基于SVM的中医心系证候分类研究*

□徐　琎**　王忆勤**（上海中药大学基础医学院　上海　201203）
邓　峰　夏春明（华东理工大学　上海　200237）
许朝霞（上海中药大学基础医学院　上海　201203）

摘　要：本文介绍了SVM支持向量机的分类技术，以中医心系503个样本为例，利用SVM进行中医心系证候分类研究，实验结果表明，该方法在证候分类中能达到较高的准确率。

关键词：SVM　证候　分类
doi：10.3969/j.issn.1674-3849.2010.05.008

中医诊断过程是一个多源信息的获取、处理和整合的过程，但传统中医诊断往往取决于医师的主观意识、经验累积，受限于当时的环境因素，缺乏客观指标，难以重复验证[1]。就现状而言，中医辨证有几个主要特点：缺乏客观标准；医生辨证技能的高低很大程度取决于经验；辨证结果受主观因素的影响很大，不同医生对同一病人所做的辨证结论往往不同[2]。中医辨证在客观化、定量化、标准化等方面的不足，限制了它的应用和发展。因此，迫切需要运用现代技术及信息科学的理论和方法，进行中医诊断信息获取与处理的理论与技术研究，促进中医诊断学的客观化、规范化。

自1992年首次提出最优边界分类器以来，支持向量机算法不仅在理论方面获得了很大的发展，而且在很多领域得到了成功的应用，如情报分析[3]，人脸检测[4]，故障检测[5]等。本文将支持向量机算法用于中医辨证证候分类研究，可以使辨证更客观，有益于促进中医推广。

一、临床资料

1. 数据来源

全部数据由上海中医药大学中医四诊信息化综合研究实验室于2007年1月~2009年4月在上海交通大学附属仁济医院、复旦大学附属中山医院、上海中医药大学附属龙华医院、曙光医院、上海市中医院及岳阳中西医结合医院6家医院的心内科住院所收集。

2. 病例纳入及排除标准

纳入属于西医内科心血管疾病患者；符合中医“心主血脉”的生理功能失调者；对调查知情者。排除神志不清及语言不清，病情叙述有困难者；兼有脑、肺、肾、肝等脏器的严重器质性疾病者；临床资料严重不全者；拒绝配合者。

收稿日期：2010-01-14
修回日期：2010-03-26
* 科学技术部国家“十一五”科技支撑计划子项目（2006BAI08B01）：中医四诊信息规范采集和融合方法的研究，负责人：王忆勤；上海市科委优秀学科带头人计划项目（09XD1403700）：中医四诊信息融合方法研究，负责人：王忆勤。
** 联系人：徐琎，博士研究生，主要研究方向：中医四诊客观化研究，E-mail：xujin52@gmail.com；王忆勤，教授，博士生导师，主要研究方向：中医四诊客观化、中医证候规范化，Tel：021-51322447，E-mail：wangyiqin2380@sina.com。

3. 数据采集

每个采集小组最少有 1 名主治医师以上职称(或具有博士学位)的专业人员组成。为了数据的客观及有效,定期对采集人员进行培训,保证四诊检测的规范性、一致性。

(1)问诊。

本研究采用自行研制的问诊量表采集,中医心系问诊量表包括基本情况(患者姓名、性别、年龄、吸烟史、饮酒史等)、重点问诊(包括心悸、胸闷、胸痛、气短、浮肿、乏力、心烦、健忘等)、一般问诊(包括寒热、汗、头身胸腹、饮食、二便、睡眠、情绪、妇女等)、既往病史(包括既往健康状况、手术病史、服用激素史、过敏史等)等,共 145 个问诊信息变量,并附有患者的望、切诊内容及中医诊断和西医诊断部分。

(2)舌诊和脉诊。

本研究采用 Z-BOX 型舌脉象分析仪采集。脉象数据包括时域特征参数和频域特征参数;舌象数据包括舌质的色、形参数和舌苔的色、质参数。

(3)面诊。

本研究采用中医面色诊仪采集数据,包括面部颜色、光泽、纹理参数。

(4)闻诊。

本研究采用本课题组与华东理工大学合作开发的声诊采集软件采集,包括声音特征参数。

4. 诊断标准

基于临床流调和专家论证,并参考 1990 年中西医结合心血管学会修订的《冠心病中医辨胸痛或胸闷、心悸、气短、乏力证标准》及中医内科学、中医诊断学七年制教材,拟定心系证型有:心气虚证、心阳虚证、心阴虚证、心血虚证、痰浊证、血瘀证、气滞证、寒凝证、心火亢盛证、心脾两虚证、心肾阳虚证、心胆气虚证、心肺气虚证、心肾不交证、心肝血虚证等,并制定了各证型的诊断标准。由资深中医专家参考诊断标准对每份例进行辨证。

5. 数据库的建立

问诊的信息根据其"有、无",分别赋值"1、0"。采用机器采集舌、面、脉、声等诊的信息,并分析仪器分析系统提供的数据,结果数据都具有离散型。建立四诊信息数据库。

二、数据分析

1. SVM 原理

支持向量机主要解决的是一个二分类问题,其用于分类的基本思想可用图 1 的两维情况说明。如图 1 所示,圆形点和方形点代表两类样本,H 为分类线,H1、H2 分别为过各类中离分类线最近的样本且平行于分类线的直线,它们之间的距离叫做分类间隔(Margin)。所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为 0),而且使分类间隔最大。

对于样本数为 l 的训练样本集 $\{(x_i,x_j),i=1,2,\cdots,l\}$,假定由二类别组成,如果 $x_i\in R^N$ 属于第一类,则标记为正($y_i=1$),如果属于第二类,则标记为零($y_j=0$)。学习的目标是构造一个分类超平面,可以将测试数据尽可能正确地分类。使得分类间隔最大。由此可以得到下面的二次优化问题:

$$\min\frac{1}{2}||w||^2+C\sum_{i=1}^{l}\xi_i \quad \text{st.}$$

$$y_i(w\cdot x_i+b)\geq 1-\xi_i,\xi_i\geq 0,i=1,2,\cdots,l$$

其中 C 为惩罚参数,C 越大表示对错误分类的惩罚越大;b 是分类阈值,可以用任一个支持向量求得;ξ 为训练样本线性不可分时引入的非负松弛变量。C 越大表示对错误分类的惩罚越大。利用 Lagrange 优化方法和 Wolfe(1961)的对偶理论,将上述分类问题转化为对偶问题,得到对偶最优化问题:

$$\max_{w,b}Q(\alpha)=\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j x_i x_i\phi(x_i)\phi(x_i)$$

$$\max\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j x_i x_i K(x_i,x_i)$$

$$\text{st.}\sum_{i=1}^{l}\alpha_i y_i=0,\quad 0<\alpha_i<C,\quad i=1,2,\cdots,l$$

其中 $K(x_i,x_i)=\phi(x_i)\phi(x_i)$称为核函数。当训练集为非线性时,通过一个非线性函数将训练集数据映射到一个高维线性特征空间,转换为高维空间中的

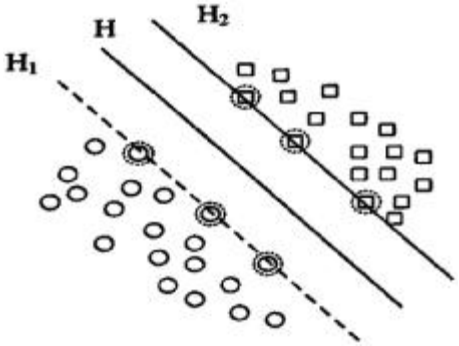

图 1 最优分类线

线性问题，然后在这个高维空间中寻求最优分类面，并得到分类器的决策函数。

最优化求解得到 a 决策函数和参数 b 分别为：

$$f(x)=\mathrm{sgn}\{(w\cdot x_i+b))\}=\mathrm{sgn}\{\sum_{i=1}^{l}\alpha_i y_i(x_i x)+b\}$$

$$b=y_i-w\cdot x_i=y_i-\sum_{x_i\in SV}\alpha_j y_j x_j\cdot x_i$$

不同的核函数将形成不同的支持向量机算法，且分类正确率和各核函数参数取值范围有很大关系，常用的核函数有多项式核函数、径向基核函数和神经网络核函数等。

2. 数据分析过程及结果

准备好样本数据以后，首先对数据进行随即划分，我们抽取其中 350 组为训练数据，另外 153 组为预测数据。由于 SVM 方法每次只能对每一个证候单独训练，因此我们构造了 16 个 2 类分类器。将 16 种证候分开计算，得到各自的训练样本和测试样本。16 个征候对应的训练样本与预测样本数据统计结果见表 2。

表 1 中分 16 个证候类型对 503 组数据进行统计，“0”与“1”分别代表“有”和“无”该证候，例如表格中第一排心气虚，表示训练样本中“0”有 178 个，“1”有 172，预测样本中“0”有 51 个，“1”有 102 个，此训练样本中“0”与“1”的分布比较均衡，说明心气虚的样本比较均衡。

本文采用 RBF 核函数，由于数据结构的不统一，首先对训练样本和预测样本都进行[0,1]之间的归一化处理，之后用 350 组样本进行训练，利用 10 倍交叉验证和 Gunplot 画图软件得出核函数最优的参数 C 和参数 Gramma（以第一个证候为例，如图 2 所示）。

图 2 得到第一个证候训练模型时得到的最优参数 C=8 和 Gamma=0.125，利用该参数的模型对检验样本进行仿真，正确率为 73.20%，在 153 个预测样本中预测正确的个数为 112。依次对其余 15 个证候训练检验，得到的正确率如表 2 所示。心血虚，寒凝，心胆气寒，心肾不交和心肝血虚达到 100%预测正确率，心火亢盛和心肺气虚正确率接近 100%，主要由于训练样本和预测样本的不平衡造成，因为表 2 中用粗体标识的证候，训练样本与预测样本带有明显的不平衡性，例如心血虚训练样本中“0”有 178 个，“1”有 172，预测样本中“0”有 153 个，没有“1”，所以造成了预测正确率达到 100%。而表 3 中，心脾

表 1 16 个证候的样本统计

证候	训练样本(个)		预测样本(个)	
	0	1	0	1
心气虚	178	172	51	102
心阳虚	223	117	126	27
心阴虚	201	149	99	54
心血虚	342	8	153	0
痰浊	268	82	73	80
血淤	282	68	117	36
气滞	337	13	130	36
寒凝	349	1	153	0
心火亢盛	338	12	152	1
心脾两虚	178	172	153	0
心肾阳虚	178	172	153	0
心胆气虚	350	0	153	0
心肺气虚	350	0	148	5
心肾不交	338	12	153	0
心肝血虚	350	0	153	0
其它	330	20	152	1

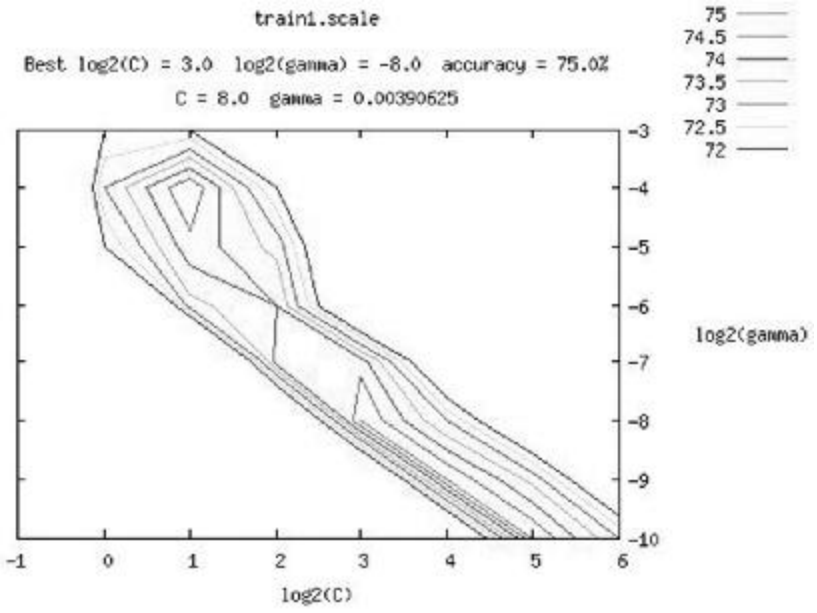

图 2 参数最优化分析曲线

表 2 16 个证候的正确率

证候	参数 C	参数 Gamma	交叉验证正确率(%)	支持向量数	正确率(%)	比例
心气虚	2	0.12500	91.14	181	73.20	112/153
心阳虚	8	0.06250	94.29	143	81.70	125/153
心阴虚	8	0.06250	90.29	168	68.63	105/153
心血虚	2	0.03125	99.14	39	100	153/153
痰浊	8	0.03125	89.14	157	50.33	77/153
血瘀	1	1.00000	92.57	197	76.47	117/153
气滞	4	0.03125	98.29	60	85.62	131/153
寒凝	0.03125	0.03125	99.71	12	100	153/153
心火亢盛	4	0.03125	99.43	47	99.35	152/153
心脾两虚	2	0.12500	91.43	181	43.14	66/153
心肾阳虚	2	0.12500	91.43	181	43.14	66/153
心胆气虚	0.03125	0.03125	100	0	100	153/153
心肺气虚	0.03125	0.03125	100	0	96.73	148/153
心肾不交	4	0.03125	99.43	50	100	153/153
心肝血虚	0.03125	0.03125	100	0	100	153/153
其它	1	1.00000	98	170	99.35	152/153

两虚和心肾两虚正确率仅为 43.14%，也是由于训练样本与预测样本的不平衡性造成。

三、讨 论

近年来国内外学者在中医四诊客观化方面进行了大量的探索性研究，结果表明要实现中医诊断的现代化及客观化，必须让“望、闻、问、切”四诊由功能想象的描述逐渐向微观的、定量的方向过渡。四诊信息的客观化是中医辨证的客观化、规范化的前提和基础。随着科学技术的发展，中医四诊信息融合方法也取得一定的进展，特别是人工神经网络、支持向量机等算法的应用，为中医辨证客观化研究带来了新的思路和方法。近年来，由于支持向量机具有完美的数学形式、直观的几何解释和良好的泛化能力，解决了模型选择与欠学习、过学习问题以及非线性问题，避免了局部最优解，有效地克服了“维数灾难”，且人为设定的参数少，便于使用，使支持向量机以其独特的优势开始在中医证候规范化研究方面得到应用，如：SVM 在中医证候分类方面，孙战全等[6]在证候分类中首先采用非线性主元分析(PCA)对症状信息进行降维处理，之后分别应用神经网络和多类 SVM 对证候进行分类，提出由于中医症状信息本身具有非线性和多维性，运用多类 SVM 比神经网络更具优势。杨晓波等[7]引入了基于先验知识的支持向量机(p-SVM)。张涛等[8]应用 v-SVM 在脏腑器官的分类上面，表明将 SVM 方法运用到中医辨证施治中可以弥补传统中医辨证中的人为不确定性，具有良好的实用价值。

本文利用 SVM 的分类方法，对心系 503 例患者的四诊信息进行分析，进行心系证候的分类研究。结果表明 SVM 方法的引入可以较大程度的提高识别率，是一条可行性很高的途径，但样本的不平衡性能造成预测正确率虚高和偏低，也是进行数据分析时重点关注的问题。

本研究显示，在有先验知识的情况下，SVM 分类的正确率有比较大的提高，提示了 SMV 算法是中医证候客观化、规范化研究中的一个较好的方法；提示统计学理论在中医证候分类中有良好的应用前景。在今后的研究中，我们需要扩大训练样本，提高算法准确率，并将其与人工神经网络、隐结构模型等方法进行比较，选择较好的适用于中医证候分类的方法，为中医诊断客观化、规范化研究做进一步的探索。

参考文献

1 王忆勤，李福凤，燕海霞，等. 中医四诊信息数字化研究现状评析. 世界科学技术-中医药现代化. 2007，9(3)：96~101.

2 张连文，袁世宏主编. 见叶斯网引论. 北京：科技出版社，2006.

3 赵天昀. 多分类 SVM 在企业竞争情报自动分类中的应用. 现代情报，2008，10(10)：184~186.

4 孟庆涛，程童. 基于 SVM 的人脸精确验证. 图形图像，2008，10：89~91.

5 翟永杰，王东风，韩璞. 基于多类支持向量机的汽轮发电机组故障诊断. 动力工程，2003，23(5)：2691~2698.

6 Zhanquan Sun, Guangcheng Xi, Jianqiang Yi. Differentiation of Syndromes with SVM. Advances in neural network, 2006：786~791.

7 Xiao-Bo Yang, Zhao-Hui Liang, Gang Zhang. A classification algorithm for TCM syndromes based on P-SVM. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, 2005：3692~3697.

8 张涛，朱维克，徐自力，等. 支持向量机在脏腑辨证中的应用. 光盘技术，2007，(4)：42~43.

Syndromes Classification of TCM Heart Diseases Based on SVM

Xu Jin[1], Wang Yiqin[1], Deng Feng[2], Xia Chunming[2], Xu Zhaoxia[1]
(1. Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China;
2. East China University of Science and Technology, Shanghai 200237, China)

Abstract: The support vector machine(SVM) is a new kind of machine learning method. Based on the structural risk minimization rule, the SVM has good generalization ability. As the SVM algorithm has been proved to be a convex quadratic optimization problem, any extremal solution is definitely a global optimal solution. This paper introduces the SVM classification techniques, and analyzes 503 cases of heart diseases using the SVM. The results show that this method may help to realize syndromes classification at high precision.

Keywords: SVM Syndrome Classification

(责任编辑:李沙沙,责任译审:张立崴)