利用患者相似性建立个性化糖尿病预测模型*

黄艳群 王 妮 张 慧 刘红蕾 陈 卉 魏 岚 费晓璐

（1 首都医科大学生物医学工程学院 北京 100069
2 首都医科大学临床生物力学应用基础研究北京市重点实验室 北京 100069）

（首都医科大学宣武医院 北京 100053）

〔摘要〕 利用患者相似性筛选不同规模的研究队列，分别建立基于 Logistic 回归、决策树和 BP 神经网络的糖尿病个性化及非个性化预测模型，探讨基于患者相似性的个性化与非个性化疾病预测模型性能差异，以及基于不同机器学习算法的个性化预测模型性能差异。

〔关键词〕 患者相似性；个性化预测模型；糖尿病

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2019.01.012

Establishing the Personalized Diabetes Prediction Models by Making Use of Patient Similarity *HUANG Yanqun, WANG Ni, ZHANG Hui, LIU Honglei, CHEN Hui, 1School of Biomedical Engineering, Capital Medical University, Beijing 100069, China, 2Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing100069, China; WEI Lan, FEI Xiaolu, Xuanwu Hospital, Capital Medical University, Beijing 100053, China*

〔**Abstract**〕 By making use of patient similarity, the paper screens study cohort of various sizes, establishes personalized and non－personalized diabetes prediction models based on Logistics Regression (LR), Decision Tree (DT) and Back Propagation (BP) neural network, discusses the difference between the performances of the personalized disease prediction model and the non－personalized one based on patient similarity, as well as the difference between the performances of personalized prediction models based on different machine learning algorithms.

〔**Keywords**〕 patient similarity; personalized prediction model; diabetes

1 引言

在临床医学中疾病诊断和分期、预后预测等属于数据挖掘中的分类和预测任务。疾病诊断预测是指以疾病的多种影响因素为基础，利用可靠的大规模临床数据建立模型，预测具有某些特征的人群发生某种疾病的概率，对疾病发生与否进行判断，从而帮助临床医生进行疾病的诊断和治疗[1]。传统的预测建模方法是使用所有可用的同一批训练样本为所有测试样本构建相同的预测模型，即“全局”预测模型。由于这种方法会忽略或丢失对特定目标患者重要的信息，得到的预测结果可能不理想。近年来一些学者提出个性化建模思想，即根据患者的历

〔修回日期〕 2018－09－07

〔作者简介〕 黄艳群，硕士研究生，发表论文 1 篇；通讯作者：陈卉，教授，博士生导师。

〔基金项目〕 国家自然科学基金项目“面向跨领域异构数据的患者相似性学习方法及应用”（项目编号：81671786）。

史信息寻找与目标患者相似的患者，利用其数据构建动态预测模型，进而获得更佳的预测性能[2-7]。在个性化预测建模过程中，患者之间的相似性决定建模所使用的训练样本，其有助于提高模型的预测性能。此外基于患者相似性的个性化建模思想应用于不同的数据挖掘模型时效果也可能不同。鉴于此，本文对不同模型在个性化预测建模任务中的应用进行探索性研究，以期对个性化预测建模中的模型选择提供一定的参考依据。

2 资料和方法

2.1 数据准备

2.1.1 数据来源 本研究的数据来源于近两年首都医科大学宣武医院的电子病历系统。经过去隐私处理，提取患者性别、年龄、疾病诊断、实验室指标共 4 大项指标作为建模特征。对完成清理的数据通过国际疾病编码第 10 版 ICD－10（International Classification of Diseases, the 10th Revision）编码随机抽取糖尿病（ICD－10 编码为 E10－E14）患者和非患者各 5 000 名数据，构成 10 000 个样本的研究队列。

2.1.2 建模特征选择 由于 ICD－10 编码庞大、过于细致，主要运用于临床的疾病细致分类，不利于进行病种分类[8]，因此选用能够对疾病进行病种分类的临床分类软件（Clinical Classifications Software, CCS）编码[9]作为特征输入。首先根据样本涉及的所有疾病诊断的 ICD－10 编码生成相应的 CCS 编码（共 191 个）。然后利用卡方检验确定在糖尿病患者及非患者之间发生率有统计学差异（$p<0.05$）的疾病诊断共 28 个作为最终输入。保留所有患者中无数据缺失的实验室指标共 77 个作为输入特征。经过特征选择，选入建模的特征共 107 个，即性别、年龄、28 个疾病诊断及 77 个实验室指标。其中性别为二值变量，28 个疾病诊断表示为 28 个是否患病的二值变量，年龄和实验室指标为连续型变量，输出为患有糖尿病的概率。

2.2 计算患者相似性

2.2.1 概述 首先计算样本各个特征（年龄、性别、疾病诊断、实验室指标）间的相似性，然后汇总为样本间的相似性。设 X 和 Y 分别表示两个样本（患者），患者特征相似性的计算方法如下。

2.2.2 年龄相似性 利用患者 X 和 Y 两者最小年龄与最大年龄之比作为年龄相似性 $S_{age}(X, Y)$。其中 AGEx 和 AGE_Y 分别表示患者 X 和 Y 的年龄，MIN 和 MAX 表示求最大值和最小值。

$$S_{age}(X,Y)\ \frac{\mathrm{MIN}(\mathrm{AGE_X AGE_Y})}{\mathrm{MAX}(\mathrm{AGE_X AGE_Y})} \quad (1)$$

2.2.3 性别相似性 患者 X 和 Y 的性别相同时性别相似性 $S_{sex}(X,Y)$ 为 1，不同时为 0。

$$S_{sex}(X,Y) = \begin{cases} 1, & SEX_X = SEX_Y \\ 0, & SEX_X \neq SEX_Y \end{cases} \quad (2)$$

2.2.4 疾病诊断相似性 利用 4 位 ICD－10 疾病编码层级结构计算患者 X 和 Y 的疾病诊断相似性 $S_{dis}(X,Y)$[10]，见图 1。其中 |A| 和 |B| 分别表示两个患者的疾病诊断的个数，A、B 分别为两个患者所有疾病诊断（按 4 位 ICD－10 编码分类）的集合，|AUB| 表示它们的并集，A \ B 表示集合 A 中包含但集合 B 中不包含的疾病集合，B \ A 表示集合 B 中包含但集合 A 中不包含的疾病集合，d(a,b) 是疾病 a 和 b 的 ICD－10 编码在树型 ICD－10 编码体系中的层级距离，它根据疾病层级自上而下计算而得。NCA(a,b) 表示当自上而下遍历疾病 a 和 b 的 ICD－10 编码 4 位编码层级结构时相同的层数，遍历直至遇到不同的层时停止；#levels 表示疾病层数，本研究中#levels = 4。若患者 X 和 Y 的疾病诊断中有相同的疾病诊断，则不计算该疾病与其他疾病的层级距离。如疾病编码为 C16.9（胃癌）和 C34.9（右支气管肺癌）只有层级结构的第 1 层“C”相同，故两者的层级距离为 1/4。

$$S_{dis}(X,Y) = 1 - \frac{1}{|AUB|}\left(\sum_{a\in A\setminus B}\frac{1}{|B|}\sum_{b\in B} d(a,b) + \sum_{b\in B\setminus A}\frac{1}{|A|}\sum_{a\in A} d(a,b)\right) \quad (3)$$

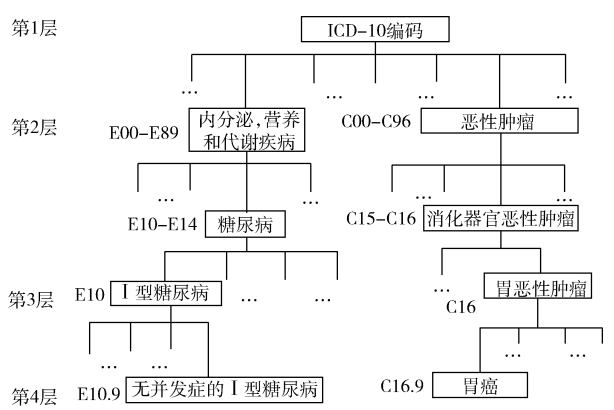

图 1 4 位 ICD－10 编码层级结构

$$d(a, b) = \frac{NCA(a, b)}{\#levels} \quad (4)$$

2.2.5 实验室指标相似性 利用欧式距离计算实验室指标相似性 $S_{lab}(X,Y)$。

$$S_{lab}(X, Y) = \sqrt{\sum_i (LAB_{Xi} - LAB_{Yi})^2} \quad (5)$$

其中 LAB_{Xi} 和 LAB_{Yi} 分别表示患者 X 和 Y 的第 i 个实验室指标值，$i = 1, 2, \cdots 77$。根据预实验对年龄、性别、疾病诊断和实验室指标相似性按照以下最佳权重加权求和得到两个样本间的相似性。

$$Similarity(X,Y) = 0.1 * S_{age}(X,Y) + 0.1 * S_{age}(X,Y) + 0.4 * S_{dis}(X,Y) + 0.4 * S_{lab}(X,Y) \quad (6)$$

2.3 预测模型

选择可输出连续值的 3 种常见机器学习模型即 Logistic 回归（Logistic Regression，LR），决策树（Decision Tree，DT），BP（Back Propagation，BP）神经网络模型进行对比。选用条件决策树构建决策树模型，能够基于显著性检验自动给决策树剪枝，有效防止决策树模型出现过拟合的现象。BP神经网络结构为 1 个输入层（包含 107 个神经元对应 107 个输入特征）、1 个隐含层（根据经验确定包含 7 个神经元）和 1 个输出层（包含 1 个神经元，输出分类概率值）。权重的初始值设置为 0 ~ 1 的随机数。为便于比较模型的性能，本研究构造一个参照模型，即利用患者相似性为待测患者抽出前 K% 个最相似的训练样本，这些训练样本中糖尿病患者所占比例作为该待测患者的预测结果。

2.4 验证与评价预测模型

本研究采用 hold - out 验证方法进行建模和验证，将整个研究队列按 9：1 的比例随机划分为训练集（9 000 个样本）和测试集（1 000 个样本）。建立个性化预测模型时，为每个测试样本抽取训练集中与该样本相似性最高的前 K% 个训练样本来建模。K 取值 1 ~ 70，即建模时训练样本的规模取 90 ~ 6 300。同时随机抽取 K% 个训练样本建立相应的非个性化模型。选用 ROC 曲线下面积（Area Under the Curve，AUC）评价模型的预测准确性。

3 结果

3.1 个性化模型与非个性化模型间的比较（表 2）

表 1 不同个性化模型和非个性化模型以及参照模型的 AUC 范围及平均值比较

ROC 曲线下面积	个性化模型			参照模型	非个性化模型		
	LR	DT	BP		LR	DT	BP
最大值	0.903	0.892	0.787	0.699	0.900	0.890	0.693
最小值	0.543	0.857	0.672	0.611	0.539	0.679	0.500
平均值	0.884	0.883	0.718	0.638	0.870	0.868	0.544
标准差	0.055	0.007	0.025	0.024	0.061	0.029	0.056

3.1.1 LR 模型 个性化与非个性化模型的 AUC 均随训练样本量的增加而增大，在训练样本量分别达到 10%（90）和 34%（3 060）之前，AUC 随训练样本量增加变化较大，随后变化减缓且基本达到最高；在训练样本量较多时两者的预测性能均属于优秀且基本保持稳定，表明不再需要更多训练样本量进行建模。整体上个性化模型性能优于非个性化模型，见图 2。

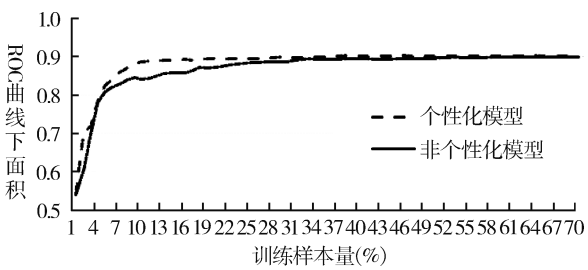

图 2 Logistic 回归模型预测性能

3.1.2 DT 模型 个性化模型的 AUC 随训练样本量的增加变化幅度较小，在 0.883 左右浮动，表明其受到训练样本量的影响较小。非个性化模型的 AUC 在训练样本量较少时（少于 4%）升高幅度较大，随后基本保持稳定。整体上个性化模型性能优于非个性化模型，见图 3。

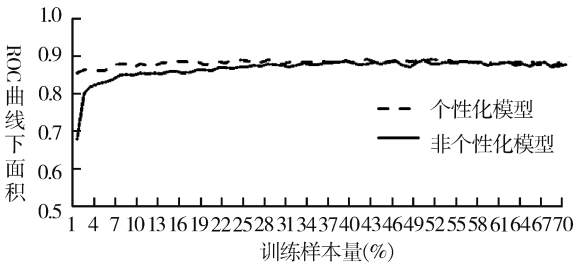

图 3 决策树模型预测性能

3.1.3 BP 模型 个性化模型的 AUC 逐渐下降，变化幅度较小，非个性化模型 AUC 变化浮动不定，规律性不强；诊断能力均较低。总体上个性化模型性能优于非个性化模型，且达到最佳预测性能时所需要的训练样本量少于非个性化模型，见图 4。

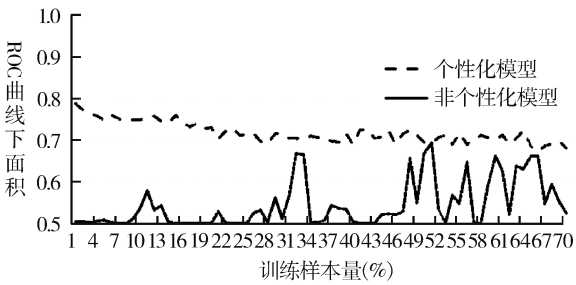

图 4 BP 神经网络模型的预测性能

3.2 不同个性化模型间的比较（图 5）

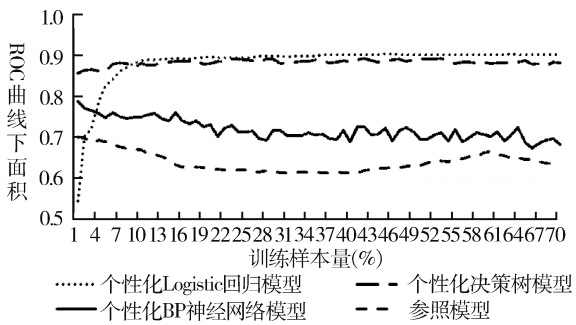

图 5 不同个性化模型预测性能比较

无论训练样本量如何变化，LR、DT 和 BP 模型的 AUC 均高于参照模型，特别是 LR 和 DT 模型的 AUC 明显高于参照模型。这说明机器学习模型在应用于疾病预测时其性能优于基于简单的投票思想的参照模型。此外，总体上 LR 和 DT 模型的 AUC 相近且较高，尤其是 DT 模型在训练样本量变化时 AUC 均维持在较高水平，预测性能较佳，BP 模型的 AUC 普遍较低。因此 LR 和 DT 模型可能更适用于疾病预测。

4 讨论

4.1 患者相似性及其在医学数据挖掘中的应用

在相似性计算方面，主要有基于 Mahalanobis 距离、欧氏距离等方法以及对相似性进行加权求和从而优化相似性的计算。Sun J 等[2]提出基于局部监督信息的患者相似性学习（Locally Supervised Metric Learning，LSML）算法，将患者的疾病诊断作为监督信息来判别该患者的相似患者，进而得到一种泛化的马氏距离来计算患者相似性。Patel A 等[3]通过对非 ICU 患者生命体征信息的相似性加权，获取最终相似性来预测 ICU 患者转入非 ICU 病房的可能性。陈婕卿等[4]基于艾滋病患者治疗前的基线资料，利用欧氏距离计算患者相似性，实现基于案例推理的治疗方案推荐。在相似性的应用方面，主要有基于患者相似性、选用不同数据挖掘预测模型等进行疾病、死亡预测等。Chan L 等[5]利用电子病历数据计算患者相似性并依此挑选相似患者，建立基于支持向量机的癌症患者个性化生存预测模型。Kenny Ng 等[6]探讨利用患者相似性进行糖尿病风险因素分析和个性化糖尿病预测的可行性。Park YJ 等[7]探索利用不同训练样本规模、临床相似性最高的患者数据构建 Logistic 回归模型来研究死亡预测准确率随建模人数变化的趋势。

4.2 研究结果及分析

本研究选用结构、算法、思想完全不同的 3 种模型，探讨利用患者相似性建立个性化糖尿病预测模型时模型本身对预测结果的影响，取得较为满意

的结果。在相似性计算方面，针对输入特征（年龄、性别、疾病诊断和实验室指标）的不同类型，采用不同的特征相似性计算方法并尝试不同的权重组合，最终得到效果最好的相似性度量结果用于筛选模型的训练样本。实验结果显示整体上个性化预测模型性能明显优于非个性化模型和参照模型，与其他研究[6]的结果类似，主要体现在个性化模型达到最佳预测性能时所需的训练样本量明显较少且在训练样本量相同时个性化模型性能更佳。此外在 3 种个性化预测模型中 LR 和 DT 模型的性能最佳，预测能力均属于优秀。特别是 DT 模型，其随训练样本量变化浮动较小，受到训练样本量的影响较小，用较少的训练样本（如 90 人）即可得到较佳的预测结果。原因可能在于本研究所选的决策树类型为条件推断树，算法本身能够根据实际情况自动剪枝，故性能稳定。BP 模型预测性能一般，原因可能是其训练过程过于依赖各个神经元的初始化赋值，但是这种初始化赋值是随机的，从而导致训练结果出现较大的随机性。

5 结语

基于患者相似性构建个性化糖尿病预测模型具有可行性且相对于传统方法预测效果有所提升；该方法运用于其他疾病的预测上可能会得到较为满意的效果。未来可从扩充样本的特征（如影像学特征）以及尝试其他相似性计算方法等方面开展相关研究。

参考文献

1 张蕊，郑黎强，潘国伟. 疾病发病风险预测模型的应用与建立［J］. 中国卫生统计，2015，32（4）：724－726.

2 Sun J，Wang F，Hu J，et al. Supervised Patient Similarity Measure of Heterogeneous Patient Records［J］. ACM Sigkdd Explorations Newsletter，2012，14（1）：16－24.

3 Patel A，Singh I，Brand L，et al. A Weighted Similarity Measure Approach to Predict Intensive Care Unit Transfers［C］. Kansas City：IEEE International Conference on Bioinformatics and Biomedicine. IEEE Computer Society，2018：1079－1084.

4 陈婕卿，杨秋英，陈卉，等. 基于案例推理在艾滋病患者个性化治疗方案推荐中的应用［J］. 北京生物医学工程，2017，36（1）：16－20.

5 Chan L，Chan T，Cheng L F，et al. Machine Learning of Patient Similarity：a case study on predicting survival in cancer patient after locoregional chemotherapy［C］. Hong Kong：IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE，2010：467－470.

6 Kenny Ng，Jimeng Sun，Jianying Hu，et al. Personalized Predictive Modeling and Risk Factor Identification Using Patient Similarity［J］. Proceedings of Amia Joint Summits on Translational Science，2015（2015）：132－136.

7 Park YJ，Kim BC，Chun S H. New Knowledge Extraction Technique Using Probability for Case－based Reasoning：application to medical diagnosis［J］. Expert Systems，2010，23（1）：2－20.

8 王妮，陈婕卿，刘文艳，等. 基于 Access 的大规模住院病案首页数据挖掘［J］. 中国医疗设备，2017，32（10）：126－128.

9 Wei WQ，Bastarache LA，Carroll RJ，et al. Evaluating Phecodes，Clinical Classification Software，and ICD－9－CM Codes for Phenome－wide Association Studies in the Electronic Health Record［J］. Plos One，2017，12（7）：e0175508.

10 Ni Wang，Yanqun Huang，Honglei Liu，et al. Electronic Medical Records－based Patient Similarity and Its Application in Building Personalized Predictive Model［C］. Sri Lanka：Proceedings of APAMI，2018.