基于集成学习融合模型的血糖预测

王荣政 廖贤艺 陈湘萍 周 凡 周 毅

（中山大学数据科学与计算机学院 广州 510006） （中山大学中山医学院生物医学工程系 广州 510080）

〔摘要〕 介绍集成学习预测方法，阐述集成学习在血糖预测中的应用，基于个体常规体检数据，使用集成学习的方法，融合线性回归、梯度提升决策树、随机森林等模型对血糖进行预测，实验结果表明该方法对血糖具有更高的预测精度并能更准确地识别血糖异常个体。

〔关键词〕 血糖预测；糖尿病；集成学习

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.01.013

Blood Glucose Prediction Based on Ensemble Learning Fusion Model

WANG Rongzheng, LIAO Xianyi, CHEN Xiangping, ZHOU Fan, School of Computer and Data Science, Sun Yat－sen University, Guangzhou 510006, China; ZHOU Yi, School of Biomedical Engineering, Sun Yat－sen University, Guangzhou 510080, China

〔**Abstract**〕 The paper introduces the ensemble learning prediction method and dilates on the application of ensemble learning in blood glucose prediction. Based on individual routine physical examination data, it predicts blood glucose through the ensemble learning method that is combined with linear regression, gradient boosted decision tree, random forest and other models. The experimental results indicate that the method boasts higher prediction precision for blood glucose and is able to identify individuals with abnormal blood glucose more accurately.

〔**Keywords**〕 blood glucose prediction; diabetes; ensemble learning

1 引言

糖尿病是威胁人类健康的 3 大慢性疾病之一，其具有多种并发症且患病人数呈逐年上升趋势。根据发病机理不同，糖尿病主要分为 1 型、2 型、其他特殊类型、妊娠、继发性糖尿病。现阶段糖尿病的治愈较为困难，因此预防与及时干预是应对糖尿病最佳措施。血糖异常检测是糖尿病预警的重要环节，金萌萌等[1]探讨血糖异常判断标准，方式一般为检测空腹或餐后血糖，当空腹血糖≥7.0mmol/L 或餐后血糖≥11.1mmol/L，即可怀疑个体患有糖尿病，应对其进行预警。目前糖尿病前期干预的一种方式是对血糖异常的个体进行预警，以促使其进行饮食、运动或药物调节从而避免患糖尿病。针对未检测血糖的个体，如果能够利用其他体检特征预测血糖将对糖尿病预警具有重要意义。

随着大量个人健康数据的积累，研究人员开始使用数据驱动的方法分析糖尿病相关问题。Rohit Prasad Bakshi 等[2]提出一种系统化的数据挖掘方法来获取最优特征以及模型，首先选择数据库中存在的最佳糖尿病相关指标，然后通过投票的方法选择现存模型中最适合的模型，结果表明这种方法的有效性；Thippa Reddy Gadekallu 等[3]通过优化搜索方

〔收稿日期〕 2018－09－04

〔作者简介〕 王荣政，硕士研究生；通讯作者：陈湘萍，助理研究员，硕士生导师，发表论文 20 余篇。

法选择有效特征，利用模糊逻辑系统预测血糖，实验表明该方法优于现有方法；Peihua Chen 等[4]采用 Boost 算法，使用 100 多项临床指标构建糖尿病预测模型，获得较高的准确率。在血糖预测分析中，往往是收集受试者一段时间的血糖值后基于时序模型进行血糖的动态预测。王延年等[5]基于 CGMS 提出一种自适应遗忘因子最小二乘 AR 模型血糖预测模型，从而动态捕捉血糖变化；Taiyu Zhu 等[6]利用 30 分钟内的血糖变化数据训练卷积神经网络来构建血糖预测模型，实验取得良好效果。在糖尿病的相关指标分析工作中，Beatriz López 等[7]利用随机森林技术搜索与糖尿病最相关的单核苷酸多态性（Single Nucleotide Polymorphisms，SNPs）属性，实验显示随机森林在这一领域的有效性。

与糖尿病预测和血糖动态分析不同，本文主要基于生理指标来预测血糖。使用个体的 5 大类共 68 维生理特征，包括血常规、肝功能、肾功能、血脂和其他类体检数据来构建血糖预测模型。在此基础上提出一种基于集成学习的预测方法，融合随机森林、梯度提升树、线性回归模型来预测血糖。在来自医院的真实数据集上对本方法进行预测准确性的验证。

2 血糖预测

2.1 概述

本研究目标是根据个体的体检数据对血糖进行预测。按常见类别将体检数据分为 5 大类：一是肝功能相关类：包括相关酶、白蛋白、球蛋白等；二是血常规相关类，包括中性粒细胞、红细胞计数等；三是血脂相关类，包括总胆固醇、甘油三酯等；四是肾功能相关类，包括尿素、尿酸、肌酐等；五是其他类，包括各种微量元素、免疫蛋白等。5 大类包含共 68 维特征，使用 5 大类来预测血糖是典型的多元回归问题。本文根据血糖预测中存在的特征维度高及共线性的特点，采用集成学习的方法，通过融合梯度提升决策树、随机森林模型、线性回归等多个模型来对血糖进行预测。

2.2 集成学习模型融合

2.2.1 概述 血糖预测任务中存在特征维度高、特征之间存在多重共线性以及体检数据噪声较大等问题，这使得单一模型往往没有较好的稳定性。为克服以上问题并获得较好的稳定性，本文提出使用集成学习的方法融合多模型来预测血糖。集成学习的方法主要包括自助法[8]（Bagging）、提升法[9]（Boosting）和叠加泛化法[10]（Stacking）等。其中 Bagging 与 Boosting 往往都只能融合相同模型，无法克服单一模型不稳定的问题。为此使用集成学习中 Stacking 方法，融合不同模型来进行血糖预测。Stacking 融合算法是指训练一个模型用于组合其他各个模型，首先训练多个差异较大的模型，然后再以这些模型作为基础模型，将其输出作为元模型的输入，通过元模型学习基础模型的结果，最后元模型的输出作为最终输出。

2.2.2 基础模型选择 针对血糖预测特点来选择模型，由于血糖预测涉及的特征维度较高，且特征之间存在多重共线性的问题，普通的线性回归不能很好地解决这一问题，因此本文引入具有良好非线性拟合能力的梯度提升树作为 Stacking 融合模型的基础模型。另外体检数据往往存在较大噪声问题，可能导致模型过拟合，而梯度决策树属于加性决策树模型，其应对过拟合的能力较弱，为此引入随机森林作为 Stacking 融合模型的另一基础模型。此外考虑到集成学习中融合的模型差异越大，其融合效果越好，将与上述模型差异较大的线性回归也作为融合模型的基础模型。梯度提升决策树（Gradient Boosting Decision Tree，GBDT）是一种迭代的决策树算法，通过构造弱学习器来对上一轮学习器的误差进行拟合，通过弱学习器的组合从而构成强学习器。随机森林是一种基于树的集成学习模型，其通过随机抽样构建多棵决策树后将其进行组合达到强学习器的效果，使其有很好的抗过拟合能力。另外由于其特征子集是随机生成的，所以在特征维度较高时也有良好的表现。线性回归是机器学习中最基础的回归模型，是对输入的一种拟合，其输出是特征的线性组合。

2.2.3 元模型选择 在引入基础模型后应选择合适的元模型。为避免融合模型过于复杂而导致过拟合，本文选择加正则项的线性回归作为元模型来融合随机森林、梯度提升树和线性回归。集成学习 Stacking 融合血糖预测模型结构，见图 1。首先从整个训练数据集中通过抽样得到各个训练子集合，每个子集合作为基础模型的输入，待基础模型学习训练后将其输出作为元模型的输入，元模型的输出即为最后输出。集成学习通过构建及合并多个学习器来完成学习任务，往往较单一学习器更有更显著优越的泛化性能。

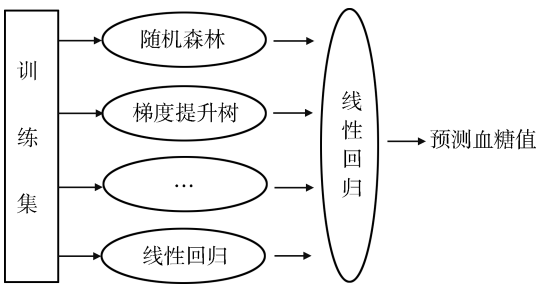

图 1 集成学习 Stacking 融合血糖预测模型结构

3 实验

3.1 数据集与评价指标

3.1.1 数据集 为达到通过个体体检数据预测血糖这一目的，从广东省某医院获取 19 802 条患者体检数据。该数据集中用户体检信息包括性别、年龄、血常规、肝功能、肾功能、血脂、尿常规、空腹血糖等个人以及体检相关数据共 68 维特征。由于部分个体只进行部分体检，因此抽取体检项 >30 的个体进行实验，最终数据集共包含 12 531 条样本。

3.1.2 评价指标 为衡量模型血糖预测效果，选择回归问题中最常见的评价指标均方根误差（RMSE）作为血糖预测的衡量指标，见公式（1）。其中 N 为样本总数，i 为第 i 个样本，y 为样本血糖真实值，y^* 为血糖预测值。

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_i}\sum_{i=1}^{N}(y_i - y_i^*)^2} \quad (1)$$

为衡量模型对血糖异常的预警能力，研究其能否正确识别血糖异常的个体。判断空腹血糖异常的标准在金萌萌的论文[1]中被探讨，美国糖尿病协会、世界卫生组织等机构对空腹血糖异常尚有分歧，但空腹血糖异常判断的阈值在 6.1mmol/L ~ 7.0mmol/L 之间。将样本集中空腹血糖 >7.0mmol/L 的个体视为异常，考虑到血糖预测往往存在误差，设置血糖预测阈值区间为 6.1 ~ 7.0mmol/L 的个体也视为异常，即当预测血糖 >6.1mmol/L 时判定为异常。将空腹血糖 >7.0mmol/L 的样本视为正类，将血糖预测值 >6.1mmol/L 判定为正类。本文选择血糖异常预测的召回率和准确率用于衡量模型的预警能力。召回率与准确率见公式（2）和公式（3）。其中 *TP* 为将正类预测为正类数，*FN* 为将正类预测为负类数，*FP* 为将负类预测为正类数，*TN* 为将负类预测为负类数。

$$\text{召回率} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{准确率} = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

3.2 实验结果

3.2.1 血糖预测分析 为探究体检特征项对血糖预测的影响，按照体检类特征数目的大小依次加入预测特征集中。初始化预测特征集为｛性别，年龄｝，随后依次加入肾功能、血脂、肝功能、血常规、其他相关项直到所有特征加入到预测特征集。实验中分别使用线性回归、随机森林、梯度提升树以及本文提出的集成学习的方法来进行预测。实验结果，见表 1。随着特征项的增加，使用不同方法预测的真实血糖值与预测值的 RMSE 都有减小的趋势，其中从 1.114 逐渐减小到 0.983，说明各大类项特征均对血糖预测具有正向影响。此外可以看出线性回归、随机森林、梯度提升树 3 个模型的预测效果并不稳定。加入血脂相关项时随机森林模型的 RMSE 为 1.144，低于梯度提升树的 1.152；而加入肝功能相关项时，随机森林模型的 RMSE 值为 1.127，高于梯度提升树；线性回归开始加入特征时表现较随机森林好，而随着特征的增多其表现逐渐较随机森林差。集成学习融合模型的 RMSE 始终低于线性回归、随机森林和梯度提升树，说明单个模型往往表现不稳定，集成学习融合模型相较于单

个模型有更强的稳定性和较高的精度。

表 1 不同模型加入体检特征后模型的预测血糖值与真实值 RMSE

模型	加入肾功能相关项	加入血脂相关项	加入肝功能相关项	加入血常规相关项	其他项（所有特征）
线性回归	1.160	1.143	1.136	1.116	1.098
随机森林	1.179	1.144	1.127	1.102	1.027
梯度提升树	1.171	1.152	1.101	1.084	1.012
Stacking 融合模型	1.114	1.121	1.034	0.994	0.983

为探究个体年龄对血糖预测的影响，将数据分为青年（20 ~ 40 岁）、中年（40 ~ 60 岁）、老年（60 ~ 80 岁）3 个子数据集。使用 5 大类特征来预测血糖，3 个子数据集的结果，见表 2。可以看出集成学习算法的效果优于单个模型的预测结果。当样本集为青年时模型预测值与血糖真实值的 RMSE 为 0.739，低于中年样本集的 1.034，也低于老年样本集的 1.276。随着样本集年龄的上升，预测血糖值与真实血糖 RMSE 会随之上升。出现这一现象的可能原因是年龄越大的个体服用药物的可能性越高。由于服用降糖药或其他相关药物影响某些指标，使得血糖预测不准确。

表 2 不同模型在不同年龄段样本集下血糖的预测值与真实值的 RMSE

模型	青年（20 ~ 40 岁）	中年（40 ~ 60 岁）	老年（60 ~ 80 岁）
线性回归	0.812	1.312	1.521
随机森林	0.767	1.177	1.418
梯度提升树	0.747	1.107	1.355
Stacking 融合模型	0.739	1.024	1.276

3.2.2 血糖异常预警分析 为衡量模型对血糖异常的预警能力，将这一问题视为模型对血糖异常个体的识别准确度。使用线性回归、随机森林、梯度提升树以及集成学习的方法进行血糖异常个体准确率与召回率判别，见表 3。实验结果显示随机森林模型的准确率最高，但其召回率最低；集成学习 Stacking 模型识别异常血糖的准确率较随机森林低 1.4%，但其召回率提高 15.6%；集成学习模型 Stacking 召回率与准确率均比梯度提升树高。因此集成学习模型具有更强的稳定性和更可靠的预测性能。

表 3 不同模型对异常血糖的预测召回率与准确率（%）

模型	召回率	准确率
线性回归	33.2	74.5
随机森林	48.0	86.2
梯度提升树	62.2	84.6
Stacking 融合模型	63.6	84.8

4 结语

为有效预测血糖，结合个体的血常规、肝功能、血脂、肾功能等生理指标，使用集成学习的方法，融合线性回归、梯度提升决策树、随机森林等模型对血糖值进行预测。实验结果显示本文提出的 Stacking 融合模型预测的血糖值与真实值之间的均方根误差为 0.98，可以根据常规体检数据进行较精确的血糖预测。此外 Stacking 融合模型预测对异常血糖值识别的准确率达到 84.8%，召回率达到 63.6%，能在一定程度上识别血糖异常个体，这为糖尿病的提前干预提供支持。实验表明血常规、肝功能、血脂、肾功能等生理指标对血糖预测都具有正向影响，这也证实人体生理指标相互联系。另外实验发现青年血糖的预测精度要远高于老年，这可能由于老年人更倾向服用药物导致某种体检指标改变从而影响血糖预测，下一步将考虑收集数据来验证该假设。此外将考虑加入更多的生理指标并对其标进行分析，从而更有效地预测血糖以及分析相关指标。

（下转第 84 页）

进而更好地促进信息产业的健康发展。对于中医药院校图书馆现代化服务发展而言，除传统的阅读宣传推广模式外还应借助新兴的网络信息技术进一步拓展业务。一方面，应有计划、分类别对中医药院校图书馆相关信息服务、管理及信息技术人员开展关于计算机、网络和现代通信技术方面培训，提升其数据挖掘、知识发现、语义网络等信息处理和交流能力；另一方面，积极创新中医药院校图书馆信息咨询的媒介、载体，努力拓宽图书馆在线借阅、在线阅读和电子阅读的渠道，通过网络信息技术、云课程等提高中医药院校图书馆信息化服务质量。

3.4 拓展差异化信息服务渠道

传统图书馆阅读推广服务由于宣传渠道狭窄、推广模式单一，往往无法扩大用户规模[4]。特别是随着互联网的飞速发展及网络阅读网站的大规模兴起，中医药院校图书馆的借阅及信息服务功能受到极大影响。网络信息时代应注重应用先进网络信息技术，挖掘阅读用户群体潜在需求，拓展服务渠道。一方面，借助网络信息技术，中医药院校图书馆可对阅读用户进行精细分类与划分，根据用户在线阅读及活动记录深入分析、了解不同层次用户的阅读需求，提供专门的图书阅读指导，利用网络新媒体工具定期推送所需图书或者为阅读用户提供专门的在线检索入口；另一方面，根据所挖掘用户阅读信息提供差异化、个性化和智能化的定制服务，提高中医药院校图书馆对用户的吸引力。

4 结语

在网络环境下借助信息技术，有助于提高中医药院校图书馆信息服务效率，为广大医学生和医务人员的学习、教学及科研工作提供全方位的信息化支持与服务。

参考文献

1 任鹏. 网络环境下医院图书馆建设与服务［J］. 医学信息学杂志，2012，33（6）：75－77.

2 崔蒙. 中医药信息学概论［M］. 北京：科学出版社，2016.

3 范雪梅. 论信息文化影响下的信息传播机制［J］. 图书馆学研究，2018（1）：13－14，12.

4 白兴勇. 关于图书馆志愿者的理论分析［J］. 图书馆杂志，2015，34（2）：37－42.

（上接第 62 页）

参考文献

1 金萌萌，潘长玉. 空腹血糖受损诊断标准的探讨［J］. 中国糖尿病杂志，2007，15（2）：125－128.

2 Prasad B R，Agarwal S. Modeling Risk Prediction of Diabetes－a preventive measure［C］. Peradeniya：International Conference on Industrial and Information Systems. IEEE，2015：1－6.

3 Gadekallu T R，Khare N. Cuckoo Search Optimized Reduction and Fuzzy Logic Classifier for Heart Disease and Diabetes Prediction［J］. International Journal of Fuzzy System Applications，2017，6（2）：25－42.

4 Chen P，Pan C. Diabetes Classification Model Based on Boosting Slgorithms［J］. Bmc Bioinformatics，2018，19（1）：109.

5 王延年，申艳蕊，张旭，等. 自适应血糖预测模型在低血糖预警中的应用［J］. 中国卫生统计，2014，31（3）：421－424.

6 Zhu T，Li K，Herrero P，et al. A Deep Learning Algorithm for Personalized Blood Glucose Prediction［C］. Stockholm：International Joint Conference on Artificial Intelligence and European Conference on Artificial Intelligence，2018：64－78.

7 López B，Torrent－Fontbona F，Viñas R，et al. Single Nucleotide Polymorphism Relevance Learning with Random Forests for Type 2 Diabetes Risk Prediction［J］. Artificial intelligence in medicine，2018（85）：43－49.

8 Breiman L. Bagging Predictors Machine Learning［J］. Machine Learning，1996，24（2）：123－140.

9 Freund Y，Schapire R E. A Desicion－theoretic Generalization of on－line Learning and an Application to Boosting［J］. Journal of Computer & System Sciences，1997，55（1）：119－139.

10 Wolpert D H. Stacked Generalization［J］. Neural Networks，1992，5（2）：241－259.