基于主成分机器学习算法的慢性肝病的智能预测新方法

常炳国[1] 李玉琴[1] 冯智超[2] 姚山虎[2]
(湖南大学信息科学与工程学院 长沙 410082)[1] (中南大学湘雅三医院 长沙 410082)[2]

摘 要 运用新一代信息技术快速预测慢性肝病的机理和特征,是提高慢性肝病诊断率的有效途径。运用主成分分析机器学习算法,对描述慢性肝病的多项指标属性项进行降维处理,结合神经网络学习,构建了慢性肝病预测模型。实验分析了125组20维慢性肝病患者的医学检验指标数据项,利用ROC(Receiver Operating Characteristic)曲线优选出13维指标项作为慢性肝病敏感度高的检验指标属性项。通过主成分分析将13维指标项降至5维综合数据项。神经网络训练115组检验指标样本集,剩余10组样本集作为测试样本。与原始20维数据作为神经网络输入相比,所提模型不仅降低了复杂度,且预测精度提高了15.07%。

关键词 慢性肝病,主成分分析,神经网络,智能预测
中图法分类号 TP311.5 **文献标识码** A

New Intelligent Prediction of Chronic Liver Disease Based on Principal Component Machine Learning Algorithm

CHANG Bing-guo[1] LI Yu-qin[1] FENG Zhi-chao[2] YAO Shan-hu[2]
(School of Information Science and Engineering, Hunan University, Changsha 410082, China)[1]
(Third Xiangya Hospital of Central South University, Changsha 410082, China)[2]

Abstract Using new information technology to predict the mechanism and characteristics of chronic liver disease is an effective way to improve its diagnosis. In this paper, we used the principal component analysis (PCA) of the machine learning algorithm to reduce the dimensional indicators of chronic liver disease, combined with neural network learning to build a new intelligent prediction of chronic liver disease (IPCLD). The experiment studied 125 data sets of 20-dimensional indicators of chronic liver disease, used receiver operating characteristic (ROC) curve to select 13-dimensional more sensitive indicators, further reduces the dimension down to 5 by PCA. The neural network is trained with 115 data sets, and the remaining 10 data sets are used as test data sets. Compared with being trained by original data, the IPCLD improves 15.07% prediction accuracy and reduces the complexity.

Keywords Chronic liver disease, Principal component analysis, Neural network, Intelligent prediction

1 引言

全球有大约4亿人遭受着慢性肝病的困扰[1]。慢性肝病的类型有很多,如慢性甲型肝炎、慢性乙型肝炎、慢性丙型肝炎、慢性酒精性肝病等[2-3]。不同类型的慢性肝病的治疗和流行的广泛程度不同。例如,世界卫生组织的报告表明,90%以上的丙型肝炎在3~6个月内可得到完全治愈[1]。此外,慢性肝病的严重程度也有多种,如慢性迁移性肝炎、慢性活动肝炎、肝硬化、肝癌等。针对每种慢性肝病的类型和严重程度,其临床结果、治疗药物以及疗效等医学行为也不尽相同。寻找形成慢性肝病的病因对于医学治疗同样有着非常重要的作用[4]。慢性肝病产生的机理存在不确定性和复杂性,并且患者个体之间的差异较大。针对这种情况,需要针对不同的患者进行个性化治疗,以提高其治疗效果[5]。由此可见,快速预测和准确诊断慢性肝病,对患者的治疗是至关重要的。

在慢性肝病的个性化治疗过程中,患者往往需要检查与之相关的多达80多项的检验指标信息,如血小板、总胆红素、丙氨酸氨基转移酶、肌酐等。在众多的信息中,如何能快速地根据患者的个体特点来精确定位哪些指标对患者的病因有着至关重要的作用,是医院对症下药的关键所在。原始指标间往往存在信息冗余的情况,此外,过多的维度会增加问题分析的复杂性。通过ROC特征曲线[6],能够初步优选出对慢性肝病敏感度较高的指标,以期降低系统的复杂性。运用主成分分析法来挖掘患者患病的严重程度对各个检验指标的影响,在不丢失绝大多数信息的情况下,能够极大地降低维度。将降维后的数据输入到人工神经网络中进行训练,并以此来诊断和智能预测慢性肝病[7-8],从而构建出慢性肝病的智能预测模型,为医院对慢性肝病的治疗提供了非常重要的技术支持。

2 数据集

2.1 数据来源

本文选择2008年3月至2016年5月入住湘雅三医院的31名患者,入选条件:1)患者均在后期确诊为不同严重程度的慢性肝病;2)所有患者均接受血常规、肝功能、肾功能、凝血功能等检查。本文将不同病情、不同病程患者的检验数据项视为独立样本。

本文受湖南省重点研发计划(2016GK2050)资助。
常炳国(1966-),男,博士,主要研究方向为机器学习、数据挖掘、粗糙集,E-mail:changbingguo@126.com(通信作者)。

2.2 数据集的建立

根据湘雅三医院提供的患者数据，将所需的检验数据和诊断结果以及患者的基本信息导入数据库中，初步选出与慢性肝病相关的 20 维检验指标共计 125 组作为本文研究样本。其中，20 维检验指标如下。

(1)血常规：白细胞(WBC)、血红蛋白(Hb)、血小板(PLT)、红细胞(RBC)；

(2)肝功能：总胆红素(TBIL)、直接胆红素(DBIL)、丙氨酸氨基转移酶(ALT)、天门冬氨酸氨基转移酶(AST)、总蛋白(TP)、白蛋白(ALB)、球蛋白(GLO)、白球蛋白比值(A/G)、总胆汁酸(TBA)；

(3)肾功能：尿素(UREA)、肌酐(Cr)、尿酸(UA)；

(4)凝血功能：凝血酶时间(TT)、凝血酶原时间(PT)、凝血酶原国际标准化比值(或国际标准化比值)(INR)；

(5)基本信息：年龄(AGE)、诊断结果(Diagnostic Result)。

2.3 数据预处理

2.3.1 数据整理

本文统一将各检验指标的结果保留两位小数，删除检验数据中的描述文字，如“复查”“已核”等。医学上“凝血酶原国际标准化比值(INR)”和“国际标准化比值”是同一指标，本文中将其统一命名为“凝血酶原国际标准化比值(INR)”。

2.3.2 诊断结果量化

根据慢性肝病从肝炎、肝硬化到肝癌的三部曲的恶化过程，将诊断结果量化为 4 种。为了预测组和期望组之间能更好地匹配和区分，使用向量来表示量化结果。“000”表示非慢性肝病；“001”表示肝炎；“010”表示肝硬化；“100”表示恶性肿瘤和肝癌。因此，“0”代表非慢性肝病，“1”代表慢性肝病。具体的量化结果如表 1 所列。

表 1 量化结果

诊断结果	量化值	向量	说明
帕森病(PD)，胆结石	0	000	非慢性肝病
肝炎		001	肝炎
病毒性肝炎肝硬化 酒精性肝硬化		010	肝硬化
肝硬化失代偿期	1		
肝脏肿块，肝癌， PLC(原发性肝癌)		100	肝癌

2.3.3 优选指标

ROC 曲线是根据一系列不同的两个子类别(或阈值)的以真阳性率(灵敏度)作为纵坐标、以假阳性率(1-特异性)作为横坐标的曲线图。通过 ROC 曲线计算其曲线下面积(AUC)，并且在 $AUC>0.5$ 的情况下，AUC 越接近 1，表明诊断效果越好。本实验中的 125 例研究样本中诊断出非慢性肝病 8 例，慢性肝病 117 例。

根据金标准，“0”代表非慢性肝病，“1”代表慢性肝病，曲线下面积(AUC)的值通过使用 SPSS19 的 ROC 曲线获得。它们分别是：AGE(0.357)，WBC(0.536)，Hb(0.542)，PLT(0.602)，RBC(0.542)，TBIL(0.587)，ALT(0.631)，AST(0.689)，DBIL(0.582)，TP(0.458)，ALB(0.421)，GLO(0.529)，A/G(0.429)，TBA(0.537)，UREA(0.347)，Cr(0.388)，UA(0.365)，PT(0.523)，INR(0.510)，TT(0.548)，括号内的数值为诊断结果的 ROC 曲线下面积(AUC)。本文选取 $AUC>0.5$ 的 13 维指标，分别为：WBC，Hb，PLT，RBC，TBIL，ALT，AST，DBIL，GLO，TBA，PT，INR，TT。

3 主成分分析

假设将一个具有一定相关性的高维向量 X，通过一个特殊的特征向量矩阵 A，映射到一个无相关性的低维向量 PC。设 $X=(X_1,X_2,\cdots,X_p)'$ 是 p 维变量，A 为特征向量矩阵，则其线性变化为：

$$\begin{aligned} PC_1&=a_1'X=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p\\ PC_2&=a_2'X=a_{12}X_1+a_{22}X_2+\cdots+a_{p2}X_p\\ &\cdots\\ PC_p&=a_p'X=a_{1p}X_1+a_{2p}X_2+\cdots+a_{pp}X_p \end{aligned} \tag{1}$$

每一个主成分都代表了原始数据的一部分信息。假设 λ_i 为 PC_i(PC_i 为第 i 个主成分)的特征值，且 $\lambda_i>\lambda_j$ $(0<i<j\leqslant p)$。通常情况下，会选取 $m(m<p)$ 个主成分来代表原始数据。主成分的数目 m 可以通过各个主成分的累计方差贡献率(Cumulative Variance Contribution Rate，CVCR)以及特征值 λ 来确定。其中，$CVCR$ 由式(2)计算得出：

$$CVCR=\sum_{k=1}^{m}\lambda_k/\sum_{i=1}^{p}\lambda_i \tag{2}$$

其中，λ 为各个主成分对应的特征值，m 为选定的主成分数目，p 为主成分总数目。

主成分数目的确定，需要考虑将原始数据降维后是否能保留绝大部分信息，因此需要保证 $CVCR$ 大于或者等于某一阈值 θ，本文选定 $\theta\geqslant 80\%$。对 ROC 曲线优选出的 13 维检验指标进行主成分解释总变量计算，结果如表 2 所列。

表 2 主成分分析解释总变量

主成分	特征值	方差比重%	CVCR%
PC1	5.024	38.649	38.649
PC2	1.850	14.232	52.881
PC3	1.416	10.892	63.773
PC4	1.242	9.558	73.331
PC5	1.016	7.814	81.145
PC6	0.851	6.544	87.690
PC7	0.645	4.958	92.648
PC8	0.351	2.699	95.347
PC9	0.258	1.988	97.335
PC10	0.213	1.639	98.975
PC11	0.112	0.862	99.837
PC12	0.018	0.136	99.973
PC13	0.004	0.027	100.000

依据表 2，第一主成分的 $CVCR$ 为 38.649%，比选择的阈值 θ(80%)小得多。为了能够进一步降低维度，需要参考主成分对应的特征值。以主成分为横坐标、相对应的特征值为纵坐标的二维图被称为碎石图(见图 1)。由图 1 可知，前 5 个主成分的特征值较大，连线较陡峭，主成分对应的特征值均大于 1；且由表 2 可得，第五个主成分的 $CVCR$ 为 81.145%。因此，提取 5 个主成分最佳，其综合了 13 维检验指标的大部分信息。

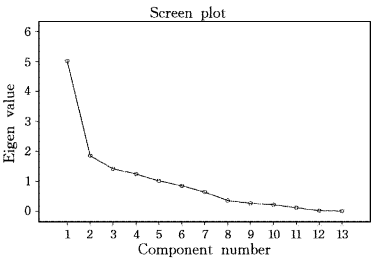

图 1 碎石图

主成分分析旋转后的成分载荷矩阵的载荷系数的绝对值与主成分的相关性成正比关系，绝对值越大，则主成分的相关性越大。载荷系数越接近 1 或越接近 0，则越能解释和命名变量。

表 3　主成分分析旋转后的成分载荷矩阵

	PC1	PC2	PC3	PC4	PC5
WBC	0.238	0.245	−0.119	0.774	−0.078
Hb	−0.278	−0.094	0.897	−0.135	−0.109
PLT	−0.208	−0.041	−0.017	0.891	0.086
RBC	−0.086	−0.221	0.897	−0.012	−0.178
TBIL	0.529	0.495	−0.011	0.278	0.546
ALT	0.067	0.889	−0.132	0.118	−0.145
AST	0.031	0.882	−0.170	0.017	0.134
DBIL	0.476	0.511	−0.040	0.297	0.561
GLO	0.108	−0.093	−0.303	−0.065	0.771
TBA	0.706	0.314	0.014	−0.035	0.422
PT	0.931	0.097	−0.281	0.009	0.003
INR	0.924	0.091	−0.246	0.049	−0.024
TT	0.493	−0.097	−0.006	−0.042	0.252

假设 D_{n*m}是代表原始数据的矩阵，其中 $m=13$。C_{m*e}是成分载荷矩阵的转置矩阵，其中 $e=5$。令 $P=D_{n*m}*C_{m*e}$，其中，矩阵 P 是原始数据矩阵通过主成分分析降维后的数据矩阵。至此，基于主成分分析的数据模型（见式(3)）已建立完成。

$$
\begin{aligned}
P_1 &= 0.238x_1 - 0.278x_2 - 0.208x_3 - 0.086x_4 + \\
&\quad 0.529x_5 + 0.067x_6 + 0.031x_7 + 0.476x_8 + \\
&\quad 0.108x_9 + 0.706x_{10} + 0.931x_{11} + \\
&\quad 0.924x_{12} + 0.493x_{13} \\
P_2 &= 0.245x_1 - 0.094x_2 - 0.041x_3 - 0.221x_4 + \\
&\quad 0.495x_5 + 0.889x_6 + 0.882x_7 + 0.511x_8 - \\
&\quad 0.093x_9 + 0.314x_{10} + 0.097x_{11} + \\
&\quad 0.091x_{12} - 0.097x_{13} \\
P_3 &= -0.119x_1 + 0.897x_2 - 0.017x_3 + 0.897x_4 - \\
&\quad 0.011x_5 - 0.132x_6 - 0.170x_7 - 0.040x_8 - \\
&\quad 0.303x_9 + 0.014x_{10} - 0.281x_{11} - \\
&\quad 0.246x_{12} - 0.006x_{13} \\
P_4 &= 0.774x_1 - 0.135x_2 + 0.891x_3 - 0.012x_4 + \\
&\quad 0.278x_5 + 0.118x_6 + 0.017x_7 + 0.297x_8 - \\
&\quad 0.065x_9 - 0.035x_{10} + 0.009x_{11} + \\
&\quad 0.049x_{12} - 0.042x_{13} \\
P_5 &= -0.078x_1 - 0.109x_2 + 0.086x_3 - 0.178x_4 + \\
&\quad 0.546x_5 - 0.145x_6 + 0.134x_7 + 0.561x_8 + \\
&\quad 0.771x_9 + 0.422x_{10} + 0.003x_{11} - \\
&\quad 0.024x_{12} + 0.252x_{13} \qquad (3)
\end{aligned}
$$

4　智能预测模型

人工神经网络能够在预先给定的训练数据集中进行自我训练，构建出用于预测测试数据集的结果的智能预测模型[7]。本文将利用 BP(Back Propagation)神经网络[8]对原始数据集和降维后的精简数据集进行训练以及预测，并将结果进行对比。

4.1　BP 神经网络

BP 神经网络是采用误差的反向传播算法（Error Back-propagation Algorithm）的多层前馈人工神经网络，其对网络拓扑结构非常敏感，不同的网络拓扑结构处理问题的能力有所不同。通过输入层到输出层的计算来完成 BP 网络，多一层隐含层虽然可以提高网络训练的速度，但是在实际中会花费更多的训练时间，因此本文采用增加隐含层节点数的方法来提高训练速度，选择一层隐含层的三层 BP 神经网络。

本文从 125 组数据中随机选取 115 组作为训练样本集，剩余 10 组作为测试样本集。设计 BP 神经网络的重要参数如下：1）网络层数为 3；2）训练次数为 500；3）期望误差为 0.0；4）训练函数为 trainlm 函数；5）链路权重调整函数为 learngdm 函数；6）性能函数为 mse 函数。

4.2　预测效果对比分析

在保持上述 BP 神经网络拓扑结构和参数不变的情况下，本文分别将样本 A（经过 ROC 优选与主成分分析降维后的精简样本集）和 B（原始 20 维样本集）作为输入，以构建出智能预测模型，对测试样本进行验证。

为了更好、更客观地评估 A 和 B 两者的预测精度，实验将会重复 1000 次，并且每一次实验时随机抽选训练样本，最后取所有结果的平均值作为最终的预测精度。图 2 是隐藏层节点分别为 8，11，14 和 17 时使用降维前和降维后的数据作为 BP 神经网络的输入所得到的准确率的对比图。

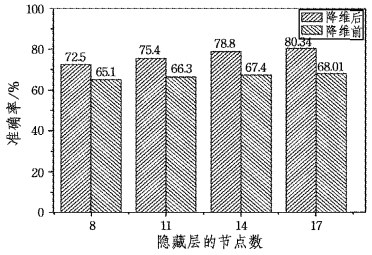

图 2　降维前和降维后的准确度对比

图 3 是采用 ROC 优选和主成分分析进行降维处理后 A 样本集的 BP 神经网络收敛图，横坐标为迭代次数，纵坐标为均方误差，虚线为最优均方误差参考线。图 4 则是同等条件下以 20 维原始数据 B 样本集为输入的 BP 神经网络收敛图。对比图 3、图 4 很容易得出：A 比 B 的收敛速度更快。

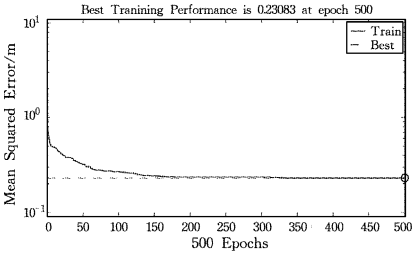

图 3　隐藏层的节点数为 17 时 A 样本集的收敛速度

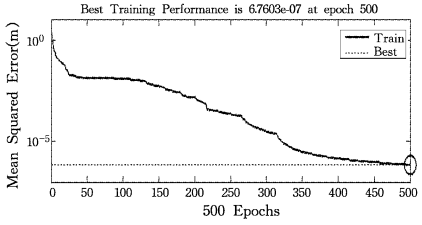

图 4　隐藏层的节点数为 17 时 B 样本集的收敛速度

（下转第 91 页）

了总体增量学习的训练速度。

结束语　本文基于局部敏感哈希在大规模高维数据中快速搜索的特性，提出了一种基于主成分分析的局部敏感哈希的 SVM 快速增量学习算法 PCA-LSH-ISVM。该算法将主成分分析应用于局部敏感哈希得到特定数量的哈希函数和上次训练得到的 SV 集样本生成哈希表，然后对新增样本进行筛选，再将筛选得到的新增样本和上次训练得到的 SV 集样本作为下次的训练样本进行训练学习。实验结果表明，在大规模高维增量学习样本中，该算法不仅能保证分类精度和良好的推广能力，而且学习训练速度也比经典的 SVM 增量学习算法快，可以快速进行增量学习。

参 考 文 献

[1] ATTAR V, SINHA P, WANKHADE K. A fast and light classifer for data streams[J]. Evol Syst, 2010, 1(3): 199-207.

[2] VAPNIK V. The Nature of Statistical Learning Theory[M]. NewYork: Springer Verlag, 1995.

[3] SYED N A, SUNG K. Handling Concept Drifts in Incremental Learning with Support Vector Machines[C]// Proc. of the 5th ACM SIGKDD International Confe-rence. 1999: 316-321.

[4] DIEHL C P, CAUWENBERGHS G. SVM Incremental Learning Adaptation and Optimization [C] // International Joint Conference on Neural Networks. IEEE, 2003: 2685-2691.

[5] 茅嫣蕾，魏赟. 一种基于 KKT 条件和壳向量的 SVM 增量学习算法[J]. 电子科技，2016，29(2)：38-40.

[6] 李妍坊，苏波，刘功申. 一种基于组合保留集的 SVM 增量学习算法[J]. 上海交通大学学报，2016，50(7)：1054-1059.

[7] 曹健，孙世宇，段修生，等. 基于 KKT 条件的 SVM 增量学习算法[J]. 火力与指挥控制，2014(7)：139-143.

[8] LUO J, PRONOBIS A, CAPUTO B. Incremental Learning for Place Recognition in Dynamic Environments[C]// IEEE International Conference on Intelligent Robots and Systems. 2007: 721-729.

[9] 张灿淋，姚明海，童小龙，等. 一种新的基于 KKT 条件的错误驱动 SVM 增量学习算法[J]. 计算机系统与应用，2014，23(1)：144-148.

[10] JHALA I S, DALAL P. Optimized Incremental SVM based Classifier for Spam Filtering using Internet Acronyms[J]. International Journal for Innovative Research in Science & Technology, 2015, 2(1): 2349-6010.

[11] CHAKROUN M, WALI A, ARIBI Y, et al. Video event detection using auto-associative neural network and incremental SVM models[C]// International Conference on Intelligent System Design & Application. 2015: 563-568.

[12] JAGTAP R V, POTEY M A. Recognition of Human Activity using Incremental SVM[J]. Imperial Journal of Interdisciplinary Research, 2016, 7(2): 2454-2462.

[13] ANDONI A, INDYK P, NGUYEN H L. Beyond locality-sensitive hash [C]// Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. 2014: 1018-1028.

[14] GIONIS A, INDYK P, MOTWANI R. Similarity search in high dimensions via hashing[C]// Proceedings of the International Conference on Very Large Databases. 1999: 25-38.

[15] DATAR M, IMMORLICA N, INDYK P, et al. Locality sensitive hashing scheme based on p-stable distributions[C]// Proceedings of the ACM Symposium on Computational Geometry. 2004: 23-36.

(上接第 67 页)

对比图 2—图 4 可以得到以下结论：1)降维后的预测精度明显高于降维前，计算图 2 所示精度平均值，结果显示平均提高了 15.07%；2)随着隐藏节点数的增加，降维前和降维后的预测精度都有所提高，但降维后的提高速度要优于前者；3)BP 神经网络在降维后收敛的速度快于降维前。

结束语　慢性肝病的成因复杂，医学检查指标多，指标之间存在信息冗余和干扰，导致准确地智能预测慢性肝病非常困难。为了提高智能预测慢性肝病的准确度并降低系统的复杂性，本文利用 ROC 特征曲线优选出 13 维敏感度较高的指标，并利用主成分分析方法基于优选指标将维度进一步降至 5 维，最后利用 BP 神经网络构建出智能预测慢性肝病模型。从 125 组数据中随机选出 115 组作为训练样本集合进行 BP 神经网络的训练，并将剩余的 10 组数据作为测试样本，用于验证智能预测模型的准确度。本文大量地重复上述训练和预测过程，以平均值作为最终的准确度，以提高智能预测模型的可信度。在同等条件下，对未经过降维处理的原始数据进行 BP 神经网络训练和预测，实验表明前者不但降低了系统的复杂性，而且提高了 15.07%的准确度。因此，针对慢性肝病，本文的智能预测模型能达到一定的精度和准确度，为医生的辅助诊断提供可行的方法。

参 考 文 献

[1] 世界卫生组织[OL]. http://www.who.int/campaigns/hepatitis-day/2016/event/zh.

[2] 王恩成，唐琳，王健，等. 慢性乙型肝炎中医症候聚类分析研究[J]. 中国中西医结合杂志，2014，34(1)：39-42.

[3] HO C Y, LAI Y C, CHEN I W, et al. Statistical Analysis of False Positives and False Negatives from Real Traffic with Intrusion Detection/Prevention Systems[J]. Communications Magazine, IEEE, 2012, 50(3): 146-154.

[4] PAXSON V, ASANOVIC K, DHARMAPURIKAR S, et al. Rethinking hardware support for network analysis and intrusion prevention[C]// Proc of the 15th USENIX Workshop on Hot Topics in Security. Berkeley, CA: USENIX, 2006: 63-68.

[5] THOMPSON K. Programming techniques: Regular expression search algorithm[J]. Communications of the ACM, 1968, 11(6): 419-422.

[6] RABIN M O, SCOTT D. Finite automata and their decision problems[J]. IBM Journal of Research & Development, 1959, 3(2): 114-125.

[7] 高秀娟. BP 神经网路在肝硬化治疗预测中的应用[J]. 数学理论与应用，2011，31(3)：21-23.

[8] 李耐萍. 基于机器算法的肝纤维化无创诊断分析与比较[D]. 长沙：中南大学，2013：30-36.