基于深度神经网络的肝硬化中医治疗预测研究*

肖 瑞 裴 卫 胡冯菊 肖 勇

（湖北中医药大学信息工程学院 武汉 430065）

〔摘要〕 以中医电子病历中肝硬化数据为数据源，运用数据清洗、主成份分析技术构建致病指标与诊断结果二元组，通过训练神经网络和支持向量机分类器模型进行预测结果对比，结果表明该方法有效可行。

〔关键词〕 中医；电子病历；神经网络；文本挖掘；肝硬化

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.05.013

Predictive Study on TCM Treatment of Liver Cirrhosis Based on Deep Neural Network *XIAO Rui, PEI Wei, HU Fengju, XIAO Yong, Hubei University of Chinese Medicine, Wuhan 430065, China*

〔**Abstract**〕 The paper selects liver cirrhosis data from Electronic Medical Records (EMR) of Traditional Chinese Medicine (TCM) as data source, builds up a tuple consisting of pathogenic indicators and diagnosis results by means of data cleansing and Principal Component Analysis (PCA), and then compares forecasting results through neural networks and Support Vector Machine (SVM) classifier model to prove feasibility of such a method.

〔**Keywords**〕 Tradional Chinese Medicine (TCM); Electronic Medical Records (EMR); neural network; text mining; liver cirrhosis

1 引言

中医病历又称医案、诊籍，是中医临床各科医生对具体患者进行辨证论治的文字记录，包括患者的生活习性、病情、诊断、治疗及预后等情况，从而成为保存、查核、考评乃至研究具体医生开展具体诊疗活动的档案资料[1]。但随着信息化、网络化的不断推进，电子病历已成为现今医疗记录的大趋势[2]。应用电子病历不仅提高就诊效率、规范中医行业术语，还为后期中医药研究提供数据资源。中医电子病历除具备一般电子病历的特征外还具有自身的特殊性。在病历内容上不仅包括四诊、辩证、立法、处方，西医检查和诊断等现代医学诊疗信息，还包括中医学辨证论治的诊疗信息；在病历结构上既要满足医疗、法律、管理的要求，还要满足中医临床信息全面、准确采集的要求并做到高度结构化，以便对四诊信息中的定性描述进行量化记录；在标准规范化上，建立统一、全面、规范的中医治疗术语词表以便对诊疗用语进行规范；在诊疗处方上，中医处方及中药的药疗医嘱与西医处方和

〔修回日期〕 2018－10－19

〔作者简介〕 肖瑞，讲师，发表论文 3 篇；通讯作者：肖勇，高级实验师。

〔基金项目〕 2017 年湖北中医药大学“青苗计划”项目“基于中医电子病历的慢性乙型肝炎诊断预测算法研究”（项目编号：2017ZZX016）；国家中医药管理局 2018 年度中医药法制化建设项目“互联网虚假违法中医医疗广告监测”（项目编号：GZY－FJS－2018－162）。

配药有很大不同，其配药流程和西医也不相同[3-4]。

肝硬化是由各种因素导致慢性肝损害的一类晚期肝纤维化疾病，肝移植是治疗肝硬化唯一有效手段，但受到供肝及费用等问题限制[5]。查阅近 10 年关于中医药治疗肝硬化腹水的相关文献可知，从病因病机及中医治疗两方面而言，肝硬化腹水的中医病机为正气亏虚，气滞、水停、血瘀 3 者错综为患，中医治疗以辨证分型施治、基本方加减、外治法为主[6]。

2 研究现状

在电子病历研究方面国内外均有一定成果。王昱[7]等基于电子病历数据进行临床接触支持研究，对电子病历数据进行挖掘。李昆[8]等利用深度学习方法结合传统机器学习方法，在电子病历匿名化、胎儿体重预测和疾病分类预测等方面进行预测模型构建的尝试。李准[9]等研究冠心病电子病历中与患者、疾病相关的指标，对冠心病进行分类，进一步探讨检查检验结果与用药之间的关联性。商金秋[10]等通过电子病历进行数据预处理和结构化提取，结合具体需求进行可视化组织与分析。蒋慧丽[11]等提出基于语义技术的电子病历信息集成框架，利用该框架解决电子病历集成及推理问题。陆奕宇[12]等通过对慢性乙型肝炎（乙肝）及肝炎后肝硬化中医证候分类进行系统生物学研究，为乙肝及肝炎后肝硬化的诊断和个体化治疗提供参考依据。本文以中医电子病历中肝硬化数据为研究基点，从中医治疗肝硬化的检查指标入手，通过对电子病历中检查数据进行主成份分析（Principal Component Analysis，PCA），提取出符合要求的致病指标（特征），构建致病指标和诊断结果二元组，将得到的致病指标与诊断结果二元组进行深度神经网络（Deep Neural Network，DNN）预测和支持向量机（Support Vector Machine，SVM）分类处理，通过对两种模型结果对比分析，对肝硬化中医电子病历中检查与诊断结果的关系进行研究。其中 SVM 是基于统计学习理论的结构风险最小化原则的分类方法[13]，是一种监督化学习分类模型。基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略是间隔最大化，最终可转化为一个凸二次规划问题的求解。基本原理是通过将非线性数据映射到高维特征空间，在这个空间构造最优分类超平面，该超平面使类别间的分类间隔最大，有效克服维数灾难和过拟合等传统算法的缺点，能处理小样本、非线性、高维数据，因而成为研究复杂系统问题的热点算法。

3 前期准备

3.1 数据来源

以某地区三甲中医院 2015 年 1 月 - 2016 年 1 月期间诊断结果为乙肝肝硬化和非乙肝肝硬化的 1 273例门诊记录的电子病历为数据来源（参照 2011 年 8 月中国中西医结合学会消化系统疾病专业委员会制定的《肝硬化中西医结合诊疗共识》[14]）。字段主要由诊疗记录中的患者基本信息（门诊号、西医诊断、性别、年龄等）、检验（首次来末次检查总胆红素、凝血酶原时间、白蛋白等）、检查（部位、时间、报告结果等）以及中医诊断信息构成。

3.2 纳入标准

该中医院属于国家重点专科医院，中医电子病历数据结构化程度较为规整，根据筛查检验检查结果，借助具有多年临床经验的医生的指导，将有明确诊断结果的数据纳入。对于检查检验指标缺少数据则不纳入使用。不影响实验的指标缺失，如个人信息，纳入使用。按此标准进行统计纳入，最终符合要求数据为 1 243 例。

3.3 数据预处理

特指对中医检查数据的预处理，主要是针对中医检查数据中的常规字段，包括对检查数据进行修正和规范化。主要是对表意不明确或有歧义的数据进行修正，主要由临床医师进行人工筛查、纠正。对检查数据的规范化主要由于检查数据中存在一种指标有多种说法或有的说法不规范，先通过模糊查找，再通过医学相关专业人员辅助核定。

3.4 特征提取

完成源数据预处理后进行特征提取，主要是通过主成份分析法对肝硬化检查指标进行分析，提取数据中的中医检查数据，重点对中医检查部位、结果等方面进行主成份分析，具体步骤为：将检查记录中各项数据按句号进行分列，人工剔除不可用或无效信息指标；规整数据，统计诊断指标总数；统计源数据中每个诊断指标出现次数，计算各诊断指标频率；将各诊断指标频率除以诊断指标总数，计算每个诊断指标占有率；通过诊断指标占有率进行指标筛选，选取诊断指标占有率高的指标，确定为主要致病指标，即为特征。按照纳入标准完成数据预处理后，利用特征构建方法对检查记录各项数据进行分列，得到共包含指标数据 4 914 条（含重复项）；对分列数据进行规整统计后共包含指标数据2 002条（不含重复项）；对规整后数据进行统计指标占有率筛选后最后得到主要用于训练模型指标数据 140 条。

4 模型构建

本研究使用的中医电子病历门诊数据中包含明确的诊断结果，对于未包含明确诊断结果的数据进行剔除处理，通过对病例特征分析得到可用特征，将可用特征与疾病的明确结果相结合，构建致病指标与诊断结果二元组，将获取的特征按照 one - hot representation 编码规则进行编码，每一病例均以特征展开而构成特征向量，以此构建特征矩阵。将构建好的特征矩阵进行神经网络预测分析和 SVM 分类器训练，其中神经网络模型中输出层和 SVM 分类器结果均定义为二维向量形式，表示电子病历中检查结果为阴性和阳性，即代表是否患病。在神经网络训练过程中对每个训练样本存在一个标准输出，即标签 y，取值为 1 或 0，使用交叉熵损失函数优化此神经网络模型，其交叉熵表达式为：

$$l = -y\ln(y') - (1-y)\ln(1-y') \quad (1)$$

其中 y 为期望输出，y' 为神经元实际输出。对于第 i 个训练样本，模型实际得到的值为 y'_i，其标签为 y_i，可求得其交叉熵数值为：

$$l_i = -y_i\ln(y'_i) - (1-y_i)\ln(1-y'_i) \quad (2)$$

对于一个训练集 S_t 来说，将其均匀划分为多个小数据集（mini - batch）：S_{ti}，每个 mini - batch 中具有 M 个训练样本，对训练集 $S_{ti} = \{x_1, x_2, \cdots, x_M\}$ 而言，交叉熵总和为：

$$l = \frac{1}{|M|}\sum_{i=1}^{|M|} l_i = -\frac{1}{|M|}\sum_{i=1}^{|M|}(y_i\ln(y'_i) + (1-y_i)\ln(1-y'_i)) \quad (3)$$

损失函数为 l，因此优化目标是尽可能地减小 l，即（min（l））。

神经网络预测模型，见图 1。图例通过 Visio 绘制，最底层为输入层，也就是构建的特征矩阵，共 140 维；最顶层为输出层，与诊断结果相对应，共 2 维，即代表肝硬化检查结果是阴性还是阳性（是否患肝硬化）。

图 1 神经网络预测模型

根据电子病历诊断信息可将诊断数据分为两类：诊断结果为阳性或阴性。构建出二分类 SVM 分类器，通过与神经网络模型相同的数据集进行训练，将结果与神经网络预测模型进行对比分析。

5 结果分析

深度神经网络预测结果，见表 1、表 2。两表分别是迭代 100 次和 1 000 次的结果，另外对训练和

测试数据进行不同比例的预测。结果表明运用本研究使用的方法预测结果准确率可达到 80%，其中训练数据和测试数据的比值在 7∶3 较为合适。

表 1 预测结果（迭代 100 次）

总数据（条）	训练数据（条）	测试数据（条）	准确率（%）
1 057	500	557	74.147 2
1 057	600	457	79.212 3
1 057	700	357	80.952 4
1 057	800	257	61.089 5
1 057	900	157	78.980 9

表 2 预测结果（迭代 1 000 次）

总数据（条）	训练数据（条）	测试数据（条）	准确率（%）
1 057	500	557	77.737 9
1 057	600	457	82.056 9
1 057	700	357	80.952 4
1 057	800	257	77.042 8
1 057	900	157	79.617 8

在进行 SVM 训练中阳性和阴性分别用 +1 和 −1 表示，通过已构建的特征向量，采用 SVM 模型进行训练，Libsvm 开源软件包，利用 n − fold 进行交叉验证，其中 n 取值为 10，通过反复试验跳转参数，最终结果，见表 3。

表 3 SVM 实验结果

类别	数值
交叉验证准确性	80.15%
惩罚因子	1
核参数	4

通过对比可以看出在两者预测准确率均达到 80% 的情况下神经网络模型准确率相对于 SVM 模型准确率要高。表明筛选出的诊断肝硬化的指标可作为诊断肝硬化核心指标，以该指标构建训练的模型可对患者进行肝硬化预测诊断，若将该模型应用于临床能够有效降低患者就医成本，提高医生诊疗效率，对临床诊断肝硬化或研究其他疾病具有一定指导意义。

6 讨论

6.1 电子病历缺陷

在互联网高速发展下电子病历普及程度越来越高，但各电子病历软件智能程度不一，特别是中医电子病历，其中的医用专业术语标准不统一且当前未形成统一规范，不同医生记录过程存在差异，在进行电子病历相关数据挖掘过程中存在各种问题，从而影响数据质量。

6.2 数据清洗

数据挖掘过程中不可或缺的重要步骤，决定后期挖掘效果和质量。由于中医电子病历中医用专业术语标准不统一、描述不规范，在进行数据清洗和预处理时需要剔除掉不可用、修改不规范、填补缺失值等，从而使得数据集减小，对模型训练有一定影响，同时由于数据预处理过程中需采用人工筛查、规整和规范化，可能造成异常或错误数据等问题，从而使得整体数据质量出现问题。

7 结语

在模型构建算法上，本文仅从神经网络模型和支持向量机分类模型出发，借鉴前人经验，缺乏其他算法的对比和对复合算法的构建。后续研究中将进行更加严格、规范化的清洗工作，以进一步提高模型准确性，采用更大、更有效的数据集进行模型训练，对更多算法进行对比，以求提出更适合肝硬化病症特点的算法进行算法复合模型训练，从多种角度进行探索，训练出准确率更高的模型，将模型投入临床试用，为中医临床提供辅助诊疗，为中医药智能化提供辅助。

参考文献

1 岳琳哲，施诚. 中医电子病历概述［J］. 中医药管理杂志，2008，16（2）：138－141.
2 邸丽，赵菁，王亚军. 电子病历对临床教学的影响与改进［J］. 中国病案，2018，19（1）：72－74.
3 刘保延，张红，倪皖东. 试论中医电子病历系统及其特殊性［J］. 医学信息，2004，17（1）：9－11.
4 赵移畛，尚东挺. 中医电子病历系统应具备的中医特点［J］. 中医药管理杂志，2007，15（12）：894－896.

（下转第 76 页）

基本设计和思路，探讨实践过程，通过认识、内化、转移流程使临床工作者熟悉病案首页数据质量控制相关知识体系和管理流程，有利于专业知识共享与交流，实现组织的协作与沟通，从不同方面弥补知识共享与交流的不足，有力推动创建学习型团队管理机制。知识服务关注焦点和最终评价不应仅仅是向用户提供所需信息，而应是通过服务解决用户所面临的问题[15]。在知识库管理过程中重视临床工作者所面临的问题需求，为其提供准确、有效的知识服务，不断提升病案首页数据质量。实践应用工作对住院病案首页数据质控的知识培训和固化需要持续地推动，调动医务人员参与积极性，不断深化新问题的分析与应对，定期评估知识服务成效与优化沟通模式，完善知识库建设，推动医院病案首页数据质控创新管理。

参考文献

1 郑大喜，马月耳. 推行按病种支付医保费用与病种成本核算的探讨［J］. 中国卫生经济，2005，24（266）：38－40.

2 李田，王彬，沈绍武，等. 近20年我国病案首页研究文献的可视化分析［J］. 医学信息学杂志，2017，38（1）：70－74.

3 孙蕾，谢志耘，李晓霞. 医院机构知识库构建［J］. 医学信息学杂志，2016，37（4）：14－19.

4 余元龙，郭茜. 数据分析与挖掘技术在医疗质量管理中的应用［J］. 医学信息学杂志，2011，32（1）：34－37.

5 曹锦丹. 基于文献知识单元的知识组织——文献知识库建设研究［J］. 情报科学，2002，20（11）：1187－1189.

6 唐慧，唐娟，周莉莉. 支持复杂疾病的临床路径知识库平台的构建及应用［J］. 中国医院管理，2015，35（6）：45－47.

7 马云，夏新，刘博，等. 基于临床决策支持系统与知识库的临床数据中心的研究与应用［J］. 中国医疗设备，2014，29（7）：61－63.

8 金蕾，杨耀芳，汤春红，等. 社区医院用不同分级管理抗菌药物知识库智能化管理系统开发［J］. 中国全科医学，2018，21（11）：1382－1386.

9 蒋勋，徐绪堪. 面向知识服务的知识库逻辑结构模型［J］. 图书与情报，2013（6）：23－31.

10 田志刚. 知识库建设的5个步骤［EB/OL］.［2018－12－27］. http：//www. kmcenter. org/zhishikujianshebuzhou－wubuzhou.

11 张玉，张文举，李娜. 构建以知识服务和知识组织为基础的医药学知识库［J］. 医学信息学杂志，2010，31（2）：26－29.

12 郑西川，于广军，吴刚，等. 面向区域医疗协同的临床路径诊疗决策知识库平台模型［J］. 中国数字医学，2009，4（5）：23－26.

13 邓盼盼，李军莲. 国外临床知识库发展现状及启示［J］. 医学信息学杂志，2017，38（3）：77－83.

14 何伟，赵慧智，马艳凯，等. 绩效考核对电子病历首页质量控制的影响［J］. 医学信息学杂志，2015，36（5）：39－41.

15 靳红，程宏. 知识管理与知识服务的实践探索——特色专题知识库建设［J］. 现代情报，2004，24（5）：119－120.

（上接第59 页）

5 张静雯，时永全，韩英. 肝硬化的治疗进展［J］. 临床肝胆病杂志，2015，31（3）：465－468.

6 王栋平，李娟梅，刘明坤，等. 肝硬化腹水的中医治疗现状［J］. 吉林中医药，2018，38（2）：240－242.

7 王昱. 基于电子病历数据的临床决策支持研究［D］. 杭州：浙江大学，2016.

8 李昆. 基于电子病历的深度神经网络预测模型研究与应用［D］. 郑州：郑州大学，2017.

9 李准. 基于冠心病电子病历的数据挖掘研究［D］. 重庆：重庆医科大学，2013.

10 商金秋，朱卫国，樊银亭，等. 基于电子病历可视分析的临床诊断模型［J］. 计算机系统应用，2016，25（12）：100－107.

11 蒋慧丽. 基于语义的电子病历数据集成［D］. 重庆：西南大学，2016.

12 陆奕宇，宋雅楠，张贵彪，等. 慢性乙型肝炎及其肝炎后肝硬化中医证候分类的系统生物学研究［J］. 世界科学技术－中医药现代化，2013，15（6）：1281－1287.

13 V. VAPNIC. 张学工译. 统计学习理论的本质［M］. 北京：清华大学出版社，2000：226.

14 刘成海，危北海，姚树坤. 肝硬化中西医结合诊疗共识［J］. 中国中西医结合消化杂志，2011，19（4）：277－279.