·综述·

支持向量机在临床医学中的应用研究进展

张会杰　李恒芬　李湘露

【摘要】 支持向量机是在统计学习理论基础上发展而来的一种新的可训练学习方法，它作为一种有效的分类工具，被广泛应用于临床诊断、医学影像、信号识别、疾病亚型区分、预后判断、基因微阵列等医学领域，并正在成为继神经网络之后新的研究热点，现对支持向量机在临床医学的应用进展进行综述，并对其在临床医学领域未来的应用趋势进行展望。

【关键词】 临床医学；　机器学习；　支持向量机；　综述

doi：10.3969/j.issn.1009-6574.2017.11.013

Application advances of support vector machine in the clinical medicine　*ZHANG Hui-jie, LI Heng-fen, LI Xiang-lu. The First Affiliated Hospital of Zhengzhou University, Zhengzhou* 450052, *China*

【**Key words**】 Clinical medicine;　Machine learning;　Support vector machine;　Review

支持向量机(Support Vector Machine，SVM)作为基于统计学习理论的新的机器学习方法，由于其在实际分类问题应用中所展示的出色学习性能和泛化能力，已在诸多领域得到广泛应用。相对而言，SVM在医疗卫生领域的应用起步较晚，但近年来随着医疗信息化工程的快速发展，医疗信息的数字化、大数据化，催生了对医疗数据分析领域的研究，从而促使SVM成为辅助医学的有效数据分析工具，在临床诊断、医学影像、信号识别、疾病类型区分、预后判断、基因微阵列等应用中发挥出愈来愈重要的作用。本文将对近年来SVM在临床医学中的应用状况进行综述，并展望其应用前景。

1　概述

Cherkassky自20世纪60~70年代开始致力于统计学习理论方面的研究，但是直到1995年，才正式提出这种基于统计学习理论的机器学习方法–SVM[1-2]。SVM是在统计学习理论(Statistical Learning Theory，SLT)[2]和VC维的基础上发展起来的一门专门研究有限样本的分类学习方法。统计学习理论脱离了传统统计学研究的样本数目趋于无穷大时的渐进性能，能有效地解决过学习问题，具有很好的泛化能力和较好的分类精准性，它在解决小样本、非线性及高维模式识别等问题中表现出许多特有的优势，在很大程度上克服了过学习及数据高维化的复杂运算的问题。

基金项目：国家自然科学基金项目(81371494)

作者单位：450052 郑州大学第一附属医院精神医学科

通讯作者：李恒芬　Email：lihengfen@sohu.com；李湘露　Email：li_xl@163.com

SVM的基本原理[3]是通过寻找一个能够满足分类要求的最大分离超平面，同时保证分类精度的情况下使分类间隔达到最大，最终使SVM问题转化为一个受约束的凸二次规划(Quadratic Programming，QP)问题。理论上，SVM能够实现对线性可分数据的最优分类。

SVM还是一种基于核函数的学习机器，其泛化能力在很大程度上依赖于所选择的核函数。SVM的学习效率取决于样本数据集的规模，为了不断提高现有的SVM对于实际问题的大规模样本数据集的训练效率，研究者们渐渐将SVM与其他学科不断融合构造出的新型支持向量，以期望降低分类误差，提高训练效率及泛化能力，并且取得了很多成果。由于SVM算法的潜在应用价值，吸引了国内外很多的学者，尤其近几年，出现了许多新的SVM算法。如2001年Inoue等[4]将模糊数学和SVM结合，提出的F–SVM模糊SVM算法，主要用来处理训练样本中的噪声数据，优化了分类效果。2005年Chen等[5]提出了v–SVM算法，参数v的引入便于在不均衡数据中集中发现孤立点，很好的简化支持向量，进而提高SVM算法的速度。可见，对于SVM算法本身的完善和改进将会促进SVM的应用研究领域的拓展。

2　SVM在临床医学中的应用

在医学研究中，有时取得足够的样本量并非一件易事，面对复杂的医学数据，SVM为我们提供了一种新的方法。2001年Dreiseitl等[6]首次将SVM用于诊断皮肤色素病变，SVM模型开始出现。随之SVM被应用于影像医学、检验学诊断，取得较好的效果。

经过最近几年的应用研究和推广，尤其是与其他分类优化模型、技术相结合，使得SVM在临床医学的应用领域更为广泛和深入，不仅扩大了临床诊断的疾病类型，提高了疾病分类的精度，提升了基因诊断与治疗的选择有效性，更在建立诊断系统、医疗大数据模型等方面做出有益尝试。基于这种新兴的机器学习方法SVM在分类问题方面表现的良好学习和泛化能力，现已被广泛应用于诸多研究领域。

2.1 临床诊断 SVM最成熟的分类应用是解决两类问题的判别分类，所以在临床诊断中的应用最早起源于肿瘤良恶性的判别，即癌症的辅助诊断研究，包括皮肤癌、脑神经胶质细胞瘤、乳腺癌、胃癌的诊断等。SVM作为有效地癌症分类方法，其分类的精度依赖于它的核心参数和输入特征，研究者为进一步提高其分类预测的准确度，常常以SVM为基础算法将其他算法与SVM相结合来进一步优化支持向量机。有研究发现[7]，SVM递归特征消除法(SVM-RFE)对每个特征得分进行排序，来逐渐剔除冗余特征，再用SVM分类算法进行卵巢癌疾病诊断，这种改进的模型得到了96.64%的准确率，94.59%的灵敏度；为了克服训练样本中的噪声数据，文献[8]也提出将模糊变换和SVM分类器相结合的疾病诊断模型，在多种疾病数据集上进行诊断应用。医疗疾病的诊断模型是常见的监督分类问题，利用SVM能自动学习的能力，有研究[9]提出以SVM学习技术为智能诊断模型的核心，将医学知识的特征提取过程与基于计算机智能的特征选择机制如朴素贝叶斯、SMO优化算法、K-邻算法相对比，SVM大大优化了诊断模型，丰富了临床诊断模型的思路和研究方向。2015年梁丽军等[10]提出的结合弹性网和SVM的疾病诊断模型，有效地去除了原始样本中的不相关特征，提高了诊断系统的泛化能力。2016年张萍萍等[11]分别应用BP神经网络算法和SVM算法建立胃癌的数学辅助诊断模型，并通过测试集评价其效果，验证支持向量机诊断模型的诊断准确率、敏感性、特异性均高于BP神经网络算法建立的诊断模型。可预见SVM在临床诊断模型对某些疾病的早期预测及诊断有重要的参考价值。

2.2 影像诊断 SVM在医学影像中的应用已显现出比传统方法更大的优势。2003年Change等[12]利用SVM对乳腺癌的超声影像进行研究，分析了250例肿瘤患者，提出了基于SVM的图像特征因子构建方法，经SVM训练和测试，总准确率85.6%，灵敏度达到了95.45%，并证实SVM的分类能力及训练速度高于多层感知的神经网络。医学影像中，脑MRI是发现脑结构和诊断脑疾病细微解剖结构的重要一环，有文献[13]提到利用SVM对大脑灰质、白质、脑脊液等脑组织进行了的图像自动分割实验，实验结果表明SVM方法在对小样本、目标边界模糊目标灰度不均及不连续等情况下图像分割，其应用也许是一种有效的方法。由于SVM在学习能力和处理非线性问题的独特优势，有学者[14]提出了一种使用SVM的SAR图像识别方法，充分利用了SVM的分类能力。2013年，张德发和何亮[15]利用粗糙集和SVM优点提出基于粗糙集和SVM相融合的图像分割算法。该方法利用粗糙集图像区域特征进行约简，以降低特征向量维数，然后采用SVM对这些特征进行学习，建立图像分割模型，从而实现图像的分割。最终不仅提高图像分割精度，缩短训练时间，也很好地满足了图像处理的实时性要求。

2.3 电生理诊断 将SVM及其与优化方法的结合应用于脑电信号的分析，获得了良好的识别效果。研究发现[16]，采用小波分析对心音信号进行降噪预处理，将提取的心音信号的Mel频率倒谱系数作为心音信号特征，采用SVM建立信号分类器，对采集心音信号数据的识别性能进行验证。结果表明，该方法可提高心音信号自动识别的正确率。脑电信号有高度的复杂性、非线性和非平稳性，信号特征会随受试者的年龄和心理状态而变化，有研究表明[17]，抑郁症患者脑电信号在节律、波形幅度和功率谱幅值等参数中存在有别于健康人的特征，该研究采用小波变换和小波包变换的能量特征提取方法及特征向量估计方法对脑电信号功率谱幅值进行特征提取，将提取的特征参数作为分类器的输入向量，然后用SVM分类器进行训练和分类测试，发现抑郁症患者和健康人的脑电信号在能量特征和功率谱特征上表现出较高的模式可分性，并取理想的分类效果，准确率达到了98%。该研究将为精神抑郁症的病理临床诊断提供一定的参考依据，将为未来研究基于自发脑电的抑郁症疾病诊断实用系统开发奠定实践基础。

2.4 判断疾病的预后 SVM在疾病预后判断中的应用已逐渐成熟，有效地提高了预测的准确率。有研究表明[18]，采用583例宫颈癌患者术后资料作为实例分别训练和建立两个SVM的模型，将Logistic回归的结果与两个SVM模型的结果进行比较。结果表明，SVM技术的结果比Logistic回归的预测结果在灵敏度、特异度、Youden指数和错分率方面均要好，SVM的技术方法值得在疾病预后的领域中进一步应用推广。此外2013年张黎明等[19]对SVM及Logistic回归模型两种方法对结肠腺瘤高级别上皮内瘤变的发生的预测结果进行比较，在应用SVM预测部分进行反复的训练及预测，得出平均预测正确率及特异度、敏

感度，相比较Logistic回归分析部分，得出应用SVM建立的预测模型在小样本的基础上对结肠腺瘤发生高级别上皮内瘤变可以获得了较好的预测效果的结论。

2.5 基因微阵列数据分析 由于基因微阵列数据通常具有容量大、高维度、小样本、非线性等特征，对其进行数据分析有较大难度，使得无法用传统的机器学习方法来做分类识别，且存在的大量与分类无关的噪声基因造成分类性能严重下降，而SVM在解决小样本、非线性、高维模式问题中具有突出的优势，因此在分析基因表达数据中得到了广泛的关注。这对获取特征基因和定位药物靶点起到重要作用。有研究发现[20]，在疾病的诊断中，通常只有少部分基因决定着疾病的发生。无论是从诊断还是从治疗的角度看，都有必要通过数据挖掘技术将少数致病基因选择出来。

利用机器学习从海量的基因表达谱数据中寻找对疾病有鉴别力或与疾病相关的特征基因，遗传算法模拟生物的自然进化过程，可以避免局部极值问题，在结合SVM自身优点，有文献[21-22]提出将遗传算法与SVM相结合的基因选择算法，进而选择出数目较少的关键诊断基因。基因在复杂疾病的病因学中可能存在着遗传效应，基因与基因之间有时存在某种相互作用，当疾病的变异是罕见的或者样本非常有限，尤其样本自变量存在非线性关系时，传统的统计方法将非常有限，基于SVM在解决非线性问题中的独特优势将提供给大家一种新的思路，有文献[23]表明，将SVM与神经网络、随机森林等其他算法进行比较后，发现对于基因表达谱数据而言，SVM的处理效果较佳。

2.6 其他领域的应用 由于SVM的优越性，其应用研究的开展目前已日渐广泛，贝尔实验室率先对美国邮政手写数字库识别研究方面应用了SVM方法，取得了较大的成功。在随后的近几年内，有关SVM的应用研究得到了很多领域的学者的重视，在文本识别、人脸图像识别、语音识别、蛋白质结构预测[24-25]及其他应用研究等方面取得了大量的研究成果。

3 展望

统计学习理论系统地研究了机器学习问题，尤其是在有限样本情况下的统计学习问题。这一理论框架下产生的SVM是一种通用的机器学习新方法，在理论和实际应用中表现出很多优越的性能，并在其理论与算法及应用研究方面取得了长足的进步，但与其理论研究相比，应用研究还比较滞后，目前只有比较有限的实验研究报道，且多属于仿真及对比试验。SVM最初是基于二分类问题提出的，而在实际的问题中，我们需要处理大量的多分类问题，因此，对多分类问题建立快速有效地算法也是一个亟待解决的问题。未来，随着相关研究方法的扩展与更新，特别是针对多分类变量数据分析方法及软件被开发后，SVM在临床医学领域的应用会更广泛，可为疾病早期诊断、后期预防、高危人群中的筛查等提供可靠的依据。

参 考 文 献

[1] Cherkassky V. The nature of statistical learning theory[J]. IEEE Trans Neural Netw, 1997, 8(6): 1 564.

[2] 张学工.关于统计学习理论与支持向量机[J].自动化学报, 2000, 26(1): 32-42.

[3] Huang GB, Zhou H, Ding X, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE Trans Syst Man Cybern B Cybern, 2012, 42(2): 513-529.

[4] Inoue T, Abe S. Fuzzy support vector machines for pattern classification[C]// International Joint Conference on Neural Networks, 2001. Proceedings. IJCNN. IEEE Xplore, 2001: 1 449-1 454.

[5] Chen PH, Lin CJ, Scholkopf B. A tutorial on v-support vector machines[J]. Appl Stoch Models Bus Ind, 2005, 21(2): 111-136.

[6] Dreiseitl S, Ohnomachado L, Kittler H, et al. A comparison of machine learning methods for the diagnosis of pigmented skin lesions[J]. J Biomed Inform, 2001, 34(1): 28-36.

[7] Wang R, Li R, Lei Y, et al. Tuning to optimize SVM approach for assisting ovarian cancer diagnosis with photoacoustic imaging[J]. Biomed Mater Eng, 2015, 26 Suppl 1: S975-S981.

[8] Li DC, Liu CW, Hu SC. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets[J]. Artif Intell Med, 2011, 52(1): 45-52.

[9] Nahar J, Imam T, Tickle KS, et al. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach[J]. Expert Syst Appl, 2013, 40(1): 96-104.

[10] 梁丽军, 刘子先, 王化强, 等.基于弹性网-SVM的疾病诊断关键特征识别[J].计算机应用研究, 2015,(5): 1 301-1 304; 1 308.

[11] 张萍萍, 张建华, 尹咪咪.血清标记物检测结合智能算法在胃癌诊断中的应用[J].郑州大学学报(医学版), 2016, 51(2): 196-200.

[12] Chang RF, Wu WJ, Moon WK, et al. Support vector machines for diagnosis of breast tumors on US images[J]. Acad Radiol, 2003, 10(2): 189-197.

[13] Yazdani S, Yusof R, Karimian A, et al. A Unified Framework for Brain Segmentation in MR Images[J]. Comput Math Methods Med, 2015, 2 015: 829 893.

[14] Wang A, Zhao C. SAR Image Feature Extraction and Target Recognition Based on Contourlet and SVM[C]// International Conference on Computer and Electrical Engineering, 2012.

[15] 张德发, 何亮.基于粗糙集和支持向量机的图像分割技术研究[J].计算机应用与软件, 2013, 30(6): 111-113; 120.

[16] 郭春璐, 岳小冰.基于支持向量机的心音信号自动识别[J].计算机与现代化, 2016(6): 36-39.

[17] Acharya UR, Sudarshan VK, Adeli H, et al. A Novel Depression Diagnosis Index Using Nonlinear Features in EEG Signals[J]. Eur Neurol, 2015, 74(1/2): 79-83.

[18] 周舒冬，张磊，叶小华，等.支持向量机技术在疾病预后中的应用和比较[J].数理医药学杂志，2007，20(6)：760-762.
[19] 张黎明，刘玉兰，康晓平.支持向量机预测结肠腺瘤高级别上皮内瘤变效果研究[J].中国实用内科杂志，2013，33(11)：872-875.
[20] Li S，Wu X，Tan M. Gene selection using hybrid particle swarm optimization and genetic algorithm[J]. Soft Computing，2008，12(11)：1 039-1 048.
[21] Luquebaena RM，Urda D，Subirats JL，et al. Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data[J]. Theor Biol Med Model，2014，11Suppl1：S7.
[22] Tong M，Liu KH，Xu C，et al. An ensemble of SVM classifiers based on gene pairs[J]. Comput Biol Med，2013，43(6)：729-737.
[23] Koo CL，Mei JL，Mohamad MS，et al. A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology[J]. Biomed Res Int，2013，2 013(1)：432 375.
[24] Nanni L，Brahnam S，Lumini A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition[J]. J Theor Biol，2014，360：109-116.
[25] Li Z，Wang J，Zhang S.A new hybrid coding for protein secondary structure prediction based on primary structure similarity[J]. Gene，201，618：8-13.

(收稿日期：2017-10-08)

·读者·作者·编者·

本刊文稿中缩略语的书写要求

在本刊发表的学术论文中，已被公知公认的缩略语在摘要和正文中可以不加注释直接使用(表1)；不常用的和尚未被公知公认的缩略语以及原词过长、在文中多次出现者，若为中文可于文中第1次出现时写明全称，在圆括号内写出缩略语，如：流行性脑脊髓膜炎(流脑)；若为外文可于文中第1次出现时写出中文全称，在圆括号内写出外文全称及其缩略语，如：阿尔茨海默病(Alzheimer Disease，AD)。若该缩略语已经公知，也可不注出其英文全称。不超过4个汉字的名词不宜使用缩略语，以免影响论文的可读性。西文缩略语不得拆开转行。

表1 神经疾病与精神卫生杂志常用缩略语

缩略语	中文全称	缩略语	中文全称	缩略语	中文全称
CNS	中枢神经系统	CSF	脑脊液	GABA	γ-氨基丁酸
IL	白细胞介素	AD	老年痴呆症(阿尔茨海默病)	PD	帕金森病
MRI	磁共振成像	CT	电子计算机体层扫描	DSA	数字减影血管造影
PCR	聚合酶链式反应	EEG	脑电图	MR	磁共振
HE	苏木素-伊红	BDNF	脑源性神经营养因子	PET	单光子发射计算机断层扫描
SOD	超氧化物歧化酶	ELISA	酶联免疫吸附剂测定	CRP	C反应蛋白
MMSE	简易精神状态检查	NIHSS	美国国立卫生研究院卒中评分	TIA	短暂性脑缺血发作
TNF	肿瘤坏死因子	WHO	世界卫生组织	HAMD	汉密尔顿抑郁量表
HAMA	汉密尔顿焦虑量表	PANSS	阳性与阴性症状量表	rTMS	重复经颅磁刺激
5-HT	5-羟色胺	SSRIs	选择性5-羟色胺再摄取抑制剂	MoCA	蒙特利尔认知评估量表
PTSD	创伤后应激障碍	ICD-10	国际疾病分类第十版	DSM	美国精神障碍诊断与统计手册
CCMD-3	中国精神障碍分类与诊断标准第3版				