第33卷，第6期
2013年6月

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

Vol. 33, No. 6, pp1566-1569
June, 2013

支持向量机算法对鼻咽癌与正常鼻咽细胞株拉曼光谱分析

孙 磊[1]，陈 阳[1*]，黄洋文[2]，欧 琳[3]，苏 颖[4]，冯尚源[3]，雷晋萍[3]

1. 福州大学至诚学院，福建 福州 350002
2. 中北大学仪器科学与动态测试教育部重点实验室，山西 太原 030051
3. 福建师范大学医学光电科学与技术教育部重点实验室，福建 福州 350007
4. 福建省肿瘤医院放射生物研究室，福建 福州 350014

摘 要 拉曼光谱技术在肿瘤与正常细胞株的鉴别方面有着广泛的应用。对一个已有的诊断模型进行可靠性验证是非常重要和有意义的工作。采用两种不同的支持向量机分类算法对鼻咽癌和正常鼻咽细胞株的拉曼光谱进行分析和识别，结果显示灵敏度和特异性均在90%以上，并且与已知的相关线性判别分析结果一致。结论表明，两种支持向量机算法都能较好地对细胞株进行鉴别，同时也表明拉曼光谱技术结合相关统计分类算法的方法可以实现对肿瘤细胞的准确鉴别，这一结果将进一步证实拉曼光谱可以作为鼻咽癌诊断的一种方式。

关键词 拉曼光谱；鼻咽癌；细胞；支持向量机；线性判别分析

中图分类号：O657.3；R318.5 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2013)06-1566-04

引 言

鼻咽癌(nasopharyngeal carcinoma，NPC)是一种恶性的肿瘤，我国南方地区的发病率最高，也是NPC的主要集中地，其他地区的发病率则比较低[1]。自从1990年Pupples率先报道单个细胞的拉曼光谱开始，拉曼光谱技术在生物医学领域内的应用越来越广泛，并且因为其非侵入式和非破坏性的检测方式使得拉曼光谱技术可以作为生化样品检测的理想方法[2-4]。拉曼光谱技术可以提供有关物质振动模式的指纹性信息，结合诸如偏最小二乘法、线性判别分析(linear discriminant analysis，LDA)、支持向量机(support vector machines，SVM)等统计分析方法，可以用于分析肿瘤细胞和正常细胞之间或者不同类型的细胞之间的不同特征并加以鉴别[5-7]。但是同时也存在一个有关判别模型的可靠性检验问题，即如何评价一个分类模型的可靠性，因为这是一个十分必要和有意义的课题，在高可靠性的前提下利用相关判别模型对肿瘤细胞与正常细胞作出正确识别能够为拉曼光谱技术的临床应用提供必要的实验依据。基于上述考虑，提出采用不同的统计分类算法对同一个判别模型进行可靠性检验，以此充分证实该模型的可靠性。

之前的相关工作已经提出了一个基于LDA的判别NPC细胞株(C666-1，CNE2)和正常鼻咽细胞株NP69的诊断模型[8]，本工作是利用支持向量分类(support vector classification，SVC)算法对NPC细胞株(C666-1，CNE2)和正常鼻咽细胞株NP69的拉曼光谱进行分析和识别。SVM最早是由Vapnik提出[9]，而其中的SVC算法是SVM的主要功能之一，用于解决分类问题。研究思路是将输入的数据映射到高维空间，并建立一个最佳超平面，使原先难以区分的两个群体在高维空间中得以区分开来，且类间距达到最大(如图1)。为了充分检验上述LDA诊断模型的可靠性，采用两种不同的SVC算法程序来进行。

1 实验部分

1.1 数据来源

所有三种细胞株C666-1，CNE2和NP69采用InVia型共聚焦显微拉曼光谱仪(Renishaw公司，英国)获取光谱数据，详见参考文献[6]。NPC细胞株C666-1，CNE2和正常鼻咽细胞株NP69分别选取30，31和46个样品数据。然后所有的拉曼光谱均进行进行荧光背景去除(Vancouver Raman Algorithm，BC Cancer Agency & University of British

收稿日期：2012-11-09，**修订日期**：2013-03-25

基金项目：国家自然科学基金项目(11104030)和福建省自然科学基金项目(2011J01153)资助

作者简介：孙 磊，女，1980年生，福州大学至诚学院讲师 e-mail：fannysun@163.com

*通讯联系人 e-mail：chzhy85@163.com

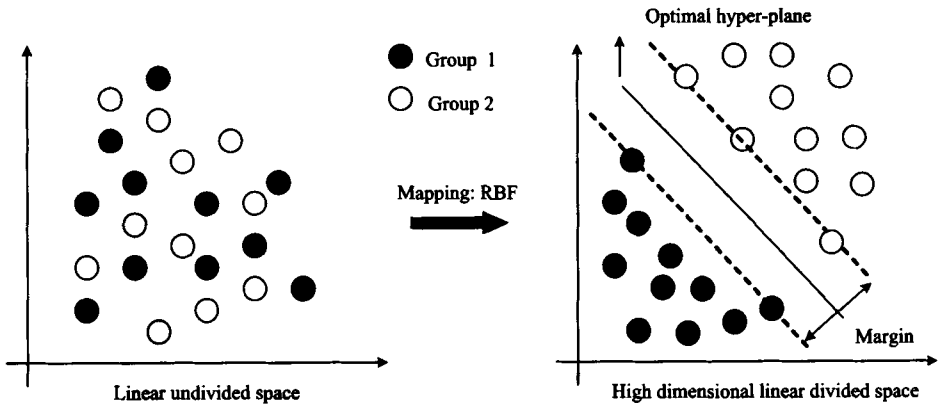

Fig. 1　Principle diagram of SVMs. The black and white circles separately represent two different classes which are maximally separated in a high-dimensional linear space

Columbia[10])，再将光谱面积作归一化处理。最后利用 SVC 分类器对所测样品数据进行训练和识别。

1.2　SVC 算法

SVC 分类器有两个：一是 LIBSVM(Machine learning and data mining group[11])。LIBSVM 集成了支持向量分类(C-SVC，nu-SVC)、回归(epsilon-SVR，nu-SVR)和分布估计(one-class SVM)，且支持多类别的分类问题，一般的支持向量机仅仅只能解决两类别的分类问题。这一特点正好符合本诊断模型，即三类样本之间的识别。另外一个 SVC 算法是自编的 SVC，其基本思想是将三类样本的识别分解成几个两分类的识别，针对支持向量机中核心参数和误差惩罚因子的选择问题，利用量子遗传算法(quantum genetic algorithm，QGA)对支持向量机模型参数进行全局寻优，克服了以往反复试验以确定其参数的缺点，训练算法采用序列最小优化算法，因为它将工作集的规模减到最小，即两个样本，其避开了复杂的数值求解优化问题的过程，更多细节详见参考文献[12]。

1.3　统计分析

利用 LDA 方法建立的分类判别模型，是分别随机选取 C666-1，CNE2 和 NP69 各 10 条拉曼光谱作为待判样本集，剩余的作为训练样本集。训练完毕后，待判样本即被送入判别模型进行识别。采用上述两个 SVC 分类器对三种细胞株的拉曼光谱数据重新训练和识别，待判样本集的设置与 LDA 模型的数据设置完全相同，剩余的光谱数据同样作为训练样本集分别送入两个 SVC 分类器并根据算法要求的操作进行训练。

1.4　估计模型的可靠性

由于 5 折交叉验证误差通常用于评估分类模型的可靠性[13]，所以上述两种分类算法的识别效果采用 5 折交叉验证误差 E 来评价，E 定义为

$$E=\frac{N_{\text{错判}}}{N_{\text{样本总数}}}=\frac{N_{\text{样本总数}}-N_{\text{对判}}}{N_{\text{样本总数}}}$$

其中 $N_{\text{错判}}$ 和 $N_{\text{对判}}$ 分别代表错误识别和正确识别的样本数，$N_{\text{样本总数}}$ 代表样本总数。

2　结果与讨论

前期工作[5]提供了基于 LDA 算法的分类模型，通过各样本对线性判别函数 LDF(linear discriminant function)1 和 2 的相关系数绘制散点图，如图 2。

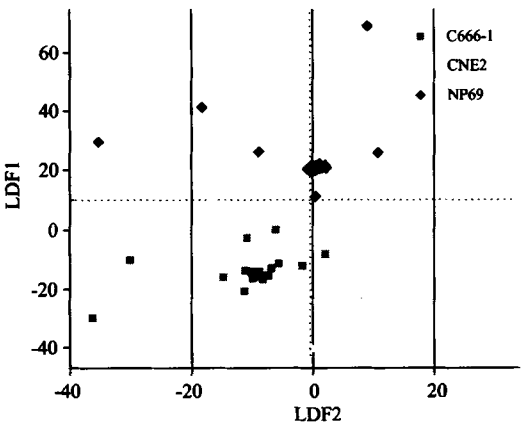

Fig. 2　Scatter plot of the corresponding efficiency of LDF1 vs. LDF2 in LDA analysis

从图中可以看出，NPC 细胞株 C666-1，CNE2 和正常鼻咽细胞株 NP69 可以由 LDF1 和 LDF2 划分为三个组群，分类的结果示于表 1 中。

Table 1　Classification by LDA model: The columns show the true class while the rows refer to the tested class

Cell line	$N_{\text{train}}/N_{\text{test}}$ #	C666-1	CNE2	NP69	Total	E/%
C666-1	20/10	9	1	0	30	3.33
CNE2	21/10	1	9	0	31	3.23
NP69	36/10	1	0	9	46	2.17

N_{train}：代表用于训练的样本数(下同)；N_{test}：代表等待判别的样本数(下同)

从表 1 可以看出，三种细胞株各有一个样本被误判，因此，误差 E 很低且大致相等。由于 C666-1 和 CNE2 同属于

NPC 细胞株，当 NP69 样品被误判为 C666-1 时，判别的特异性就降到了 90%（特异性和灵敏度见表 2）。

Table 2 Sensitivity and specificity of classification results executed by the three algorithms

Algorithm	Sensitivity/%	Specificity/%
LDA	100	90
LIBSVM	100	100
SVC	100	90

在执行 LIBSVM 和自编 SVC 算法的过程中，建立的识别模型也包括训练样本集和待判样本集，数据的设置与 LDA 判别模型的完全相同。采用 SVM 中 C_SVC 算法的 RBF 核函数。对于执行 LIBSVM，核函数的参数的自动优化已经集成在算法中，由此建立的识别模型用于预测未知样品；而对于执行自编 SVC 算法，训练样本集先被投入算法中进行训练和建模，训练完成后，再将待判样本集投入算法中来对各样本的归属进行预测。在此基础上进行识别的效果可由表 2 的特异性和灵敏度两个指标进行评价。同时，表 3 和表 4 中也分别列出了 LIBSVM 和自编 SVC 两种算法的的识别结果。

Table 3 Classification by SVC model of LIBSVM: The columns show the true class while the rows refer to the tested class

Cell line	N_{train}/N_{test}	C666-1	CNE2	NP69	Total	E/%
C666-1	20/10	10	0	0	30	0
CNE2	21/10	0	10	0	31	0
NP69	36/10	0	0	10	46	0

Table 4 Classification by self-programming SVC model: The columns show the true class while the rows refer to the tested class

Cell line	N_{train}/N_{test}	C666-1	CNE2	NP69	Total	E/%
C666-1	20/10	1	0	0	30	0
CNE2	21/10	0	10	0	31	0
NP69	36/10	1	0	9	46	2.17

从表 2 的结果上看，采用 LIBSVM 的 SVC 识别模型和自编 SVC 识别模型都可以实现较高的灵敏度和特异性。在执行 LIBSVM 操作中，特异性和灵敏度均达到了 100%，这就意味着没有样本被误判；而在执行自编 SVC 操作中，误判的样本存在于 NP69 细胞株，此时仅有一个 NP69 样本被误判为 C666-1。由于这些算法的数据构成及训练集(和待判集)与之前的 LDA 判别模型的数据设置完全一致，因此可以说明之前的 LDA 判别模型能够得到这两种 SVC 算法的验证，且都具有较高的灵敏度和特异性。与此同时，在自编 SVC 模型中误判的 NP69 样本与 LDA 模型中误判的 NP69 样本为同一样本，通过观察，其原因是该样本的光谱信噪比较差。然而，从另一个指标，即 5 折交叉验证误差 E 的值上看，情况便稍有不同，如图 3。

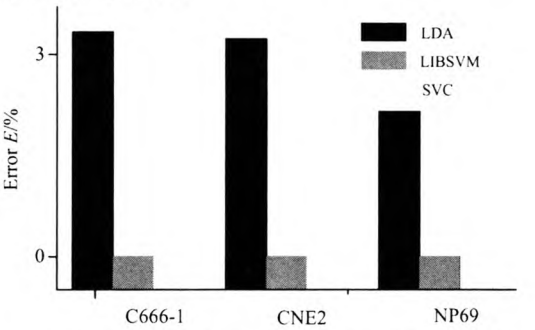

Fig. 3 Comparison of E using different classification algorithms

从图 3 上看，不仅不同细胞株的识别误差不同，而且不同算法的识别误差也不同。与 LIBSVM 的 SVC 算法和自编 SVC 算法的准确率(即 1-E)相比，之前的 LDA 判别模型的准确率是比较低的(约低 2～3 个百分点)。这一现象可能说明利用 SVC 算法(来自 LIBSVM 或自编的)对 NPC 细胞株 C666-1，CNE2 和正常鼻咽细胞株 NP69 进行识别具有一定的优势，这与 SVC 算法自身的特点是息息相关的。因为该判别模型属于小样本，采用 SVC 算法在结构风险最小化原则下进行机器学习，在一定条件下，基本不受非支持向量的样本的影响，这就在一定程度上增强了算法的鲁棒性，因而采用 SVC 算法对鼻咽癌细胞株进行识别使得识别准确率有所提高。但同时也表明与 LIBSVM 的 SVC 算法相比，采用的自编 SVC 算法应该在正确识别信噪比较差的光谱上作一些改进。本工作与前期工作结论一致，并且进一步验证了已有的基于 LDA 算法的有关 NPC 细胞株 C666-1，CNE2 和正常鼻咽细胞株 NP69 的鉴别模型，达到了更高的可靠性。

3 结 论

拉曼光谱技术结合各种多元统计分析算法用于区分正常细胞和肿瘤细胞已被认为是可行的，但是采用不同的算法来验证现有的分类模型其的可靠性也非常重要。采用两种不同的 SVC 算法对 NPC 细胞株和正常鼻咽细胞株的拉曼光谱进行分析和识别，并对之前建立的 LDA 判别模型进行验证。结果表明，两种不同的 SVC 算法都能准确地区分出 NPC 细胞株和正常鼻咽细胞株，具有较高的灵敏度和特异性，从而进一步说明拉曼光谱技术结合各种统计分析方法能够有效的对肿瘤细胞进行鉴别，也为拉曼光谱技术用于鼻咽癌的临床诊断提供有益的参考和实验依据。

References

[1] Parkin D M, Whelan S L, Ferlay J, et al. IARC Scientific Publication, France Lyons. 2003, 155, Lyon.

[2] Puppels G J, de Mul F F M, Otto C, et al. Nature, 1990, 347(6290): 301.

[3] Feng S, Chen R, Lin J, et al. Biosens Bioelectron, 2010, 25(11): 2414.

[4] Feng S, Chen R, Lin J, et al. Biosens Bioelectron, 2011, 26(7): 3167.

[5] Schmid U, Roesch P, Krause M, et al. Chemometr Intell. Lab., 2009, 96(2): 159.

[6] Bensalah K, Fleureau J, Rolland D, et al. Eur. Urol., 2010, 58(4): 602.

[7] LIU You, HUANG Li-qing, WANG Jun, et al(刘　悠, 黄丽清, 王　军, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2012, 32(2): 386.

[8] YE Yu-huang, CHEN Yang, LI Yong-zeng, et al(叶宇煌, 陈　阳, 李永增, 等). Chinese Journal of Lasers(中国激光), 2012, 39(5), 0504003-1.

[9] Vapnik V N. The Nature of Statistical Learning Theory. 1st ed. New York: Springer-Verlag, 1995.

[10] Zhao J, Lui H, Mclean D I, et al. Appl. Spectrosc., 2007, 61(11): 1225.

[11] Chang C, Lin C. ACM Transaction on Intelligent Systems and Technology, 2011, 2(27): 1.

[12] Wang H, Huang Y, Ding H. Proc. SPIE, 2010, 7820: 78201O.

[13] Balabin R M, Safieva R Z, Lomakina E I. Anal. Chim. Acta, 2010, 671(1/2): 27.

Raman Spectral Analysis of Nasopharyngeal Carcinoma and Nasopharyngeal Normal Cell Lines Based on Support Vector Machines

SUN Lei[1], CHEN Yang[1]*, HUANG Yang-wen[2], OU Lin[3], SU Ying[4], FENG Shang-yuan[3], LEI Jin-ping[3]

1. Zhicheng College, Fuzhou University, Fuzhou 350002, China
2. Key Laboratory of Instrumentation Science & Dynamic Measurement(North University of China), Ministry of Education, North University of China, Taiyuan 030051, China
3. Key Laboratory of Optoelectronic Science and Technology for Medicine, Ministry of Education, Fujian Normal University, Fuzhou 350007, China
4. Laboratory of Radiobiology, Fujian Provincial Tumor Hospital, Fuzhou 350014, China

Abstract In the present work, two algorithms of support vector classification (SVC) were utilized to analyze and classify Raman spectra of nasopharyngeal cell lines C666-1, CNE2 and nasopharyngeal normal cell line NP69, and achieved great sensitivity and specificity which are all up to 90%. This is coincident with our previous LDA classification model. The final results show that both of these two SVC algorithms can well classify the cell lines, and meanwhile may be helpful to the realization of Raman spectroscopy to be one of diagnostic techniques of nasopharyngeal carcinoma.

Keywords Raman spectroscopy; Nasopharyngeal carcinoma; Cell; Support vector machines; Linear discriminant analysis

(Received Nov. 9, 2012; accepted Mar. 25, 2013)

* Corresponding author