文章编号：1672-8785(2012)09-0041-05

基于先验信息的SVM红外光谱定性分析方法

姜　安[1,2]　胡　勇[3]　彭江涛[1,2]　谢启伟[3]　彭思龙[3]

(1. 南京财经大学管理科学与工程学院，江苏 南京 210046；
2. 江苏省质量安全工程研究院，江苏 南京 210046；
3. 中国科学院自动化研究所集成中心，北京 100190)

摘　要：通过将类不变性先验信息融入到支持向量机 (Support Vector Machine, SVM) 算法的目标函数中，提出了一种基于漂移约束的 SVM 红外光谱定性分析算法。该算法将红外光谱的漂移项模拟成一个低阶多项式，并在 SVM 优化目标中要求决策面的法向量与漂移方向垂直，从而使分类器能够消除样本漂移影响。详细讨论了波段选择和正则化参数对分类准确率的影响，并对比了各种变形 SVM 算法的分类效果。实验结果表明，与标准的 SVM 算法及其各种变形算法相比，本文提出的 DCSVM 算法具有更高的分类准确度。

关键词：先验信息；支持向量机；红外光谱；定性分析

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3969/j.issn.1672-8785.2012.09.008

SVM Infrared Spectroscopic Qualitative Analysis Based on Prior Information

JIANG An[1,2], HU Yong[3], PENG Jiang-tao[1,2], XIE Qi-wei[3], PENG Si-long[3]

(1. School of Management Science & Engineering, Nanjing University of Finance & Economics, Nanjing 210046, China;
2. Jiangsu Province Institute of Quality and Safety Engineering, Nanjing 210046, China;
3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: By incorporating class-invariant prior information into the object function of a Support Vector Machine (SVM) algorithm, a SVM infrared spectroscopic qualitative analysis algorithm based on drift constraint is proposed. Because the algorithm simulates the drift term of infrared spectrum into a low order polynomial and requires the normal vector of the decision surface to be perpendicular to the drift direction, the classifier can remove the effect of sample drift. The influence of band selection and regularization parameters on classification accuracy is described in detail and the classification results of different SVM algorithms are compared. The experimental result shows that compared with the standard SVM algorithm and other similar algorithms, the DCSVM algorithm has a higher classification accuracy.

Key words: prior information; SVM; infrared spectroscopy; qualitative analysis

0 引言

20 世纪 90 年代，Vapnik 等人在统计学习理论和结构风险最小化原则的基础上，建立了一种通用的机器学习算法——支持向量机 (Support Vector Machine, SVM)[1−2]。近年来，SVM 算法被广泛应用于红外光谱定性分析，如中药材产地鉴别、混合气体种类分析和茶叶真伪鉴别等。然

收稿日期：2012–03–07
基金项目：国家自然科学基金项目 (61101219; 71141020; 61032007)
作者简介：姜安 (1979-)，男，安徽东至人，博士，主要研究方向为红外光谱技术及模式识别。E-mail: ja0831@163.com

而在这些定性分类应用中，人们均是直接采用标准的 SVM 算法，但在算法设计上却没有考虑红外光谱自身的物理特性等先验信息。在红外光谱采集过程中，由于受系统偏差、样品散射、外界温度、湿度以及人为因素的影响，光谱中会存在过多的噪声，同时也会发生谱图漂移和干涉条纹等现象。然而即使一些样本发生了这些变化，也出现了不同程度的漂移，但是它们本身所属的类别是不变的。如果直接利用 SVM 对样本进行学习，那么得到的分类器只能反映训练样本自身的信息，而不会考虑由训练样本的漂移等因素所带来的潜在影响，因此该分类器并不适用于未知测试样本。

如何将先验信息与 SVM 训练模型进行有效结合是人们近年来的研究热点之一。目前已经有很多学者将先验知识应用到了 SVM 分类模型中，主要包括以下两个研究方向：一是如何选择特定应用背景下的核函数。 Scholkopf 指出，通过构造合适的核函数，可以有效地利用数据先验信息，从而将不变性作为先验信息融入到 SVM 中，这样便可有效提高学习性能 [3−4] 。二是直接把先验信息引入到 SVM 分类器中，包括直接在训练样本上引入先验知识属性，或者对分类的目标函数重新进行定义，使之成为在某种意义下的最大间隔 [5] 。本文针对红外光谱定性分析这一特定应用背景，充分考虑红外光谱自身的漂移影响，将漂移类别不变性作为先验信息加入到 SVM 训练模型中，从而提出一种基于先验信息 SVM 的红外光谱定性分析算法。

1 算法模型

尽管目前存在很多变形的 SVM 算法，如 ν-SVC [6] 、 TSVM [7] 和 LS-SVM [8] 算法等，但是这些算法均是针对标准的 SVM 算法作了一些改动。这些算法主要还是关注 SVM 算法的本身，如参数直观意义和计算速度等，并没有结合实际应用问题。因此， SVM 及其变形算法在红外光谱定性分析应用中将会不可避免地受到样本谱图变化所带来的影响。基于谱图的先验信息，挖掘谱图的内在规律，并通过将数据自身的特性与 SVM 算法相结合设计适合具体谱图分类问题的 SVM 算法，是本文考虑的一个方向。本文将样本的漂移类不变特性作为先验信息加入到原始的 SVM 学习模型中，提出漂移约束 SVM (Drift Constraint SVM, DCSVM) 算法。

红外光谱样本采集通常会由于受到仪器以及外界环境条件的影响而出现漂移现象。将样本的漂移模型记为

$$Tx = x + dx, \quad dx = a_0 + a_1\lambda + a_2\lambda^2 \tag{1}$$

式中，dx 为样本 x 所对应的漂移项，λ 为波长向量。一般说来，通常可以将红外光谱的漂移模拟成一个关于波长的低阶多项式 (通常为二次多项式) 。另外， T 为样本上所施加的某种变换，这相当于是为样本附加一个漂移基线； Tx 则表示漂移谱图。

为了使样本在漂移条件下保持类别不变，则要求对于分类函数 f 有：

$$f(x) = (\omega * x), \quad f(Tx) - f(x) = (\omega * dx) \to 0 \tag{2}$$

基于 SVM 将上面的约束条件加入优化目标函数，求解以下优化问题：

$$\arg\min(1-\gamma)(\omega * \omega) + \gamma \sum_{i=0}^{k} (\omega * \lambda^i)^2 \tag{3}$$

$$s.t. \quad y_i[(\omega * x_i) + b] \geq 1, \quad i = 1, \cdots, l$$

式中， $0 \leq \gamma \leq 1$ 为正则化参数。式 (3) 中，目标函数要求在最大化分类间隔的同时还兼顾到漂移类别不变性。实际上，该算法要求 SVM 权向量对谱图漂移不敏感。将漂移项模拟成一个多项式，并要求权向量与多项式正交。在这种约束条件下，分类器能够在一定程度上消除光谱的基线漂移影响，从而反映样本内在的变化规律。参数 γ 在最大化间隔与消除漂移影响之间寻求折中。当参数 $\gamma = 0$ 时，式 (3) 即为标准的 SVM 算法；当 $\gamma = 1$ 时，优化目标要求漂移量与分类面法线方向正交，即分类器能够消除样本漂移的影响，分类面对样本漂移不敏感。

定义：$C_g = (1-\gamma)I + \gamma\sum_{i=0}^{k}\lambda^i(\lambda^i)^T$，记 $\tilde{C}_g = C_g^{1/2}$，$\tilde{\omega} = \tilde{C}_g\omega$，$\tilde{x}_i = \tilde{C}_g^{-1}x_i$，有：

$$(\tilde{\omega} * \tilde{x}_i) = \tilde{\omega}^T\tilde{x}_i = (\tilde{C}_g\omega)^T\tilde{C}_g^{-1}x_i$$

$$= \omega^T\tilde{C}_g^T\tilde{C}_g^{-1}x_i = \omega^T x_i = (\omega * x_i)$$

则式 (3) 可简化为

$$\arg\min(\tilde{\omega} * \tilde{\omega}) \tag{4}$$

$$s.t.\ \ y_i[(\tilde{\omega} * \tilde{x}_i) + b] \geq 1,\ \ i = 1, \cdots, l$$

求解式 (4) 后可得决策函数：

$$f(x) = (\tilde{\omega}^* * \tilde{x}) + b^* \tag{5}$$

在对测试样本 x_{un} 进行预测时，首先利用 $\tilde{C}_g^{-1}$ 对测试样本进行线性变换 ($\tilde{x}_{un} = \tilde{C}_g^{-1}x_{un}$)，然后利用 SVM 决策函数对变换后的样本 $\tilde{x}_{un}$ 进行预测，得到

$$f(x_{un}) = (\tilde{\omega}^* * \tilde{x}_{un}) + b^*$$

$$= \sum_{i=1}^{l}\alpha_i^* y_i\tilde{C}_g^{-2}(x_i * x_{un}) + b^* \tag{6}$$

DCSVM 算法的流程如下：

已知训练集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$

(1) 选择合适的参数 γ，求取矩阵 $\tilde{C}_g$，并实现数据变换；

(2) 选择核函数及其对应的核参数；

(3) 求解式 (4)，得到最优解 α^* 和 b^*；

(4) 构造式 (5) 的决策函数 $f(x) = (\tilde{\omega}^* * \tilde{x}) + b^*$；

(5) 采用 $\tilde{C}_g^{-1}$ 对未知测试样本进行变换，并按式 (6) 进行预测。

为了更好地说明这个问题，下面给出一张示意图。在图 1(a) 中，两类样本点完全线性可分。直接用 SVM 分类便可得到图中分界线 y=3。当数据有一个向右上方的线性偏移时，如图 1(b) 所示，加漂移约束的 SVM 算法 (DCSVM) 学习得到的决策线为图中实线所示。从图 1 中可以看出，决策线的法方向与数据偏移趋势方向近乎垂直，因此所得到的分类函数能够很好地消除这种漂移的影响。特别是，如果图 1(a) 中的数据为训练样本，其漂移后的样本为测试样本 (即图 1(b) 右边 6 个样本)，则训练样本所得到的分类线 ((b) 中虚线) 无法完全正确地对存在漂移趋势的测试样本进行分类。因此，在已知数据会存在某种偏移时，据此先验知识，可以设计出更加合理的分类器。DCSVM 算法正是考虑到样本会存在多项式偏移这个特殊的物理背景，在 SVM 训练时尽可能地消除谱图漂移所带来的影响。

2 实验结果及分析

2.1 实验数据

实验所用白酒均为某酱香型酒样，其红外谱图由 Perkin–Elmer Spectrum GX FTIR 光谱仪加

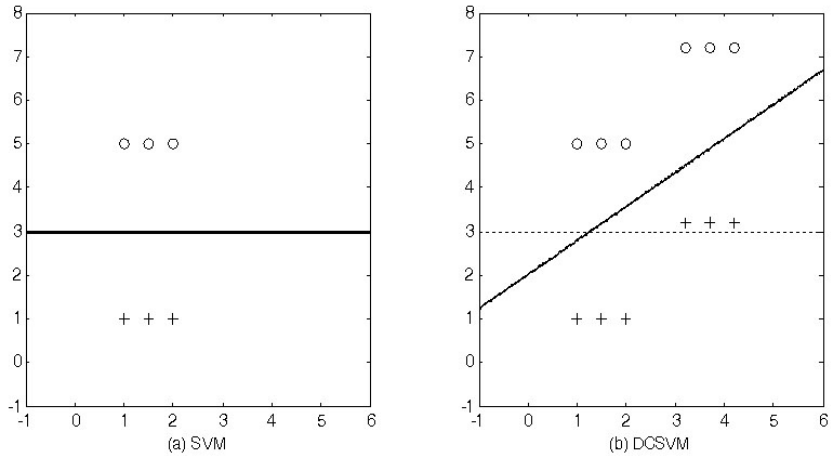

图 1　加漂移约束的 SVM 算法 (DCSVM) 的示意图

上 ATR 附件 (ZnSe cell) 通过采样得到。谱图采集区域为 4000 ~ 650 cm^{-1}，光谱仪的分辨率为 4 cm^{-1}，每张谱图扫描 16 次。

数据集 1 为 2009 年第四轮次窖中和 2009 年第五轮次窖中，分别有样本 250 个和 508 个。数据集 2 为 2010 年第一轮次醇甜和第二轮次醇甜，分别有 313 个和 361 个样本。数据集 3 为 2008 年第三轮次一等酒和二等酒，分别有 121 个和 120 个样本。数据集 4 为 2009 年第四轮次窖面和 2009 第五轮次窖面，分别有样本 116 个和 120 个样本。

2.2 实验分析

实验中考虑了线性漂移的影响，对标准 SVM 分类方法、变形 SVM 算法以及加漂移约束的 SVM 方法 (DCSVM) 进行了比较。在这些 SVM 算法中，核函数选择 RBF 核，并采用交叉验证方法选择核参数。在分类器训练过程中，随机选取 75 % 的样本作为训练集，并选取 25 % 的样本作为测试集，重复十次，计算平均准确率。

2.2.1 波长选择对分类结果的影响

波长选择一方面可以简化模型，另一方面还可以剔除不相关或非线性变量。在白酒定性分析中考虑 1500 ~ 1200 cm^{-1} 指纹区，选择标准 SVM 作为分类器，对比选择指纹区和全谱的分类结果 (见表 1)。

表 1 不同波段对分类结果的影响

所选波段	数据集 1	数据集 2	数据集 3	数据集 4
全谱	0.8942	0.9762	0.8667	0.9492
指纹区	0.9048	0.9881	0.8833	0.9661

从表 1 中可以看出，指纹区的分类结果略强于全谱。指纹区主要是由一些单键 C–O、C–N 和 C–X (卤素原子) 等的伸缩振动及 C–H、O–H 等含氢基团的弯曲振动以及 C–C 骨架振动产生的。当分子结构稍有不同时，该区的吸收就会有细微的差异。指纹区对于区别结构类似的化合物很有帮助。以下的实验都是针对指纹区进行讨论的。

2.2.2 参数 γ 的影响

将训练集中的 75 % 用于建模，将其余的 25 % 用于验证，然后比较 DCSVM 算法中参数 γ 的不同取值对分类准确率的影响。其中，参数 γ 从 0.1 变化到 1，步长为 0.05，如图 2 所示。

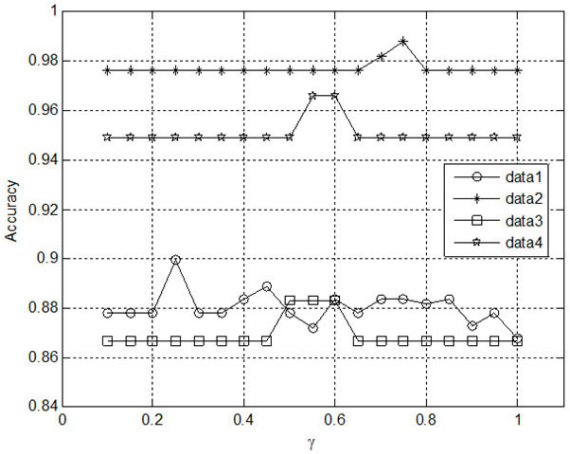

图 2 参数 γ 对分类结果的影响

参数 γ 在最大化间隔与消除漂移影响之间寻求折中。当 γ 较大时，更倾向于对基线的约束，希望分类超平面对基线漂移不敏感；当 γ 较小时，倾向于寻求具有最大间隔分类超平面，此时分类结果趋近于标准 SVM 算法。从图 2 中可以看出，实验所用的四个数据集受不同参数 γ 的影响基本一致：随着 γ 的增大，分类准确率逐渐变大；但是当 γ 超过一定阈值时，分类准确率又逐渐变小而趋于稳定。在下面的实验中，对于四个不同的数据集，分别将 γ 的大小取为 0.25、0.75、0.55 和 0.60。

2.2.3 不同分类方法的对比

对于原始谱图，考虑标准 SVM、去基线后再作 SVM、ν-SVC、TSVM、LS-SVM 以及加基线约束的 DCSVM 等六种分类方法，其分类结果的对比情况见表 2。

表 2 各种 SVM 算法的分类准确率

分类方法	数据集 1	数据集 2	数据集 3	数据集 4
SVM	0.9048	0.9881	0.8833	0.9661
去基线 +SVM	0.9259	0.9881	0.8833	0.9831
ν-SVC	0.9048	0.9821	0.9000	0.9661
TSVM	0.9048	0.9881	0.8833	0.9661
LS-SVM	0.9101	0.9821	0.8833	0.9831
DCSVM	0.9524	0.9940	0.9167	0.9831

从表 2 中可以看出，与其它的 SVM 方法相比，本文提出的 DCSVM 方法具有更高的分类准确率。在标准 SVM 方法中，分类超平面未考虑训练样本中潜在的基线漂移。 ν-SVC 方法通过引入新参数 ν 来控制支持向量的数目和误差。LS-SVM 方法将不等式约束变为等式约束，通过求解线性方程组来提高运算速度。TSVM 算法在未标记样本较多时可能会起到一定的作用。总的说来，这些变形的 SVM 算法都没有从本质上引入谱图漂移的先验信息。将谱图去除基线漂移之后，再进行 SVM 分类，其分类效果相对于标准 SVM 得到了一定的改进，但仍稍差于 DCSVM 。

3 结论

本文通过将类不变性这个先验信息融入到 SVM 算法的目标函数中，提出了一种基于漂移先验信息的 SVM 红外光谱定性分析算法，并详细讨论了波段选择、参数 γ 以及各种分类方法对分类准确率的影响情况。实验结果表明，与 SVM 算法及其各种变形算法相比，本文提出的 DCSVM 算法具有更好的推广性能。

参考文献

[1] Vapnik V N. Statistical Learning Theory [M]. New York: Wiley, 1998.

[2] Vapnik V N. An Overview of Statistical Learning Theory [J]. *IEEE Transactions on Neural Networks*, 1999, **10**(5): 988–999.

[3] Scholkopf B, Simard P, Smola A. Prior Knowledge in Support Vector Kernels [C]. *Advances in Neural Information Processing Systems*, 1998: 640–646.

[4] Scholkopf B, Tsuda K, Vert J P. Kernel Methods in Computational Biology [M]. Boston: The MIT Press, 2004.

[5] Scholkopf B, Burges C, Vapnik V. Incorporating Invariances in Support Vector Learning Machines [C]. *Conf on Artificial Neural Networks*, 1996: 47–52.

[6] Scholkopf B, Smola A J, Williamson R C, et al. New Support Vector Algorithms [J]. *Neural Computation*, 2000, **12**(5): 1207–1245.

[7] Wu D, Bennett K P, Cristianini N, et al. Large Margin Trees for Induction and Transduction [C]. *ICML*, 1999: 474–483.

[8] Suykens J, Vandewalle J. Least Squares Support Vector Machines Classifiers [J]. *Neural Processing Letters*, 1999, **9**(3): 293–300.

（上接第 13 页）

[65] Redfern D A, Fang W, Ito K, et al. Investigation of Laser Beam-induced Current Techniques for Heterojunction Photodiode Characterization [J]. *Journal of Applied Physics*, 2005, **98**(19): 043501.

[66] Sun T, Li Y, Chen X, et al. The Dark Current Mechanism of HgCdTe Photovoltaic Detector Passivated by Different Structure [C]. *SPIE*, 2005, **5640**: 26–33.

[67] Qiao H, Hu W, Ye Z, et al. Influence of Hydrogenation on the Dark Current Mechanism of HgCdTe Photovoltaic Detectors [J]. *Journal of Semiconductors*, 2010, **31**(19): 036003.

[68] Bhan R K, Srivastava V, Saxena R S, et al. Improved High Resistivity ZnS Films on HgCdTe for Passivation of Infrared Devices [J]. *Infrared Physics & Technology*, 2010, **53**(5): 404–409.

[69] Radford W A. Photovoltaic Detector with Integrated Dark Current Offset Correction: US, 5663564 [P]. 1997-09-02.

[70] Dreiske P D, Turner A M, Forehand D I. Method of Making Photodiodes for Low Dark Current Operation Having Geometric Enhancement: US, 5593902 [P]. 1997-01-14.

[71] Dreiske P D, Turner A M, Forehand D I. Geometric Enhancement of Photodiodes for Low Dark Current Operation: US, 5665998 [P]. 1997-09-09.

[72] Klipstein P. A Unipolar Semiconductor Photodetector with Suppressed Dark Current and Method for Producing the Same: EP, 2249400 [P]. 2010-11-23.

[73] Jurgen M. Dark Current Reduction for Large Area Photodiodes: WO, 10009462 [P]. 2010-01-21.

[74] Pradip M, Beck J D, Skokan M R. A Radiation Detector Having a Bandgap Engineered Absorber: EP, 2284895 [P]. 2011-02-10.

[75] Klipstein P. Depletion-less Photodiode with Suppressed Dark Current and Method for Producing the Same: WO, 05004243 [P]. 2005-01-13.

[76] Maimon S. Reduced Dark Current Photodetector: WO, 07107973 [P]. 2010-03-30.

[77] Velicu S, Grein C, Rafol S B. Photodetector with Dark Current Reduction: US, 07820971 [P]. 2010-10-26.