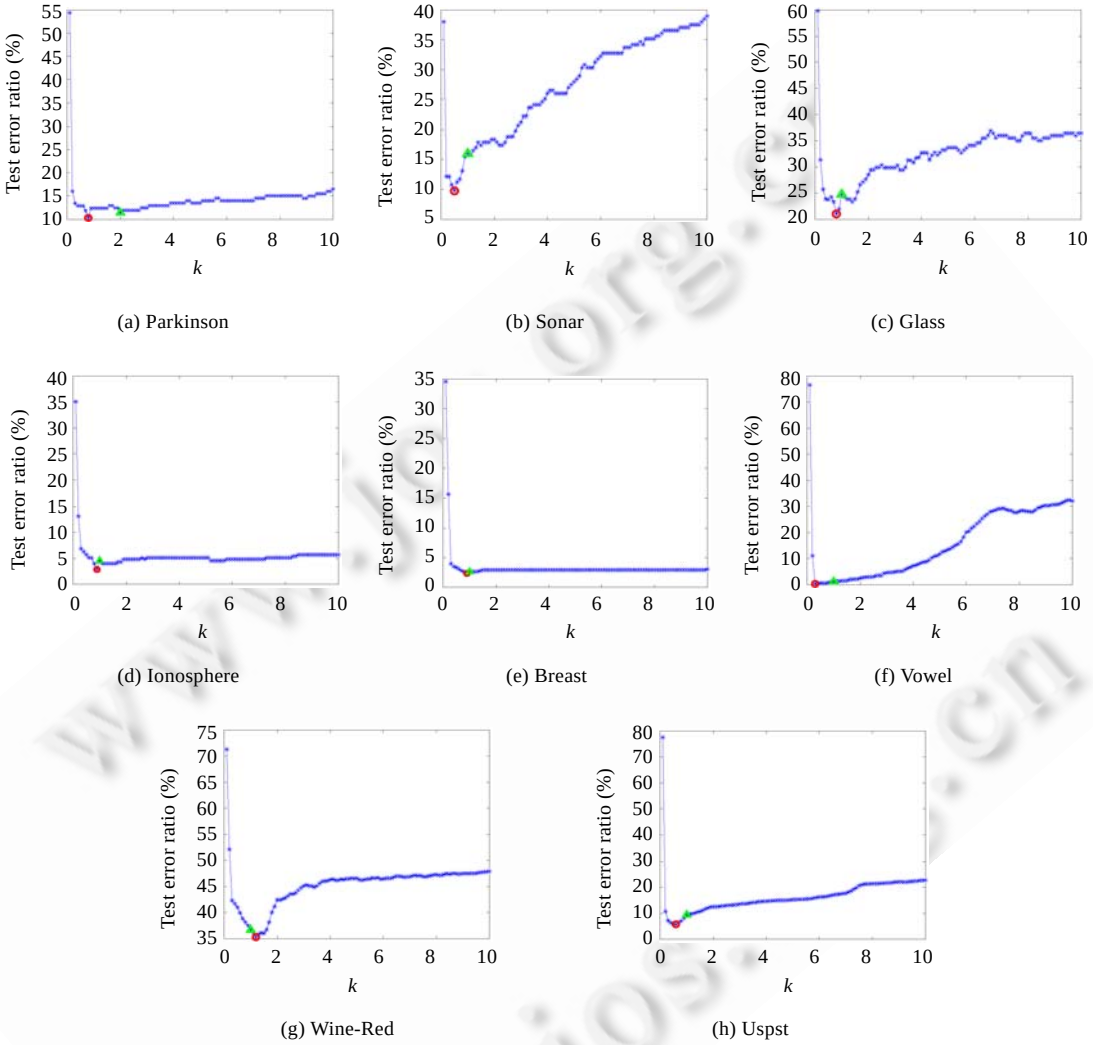


Fig.5 Learning curves of the relationship between k and the test error ratios of S^{3}T on the 8 UCI data sets

图 5 S^{3}T 在 8 个 UCI 测试数据集上的关于 k 的分类错误率曲线

4.3.2 权衡因子α

为了研究参数α对 S^{3}T 算法分类性能的影响,我们分别设定了 5 个不同的 k 参数值{0.1,0.5,1,5,10},然后对α∈[0,1](精度为 0.1)的分类错误率进行了比较.这里选择 k={0.1,0.5,1,5,10}的原因是在第 4.3.1 节中我们已经发现,在 8 个 UCI 测试数据集上的最优参数 k 值更集中于[0,1]范围内,因此,采用这种具有递增间隔的 k 设置方法可能更能体现参数α对分类错误率的影响.图 6 演示了 S^{3}T 在 8 个 UCI 数据集上关于参数α的分类错误率曲线.

由组合相似度矩阵的定义可知,α值表示条件相似度矩阵的权重,(1−α)值则表示类别相似度矩阵的权重,α

值越小,监督信息的权重就越大.当α=0时,组合相似度矩阵就等价于类别相似度矩阵;而当α=1时,组合相似度矩阵就是没有任何监督信息的条件相似度矩阵对应先用谱方法降维后再进行分类的分类算法.在图 6 中,关于参数α的分类错误率曲线明显要比图 5 中关于参数 k 的分类错误率曲线平坦得多,唯一的例外就是当α=1 时.这说明,只要有监督信息被加入到组合相似度矩阵的构建当中(α=0),在参数 k 固定的情况下,不同的α值对分类错误率的影响并不是很大;而α=1 所对应的高分类错误率则验证了先采用谱方法无监督降维再分类的算法远不如本文所提出的监督式谱空间分类器,而前面第 4.2 节中讨论比较的流形学习分类器 KLPP+Linear SVM 也同样验证了采用有监督降维的 S^3T 算法的确优于采用无监督降维的一般分类算法.另一方面,在α=0(即完全基于类别相似度矩阵的监督式谱空间映射)时,S^3T 的分类测试错误率与其他 $0<\alpha<1$ 时并没有很大的差异,这也进一步说明了监督信息的融入对提高 S^3T 算法的分类性能的确起着至关重要的作用.

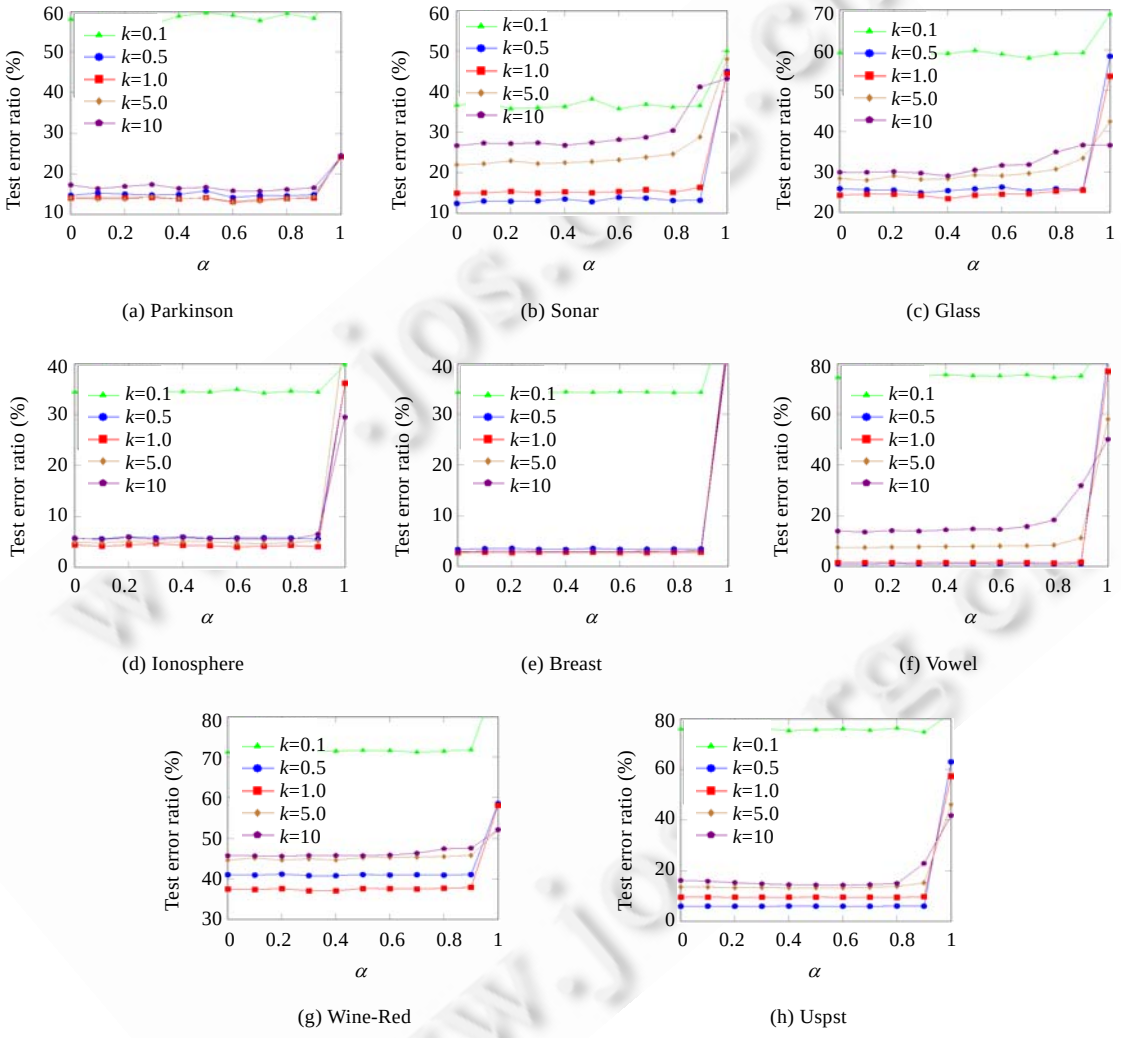


Fig.6 Learning curves of the relationship between α and the test error ratios of S^3T on the 8 UCI data sets

图 6 S^3T 在 8 个 UCI 测试数据集上的关于α的测试错误率曲线

5 结 论

本文提出了一种监督式谱空间分类器(S^3C),它通过将输入数据映射到低维的监督式谱空间中,在该监督式

谱空间中构建最大化异类数据之间几何间隔的最优分割超平面,从而间接获得在原始空间内的非线性分类决策面.S^3C 允许研究者直接观察到在监督式谱空间内的训练数据和测试数据,不但使数据的空间变换和决策面构建变得更加易于理解,而且也有利于我们对算法的评价和对参数的选择.在 S^3C 的基础上,本文还进一步提出了一种改进的监督式谱空间变换多分类算法(S^3T),它通过将监督式谱空间内的映射数据再变换到指定的类别指示空间中去,从而获得关于测试数据的类别指示矩阵用于类别预测.S^3T 除了可以直接应用于多分类问题而无需任何组合策略之外,还对噪声数据有较好的抵抗能力.大量的实验结果表明,S^3C 和 S^3T 算法的分类性能优于其他多种经典的分类算法,且 S^3T 的分类性能更优于 S^3C.

References:

[1] Zhou DL, Gao W, Zhao DB. Face recognition based on singular value decomposition and discriminant KL projection. Journal of Software, 2003,14(4):783−789 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/783.htm

[2] Su JS, Zhang BF, Xu X. Advances in machine learning based text categorization. Journal of Software, 2006,17(9):1848−1859 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1848.htm [doi: 10.1360/jos171848]

[3] Sun XD, Huang RB. Prediction of protein structural classes using support vector machines. Amino Acids, 2006,30(4):469−475. [doi: 10.1007/s00726-005-0239-0]

[4] Guo SQ, Gao C, Yao J, Xie L. An intrusion detection model based on improved random forests algorithm. Journal of Software, 2005,16(8):1490−1498 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1490.htm [doi: 10.1360/jos161490]

[5] Mika S, Rätsch G, Weston J, Schölkopf B, Müller KR. Fisher discriminant analysis with kernels. In: Proc. of the '99 IEEE Signal Processing Society Workshop. IEEE, 1999. 41−48. [doi: 10.1109/NNSP.1999.788121]

[6] Ralf H. Learning Kernel Classifiers: Theory and Algorithms (Adaptive Computation and Machine Learning). Cambridge, MA: The MIT Press, 2001.

[7] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proc. of the 5th Annual ACM Workshop on COLT. 1992. 144−152. [doi: 10.1145/130385.130401]

[8] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000,290(5500):2319−2323. [doi: 10.1126/science.290.5500.2319]

[9] Rowies ST, Sual LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326. [doi: 10.1126/science.290.5500.2323]

[10] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems, Vol.14. MIT Press, 2002. 585−591. http://citeseerx.ist.psu.edu/viewdoc/summary? doi=10.1.1.19.9400

[11] He XF, Cai D, Yan SC, Zhang HJ. Neighborhood preserving embedding. In: Proc. of the 10th IEEE Int'l Conf. on Computer Vision. IEEE Computer Society, 2005. 1208−1213. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1544858 [doi: 10.1109/ICCV.2005.167]

[12] He XF, Niyogi P. Locality preserving projections. In: Thrun S, Saul LK, Scholkopf B, eds. Advances in Neural Information Processing Systems, Vol.16. MIT Press, 2003. 585−591. http://people.cs.uchicago.edu/~xiaofei/conference-24.pdf

[13] von Luxburg U. A tutorial on spectral clustering. Statistics and Computing, 2007,17(4):395−416. [doi: 10.1007/s11222-007-9033-z]

[14] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, Vol.14. 2011. 849−856. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.8100

[15] Zelnik-Manor L, Perona P. Self-Tuning spectral clustering. In: Advances in Neural Information Processing Systems, Vol.17. 2005. 1601−1608. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.7940

[16] Zhu XJ, Kandola J, Ghahramani Z, Lafferty J. Nonparametric transforms of graph kernels for semi-supervised learning. In: Proc. of the Neural Information Processing Systems, Vol.17. 2005. 1641−1648. http://mlg.eng.cam.ac.uk/zoubin/papers/ZhuKanGhaLaf04.pdf

[17] Liu W, Qian BY, Cui JY, Liu JZ. Spectral kernel learning for semi-supervised classification. In: Proc. of the 21st Joint Conf. on Artificial Intelligence. 2008. 1150−1155. http://mlg.eng.cam.ac.uk/zoubin/papers/ZhuKanGhaLaf 04.pdf

[18] Johnson R, Zhang T. Graph-Based semi-supervised learning and spectral kernel design. IEEE Trans. on Information Theory, 2008, 54(1):275−288. [doi: 10.1109/TIT.2007.911294]

[19] Li WY, Ong KL, Ng WK, Sun AX. Spectral kernels for classification. In: Proc. of the 7th Int'l Conf. on Data Warehousing and Knowledge Discovery, Vol.3589. 2005. 520−529. http://www-rcf.usc.edu/~wel/papers/TR/trc0505.pdf

[20] Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation, 2001,13(3):637−649. [doi: 10.1162/089976601300014493]

[21] Chung FRK. Spectral Graph Theory. Providence: American Mathematical Society, 1997.

[22] Shi JB, Malik J. Normalized cuts and image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Vol.22. 1997. 731−737. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=868688 [doi: 10.1109/34.868688]

[23] He P, Chen L, Xu XH. Fast C4.5. In: Proc. of the Int'l Conf. on Machine Learning and Cybernetics. 2007. 2841−2846. http://code.google.com/p/fastc45/downloads/list

[24] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm

附中文参考文献:

[1] 周德龙,高文,赵德斌.基于奇异值分解和判别式 KL 投影的人脸识别.软件学报,2003,14(4):783−789. http://www.jos.org.cn/1000-9825/14/783.htm

[2] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展.软件学报,2006,17(9):1848−1859. http://www.jos.org.cn/1000-9825/17/1848.htm [doi: 10.1360/jos171848]

[4] 郭山清,高丛,姚建,谢立.基于改进的随机森林算法的入侵检测模型.软件学报,2005,16(8):1490−1498. http://www.jos.org.cn/1000-9825/16/1490.htm [doi: 10.1360/jos161490]

何萍(1983－),女,江苏太仓人,博士生,主要研究领域为机器学习,数据挖掘.

徐晓华(1979－),男,博士,讲师,主要研究领域为机器学习,数据挖掘,并行计算,生物信息学.

陈崚(1951－),男,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,数据挖掘,系统优化.