ADST:用机器学习方法鉴别结节病和肺结核

陈蔼祥[1] 陈智锋[2]
(广东财经大学数学与统计学院 广州 510320)[1] (肇庆市第一人民医院 肇庆 526021)[2]

摘 要 结节病和肺结核的临床鉴别诊断目前仍然是困难的。搜集了 106 例结节病和肺结核的对比资料,并筛选出对分类有意义的临床指标作为特征,将其进行必要的量化和缩放形成训练数据,然后分别用支持向量机(SVM:Support Vector Machine)、决策分类树(DCT:Decision Classification Tree)、朴素贝叶斯(NB:Naïve Bayes) 3 种不同的方法进行训练,并用 5 倍交叉验证评估各种不同的模型的有效性。实验结果表明,这 3 种方法在识别结节病时对应的 ROC 曲线下的面积分别为 0.978,0.96,0.690,得到的测试精度分别达到 100%,96.15%,96.15%,训练精度分别为 95.28%,90.57%,92.38%。用这 3 种方法得到的分类器对 19 例临床未能确诊的病患进行预测,DCT 方法的预测结果与 SVM 方法的结果高度吻合(19 例中仅 1 例预测结果不同),而 NB 方法预测结果稍差(19 例中有 3 例与 SVM 预测结果不一致)。实验结果表明,3 种方法中,SVM 方法的分类能力和分类精度最高。临床实验结果表明,19 例临床未能确诊的病患按照 SVM 算法预测的结果进行治疗均得到了康复。

关键词 结节病,肺结核,DCT,NB,SVM
中图法分类号 TP181 **文献标识码** A

ADST:Approache of Automated Differentiating Sarcoidosis from Tuberculosis Based on Statistical Learning Theory

CHEN Ai-xiang[1] CHEN Zhi-feng[2]
(School of Mathematics and Statistics,Guangdong University of Business Studies,Guangzhou 510320,China)[1]
(Zhaoqing No. 1 People's Hospital,Zhaoqing 526021,China)[2]

Abstract Differentiating sarcoidosis from tuberculosis is still difficult. The support vector machine is a powerful tool in statistical learning. In this paper,we collected 106 cases of sarcoidosis and tuberculosis,used an SVM to build a disease classifier named ADST(Automated Differentiating Sarcoidosis from Tuberculosis). In order to get the raw medical data into a form usable by SVM,we extracted feature vectors of the raw medical data by turnning the qualitative feature into digital one and dropping the features that do not have much classification value. Then ADST conducts simple scaling on the data,uses cross-validation to find the best parameter of model,uses the best parameter to train the whole training set to obtain the SVM model. Finally ADST uses the resulted SVM model to predict a new patient case. The experiment result shows that the ROC areas of SVM,DCT and NB are 0.978,0.96,0.690 respectively,and the training accuracy is 95.28%,90.57%,92.38%,and test accuracy is 100%,96.15%,96.15%. Clinical pratice shows that the classification result is correct:19 cases of undiagnosed patients are recovered after treatment according to the results of the diagnosis of ADST.

Keywords Sarcoidosis,Tuberculosis,Statistical learning,SVM

1 引言

医学上,结节病是一种病因未明、可同时或相继侵犯多系统的非干酪坏死性肉芽肿性疾病[1]。一方面,由于结节病与结核病在临床、病理、免疫特点方面极为相似,且两者均多发于肺部,这导致两者的鉴别诊断非常困难,临床上经常可见类似"组织内可见上皮样肉芽肿结节,请结合临床判断为结节病或结核病"或"请结合临床排除结节病后考虑结核"等这样模棱两可的病理报告。但另一方面,对这两种疾病的治疗却又截然不同[2,3]:前者的治疗以类固醇激素为代表的免疫抑制制药物为主,而后者则采用抗结核治疗。因此,对结节病和结核病的鉴别诊断困难但又非常重要。

国内有大量的关于结节病的某些临床指标的分析研究,例如邹兰芳等人[4]对结节病纵隔、肺门淋巴结氟-氟代脱氧葡萄糖(F-FDG)代谢显像的特征表现进行了分析。黄燕等[5]对不同时期结节病患者外周血淋巴细胞亚群 Th7 细胞的变化

本文受国家自然科学(60773201),广东省自然科学基金(10451032001006140,2012040006785),广州市科技和信息化局应用基础研究项目(10C12140131),广东省教育厅普通高校育苗工程(LYM10081),肇庆市科技创新计划(2011E241)资助。
陈蔼祥(1978-),男,博士,副教授,主要研究方向为统计机器学习、智能规划,E-mail:cax413@163.com;陈智锋(1978-),男,主要研究方向为呼吸道疾病的临床诊治。

表达进行了研究。刘长军等[6]对结节病的CT诊断进行了研究。也有很多学者对结节病和结核病的鉴别诊断进行了系统性的研究。例如叶秋月等[7]则综合分析肺结节病和肺结核的临床表现、胸部CT表现、实验室检查、支气管镜检查、支气管肺泡灌洗液、结核菌素实验、T-SPOT.TB等，得到这两种疾病具有统计学差异的指标；李秋红、周瑛、李惠萍等[8-11]对实时萤光定量PCR检测结核分枝杆菌DNA(TB-PCR)技术进行了研究，并通过选取有统计学意义上的指标，建立了一套临床-病理-影像综合评分系统来对结节病与菌阴性结核病进行鉴别诊断。美国[12]和荷兰等[13]相继进行了由多中心参与的关于结节病的大规模的、系统的研究。然而，无论是国内还是国外学者的这些研究，都未能提出结节病区别于其他疾病的鉴别要点。

目前临床上结节病的诊断主要建立在临床/影像符合，结合病理发现非干酪样上皮样肉芽的基础上，同时参考血清血管紧张素转化酶(SACE)、结核菌素试验(TST)、支气管肺泡灌洗液(BALF)中T淋巴细胞及其亚群结果等指标综合诊断。由于结节病和结核病两者高度相似，且需要综合考虑的指标太多，要临床医生肉眼识别这些指标上的差异并作出准确判断是困难的。总体而言，综合诊断的过程繁琐，可操作性差，有时需要诊断性治疗。

近年来，随着统计机器学习研究领域的进展，通过统计机器学习技术理解某种疾病与临床指标之间的相关性越来越受到相关研究人员的重视。尤其是面对病因未明、传统诊断方法效果不佳甚至失效的情况，统计学习技术不失为一种有效的补充手段[21]。Trevor Hastie等[14]给出了统计机器学习技术在医学信息领域中的两个应用例子，一个是在前列腺特异性抗原水平与肿瘤大小、肿瘤重量、前列腺增生程度等一序列临床指标之间建立回归模型，实现对前列腺癌的识别和预测。另一个则是将统计学习技术应用于基因信息分析，试图识别(排除)某些肿瘤癌细胞高(低)表达基因。刘伟等用统计机器学习方法发现了近1976个潜在癌症特征[21]。

从统计机器学习的角度看，结节病和结核病的鉴别诊断本质上是一分类问题，即从一序列指标构成的样本集中进行有效分类。目前，已开始有关于用统计机器学习对结节病某一特定指标进行分类识别的研究。例如曹蕾等[15]在已实现的疑似肺结节图像分割的基础上，提取肺结节图像多维特征，应用LDA和SVM统计分类器，通过对大量样本的训练，实现对肺结节CT图像的自动检测和诊断。张婧等[16]为识别CT图像中的肺结节，提出了一种结合规则和支持向量机(SVM)的识别方法，来对分割出来的感兴趣区域(ROI)进行分类，试图提高分类的正确率。但在未有结节病区别于其他疾病的鉴别要点之前，这种对单一指标进行分类识别的方法对结节病本身的识别效率是非常有限的。

本文试图通过搜集结节病和结核病病人的相关数据，包括病人的主诉、临床症状、病理、医学影像资料，以及实验室检查、支气管镜检查、支气管肺泡灌洗、结核菌素试验等各种检查数据在内的资料，构成训练样本，采用支持向量机[1)](SVM)、朴素贝叶斯、决策分类树3种统计机器学习算法分别进行训练。通过ROC曲线(Receiver Operating Characteristic)下面积[2)]、训练精度、测试精度这3个重要指标衡量算法结果的可靠性。对比3种不同方法的预测结果，选择最可靠的结果进行临床验证。实验结果和临床检验结果均表明了我们方法的有效性。

本文第2节将从某医院的电子医疗记录中搜集被确诊为结节病和结核病的病人的相关资料，构成训练样本/医疗知识库，同时，获得一些疑似结节病或结核病的病人资料。并将这些原始医疗数据转换成能被学习算法接受的训练数据。第3节介绍本文中使用的3种学习算法NB、DCT、SVM。第4节给出实验结果。最后我们对结果进行了总结和讨论，并指出未来进一步研究方向。

2 训练数据

统计机器学习方法的应用面临的首要问题是如何获得高质量的足够的训练样本数据。应用领域的数据往往比较杂乱，很多是文字、图象等形式的资料，有时会面临重要资料缺失，而有时又存在过多无用的资料，对很多原始资料的理解是领域相关的工作。同时，由于一般学习算法对输入数据有相应的格式，例如，SVM能接受的数据格式是实数形式的数据，因此，我们还需要将原始数据作相应的编码和转化，使之能满足算法对输入数据的要求。下面将从原始数据搜集和准备、数据变换和特征选择、训练数据集的表示几个方面讨论如何获取必要的训练数据。

2.1 原始数据的搜集和准备

为了获得足够的训练样本，我们有选择地搜集了某医院的结节病和结核病的原始临床医疗数据。由于两种疾病的高度相似性，我们搜集这些临床数据时采取宁缺勿滥的原则，对于只要确诊为结节病或结核病的病患，我们尽可能搜集该病患的详细资料，包括年龄、性别、既往病史、有无复发记录、并发症、病人主诉、病程、职业以及各种详尽的临床诊断检查数据。这些原始数据有些是文字性的，比如病人的主诉通常用文字语言表述为“咳嗽、活动后气短3月、发热20天”。有些是数字性的，比如体温、年龄等。有些则是医学影像资料，比如CT图像。我们一共搜集了125例病患资料，其中106例病患已被确诊为结节病或结核病，有19例为临床医生无法确诊的疑似病例。

2.2 数据的变换和特征选择

在搜集的125例病患原始资料中，我们在与临床医生充分讨论的基础上，排除了一部分对临床鉴别完全没有意义的指标，例如“年龄”这个因素对区分结节病还是结核病意义不大，因此我们舍去“年龄”这个指标。又比如，在106例被确诊的病患中，无论是结核病，还是结节病，“六胺银染色阳性”这一指标均取相同值，这表明该指标对鉴别诊断意义不大，因此，我们排除这些指标。但是对于那些无法取舍的指标，我们均予以保留。

此外，由于原始数据中，存在一些文字性的特征描述，我

1) SVMs是目前统计学习领域最佳的监督学习算法(大多数研究人员相信SVMs是事实上最佳的监督学习算法)。

2) ROC曲线下面积(the area under ROC curve，AUC)衡量分类器的整体性能，曲线下面积越接近1，说明该测试的分类效果越好。

们需要对这一部分指标进行处理，把它转换成相应的特征。例如，病人的“主诉”往往用文字描述为“咳嗽、痰多、憋喘 2 个月”，则该病人的“主诉”将被转换成“咳嗽”、“痰多”、“憋喘”、“病程” 4 个特征，其中前 3 个是二值变量，用“是”或“否”表示对应的特征是否出现，而第 4 个特征则可直接用数字进行量化。

原始数据经过如此处理后，我们得到了近 100 多个特征，考虑到临床医学读者的需要，列出了我们使用的这些特征如下：

气短，咳嗽，痰多，咯血，乏力，盗汗，皮疹症状，发热，病程，Tmax，胸痛，喘憋，体重下降，胸闷，淋巴结肿大，结节红斑，皮下结节，肺部爆裂音，湿罗音，哮鸣音，呼吸音减低，PPD test 阳性，CRP(mg/L)，SACE，抗 TB 抗体阳性，ESR(mm/1h)，肝功能异常，肾功能异常，心功能异常，T 细胞亚群异常，Ig 异常，蛋白电泳异常，ESAT-6(SFCs/10^6PBMC)，CFP-10(SFCs/10^6PBMC)，阻塞性通气功能障碍，限制性通气功能障碍，FEV1%，FEV1/FVC%，肺功能正常，粘膜充血水肿，管腔狭窄，管腔堵塞，结节，粟粒样，分嵴增宽，粘膜粗糙，脓性分泌物，脓性分泌物，碳末沉积，病变位置——左上叶，病变位置——左下叶，病变位置——左舌叶，病变位置——右上叶，病变位置——右中叶，病变位置——右下叶，病变位置——右中间段，病变位置——正气道，病变位置——隆突，病变位置——左主支气管，病变位置——右主支气管，病变位置——广泛，结节影，粟粒状，片状影，纤维索条影，毛玻璃影，网格影，钙化灶，胸腔积液，胸膜增厚，肺内空洞，斑点影，弥漫性间质改变，沿支气管血管束分布，胸膜下分布，影像学——肺内病变. 其他，纵隔，单侧肺门，双侧肺门，伴有钙化，痰抗酸染色，BALF 抗酸染色，其他找到抗酸杆菌，灌洗位置——右中叶，灌洗位置——左舌叶，灌洗量 ml，回收量 ml，细胞分类——总数，细胞分类——淋巴%，细胞分类——中性%，细胞分类——嗜酸%，T 亚群——CD3%，T 亚群——CD4%，T 亚群——CD8%，T 亚群——CD4/CD8，材料——TBLB，材料——CT 引导下肺活检，材料——淋巴结活检，材料——结节活检，材料——开胸肺活检，材料——外科切除组织，上皮样肉芽肿，干酪性坏死，抗酸染色阳性，抗酸染色阴性，病理结果(融合)，普通坏死，多核巨细胞。

根据这些特征，我们制作了结节病和结核病的对比资料，形成了“结节病和结核病的对比资料库”。这一阶段形成的数据将作为我们的 ADST(Automated Differentiating Sarcoidosis from Tuberculosis)系统的输入数据。因此，一旦得到了“结节病和结核病的对比资料库”后，后面所有工作步骤不再需要任何形式的人工干预，所有数据的进一步处理、训练、预测等工作都由 ADST 系统自动完成。

2.3 训练数据集的表示

由于大多数学习算法只能接受数字表示的训练数据，我们需要在“结节病和结核病的对比资料库”的基础上，进一步将之量化成数字特征，形成学习算法能接受的训练数据。在 ADST 系统中，这部分的工作由数据预处理模 preprocess 块完成。

我们将用一行向量来表示一个病患，向量的长度取决于我们使用的特征的个数。具体地，如果一个病人出现了特征库中第 i 个特征所表示的症状，则 $x_i=1$，否则 $x_i=0$。例如，下列向量：

$$x^T=\begin{bmatrix}1\\0\\0\\\vdots\\39\\\vdots\\0\end{bmatrix}\begin{matrix}\text{气短}\\\text{咳嗽}\\\text{痰多}\\\vdots\\T_{\max}\\\vdots\\\text{多核巨细胞}\end{matrix}$$

被用来表示某一病人有“气短”症状，“$T_{\max}$”取值为 39，但未见“咳嗽”、“痰多”、“多核巨细胞”等临床症状。

由于我们的 ADST 系统是一个二分类系统，因此，我们用布尔变量 y 表示分类信息，当 $y=1$ 时，表示该病人为结节病患，当 $y=0$ 时，表示该病人是结核病患。

约定行向量 $x^{(i)}$ 表示第 i 个病患的资料，$y^{(i)}$ 表示该病患所患疾病类别(1 结节病，0 结核病)，则 m 个病人的训练数据可表示成以下矩阵形式：

$$T=[X\quad \vec{y}]$$

其中，$X=\begin{bmatrix}x^{(1)}\\x^{(2)}\\\vdots\\x^{(m)}\end{bmatrix}$，$\vec{y}=\begin{bmatrix}y^{(1)}\\y^{(2)}\\\vdots\\y^{(m)}\end{bmatrix}$

因此，数据预处理模块 preprocess 的功能主要是通过读取“结节病和结核病的对比资料库”，产生数值格式的训练样本集 T。

3 方法

有了训练数据后，我们希望使用统计机器学习方法自动识别结节病和结核病。具体地，给定训练集 T，我们希望能学习从特征空间 χ 到类别空间 γ 中的映射函数(识别/分类函数) $h:\chi\to\gamma$，使得当给定特征空间中的某一样本点 x 后，$h(x)$ 能预测 x 所属类别 y。学习和预测的过程如图 1 所示。

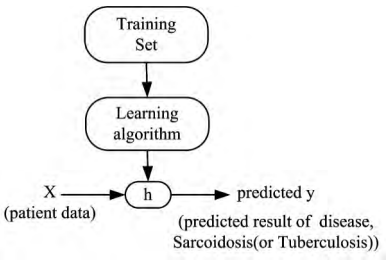


图 1 ADST 系统工作流程

本文将使用 DCT、NB、SVM 3 种方法构造结节病和结核病的分类识别器。

3.1 朴素贝叶斯方法

朴素贝叶斯方法由于简单而易于实现，常常是首先考虑使用的一种方法。在第 2 节中得到的数据，有些是连续型数据(例如属性“$T_{\max}$”)，因此，在使用 NB 方法之前，我们首先需要对第 2 节中得到的连续型数据进行离散化。表 1 给出了这些连续型数据的离散化方案。根据表 1，如果某连续型特征取值落到某个区间，则该特征的取值就被设定为该区间所对应的类别。例如，如果某一病人，其 $T_{\max}$ 取值为 39，则相应

的特征将被设为 2(39 落在 $T_{\max}$ 的第 2 个区间),见图 2。

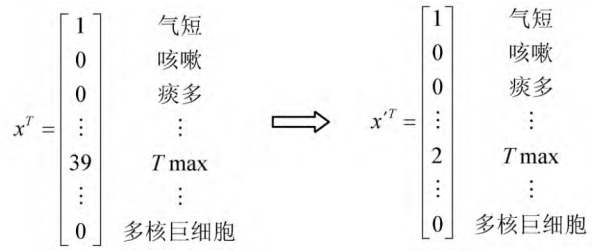


图 2　对 x 中的连续型分量离散化得到 x'

后文如无特别说明,T 表示带连续变量的训练数据,而将连续变量按表 1 的方案离散化后得到的训练数据我们用 T' 表示。NB 方法是基于 T' 作为训练数据的。

使用 NB 方法进行预测首先需要对 $p(x|y)$ 进行建模。对 $p(x|y)$ 的建模建立在一个叫条件独立的假定上:即 x 中的某个特征 x_i 出现与否是关于 y 条件独立的(这一假定被称为朴素贝叶斯假定)。例如,假如 $y=1$ 表明该病是结节病,"fever"是第 8 个特征(x_8),"BALF"是第 84 个特征(x_{84}),那么如果已知 $y=1$(结节病),则 x_8 是否出现(病人"发烧"还是不发烧)对 x_{84} 的出现与否没有影响。即 $p(x_{84}|y)=p(x_{84}|y,x_8)$。

显然根据条件独立性,我们有:

$$\begin{aligned}P(x_1,\cdots,x_{125}|y)&=p(x_1|y)p(x_2|y,x_1),\cdots,p(x_{125}|y,\\&\quad x_1,\cdots,x_{124})\\&=p(x_1|y)p(x_2|y),\cdots,p(x_{125}|y)\\&=\sum_{i=1}^{n}p(x_i|y)\end{aligned}$$

表 1　连续属性离散化方案

class	1	2	3	4	5	6	7
T_{max}	<5	35～45					
PPD test 阳性程度	0	1	2	3	4		
CRP(mg/L)	−1	0～20	20～60	60～100	>100		
SACE	−1	0	0～50	50～100	>100		
ESR(mm/1h)	−1	0	0～20	20～40	40～60	60～80	>80
SAT-6(SFCs/10∧6PBMC)	0	0～200	200～600	>600			
FP-10(SFCs/10∧6PBMC)	0	0～500	500～1500	>1500			
FEV1%	0	0～60	60～80	>80			
FEV1/FVC%	0	0～100	>100				
弥散量%	0	0～50	50～80	>80			
灌洗量 ml	0	0～60	60～100	>100			
回收量 ml	0	0～20	20～40	40～60	>60		
细胞分类-总数	0	0～20000000	20000000～40000000	>40000000			
细胞分类-吞噬%	0	0～20	20～40	40～60	>60		
细胞分类-淋巴%	0	0～20	20～40	40～60	>60		
细胞分类-中性%	0	0～5	5～10	>10			
细胞分类-嗜酸%	0	0～2	2～4	4～8	>8		
T 亚群-CD3%	0	0～80	>80				
T 亚群-CD4%	0	0～40	40～70	>70			
T 亚群-CD8%	0	0～10	10～30	>30			
T 亚群-CD4/CD8	0	0～3	3～10	>10			

NB 模型中的参数主要有:$\phi_{i|y=1}=p(x_i=1|y=1)$,$\phi_{i|y=0}=p(x_i=1|y=0)$,以及 $\phi_y=p(y=1)$。给定训练集 $\{(x^{(i)},y^{(i)});i=1,\cdots,m\}$,我们可以得到联合似然函数:$L(\phi_y,\phi_{j,k|y=0},\phi_{j,k|y=1})=\prod_{i=1}^{m}p(x^{(i)},y^{(i)})$

对上式求极大值,我们可以得到 $\phi_{i|y=1}=p(x_i=1|y=1)$,$\phi_{i|y=0}=p(x_i=1|y=0)$,以及 $\phi_y=p(y=1)$ 极大似然估计:

$$\phi_{j,k|y=1}=\frac{\sum_{i=1}^{m}1\{x_j^{(i)}=k\wedge y^{(i)}=1\}}{\sum_{i=1}^{m}1\{y^{(i)}=1\}},k=1,\cdots,k_j$$

$$\phi_{j,k|y=0}=\frac{\sum_{i=1}^{m}1\{x_j^{(i)}=k\wedge y^{(i)}=0\}}{\sum_{i=1}^{m}1\{y^{(i)}=0\}},k=1,\cdots,k_j$$

$$\phi_y=\frac{\sum_{i=1}^{m}1\{y^{(i)}=1\}}{m}$$

式中,符号"∧"表示"且"的意思。符号 $1\{.\}$ 表示当花括号中的表达式成立时为 1,否则取 0。上述参数的含义是很直观的,例如 $\phi_{j,k|y=1}$ 表示结节病患出现特征 j 且特征 j 值为 k 所占比例。

得到上述极大似然估计后,如果有新的病患资料 x,我们可以根据下式计算:

$$p(y=1|x)=\frac{p(x|y=1)p(y=1)}{p(x)}=\frac{\prod_{i=1}^{n}\prod_{k=1}^{k_i}p(x_i=k|y=1)p(y=1)}{\prod_{i=1}^{n}\prod_{k=1}^{k_i}p(x_i=k|y=1)p(y=1)+\prod_{i=1}^{n}\prod_{k=1}^{k_i}p(x_i=k|y=0)p(y=0)}$$

并选择后验概率最大的作为诊断结果。

3.2　决策分类树

我们使用的第 2 种方法是决策分类树 DCT。决策分类树是一种直观且有力的工具。图 3 是一个决策分类树在临床医学领域中应用的例子[22],根据图 3,我们很容易得到以下两条规则:

(1)如果病人有心脏收缩杂音,MCI 且有心脏畸形,则提示有可能脱垂;

(2)如果病人有心脏收缩杂音,MCI 但无心脏畸形,则排除脱垂。

quinlan[23] 给出了如下构造分类树的方法:

给定训练集 T、k 个分类类别 $C=\{c_1,c_2,\cdots,c_k\}$,则:

·如果 T 中有 1 个或多个实例被划分到同一类别 c_j 中,则得到决策树的一个标识为 c_j 的叶子节点。

·如果 T 中不再包含任何实例，则同样得到一个由 T 之外的其他实例提供的信息确定的叶子节点。

·如果 T 中包含的实例属于混合类别（存在某一实例 a，根据某一性质 i，a 属于类别 c_i，而根据性质 j，a 又属于 c_j），则递归地选择某一性质进行测试，根据测试结果 $\{o_1,o_2,\cdots,o_n\}$，将 T 划分成 $T_1,T_2,\cdots,T_n$ 个子集，其中 T_i 包含 T 中所有测试结果为 o_i 的实例。

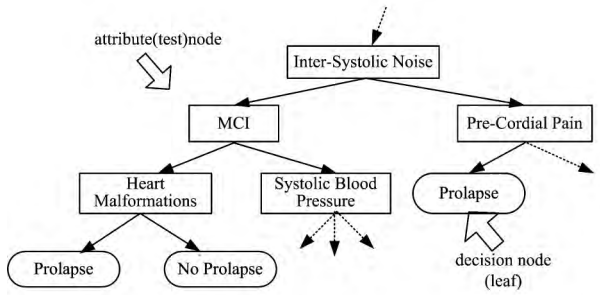


图 3　一个用于诊断心脏疾病的决策树[22]

一般地，优先选择哪一个属性进行测试对构造决策树非常重要，不同的测试顺序会导致构造出来的决策树的深度不同。我们总是希望构造出来的决策树的平均深度最小，平均深度最小的决策树意味着快速决策过程以及好的泛化能力。然而，从计算角度来看，寻找最小平均深度的决策树是困难的。为此，quinlan[23]给出了一个宽度优先的启发式搜索技术 ID3 算法，它可以较快速度返回“不算太深”的决策树。他的启发式技术基于信息论中熵的概念。

所谓熵是信息论中用来刻画不确定程度的一个概念，其形式定义为：

$$\varepsilon(P)=-\sum_{i=1}^{n}p_i\log_2 p_i$$

式中，$P=\{p_1,\cdots,p_n\}$ 为概率分布。

ID3 算法每次总是选择使分类具有最小不确定度的属性进行测试。为说明这一过程，我们需要引入以下符号。

$T_{|c}=\{t\in T|$ 实例 t 被划分到类别 $c\}$

对于类别 $C=\{c_1,c_2,\cdots,c_k\}$，$T_{|C}=\{T_{|c1},\cdots,T_{|ck}\}$ 构成对 T 的一个划分。对每一 $c_i(C,i=1,\cdots,k$，定义 c_i 关于 T 的概率为：$P(c_i;T)=\dfrac{|(T|c_i)|}{|T|}$，从而 T 的熵为：$\varepsilon(T)=-\sum_{i=1}^{n}p(c_i;T)\log_2 p(c_i;T)$。

所有决策都是等可能的，则 $\varepsilon(T)=\log_2 n$ 为 n 个决策的最大不确定度。如果所有决策中有一个概率为 0，则 $\varepsilon(T)=0$，具有最小的不确定度。

假定某一属性 o 具有 $o_1,\cdots,o_k$ 共 k 个可能的取值，这 k 个可能取值将 T 划分成 k 个子集 $T_{|o}=\{T_{|o\to o1},\cdots,T_{|o\to on}\}$，其中 $T_{|o\to oi}=\{t\in T|t$ 中属性 o 具有属性值 $o_i\}$。假定构造决策树时我们选择属性 o，并从子集 $T_{|o\to oj}$ 出发构造关于属性值 o_i 的子树。显然，越靠近叶子节点，$T_{|o\to oj}$ 的不确定度越低。因此，我们可定义 $T_{|o\to oj}$ 的不确定度为：

$$((T_{|o\to oj})=-\sum_{i=1}^{n}p(c_i;T_{|o\to oj})\log_2 p(c_i;T_{|o\to oj})$$

最后，有了 $T_{|o\to oj}$ 的不确定度后，我们可以得到 o 的平均不确定度：

$$\varepsilon(o;T)=\sum_{j=1}^{n}p(o\to o_j)\varepsilon(T_{|o\to oj})$$

式中，$p(o\to o_j)=\dfrac{|(T_{|o\to oj})|}{|T|}$

对于属性集 O，ID3 算法每次总是优先选择具有最小不确定度的属性进行决策树的构造。

决策树被构造出来后，若有新的病患资料 x，则只从决策树的根节点出发，沿着决策树的属性节点，逐一检查 x 的属性值，看它落在决策树的哪一分支，并一直到叶子节点。本文中决策树只有两个叶子节点，阳性代表结节病，阴性代表结核病。

3.3　支持向量机

支持向量机技术最初由 Vapnik 等[17,18]提出。由于 SVM 采用了极大化间隔原理、对偶理论和核函数技巧，并以统计学习理论和最优化方法为基础，使得 SVM 成为最有效的且有力的数据分析处理工具。SVM 自提出以来不断得到发展，目前已有多种不同的变种，C-SVM 是标准的 SVM，这里将对 C-SVM 进行简要介绍。

C-SVM 是一种标准的二分类判决器。所谓二分类问题可被表示成如下的形式：

$$T=\{(x_1,y_1),\cdots,(x_m,y_m)\}\in(R^n\times Y)^m \tag{1}$$

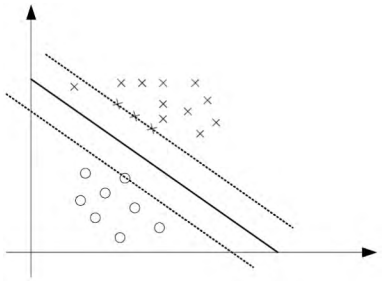
式中，$x_i\in R^n$，$y_i\in Y=\{1,-1\}$，$i=1,\cdots,m$，要求在 R^n 中寻找一实值函数 $g(x)$，使得给定 x，便能通过判决函数 $f(x)$ 得到 y 的输出值，其中判决函数 $f(x)=\mathrm{sign}(g(x))$。

C-SVM 将上述问题表示成如下凸二次规划问题：

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i\\&\text{s.t. } y_i((w\cdot x_i)+b)\geqslant 1-\xi_i,\ i=1,\cdots,m\\&\qquad\xi_i\geqslant 0,\ i=1,\cdots,m\end{aligned} \tag{2}$$

式中，$\zeta=(\zeta_1,\cdots,\zeta_m)^{\mathrm{T}}$，$C>0$ 为惩罚因子。

上述凸二次规划模型可通过图 4 加以解释。图 4 中每一个样本点到中间的实线都有一个几何距离，这些几何距离中的最小值构成所谓的几何间隔。将式(2)中的目标函数的第一项最小化，相当于寻找一超平面（例如图 4 中的实线），使得训练样本的几何间隔最大。而根据最优解的 KKT 互补松弛定理，可知落在图 4 中虚线上的点是离分类超平面最近的样本点。正是这些离分类超平面最近的样本点，决定了分类超平面的位置，称这些样本点为支持向量。一般而言，支持向量的个数远少于训练样本个数，这使得 SVM 具有很高的效率。



中间的实线代表分类超平面，落在虚线上的点就是支持向量

图 4　SVM 几何间隔示意图

为了避免训练样本中极少量样本点的出现导致分类超平面的剧烈变化，我们容许少量样本点落在支持向量与分类超平面之间（例如图 4 中实线和虚线之间存在一个样本点），此时，对应的样本点的 ξ_i 大于 0。因此，式(2)中第二项是一惩罚项，表示对于所有违规的样本点给予一定惩罚。

模型(2)代表的凸二次规划问题并非 C-SVM 直接求解的问题，C-SVM 求解的是该问题的拉格朗日对偶问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{j=1}^{m}\alpha_j$$

$$\text{s.t.}\sum_{i=1}^{m}y_i\alpha_i=0 \qquad (3)$$

$$0\leqslant\alpha_i\leqslant C,i=1,\cdots,m$$

式中，$K(x,x')$是核函数，$\alpha_i(i=1,\cdots,m)$为拉格朗日系数。

模型(3)构成了原问题的对偶问题。C-SVM 通过求解该对偶问题，得到该对偶问题最优解 $\alpha^*=(\alpha_1^*,\cdots,\alpha_m^*)$，即可通过下式确定分类超平面：

$$w^*=\sum_{i=1}^{m}\alpha_i^* y^{(i)}x^{(i)}$$

$$b^*=-\frac{\max_{i;y^{(i)}=-1}w^{*\mathrm{T}}x^{(i)}+\max_{i;y^{(i)}=1}w^{*\mathrm{T}}x^{(i)}}{2} \qquad (4)$$

有了式(4)所代表的分类超平面后，假设有新样本 x 需要进行预测，我们只需要将新样本代入如下函数进行计算即可：

$$w^{*\mathrm{T}}x+b^*=(\sum_{i=1}^{m}\alpha_i^* y^{(i)}x^{(i)})^{\mathrm{T}}x+b=\sum_{i=1}^{m}\alpha_i^* y^{(i)}\langle x^{(i)},x\rangle \qquad (5)$$

式(5)告诉我们，预测的结果只依赖于 x 与训练样本之间的内积。而 KKT 互补松弛定理告诉我们，除少量的支持向量对应的 α_i^* 非零外，其他大部分训练样本对应的 α_i^* 均等于0，这意味着计算式(5)只需要计算 x 与少数 α_i^* 非零的样本的内积。这使得 SVM 算法能以很高的效率工作。

4 实验结果

ADST 系统使用了 DCT、NB、SVM 3 种方法来检验系统的分类能力。DCT 和 NB 方法使用的数据集是经过离散化后训练数据 T'，而 SVM 方法则是训练数据 T，即含有连续型数据的训练数据。DCT 和 NB 方法直接采用我们得到的训练数据 T' 进行训练，而 SVM 方法则采用台湾大学林智仁(Lin Chih-Jen)教授等开发设计的 LIBSVM[19]，并按照标准 LIBSVM 的做法：1)将训练数据 T 进行简单的缩放操作，使之落在[0 1]范围内；2)考虑选用 RBF 核函数；3)采用交叉验证选择最佳参数 C 与 g(交叉验证选得的最佳参数 $C=2,g=0.0625$)；4)采用最佳参数 C 与 g 对整个训练集进行训练获取支持向量机模型；5)利用获取的模型进行测试与预测。

实验过程中，我们将前面获得的 106 例确诊的病患资料随机抽取 26 例，作为测试数据。再在余下的 80 例中随机抽取 2、4、6、8、10、20、40、60、80 例作为训练数据，分别对算法进行训练，然后用测试数据检验训练后得到的模型的预测效果。表 2 记录了实验过程中观察到的详细数据，包括训练误差、测试误差，以及 SVM 方法所需要的迭代次数、训练时间、优化目标函数值、支持向量数等。

表 2 ADST 系统实验数据

	m (train set size)	2	4	6	8	10	20	40	60	80	106
DCT	train error	*	*	*	*	0	0	0.025	0.0667	0.05	0.0472
	test error	*	*	*	*	0.4231	0.2692	0.3077	0.2692	0.1154	0.0385
NB	train error	0	0	0	0	0	0.1	0.125	0.1167	0.0875	0.0943
	test error	0.5385	0.1154	0.0769	0.0769	0.0769	0.0769	0.1154	0.0385	0.0769	0.0385
SVM	train error	0	0	0	0	0	0.05	0.075	0.05	0.0875	0.0762
	test error	0.2692	0.2692	0.1923	0.1538	0.1154	0.0769	0	0	0	0
	#iter	1	2	3	4	5	10	20	36	43	60
	obj	−1.89989	−3.67704	−5.46173	−6.98785	−8.56489	−15.8271	−29.0391	−34.7215	−45.8301	−52.5704
	rho	0	0.018558	0.00727	0.003543	−0.02291	0.011903	0.087256	0.20031	0.060512	0.083315
	nSV	2	4	6	8	10	20	40	53	66	78
	nBSV	2	4	6	8	10	20	38	44	59	66

由表 2 可知，3 种方法的训练误差和测试误差均比较小，这说明模型并没有出现过配(overfitting)或失配(underfitting)的情况，我们选择的这些特征用来识别是结节病还是结核病是合适的。DCT、NB 和 SVM 3 种方法的训练精度分别达到 95.28%，90.57%，92.38%，而测试精度分别为96.15%，96.15%，100%，均是令人满意的。

同样，由表 2 可以看出，当训练样本比较小时，NB 方法具有较快的学习和识别能力，但 NB 方法具有较高的渐近误差。相反，在训练样本比较小时，SVM 方法具有较高的误差，但 SVM 的渐近误差较小。DCT 方法则在训练样本不足时(样本数少于 10 时)，不具备分类能力。

图 5 给出了 DCT、NB、SVM 方法的 ROC 曲线，SVM、DCT、NB 方法的 ROC 曲线下的面积分别为 0.9706，0.955，0.6968，这表明 SVM 方法具有最好的分类能力，而 DCT 方法次之，最差的是 NB 方法。朴素贝叶斯方法非常接近人的经验判断，这一结果在一定程度上解释了为何根据临床医生的经验对结节病和结核病进行鉴别是困难的。

此外，我们还记录了 SVM 方法训练所需要的迭代次数和训练时间(见图 6)随训练样本的变化情况。SVM 方法的训练效率是很高的，80 个训练样本只需要不到 0.6 秒，经过 43 步迭代即可收敛到最优解，这还包含了交叉验证的时间。

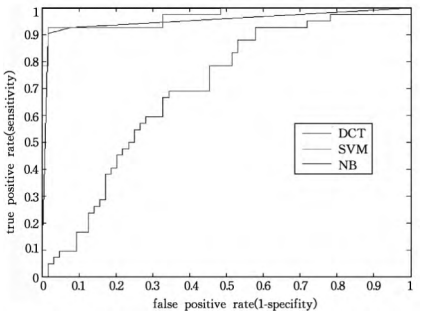


图 5 DCT，NB，SVM 关于结节病的 ROC 曲线

图 7 给出了支持向量数 nSV 和边界支持向量数 nBSV 随训练样本增加而变化的情况。当训练样本比较少时(样本数不超过 20 时)，所有支持向量都是边界支持向量。当训练样本进一步增大时，为了确保最佳分类质量，允许少量支持向量违反几何间隔要求，此时就出现了支持向量数大于边界支持向量个数的情况。

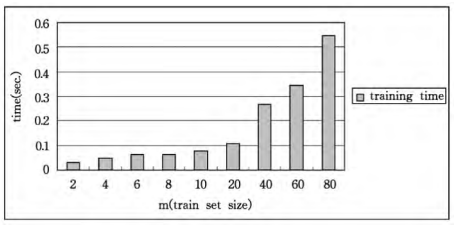


图 6　SVM 训练时间关于样本规模 m 的变化直方图

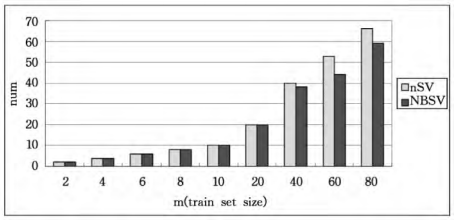


图 7　SVM 在不同训练样本时产生的支持向量数 nSV，以及边界支持向量数 nBSV

本文的训练数据 T 或 T' 中，一共使用了 125 个特征，其中有些特征对疾病的分类具有显著意义。本文尝试使用以下公式选择前 5 个最具有分类意义的特征：

$$\log\left(\frac{p(x_j=i\mid y=1)}{p(x_j=i\mid y=0)}\right)=\log\left(\frac{p(\text{特征}_i\mid\text{结节病})}{p(\text{特征}_i\mid\text{结核病})}\right)$$

根据 NB 方法的参数，我们得到前 5 个最具有分类意义的特征：PPD 测试阳性程度达到 4，盥洗回收量超过 60ml，淋巴结肿大，T 亚群-CD3%超过 80，盥洗量少于 60ml。

而根据 DCT 方法的参数，我们得到前 5 个最具有分类意义的特征：淋巴结肿大，PPD 测试阳性程度达 4，SACE 增高，抗酸染色阴性，T 亚群-CD3%超过 80。

遗憾的是，对于分类能力最好的 SVM 方法，由于 SVM 方法的黑箱性，我们没法得到最具有分类意义的指标。

为了进一步检验 ADST 系统的分类准确度，我们将所搜集的 19 例临床无法确诊的病例的资料输入到 ADST 系统中，让 ADST 分别用 DCT、NB 和 SVM 方法进行预测。结果如表 3 所列。根据表 3 的结果，DCT 方法和 SVM 方法的预测结果高度相似，只有第 27 号病患预测结果不一致。而 NB 方法和 SVM 方法的预测结果有 3 例不一致，分别是编号为 61、75、81 的病患。而 NB 方法和 DCT 方法的预测结果则有 4 例不一致，分别是编号为 27、61、75、81 的病患。根据 ADST 的诊断结果，我们邀请结节病和肺结核的一线临床医生针对 ADST 的诊断结果进行讨论，经过他们的评估后，结合 SVM 具有最强的分类能力这一实验结果，我们根据 SVM 方法得到的诊断结果对病患做结节病或肺结核的治疗。治疗的结果是令人满意的，19 个病患均得到了康复，治愈率达到 100%。临床结果与实验结果是吻合的。

表 3　DCT，NB，SVM 3 种方法对临床无法确诊的 19 病例的诊断结果对比

No. of patient		3	16	27	28	39	34	49	52	57	59	61	68	69	72	75	81	97	100	103
classifier	DCT	1	1	0	1	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0
	SVM	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0
	NB	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	0	0

注："1"代表被诊断为"结节病"，"0"代表被诊断为"结核病"。

结束语　结节病和肺结核的临床鉴别诊断目前仍然是困难的。本文的 ADST 系统在搜集这两类病患资料的基础上，通过特征筛选、原始资料量化得到训练数据，然后分别用 DCT、NB、SVM 算法进行训练得到模型。我们的研究结果可总结为以下几点：1）本文使用 125 个特征作为结节病和结核病的鉴别诊断是合适的，DCT、NB、SVM 3 种方法都达到了令人满意的精度；2）SVM 方法具有最好的分类能力，DCT 方法次之，最差的是 NB 方法；3）NB 方法在 3 种方法中的分类能力最差这一结果解释了凭临床医生的经验对结节病和结核病进行鉴别是困难的；4）根据 DCT 方法得到鉴别是结节病还是结核病的 5 个最有意义的指标分别是：淋巴结肿大，PPD 测试阳性程度达 4，SACE 增高，抗酸染色阴性，T 亚群-CD3%超过 80；5）根据 NB 方法得到鉴别是结节病还是结核病的 5 个最有意义的指标分别是：PPD 测试阳性程度达到 4，盥洗回收量超过 60ml，淋巴结肿大，T 亚群-CD3%超过 80，盥洗量少于 60ml；6）SVM 方法的分类结果经临床检验是正确的。

目前 ADST 系统只实现了对结节病和肺结核的二分类识别，对 ADST 能力进行扩展，使之能实现对肺炎、支气管炎、肺泡蛋白沉积症等其他呼吸道疾病的分类识别能力是本文下一步的工作。

此外，由于通过 SVM 学习得到的知识以分类判决函数的形式存在，导致 SVM 得到的模型难以被理解。如何提高 SVM 模型的"可解释性"是一个重要而富有挑战性的工作。Barakat N[20] 等人研究了如何从支持向量中提取规则。类似地，我们下一步尝试在 SVM 模型的理解方面开展一些工作，希望能从 ADST 产生的支持向量中寻找某些线索，尝试给出结节病与结核病的临床鉴别要点，这在临床上是非常有意义的。

参 考 文 献

[1] Iannuzzi MC, Fontana JR. Sarcoidosis: Clinical Presentation, Immunopathogenesis, and Therapeutics[J]. JAMA, 2011, 305(4): 391-399

[2] 中华医学会呼吸病学会结节病组. 结节病诊断及治疗方案(第三次修订稿案)[J]. 中华结核和呼吸杂志, 1994, 17(1): 9-10

[3] 中华医学会结核学分会. 肺结核诊断和治疗指南[J]. 中华结核和呼吸杂志, 2001, 24(2): 70-74

[4] 邹兰芳, 杨吉刚, 李春林, 等. 结节病 18F-FDG 符合线路显像胸部淋巴结的特征表现[J]. 临床和实验医学杂志, 2013, 12(3): 169-170

[5] 黄燕, 陆聪哲, 王彩彩, 等. 结节病患者外周血 Th7 细胞表达及临床意义[J]. 中国呼吸与危重监护杂志, 2013, 12(2): 173-176

[6] 刘长军, 李洪松. 64 排螺旋 CT 在肺结节病变经皮穿刺活检中的临床应用研究[J]. 实用医学影像杂志, 2012, 13(6): 367-370

[7] 叶秋月. 肺结节病与肺结核鉴别诊断的临床分析[D]. 北京: 北京协和医学院(中国医学科学院), 2011

[8] 李秋红. 结节病与不典型结核病鉴别诊断方法的研究[D]. 苏州: 苏州大学, 2007

(下转第 138 页)

图 5 为本文特征量加上文献[9]特征量，一并送入支持矢量机中训练后的鉴别结果。从图 3－图 5 中可以看出，本文提取的特征对合成图像能有效鉴别。

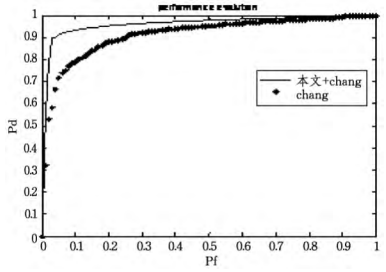


图 5　本文特征加上文献[9]后性能比较

结束语　本文对图像进行树图分割，提取物体轮廓线，然后采用最小二乘法线性预测图像像素值，得到预测误差图，利用分割结果对预测误差图内部、边界区域分别计算误差均值、方差等特征量，结合 4 个方向上共生矩阵的特征统计特性，送入支持矢量机获得取证结果，鉴定图像是否经过拼接处理。本文特征量可以单独使用，也可以和其它方法中的特征量一起使用。

参 考 文 献

[1] Nguyen H C. Detection of copy-move forgery in digital images using radon transformation and phase correlation[C]// Proceedings of Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. 2012:134-137

[2] Popescu A C, Farid H. Exposing digital forgeries by detecting duplicated image regions[OL]. http://www.cs.dartmouth.edu/farid/downloads/publications/tr04.pdf, department of computer. science, dartmouth college, 2012

[3] Kee E, Farid H. A perceptual metric for photo retouching [J]. Proceedings of the national academy of sciences, 2011, 108(50): 19907-19912

[4] O'Brien J, Farid H. Exposing photo manipulation with inconsistent reflections[J]. ACM transactions on graphics, 2012, 31(1): 1-11

[5] Kee E, Johnson M K, Farid H. Digital image authentication from jpeg heads[J]. IEEE transactions on information forensics and security, 2011, 6(3): 1066-1075

[6] Zheng Qian-ru, Sun Wei, Lu Wei. Digital spliced image forensics based on edge blur measurement[C]// Proceedings of IEEE International Conference on Information Theory and Information Security. Dec. 2010: 399-402

[7] Gou Hong-mei, Swaminathan A, Wu Min. Noise features for image tampering detection and steganalysis[C]// Proceeding of IEEE Conference on Image Processing. 2007, 6: 97-100

[8] Shi Jian-bo, Malik J. Normalized cuts and image segmentation [J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(8): 888-905

[9] Hsu Yu-feng, Chang Shih-fu. Camera response functions for image forensics: an automatic algorithm for splicing detection[J]. IEEE transactions on information forensics and security, 2010, 5(4): 816-825

[10] Chen Ying, Wang Yu-ping. Exposing digital forgeries by detecting traces of smoothing[C]// Proceeding of 9th International Conference for Young Computer Scientists. 2008: 1440-1445

[11] Libsvm[OL]. http://www.csie.ntu.edu.tw/～cjlin/

(上接第 109 页)

[9] 李秋红，赵兰，李惠萍，等. 实时定量聚合酶链反应技术在鉴别结节病与增殖性结核病中的应用[J]. 中华结核和呼吸杂志，2007，30(9)：686-690

[10] 沈璨，周瑛，李秋红，等. 实时荧光定量 PCR 在结节病与不典型结核鉴别诊断中的临床应用[J]. 同济大学学报：医学版，2010，31(6)：46-50

[11] 周瑛. 结节病与不典型结核病鉴别诊断方法的研究[D]. 上海：同济大学医学院，2009

[12] ACCESS Research Group. Design of A Case Control Etiologic Study of Sarcoidosis(ACCESS)[J]. J Clin Epidemiol, 1999, 52(12): 1173-1186

[13] Wirnsberger RM, Vries J de, Wouters EFM, et al. Clinical presentation of sarcoidosis in the Netherlands An epidemiological study[J]. Netherlands Journal of Medicine, 1998, 53: 53-56

[14] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)[M]. Springer, February 2009

[15] 曹蕾，黎维娟，冯前进，等. 基于 LDA 和 SVM 的肺结节 CT 图像自动检测与诊断[J]. 南方医科大学学报，2011，31(2)：324-328

[16] 张婧，李彬，田联房，等. 结合规则和 SVM 方法的肺结节识别[J]. 华南理工大学学报：自然科学版，2011，39(2)：125-129

[17] Cortes C, Vladimir V. Support-Vector Networks [J]. Machine Learning, 1995, 20(3): 273-297

[18] Vapnik V, Golowich S E, Smola A J. Support Vector Method for Function Approximation, Regression Estimation and Signal Processing[C]// NIPS. 1996: 281-287

[19] Chang Chih-chung, Lin Chih-jen. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 27(2): 1-27

[20] Barakat N, Bradley A P. Rule extraction from support vector machines: A review[J]. Neurocomput, 2010, 74(2), 1-3

[21] 刘伟，谢红卫. 整合网络属性、序列特征和功能注释预测潜在的癌基因[J]. 中国科学生命科学，2013，43(7)：589-595

[22] Podgorelec V, Kokol P, Stiglic B, et al. Decision Trees: An Overview and Their Use in Medicine[J]. J. Med. Syst., 2002, 26(5): 445-463

[23] Quinlan J R. Induction of Decision Trees[J]. Mach. Learn., 1986, 1(1): 81-106