第27卷, 第3期
2007年3月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 27, No. 3, pp602-605
March, 2007

核覆盖算法在光谱分类问题中的研究

杨金福[1], 许 馨[1], 吴福朝[1], 赵永恒[2]

1. 中国科学院自动化研究所国家模式识别实验室, 北京 100080
2. 中国科学院国家天文台, 北京 100012

摘 要 针对光谱分类, 提出了一种基于核技巧的覆盖算法——核覆盖算法。该算法将核技巧与覆盖算法相结合, 并在特征空间中抽取支持向量。实验表明核覆盖算法在光谱分类中的精度与SVM相差不大, 但是它只涉及距离的计算, 不必象SVM那样求解二次规划问题, 对于核宽的选择也不象SVM那样非常敏感。核覆盖算法与覆盖算法相比分类性能相当, 它的优势在于引入的非线性映射Φ改变了样本集在特征空间中之间的距离关系, 使得核覆盖算法得到的支持向量个数大大少于覆盖算法。

关键词 光谱分类; 核覆盖算法; 支持向量机; 核技巧
中图分类号: TN911.7 文献标识码: A 文章编号: 1000-0593(2007)03-0602-04

引 言

目前, 我国正在建设"大天区面积多目标光纤光谱望远镜"(LAMOST)[1, 2]的目标是观测10^7恒星光谱、10^7的星系光谱和10^5的类星体光谱, 极限星等为20^m.5。LAMOST项目要求对观测得到的海量天文光谱数据进行处理和分析。首先将观测光谱分成恒星、星系和类星体三大主要类别, 然后进一步求出恒星的温度、重力加速度、化学丰度, 星系、类星体的红移等物理参数。由此建立一个庞大的天文数据库, 将为天文学家做前沿课题研究提供丰富的资源。

目前, 在模式识别领域中的一些先进分类识别技术都是基于核方法完成的, 如支持向量机(SVM)[3], 正则网络[4]和高斯过程分析[5], 核Fisher判别分析[6], 核主成分分析[7]。这些基于核的算法在许多领域得到了应用, 并有了相当好的结果。

覆盖算法(covering algorithm, CA)[8]的思想来源于文献[9], 他们采用球面投影函数作为非线性映射, 完成样本点的分类问题。我们利用核技巧, 针对光谱分类问题, 提出一种简单的在特征空间中寻找集覆盖的方法, 相应地得到支持向量, 整个计算实际上是在原始输入空间完成, 这一思想与SVM[3]中的核技巧是一致的, 我们称之为核覆盖算法(kernel covering algorithm, KCA)。并分别与SVM和CA进行了比较, 实验表明KCA在光谱分类中的精度与SVM相差不大, KCA较SVM算法有两个优点: 一是它只涉及距离的计算, 不必像SVM那样求解二次规划问题; 二是SVM的分类效率过于依赖核宽的选择, 在某个核宽分类错误率相当低, 大于或者小于这个核宽, 分类错误率变得相当高, 而KCA对于核宽的选择不敏感。KCA与CA相比分类错误率相当, 它的优势在于引入的非线性映射Φ使得特征空间中的数据的分布更加有利于分类, 从而得到的支持向量个数大大少于覆盖算法, 适于处理大规模数据。这也说明核技巧与覆盖算法结合确实是可行的。

1 核覆盖算法

1.1 核技巧

模式识别中的大多数问题属于非线性问题, 对于此类问题, 我们解决的常用手段是用线性模型近似建模。由于采用线性逼近的方法, 经常得不到好的结果, 因此, 寻找有效的非线性方法就成了迫切的要求。核技巧的出现及应用使我们面临的困难出现了转机。核技巧将非线性可分问题化为高维特征空间的线性可分问题, 从而提供了一个非常巧妙的非线性问题处理方法。它的贡献在于可以通过原始空间中的核运算直接得到特征空间中向量的内积, 而不必知道相应的特征映射[3]。

给定数据集$x_1, x_2, \cdots, x_M \in R^N$和Mercer核$k(x_i, x_j)$, 存在$R^N$到高维特征空间$F$的映射

$$\Phi: R^N \rightarrow F, \quad x \mapsto \Phi(x)$$

于是样本集被映射为: $\Phi(x_1), \Phi(x_2), \cdots, \Phi(x_M)$, 它们在空

收稿日期: 2005-12-06, 修订日期: 2006-03-28
基金项目: 国家高技术研究发展计划(863)(2003AA133060)和国家重大科学工程LAMOST计划资助项目
作者简介: 杨金福, 1977年生, 中国科学院自动化研究所国家模式识别实验室博士生 e-mail: yangjf@nlpr.ia.ac.cn

间 F 中的点积为

$$k(x_i, x_j) = [\Phi(x_i) \cdot \Phi(x_j)]$$

所以，只要选择适当的核函数，我们不必知道隐函数 Φ 的显示形式，就可以在输入空间完成高维特征空间的点积计算。

常用的核函数有多项式核、高斯核(径向基函数)、sigmoid 核等:

$$k(x, x') = [(x \cdot x') + 1]^d,$$

$$k(x, x') = \exp\left(-\frac{\| x - x' \|^2}{\sigma^2}\right),$$

$$k(x, x') = \tanh[(x \cdot x') + b]$$

1.2　覆盖算法

覆盖算法(CA)将分类问题转换为一个集覆盖问题。它的优化目标是最少数目的覆盖集，而不是类似线性 SVM 方法的最大边缘，因此对特征空间的线性可分性要求不严格。它将计算分类面的问题转换为基于样本点之间距离的覆盖问题。由于覆盖算法是构造性的，它避免了收敛性和收敛速度问题，计算更简单。

不失一般性，我们假设在 R^N 中有两类训练数据: $A = \{x_1, x_2, \cdots, x_{M_A}\}$，$B = \{y_1, y_2, \cdots, y_{M_B}\}$。

若存在 $A^{sv} = \{x_j^{sv}\} \subset A$，$B^{sv} = \{y_j^{sv}\} \subset B$，和一组正数 $\{d_j^A\}$，$\{d_j^B\}$，使得

$$S^A = \bigcup_{j=1}^{m_A} \{d(x, x_j^{sv}) < d_j^A\} \supset A;$$

$$S^B = \bigcup_{j=1}^{m_B} \{d(y, y_j^{sv}) < d_j^B\} \supset B$$

并且 $S^A \cap S^B = \Phi$，则称 S^A，S^B 为训练样本的覆盖集，集合 A^{sv} 和 B^{sv} 称为相应的支持向量集。

对于任意的测试样本 z，可以根据指示函数 $f(z) = d(z, A^{sv}) - d(z, B^{sv})$ 来作出类别判断。因此，分类问题归结为求训练样本集的支持向量集。

1.3　核覆盖算法

将核技巧与覆盖算法相结合，我们就得到所谓的核覆盖算法。假设非线性映射 Φ 将原始输入空间中的点集 A，B 映射到高维的特征空间 F，记为 $\Phi(A)$ 和 $\Phi(B)$，相应的点记为 $\Phi(x)$ 和 $\Phi(y)$。由于覆盖算法中涉及的运算均为距离计算，由 Mercer 定理，$k(x, y)$ 是 F 上的内积。因此，F 空间中点到集合的距离定义如下:

$$d[\Phi(x), \Phi(B)] = \min\{d[\Phi(x), \Phi(y)] \mid \Phi(y) \in \Phi(B)\}$$

利用距离的定义及核函数的性质:

$$\begin{aligned} d[\Phi(x), \Phi(y)] &= \| \Phi(x) - \Phi(y) \|^2 \\ &= k(x, x) - 2k(x, y) + k(y, y) \end{aligned} \quad (1)$$

相应的也可以得到核函数形式的 $d[\Phi(y), \Phi(A)]$ 和 $d[\Phi(B), \Phi(A)]$。这意味着覆盖算法可以在核技巧的框架内被实现。在这一点上，核覆盖算法与 SVM 中的核技巧是一致的。

将核技巧应用于覆盖算法即得到核覆盖算法，覆盖算法的详细步骤参见文献[8]，由于篇幅原因，这里不再详细叙述。

1.4　核覆盖算法与 SVM 的对比

核覆盖算法与 SVM 方法都选择了利用非线性变换将原始空间上不可分的样本映射到高维特征空间，使其在特征空间中可分。根据 Mercer 定理，选择合适的非线性变换等价于选择满足 Mercer 条件的核函数。

SVM 在特征空间上用一个最优超平面对样本集进行线性划分，如果样本集不能被线性划分，允许存在分类误差。支持向量是距离最优超平面最近的样本点，并且同一类的支持向量离最优超平面距离相等。SVM 求解支持向量和最优超平面需要解二次规划问题，计算量和内存开销较大。

核覆盖算法在特征空间上使用多个超球面对样本集进行划分，要求分类不能有误差，是以丧失最大边缘意义下的最优超平面为代价的，可同时应用于线性可分与不可分的情形。支持向量是每个超球面的中心样本点。这种构造性算法比较直观，计算简单，避免了收敛性和收敛速度问题。

2　实验分析

我们将核覆盖算法用于恒星、星系和类星体的识别中。共有数据集恒星: 1 878 个，类星体: 3 026 个，星系: 2 529 个，全部来自 SLOAN 光谱数据库。每条光谱的波长范围为 380~ 900 nm，插值后每条光谱有 521 个点。恒星由于在银河系内，它们的视向速度非常小，与星系相比都可以认为是零，信噪比较高。而正常星系的红移值均较大，在 SLOAN 的数据库中，有 98 78% 的星系红移在 0.01~ 0.5 之间，信噪比次之；类星体的红移最大，有 92 93% 的类星体红移在 0.2~ 3 之间，信噪比最低。红移会导致光谱发生连续谱和谱线的变化，在不同的红移值下，光谱的特性是不一样的，这导致了识别的困难，另外由于红移越大，信噪比也越低，也增加了识别的难度。我们的数据来自实测光谱数据库，信噪比的水平参差不齐，为了尽量消除噪声的影响，我们用 PCA 方法对所有数据处理，取前20 维的特征矢量作为输入，进行识别器的设计。分类过程分两步，首先识别出恒星，然后再将星系与类星体分开。

2.1　恒星与其他天体的分类

数据集一半用于训练，一半用于测试。核函数选择高斯核: $k(x, y) = \exp(-\| x - y \|^2/\sigma)$。核宽的选择采用 5 倍的交叉验证的方法得到。高斯核的窗口宽度对恒星和其他天体光谱分类结果的影响如图 1 所示，核宽取 0.01~ 2。

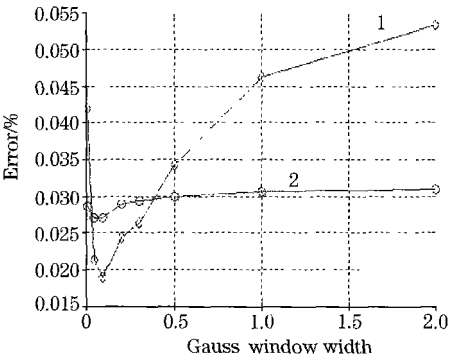

Fig 1　The classification error rate based on different kernel width of KCA and SVM respectively

1: SVM; 2: KCA

由图 1 可以看出，核覆盖算法在每个核宽上的分类性能

都相差不大。而 SVM 对于核宽的选择相当敏感。在核宽为 0.1 时，SVM 到达其分类错误的最小值为 1.9%。但是在核宽变小或者增大时，错误率升高很快。核覆盖算法在核宽较小时的分类性能稍优于核宽大时的性能。核覆盖算法在核宽为 0.01 时取得分类最小错误率为 2.7%。

如图 2 所示是核覆盖算法和 SVM 算法在支持向量个数上的比较。核覆盖算法的支持向量随着核宽的增大而降低，而且支持向量的权值在核宽较小的时候也较大，在核宽较大时其权值显著降低。SVM 的支持向量的最低值对应其分类的次最大精度，而且不论核宽是增大还是减小，支持向量都显著增加。

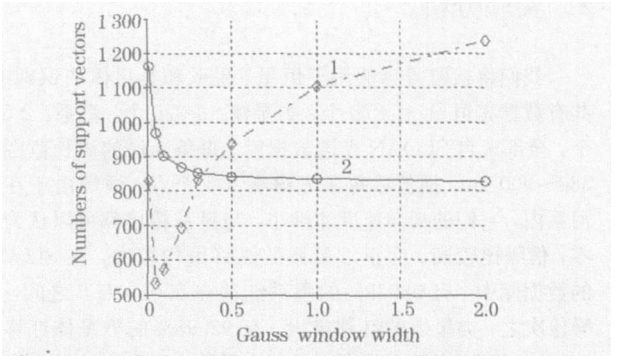

Fig 2 The number of support vectors based on different kernel width of KCA and SVM respectively

1: SVM; 2: KCA

在测试数据集上，核覆盖算法的分类错误率为 3.17%；SVM 为 1.28%；以 SVM 的分类错误率最小。

表 1 和表 2 是这些算法在选择最优参数时的分类性能和支持向量个数对比。从表 1 可以看出在训练数据集上，核覆盖算法和 SVM 的分类错误率较低，核覆盖算法比覆盖算法稍好一些。在测试数据集上，SVM 的分类效果最好。SVM 算法的支持向量个数是最少的；覆盖算法的支持向量个数是最多的，而核覆盖算法比之少了近 1/3 的支持向量。

Table 1 Comparison of classification error rate

	覆盖算法	核覆盖算法	SVM 算法
训练数据集	3.13%	2.7%	1.9%
测试数据集	3.53%	3.17%	1.28%

Table 2 Comparison of the number of support vectors

	覆盖算法	核覆盖算法	SVM 算法
训练数据集	1 305	906	572

2.2 星系与类星体的识别

核函数为高斯核，核宽的选择由 5 倍交叉验证得到，核宽范围 0.01~2。两种方法的对比如图 3 和图 4 所示。

从图 3 中可以看出，核覆盖算法同样是在核宽 0.01 处取得最小分类错误率，3.33%。SVM 依然是在固定的核宽分类精度最高，3.15%。大于或者小于这个核宽，分类错误率上升很大。由图 4 支持向量的对比，SVM 的支持向量个数是最少的，核覆盖算法的支持向量较多。这是由于 SVM 找到的是最优分类面上的向量作为支持向量，而核覆盖算法中的支持向量还包括一些距离分类面较远的向量，而且它的分类面并不是最优分类面。在测试数据集上，核覆盖算法的分类错误率为 4.59%；SVM 为 3.85%。

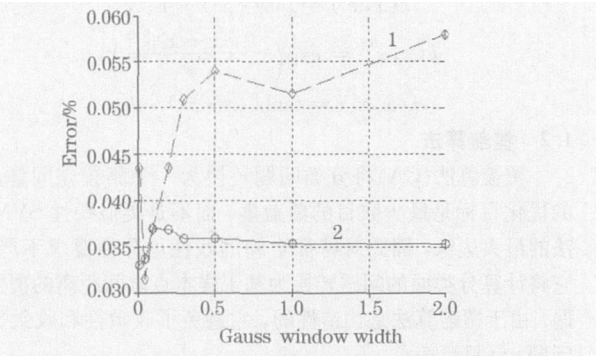

Fig 3 The classification error rate of galaxy and QSO based on different kernel width of two algorithms

1: SVM; 2: KCA

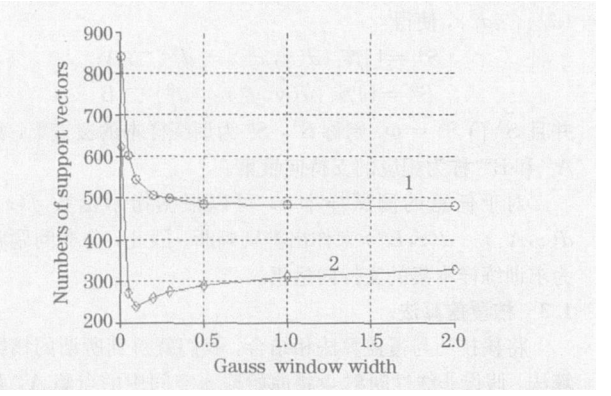

Fig 4 The number of support vectors of galaxy and QSO based on different kernel width of two algorithms

1: SVM; 2: KCA

表 3 和表 4 是这些算法在选择最优参数时对于星系和类星体的分类性能和支持向量个数对比。从表 3 可以看出，SVM 的分类错误率较低，核覆盖算法与覆盖算法分类结果相当。SVM 算法的支持向量个数是最少的；覆盖算法的支持向量个数依然是最多的，而核覆盖算法比之少了 303 个支持向量。

Table 3 Comparison of classification error rate

	覆盖算法	核覆盖算法	SVM 算法
训练数据集	3.3%	3.33%	3.15%
测试数据集	4.55%	4.59%	3.85%

Table 4 Comparison of the number of support vectors

	覆盖算法	核覆盖算法	SVM 算法
训练数据集	847	544	281

3 总 结

本文针对光谱分类问题，提出了核覆盖算法，将核技巧与覆盖算法结合，在特征空间中抽取支持向量。将核覆盖算法与 SVM 进行比较，两者同样都是利用了核技巧，SVM 是在特征空间中寻找最优分类面，核覆盖算法将计算分类超平面的问题转化为计算样本间距离的覆盖问题，这种构造性的算法不需要像 SVM 算法那样求解二次规划问题，不涉及收敛性及收敛速度问题。实验表明，核覆盖算法不像 SVM 那样对于核宽的选择过于敏感。

核覆盖算法与覆盖算法在分类精度上大致相当，由于引入的非线性映射 Φ 使得样本集在特征空间中的分布更加利于分类，使得核覆盖算法得到的支持向量个数大大少于覆盖算法，适于处理大规模数据。核技巧的使用确实可以提高分类精度或者降低支持向量的个数，对于核覆盖算法则表现为后者，这是由其分类策略决定的。由于覆盖算法要求分类不能有误差，所以在输入空间中的支持向量必然要比较多才能保证这一点；而在特征空间中，经过非线性映射 Φ，数据集分布更加利于分类，求得的支持向量的个数大大减少。这表明将核技巧与覆盖算法结合是可行的。核覆盖算法用于天体光谱的分类问题，得到了较高的分类精度。

参 考 文 献

[1] LIU Rong, DUAN Fu-qing, LUO A-li(刘　蓉, 段福庆, 罗阿理). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(7): 1155.

[2] QIN Dong-mei, HU Zhan-yi, ZHAO Yong-heng(覃冬梅, 胡占义, 赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(1): 182.

[3] Cortes C, Vapnik V N. Machine Learning, 1995, 20: 273.

[4] Girosi F, Jones M, et al. Neural Computation, 1995, 7(219): 269.

[5] Seeger Matthias. International Journal of Neural Systems, 2004, 14(2): 1.

[6] Baudat G, Anouar F. Neural Computation, 2000, 12(10): 2385.

[7] Scholkopf B, Smola A J, Müller K R. Neural Computation, 1998, 10: 1299.

[8] YANG Jin-fu, WU Fu-chao, LUO A-li, et al(杨金福, 吴福朝, 罗阿理, 等). Pattern Recognition and Artificial Intelligence(模式识别与人工智能), 2006, 19(3): 368.

[9] Zhang Ling, Zhang Bo. IEEE Transactions on Neural Networks, 1999, 10(4): 925.

Studies of Spectra Classification Based on Kernel Covering Algorithm

YANG Jin-fu[1], XU Xin[1], WU Fu-chao[1], ZHAO Yong-heng[2]

1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing　100080, China
2. National Astronomical Observatory, Chinese Academy of Sciences, Beijing　100012, China

Abstract　A kernel based covering algorithm, called the kernel covering algorithm (KCA), is proposed for the classification of celestial spectra. This algorithm is a combination of kernel trick with the covering algorithm, and is used to extract the support vectors in feature space. The experiments show that the classification result based on KCA is a little less than that based on SVM. However, KCA only involves the distance computation without the need to solve the quadratic programming problem. Also, KCA is insensitive to the width of gauss window. Although KCA has a comparable classification performance with the covering algorithm, it changes the distance between samples in feature space by the nonlinear mapping such that the distribution of samples is more adaptable to classify. Therefore, the number of KCA's resulting support vectors is significantly smaller than that of the covering algorithm.

Keywords　Spectra classification; Kernel covering algorithm; SVM; Kernel technique

(Received Dec. 6, 2005; accepted Mar. 28, 2006)