林　鑫,李　晋,刘　蕾,等. 二型糖尿病肾病风险预测模型的比较[J]. 中华医学图书情报杂志,2019,28(4):41-45.

DOI:10.3969/j.issn.1671-3982.2019.04.007

·情报研究与方法·

二型糖尿病肾病风险预测模型的比较

林　鑫[1],李　晋[2],刘　蕾[3],梁　晨[1],任慧玲[1]

[摘要]目的:选择相应的机器学习算法构建二型糖尿病肾病风险预测模型,为疾病的早期预防提供科学依据。方法:基于解放军总医院提供的糖尿病数据集,通过对缺失值、异常值等进行一系列预处理,得到 894 条二型糖尿病患者数据。利用单因素逻辑回归筛选出 24 个有效检查指标作为特征,并基于随机森林、BP 神经网络、支持向量机分别构建二型糖尿病肾病风险预测模型,同时对其查准率、召回率进行对比,以验证其应用性能。结果:随机森林预测模型的总体性能最优,3 种算法的训练效果均较好。结论:二型糖尿病肾病风险预测模型能为疾病早期预防控制提供参考依据。

[关键词]二型糖尿病肾病;风险预测模型;随机森林;BP 神经网络;支持向量机

[中图分类号]R587.2　　[文献标志码]A　　[文章编号]1671-3982(2019)04-0041-05

Risk prediction models of type 2 diabetic nephropathy

LIN Xin[1], LI Jin[2], LIU Lei[3], LIANG Chen[1], REN Hui-ling[1]

(1. Institute of Medical Information, Chinese Academy of Medical Sciences/Beijing Union Medical College, Beijing 100020, China; 2. Qinghua University, Beijing 100086, China; 3. Affiliated Dongzhimen Hospital of Beijing University of Traditional Chinese Medicine, Beijing 100700, China)

Corresponding author: REN Hui-ling

[**Abstract**] **Objective** To provide the scientific evidence for the risk prediction models of type 2 diabetic nephropathy established by different machine learning algorithms for the early prevention of diseases. **Methods** Eight hundred and ninety-four data of type 2 diabetic patients were obtained by preprocessing the missing and abnormal values based on the diabetic data set provided by the Chinese PLA General Hospital. Twenty-four effective examination indicators screened from the 894 data by univariate logistic regression analysis were used as the characteristic indicators. Three risk prediction models of type 2 diabetic nephropathy were established using the random forest algorithm, BP neural network algorithm and support vector machine algorithm respectively. Their precision ratio and recall ratio were compared to verify their applicability. **Results** The overall performance of the risk prediction model of type 2 diabetic nephropathy established by random forest algorithm was the better than that of those established by BP neural network algorithm and support vector machine algorithm. The training effect of the three algorithms was good. **Conclusion** The three risk prediction models of type 2 diabetic nephropathy can provide reference and evidence for the early prevention of diseases.

[**Key words**] Type 2 diabetic nephropathy; Risk prediction model; Random forest; BP neural network; Support vector machine

[基金项目]中国医学科学院医学与健康科技创新工程项目"中文临床医学术语系统构建研究"(2017-I2M-3-014);国家自然科学基金项目"基于自然语言处理的内分泌常用药物不良反应监测数据库的构建"(NSFC91846106)

[作者单位]1. 中国医学科学院/北京协和医学院 医学信息研究所,北京　100020;2. 清华大学,北京　100086;3. 北京中医药大学东直门医院,北京　100700

[作者简介]林　鑫(1995-),女,山东烟台人,在读硕士研究生,研究方向为情报学。

[通讯作者]任慧玲(1971-),女,河南周口人,硕士,研究馆员,研究生导师,研究方向为知识组织。E-mail:ren.huiling@imicams.ac.cn

随着人口老龄化和人们生活方式的转变,糖尿病患病率呈直线上升趋势,目前我国已成为全球糖尿病第一大国[1-2]。其中,二型糖尿病患者作为我国糖尿病人群的主体,其临床发病率呈逐步上升趋势,其并发症发生率也相对较高。目前我国大约有 20% ~40% 二型糖尿病患者并发肾病,现已成为慢性肾脏病和终末期肾病的重要原因[3]。二型糖尿病肾病的临床特征主要以蛋白尿排泄异常为主,严重时常合并肾功能衰竭,一旦发展至终末期,将会比其他肾脏疾病的治疗更加棘手[4]。因此有效的早期预测及相关的风险预测模型研究对于二型糖尿病肾病的早期预防和降低并发率具有重要的意义。

目前临床上对糖尿病肾病进行诊断的依据主要包括实验室检查、病理学诊断、糖尿病视网膜病变等,诊断过程繁琐且耗时[5]。本文以二型糖尿病肾病风险预测为目的,对解放军总医院提供的糖尿病数据集进行预处理后,依据数据集中已有的各项临床检查指标,选用随机森林、BP 神经网络、支持向量机 3 种较为成熟的算法建立风险预测模型,并利用查准率、召回率等指标对三种模型的性能进行比较,以选出在二型糖尿病肾病风险预测方面更具优势的算法,实现对二型糖尿病肾病的发生风险进行简便快捷的预测。

1 数据与方法

1.1 数据来源

本文数据来自国家人口与健康科学数据共享服务平台临床医学科学数据中心(中国人民解放军总医院)提供的 2009-2010 年糖尿病数据集[6]。该数据集包含诊断表、尿常规检查表、生化检查表等记录了患者的 ID 号、诊断结果、各项身体指征,以及包括尿白细胞、直接胆红素、血清白蛋白等在内的多项检查结果。每个表通过患者的唯一 ID 号进行关联,并对检查表中各项检查的正常值进行了说明。

1.2 数据预处理

由于原始数据被分为多个表格,且存在缺失值、异常值等噪声数据,故需要对原始数据进行预处理,以控制数据的完整性和准确性,保证结果的准确性。本文所做预处理步骤如下。

数据整合:由于原始数据被划分为诊断、尿常规和生化等多个表格,故首先依据诊断表中的信息筛选出单纯二型糖尿病及二型糖尿病并发肾病的记录,依据这些记录对应的患者 ID 以及诊断时间从诊断、尿常规、生化检查等表格中提取距离诊断时间最近的一次患者检查信息,利用 Excel 的 lookup 和 min 函数对诊断、尿常规和生化等多个表格中的检查数据进行整合。

缺失值处理:数据的缺失会增加分析过程的难度,造成分析结果的偏倚,降低结果的准确性。由于均值插补法计算量相对较小,可高效快速地对缺失值进行处理[7],对整合后的数据进行整理,然后分别求各列数据平均值后对空缺数据进行填补。

异常值处理:在处理异常值时,利用拉依达准则[8],即以给定的置信概率 99.7% 为标准,以 3 倍数据列标准差为依据,凡大于 3 倍标准差的误差则认为是粗大误差,即异常值,删除筛选出的异常值。

经过对数据集的预处理,共得到 472 条二型糖尿病并发肾病数据和 422 条单纯二型糖尿病数据。经过预处理后得到的数据集部分截图如图 1 所示。

ID	年龄	丙氨酸氨基转移酶	天冬氨酸氨基转移酶	总蛋白	血清白蛋白	总胆红素	直接胆红素	尿白细胞	尿比重测定	尿红细胞
C0D128337190	45	14.80	23.50	50.30	22.60	10.70	1.50	9.4	1.014	31.7
C0D324825190	62	13.10	16.80	51.50	24.00	6.70	1.40	6.2	1.014	3.2
C0D330314190	53	12.20	19.60	69.10	26.80	4.10	2.70	1597.8	1.014	119.3
C0D335698190	68	11.80	11.80	52.70	30.70	8.79	2.91	14.7	1.011	25.2
C0D344463190	55	21.26	24.49	65.32	35.65	8.79	2.91	4.5	1.024	2.3
C0D442923190	63	20.20	26.40	60.00	32.30	6.10	1.40	3.8	1.009	5
C0D470058190	63	8.60	12.80	77.40	37.00	14.10	4.70	17.3	1.02	5.3
C0D501170190	46	10.00	10.20	62.80	39.00	6.20	2.50	2.3	1.01	7.4
C0D523860190	61	16.70	16.50	76.80	36.90	12.80	3.70	8.4	1.012	4.5
C0D799919190	51	9.30	13.10	71.80	39.60	10.30	3.30	25.7	1.015	3.8
C0E835052190	41	9.40	9.20	53.90	29.70	7.20	2.20	11.2	1.01	23.5
C1E329656201	40	18.40	16.10	48.00	22.00	4.60	0.80	16.2	1.017	2.4
C1E439979201	64	7.70	11.00	55.10	32.50	5.50	1.90	19.7	1.012	8
C1E502653201	51	21.26	24.49	65.32	35.65	8.79	2.91	50.8	1.009	5.5
C1E502932201	38	84.30	34.30	70.10	42.50	9.80	3.30	7.5	1.015	0.7
C1E503841201	57	24.90	16.20	65.00	47.20	14.20	5.60	4.3	1.021	3.2
C1E558544201	67	28.60	22.70	56.10	31.60	6.70	2.30	3.1	1.013	3.5
C1E560370201	44	12.50	13.70	67.70	38.80	9.30	2.40	32.7	1.013	1.3

图 1 预处理后的数据截图(部分)

2　风险预测模型构建

2.1　单因素逻辑回归分析

将整合各检查表得到 38 个检查指标，赋值后，利用 SPSS 19.0 进行单因素逻辑回归分析。部分赋值情况见表 1。最终筛选出 24 个具有统计学意义的指标（$P<0.05$），分别为年龄、尿比重、尿胆原、尿红细胞、尿糖、尿液结晶、尿液颜色、尿蛋白、总蛋白、血清白蛋白、总胆红素、直接胆红素、尿素、谷氨酰基转移酶、肌酐、葡萄糖、血清尿酸、总胆固醇、肌酸激酶、乳酸脱氢酶、钙、钾、氯化物以及无机磷。

表 1　数据集字段赋值对应表（部分）

数据集字段名	变量名	赋值详情
诊断结果	Diagnose	单纯二型糖尿病=0；二型糖尿病并发肾病=1
性别	Sex	女=0；男=1
年龄	Age	18～40=1；41～65=2；≥66=3
尿浊度	UT	清亮=1；微混=2；微浊=3；混浊=4
尿酮体试验	UABT	阴性=0；阳性=1
尿白细胞	NWBC	正常（0～40）=0；异常（≥40）=1
尿比重测定	SG	正常（1.015～1.025）=0；异常（≥1.025 or <1.025）=1
尿蛋白定性试验	PRO	阴性=0；阳性=1
肌酸激酶	CK	正常（2～200）=0；异常（≥200 or <2）=1
乳酸脱氢酶	LDH	正常（40～250）=0；异常（≥250 or <40）=1

2.2　模型算法选择及参数设置

在明确具有统计学意义的检查指标后，运用机器学习中的监督学习方法构建疾病风险预测模型。其中，随机选择数据集的 70%（共 626 条）作为训练集，30%（共 268 条）作为测试集。结合不同算法特点选择的机器学习算法及相应的参数设置如下。

随机森林（Random Forest，RF）是一种基于集成学习的思想将多棵决策树进行组合从而对数据进行分类的机器学习算法[9]。最后的分类结果是由所有决策树进行投票来决定的，其分类结果比 C5.0 决策树模型更加精确，且具有更少的过拟合倾向[10]。本文利用 R 语言中的 RandomForest 函数进行模型构建，由于随机森林对参数并不敏感，因此使用默认参数。

BP 神经网络是一种按照误差逆传播算法训练的多层前馈网络，由输入层、隐藏层和输出层组成，是目前应用最为广泛的神经网络模型之一[11]。该模型对自变量的要求比较低，可以是离散型，也可以是连续型。在 BP 神经网络模型的构建中，加载 R 语言中的 nnet 包，利用 nnet 建立 BP 神经网络模型。设定神经网络的输入节点数为 24，输出节点数为 1，权值的衰减参数为 0.05，通过不断试验改变隐藏节点个数，不断优化神经网络模型，最终在中间隐藏节点数为 10 时，模型效果达到最优。

支持向量机（Support Vector Machine，SVM）是基于统计学习理论、VC 维理论以及结构风险最小化原理的一种机器学习方法[12]，在小样本、非线性以及高维模式下具有很大优势[13]。运用 SVM 算法对数据进行处理时，利用 R 语言中的 ksvm 函数和高斯 RBF 核函数，对数据进行训练和预测。

2.3　模型预测结果

利用 R 语言中相关函数包分别建立随机森林、BP 神经网络和支持向量机 3 种风险预测模型，并依据 7∶3 的比例将数据集随机划分为训练集和测试集，分别用于训练和对二型糖尿病并发肾病的预测。3 种模型对于二型糖尿病并发肾病及单纯二型糖尿病的预测结果如表 2 至表 4 所示。

表 2　随机森林预测结果

结果类型	二型糖尿病并发肾病/例	单纯二型糖尿病/例
实际结果	152	116
正确预测结果	136	108
全部预测结果	144	124

表 3 BP 神经网络预测结果

结果类型	二型糖尿病并发肾病/例	单纯二型糖尿病/例
实际结果	166	102
正确预测结果	128	78
全部预测结果	152	116

表 4 支持向量机预测结果

结果类型	二型糖尿病并发肾病/例	单纯二型糖尿病/例
实际结果	140	128
正确预测结果	104	124
全部预测结果	108	160

3 模型性能评价与比较

3.1 模型评价指标

本文选择查准率(Precision)、召回率(Recall)、正确率以及 F_1 值等 4 个度量值对各个模型的性能进行评价,以检测模型预测结果与真实结果之间的差异,为模型的选择提供依据。其中,查准率越高,算法的敏感性就越高;召回率越高,算法的特异性就越高;正确率越高,算法的精确度越好;而 F_1 度量值越高则可确保召回率和查准率都越高,算法的总体性能越好[14-15]。这 4 个度量值的公式如下。

$$查准率=\frac{算法正确预测的患者数量}{预测为此类别的患者数量}$$

$$召回率=\frac{算法正确预测的患者数量}{实际为此类别的患者数量}$$

$$正确率=\frac{算法正确预测的患者数量}{测试样本中患者总数}$$

$$F_1=\frac{2\times查准率\times召回率}{查准率+召回率}$$

除此之外,本文还引入 ROC 曲线对模型进行评估。在 ROC 曲线中,横轴为假阳性率,纵轴为真阳性率[16],ROC 曲线下面积在 0.5 ~ 0.7 之间的准确度较低,在 0.7 ~ 0.9 之间的准确度一般,在 0.9 以上的准确度较高,小于 0.5 则不符合真实情况[17]。

3.2 模型比较结果

依据预测结果及 R 语言的 ROCR 包,分别计算这 3 种算法的查准率、召回率、正确率、F_1 及 ROC 曲线下面积。结果如表 5 所示。

由表 5 可知,对于 ROC 曲线下面积,随机森林效果最优;对于正确率,随机森林效果最优;对于查准率,支持向量机效果最优;对于召回率,随机森林效果最优。但由于查准率和召回率是一组此消彼长的评价指标,仅用单个指标无法对算法的效果进行总体评价[14]。因此可以用 F_1 值对这 3 种算法的综合性能进行评价,其结果为:随机森林>支持向量机>BP 神经网络。综合上述指标来看,随机森林性能最优,这 3 种算法的训练效果均较好。

表 5 3 种算法的结果

算法	查准率/%	召回率/%	F_1	正确率/%	ROC 曲线下面积
随机森林	94.44	89.47	0.9189	91.04	0.9129
BP 神经网络	84.21	77.11	0.8050	76.86	0.8949
支持向量机	96.29	74.29	0.8387	85.07	0.8558

4 讨论

由于糖尿病肾病是二型糖尿病患者常见的并发症,目前国内外针对二型糖尿病肾病风险预测模型的建模已有相关尝试,常用的建模方法主要包括逻辑回归、分类与决策树模型、支持向量机、神经网络模型等[18]。本文在综合前人研究成果的基础上,结合不同算法的特点及优缺点,充分考虑所选数据集的实际记录情况,最终选取随机森林、BP 神经网络以及支持向量机这 3 种机器学习算法,并综合多种指标验证其应用于二型糖尿病肾病风险预测时的性能,为算法的选择提供依据。本文结合单因素逻辑回归进行指标筛选,再选择相关算法建立预测模型的方法具有普适性,其不仅可用于对二型糖尿病其他并发症进行预测,也可用于其他疾病的风险预测。

本文也存在局限性。考虑到数据集的数量和质量对模型的可靠性和可扩展性具有重要影响[19],本文采用的数据共 894 例,这些数据对于建模来说还相对较少,会直接影响到模型的效果。未来将进一步扩大数据量并对相关算法进行改进,使模型综合性能得以提高。

5 结语

二型糖尿病肾病由于存在复杂的代谢紊乱,一旦发展至终末期,其治疗将会更加棘手,因此早期对二型糖尿病肾病进行风险预测具有十分重要的意义。本文利用解放军总医院提供的 2009 - 2010 年

度糖尿病数据集,采用均值插补法及拉依达准则对原始数据进行预处理,得到 894 条数据。利用单因素逻辑回归从原数据集的 38 个检查指标中筛选出 24 个有效指标并构建训练数据集和测试数据集,同时基于随机森林、BP 神经网络、支持向量机 3 种算法分别构建二型糖尿病肾病风险预测模型。通过利用查准率、查全率、正确率、F_1 值以及 ROC 曲线下面积等 5 个度量值分别对这 3 种模型的性能进行比较,发现基于随机森林算法构建的风险预测模型性能最佳。本文结果可为二型糖尿病肾病的早期筛查及相关风险预测模型的算法选择提供参考及帮助。

【参考文献】

[1] Yang W, Lu J, Weng J, *et al*. Prevalence of diabetes among men and women in China[J]. New England Journal of Medicine, 2010, 362(12):1090-1101.

[2] 赵海燕,隋树杰,徐龙猛. 2 型糖尿病患者自我管理行为与心理一致感的相关性[J]. 现代临床护理,2015,14(2):13-16.

[3] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2017 年版)[J]. 中国糖尿病杂志,2018,10(1):4-67.

[4] 罗加凯,李志红,尹　飞,等. 糖尿病肾病的临床特征及其危险因素研究[J]. 临床荟萃,2017,32(9):763-766.

[5] 中华医学会内分泌学分会. 中国成人糖尿病肾脏病临床诊断的专家共识[J]. 糖尿病天地(临床), 2016,10(6):243-253.

[6] 国家人口与健康科学数据共享服务平台. 糖尿病数据集[EB/OL]. [2019-03-23]. http://www.ncmi.cn/info/lcyx_jb/13815.

[7] 李　杰,张晓玲. 随机试验设计中缺失值插补方法研究[J]. 大理学院学报,2013,12(10):1-5.

[8] 张　敏,袁　辉. 拉依达(PauTa)准则与异常值剔除[J]. 郑州工业大学学报,1997,18(1):84-88.

[9] 汪桂金. 随机森林算法的优化改进及其并行化研究[D]. 南昌:南昌大学,2019.

[10] Piccinini G. The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity"[J]. Synthese, 2004, 141(2): 175-215.

[11] 刘天舒. BP 神经网络的改进研究及应用[D]. 哈尔滨:东北农业大学, 2011.

[12] 丁世飞,齐丙娟,谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报,2011,40(1):2-10.

[13] 张丽娜,李国春,周学平,等. 基于支持向量机的急性出血性脑卒中早期预后模型的建立与评价[J]. 南京医科大学学报:自然科学版, 2016,36(1):80-84.

[14] 牟冬梅,任　珂. 三种数据挖掘算法在电子病历知识发现中的比较[J]. 现代图书情报技术,2016,32(6):102-109.

[15] Manning C D,Schutze H, Raghavan P. 信息检索导论[M]. 王斌,译. 北京:人民邮电出版社,2010:105-107,196-200.

[16] Hanley J A,Mcneil B J. The Meaning and Use of the Area under a Receiver Operating(ROC) Curve [J]. Radiology, 1982, 143(1): 29-36.

[17] 李　星. 基于复杂网络的症状基因预测方法研究[D]. 北京:北京交通大学, 2014.

[18] 李　攀. 基于神经网络的 2 型糖尿病并发症预测模型的研究[D]. 广州:广州中医药大学, 2016.

[19] 王　萍. 基于电子病历数据的疾病预测模型构建研究[D]. 长春:吉林大学, 2017.

[收稿日期:2019-03-30]

[本文编辑:孙伟娟]