文章编号：2095-2163(2020)02-0107-04 中图分类号：TP181 文献标志码：A

基于 Stacking 的糖尿病预测方法研究

章 权[1]，周梁琦[1]，邹 琪[1]，喻新民[2]
（1 东华理工大学 信息工程学院，南昌 330013；2 东华理工大学 软件学院，南昌 330013）

摘 要：近些年，慢性病发展迅速，且危害越来越大。其中糖尿病是一种具有遗传特性的慢性疾病，危害极大。因此，针对糖尿病预测具有非常重要的研究意义。随着人工智能技术不断发展，已经越来越多的机器学习和深度学习方法被用于对疾病发生进行预测。据此，本文提出一种基于集成学习的糖尿病预测方法。该方法采用了 Stacking 的集成学习方法，利用支持向量机、随机森林、人工神经网络等 3 种机器学习方法作为初级学习器，使用逻辑回归作为次级学习器建立糖尿病预测模型。本文以 UCI 中的皮马印第安人糖尿病数据集作为实验数据，通过实验分析，本文提出模型融合方法能取得更好的预测效果。

关键词：糖尿病预测；支持向量机；人工神经网络；随机森林；集成学习

Research on diabetes prediction method based on Stacking

ZHANG Quan[1], ZHOU Liangqi[1], ZOU Qi, YU Xinmin[2]
(1 School of Information Engineering, East China University of Technology, Nanchang 330013, China;
2 School of Software, East China University of Technology, Nanchang 330013, China)

【Abstract】 In recent years, chronic diseases have developed rapidly and are becoming more and more harmful. Among them, diabetes is a chronic disease with genetic characteristics, which is extremely harmful. Therefore, it has very important research significance for diabetes prediction. With the continuous development of artificial intelligence technology, more and more machine learning and deep learning methods have been used to predict disease occurrence. Therefore, this paper proposes a diabetes prediction method based on ensemble learning. In this method, the ensemble learning method of stacking is adopted, three machine learning methods including Support Vector Machine, Random Forest and Artificial Neural Network are used as the primary learner, and logistic regression is used as the secondary learner to establish the diabetes prediction model. In this paper, pima Indian diabetes data set in UCI is taken as experimental data. Through experimental analysis, the model fusion method proposed in this paper can achieve better prediction effect.

【Key words】 diabetes prediction; Support Vector Machine; Artificial Neural Network; Random Forest; ensemble learning

0 引 言

研究可知，糖尿病是一种危害性非常大的慢性疾病，也是一种具有的遗传特性的代谢性疾病，典型特征为多尿、多饮、多食和体重减轻。近些年来，全球糖尿病的患病人数增加较快，根据国际糖尿病联盟调查，截至到 2017 年，全球糖尿病患者已超过 4 亿，并且在报告中指出预计到 2045 年，糖尿病的患病总人数达到 6 亿多[1]。实际上糖尿病不仅是世界性问题，中国糖尿病患者的规模是全球最大的，占全球患病总人数的四分之一还多，患病人数已到 1.14 亿[2]。

由于慢性病难于治愈特点，在慢性病的治疗上需要大量的医疗投入。目前，中国每年花费超三千亿的医疗支出在糖尿病中[3]。糖尿病还具有不容易被发现的特点，据资料了解，糖尿病患者，仅仅有一半知道自己得病，剩下的一半还以为自己是正常人，因此多数患者均在不知道自己患病情况下，未能及时接受治疗，导致病情发展更加迅速[4]。

综前论述可知，目前糖尿病预测已然成为迫切需要解决的研究课题。因此，建立有效的糖尿病预测模型在糖尿病防治当中非常重要。近些年，机器学习方法在糖尿病预测领域得到广泛应用，但是大多方法基于单一学习方法建模，泛化能力较差。本文即针对单一方法准确性不高，泛化能力不强的问题，提出一种基于集成学习思想的糖尿病预测方法。该方法采用 Stacking 的设计思路，使用支持向量机、随机森林、人工神经网络等分类方法作为初级分类器，使用逻辑回归作为次级分类器建立糖尿病预测模型。

1 研究现状

在早期的糖尿病预测模型研究中，研究人员主

基金项目：江西省放射性地学大数据技术工程实验室开放项目(JELRGBDT201802)。

作者简介：章 权(1994-)，男，硕士研究生，主要研究方向：机器学习、进化算法；周梁琦(1993-)，男，硕士研究生，主要研究方向：数据挖掘、大数据处理；邹 琪(1996-)，男，硕士研究生，主要研究方向：数据挖掘、机器学习；喻新民(1995-)，男，本科生，主要研究方向：机器学习。

收稿日期：2019-11-08

要使用回归模型来建立预测模型,近期以来,随着机器学习方法和深度学习方法的快速发展,机器学习和深度学习中的模型和方法逐渐被应用于糖尿病预测模型的研究和设计。

通常在糖尿病的预测中,以多元回归模型或 Cox 回归这两种方法为主。基于多元回归的预测模型研究中,Mehlsen 等人[5]利用该方法建立模型用于对糖尿病中的视网膜病变进行预测。基于 Cox 回归预测模型的研究中,如张红艳等人[6]采用 Cox 回归模型建立基于中国农村人群的非侵袭性 2 型糖尿病风险预测模型,该模型的灵敏度为 65.96%,特异度为66.47%。周先锋等人[7]则用 Cox 回归模型研究不同程度的 C 反应蛋白与患有糖尿病之间的联系。但是,多元回归模型有着准确度较差、精度不高的缺点,而 Cox 回归模型对数据要求高且成本大。

Anuja 等人[8]提出了一种基于 SVM 的糖尿病预测模型,该模型是使用皮马印第安人糖尿病数据集作为模型的验证码数据集,模型的准确性达到了 78%;Aiswarya 等人[9]使用 J48 决策树和朴素贝叶斯作为分类器,对糖尿病的诊断进行分类。该系统使用了皮马印第安人糖尿病数据集,J48 决策树和朴素贝叶斯的分类结果分别为 74.8%,79.5%;Ahmad 等人[10]在研究工作中,通过实验发现怀孕次数这一属性与是否患糖尿病的可能性之间关联性较弱,故在去掉这一属性之后,剪枝 J48 树进一步提高了准确性,达到了 89.7%。江燕等人[11]将主成分分析和最小乘向量机结合,采用径向基核函数,对于血糖水平预测可以达到 94.82%准确率。

李飞等人[12]考虑到高血糖患者具有皮肤组织荧光特性,在此基础上利用神经网络建立糖尿病无创评估模型,整体准确率达到 74.9%;Ramesh 等人[13]使用递归神经网络(RNN)预测 2 种糖尿病。Ashiquzzaman 等人[14]使用深度神经网络(DNN),该神经网络由多层感知器(MLP)、广义回归神经网络(GRNN)和径向基函数(RBF)组成。该方法的评估基于 Pima 印度数据集,正确率为 88.41%。深度学习方法对非线性数据友好,具备很强的记忆功能、自学本领等,但是解释性很差,并且模型建立需要的数据量较大,这样限制其在糖尿病的一些小的数据集上应用。

2 模型及方法

2.1 数据分析

本文所用的数据来自 UCI 数据集里的皮马印第安人糖尿病数据集,该数据集最初来自美国国家糖尿病/消化/肾脏疾病研究所。数据集的目标是基于数据集中包含的某些预测变量来预测患者是否患有糖尿病。

该数据集具有一些明显的约束条件,数据集中的患者都是 Pima 印第安至少 21 岁的女性。数据集由多个医学预测变量和一个目标变量 *Outcome* 组成,当 *Outcome* 的值为 1 时代表患有糖尿病,当 *Outcome* 的值为 0 时表示未患糖尿病。预测变量包括患者的怀孕次数、BMI、胰岛素水平、年龄等,具体见表 1。

表 1 Pima 印第安数据集中预测变量

Tab. 1 Predictive variables in Pima Indian dataset

特征名	特征描述
Pregnancies	怀孕次数
Glucose	血糖值
BloodPressure	舒张压
SkinThickness	三头肌皮褶厚薄程度
Insulin	胰岛素含量
BMI	体重指数
DiabetesPedigreeFunction	糖尿病谱系功能,简称为遗传指数
Age	年龄

2.2 基于 SVM 的预测模型

SVM 是一种基于统计学习理论的机器学习方法,并不是采用经验风险最小的方式,而是采用结构化风险最小,因此有着较好的泛化能力,并在解决小样本、非线性及高维模式识别中表现出较大优势。本文使用的数据集样本量较小,因此在研究中采用 SVM 作为初级学习器符合数据分析的结论。

支持向量机的主要思想是找到一个合适的分类函数对未知样本进行预测。这个分类函数通过核函数以及惩罚因子来确定,而相关参数的确定对模型的准确度有着很大的影响。经典的支持向量机方法是一种二分类的算法,其最基本的思想是基于训练集在样本空间中找到一个划分超平面,将不同类别的样本分开,如图 1 所示。对于非线性问题,使用直线已经不能很好地将样本进行分类,模型使用核函数把样本数据从低维度不可分的空间映射到高维度可分空间,并以此找出分类平面。

SVM 方法关键就是选择一个合适的核函数,通过对不同核函数进行分析,在本文 SVM 模型的构建中,选择的核函数是 linear 核函数。

2.3 基于随机森林的预测模型

本文在初级学习器中选择随机森林回归模型,该模型能够处理高维度数据,并且不用做特征选择,

另外，在训练完成后，还能给出比较重要的那些特征。该模型的泛化能力比较强，在训练过程中能够检测特征间的互相影响。

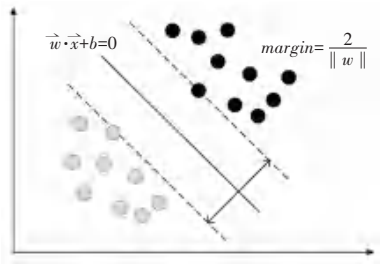

图 1　SVM 最优分类图

Fig. 1　SVM optimal classification diagram

随机森林中最重要参数就是决策树的个数，随机森林中的棵树太多或者太少都会影响模型的结果。因此，本文采用了网格搜索的方法，将森林规模确定为 2 030 棵树时，模型效果最好。

2.4　基于人工神经网络的预测模型

本文使用的人工神经网络（ANN），就是多层感知机。ANN的每层神经元与下一层神经元全互连，神经元之间不存在同层连接，这种结构通常被称为多层前馈神经网络模型。训练时要经过前向传播和误差反馈传播两个过程。

人工神经网络具有较好的自适应学习能力，现如今已广泛应用于模式识别、非线性等课题研究中。因此，本文通过向 ANN 输入 Pima 印第安数据集中预测变量数据，再加以训练，并由 ANN 输出最终结果。这些结果用于糖尿病的预测。

2.5　模型融合

模型融合主要分为 Stacking、Blending 和 Voting 等方法，是一种通过增加算法的多样性来减少泛化误差，从而提高模型准确率的有效、且实用的技术。模型融合有 2 个基本要素，分别是：单一模型之间的相关性要尽可能小，单一模型之间的性能表现相差不大。本文拟采用 Stacking 方法设计预测模型，Stacking 的基本思想是使用大量基分类器，再使用另一种顶层分类器来融合基分类器的预测，旨在降低泛化误差。本文中模型融合的整体流程如图 2 所示。

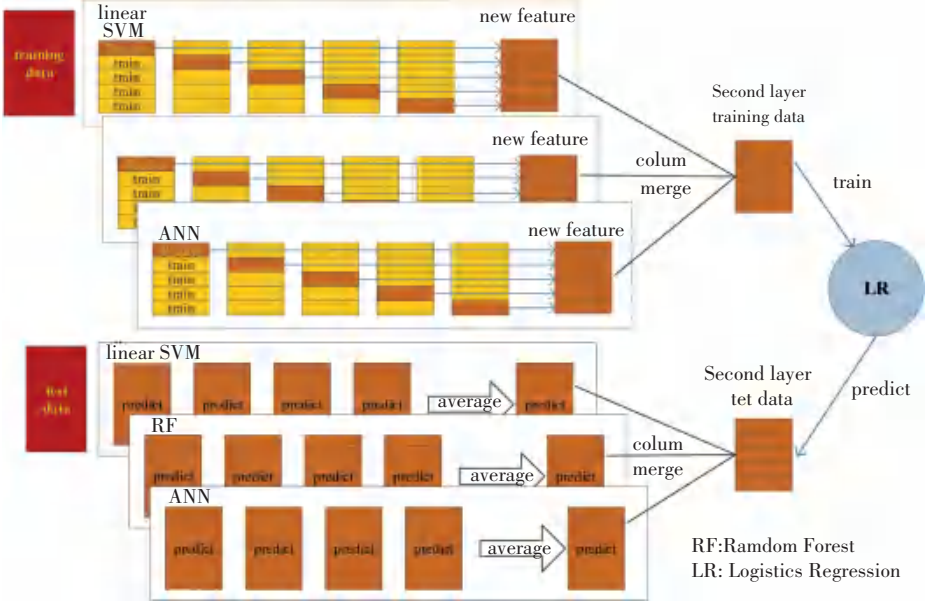

图 2　模型融合流程图

Fig. 2　Model fusion flow chart

3　模型评估

本文的模型均采用 Scikit-Learn 机器学习包来实现，实验环境为 Intel(R) Core(TM) i7-8700K CPU @ 3.70 GHz、16 GB RAM 设备。同时，采用了 5 折交叉验证方法依次对 ANN、随机森林、支持向量机三种糖尿病预测模型进行训练，经由训练得到数据将形成新的数据集，并基于 Stacking 的思想对逻辑回归模型进行训练，对训练后的模型在测试集上进行预测，不同模型的预测结果和由其它相关工作获得的模型结果对比见表 2，而由本次研究得到的不同预测模型评估分析结果，见表 3。

本文采用的评价标准除了准确率（*accuracy*）外，还有精确率（*precision*）、召回率（*recall*）和 F_1。其中，准确率（*accuracy*）是指被分类正确的样本数

占总样本数的比值;*precision* 表示预测为正的样本中,实际为正的所占的比例;*recall* 表示实际为正的样本中,被预测为正的所占的比例;F_1值是精准率和召回率的加权调和平均值。综上各评价指标的数学公式可分别表示为:

$$accuracy = \frac{TN + TP}{TN + TP + FN + FP}, \quad (1)$$

$$precision = \frac{TP}{FP + TP}, \quad (2)$$

$$recall = \frac{TP}{FN + TP}, \quad (3)$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}. \quad (4)$$

其中, P 表示阳性样本总数;TP 表示预测正确的阳性样本数;TN 表示预测错误的阳性样本数。

表 2 不同模型及其它预测方法准确率对比表

Tab. 2 Comparison of accuracy of different models and other prediction methods

模型/方法	准确率/%
SVM	85.7
随机森林	88.3
ANN	86.4
文献[10]	89.7
文献[14]	88.4
本文	92.2

表 3 不同模型及其它预测方法评价指标对比表

Tab. 3 Comparison of evaluation indexes of different models and other prediction methods %

模型	*accuracy*	*precision*	*recall*	F_1
SVM	85.7	83	83	83
随机森林	88.3	86	89	87
ANN	86.4	82	83	81
融合模型	92.2	88	89	88

4 结束语

本文提出了一种基于 Stacking 的糖尿病预测模型,以 SVM、人工神经网络和随机森林构建的模型为基础,采用了逻辑回归分类器对以上三种模型训练结果构建的新数据集进行了二次训练,得到的预测模型在验证集上正确率高于单个分类器,取得了较好的效果,在准确率和召回率上得到了较大提升,在验证集上的准确率达到了 92.2%,而且整个预测模型也表现出较强的泛化能力。

参考文献

[1] CHO N H, SHAW J E, KARURANGA S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045[J]. Diabetes research and clinical practice, 2018, 138: 271.

[2]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]. 中国实用内科杂志,2018,38(4):292.

[3] 白碧玉,于琦,苏闫兵,等. 中国糖尿病研究论文合作分析[J]. 中国药物与临床,2017,17(11):1619.

[4] JUNG H S. Prediction of diabetes using serum c-peptide[J]. Endocrinology and Metabolism, 2016, 31(2): 275.

[5] MEHLSEN J, ERLANDSEN M, POULSEN P L, et al. Individualized optimization of the screening interval for diabetic retinopathy: A new model[J]. Acta ophthalmologica, 2012, 90(2): 109.

[6] 张红艳, 石文惠, 张明, 等. 基于中国农村人群的非侵袭性 2 型糖尿病风险预测模型的建立[J]. 中华预防医学杂志, 2016, 50(5):397.

[7] 周先锋, 阮晓楠, 于思雨, 等. 血清 C 反应蛋白与糖尿病发病风险的关系[J]. 中华糖尿病杂志, 2017(2):106.

[8] ANUJA K A, CHITRA R. Classification of diabetes disease using Support Vector Machine[J]. International Journal of Engineering Research and Applications, 2013, 3(2): 1797.

[9] AISWARYA I, JEYALATHA S, SUMBALY R. Diagnosis of diabetes using classification mining techniques[J]. International Journal of Data Mining & Knowledge Management Process, 2015, 5(1): 1.

[10]AHMAD A, MUSTAPHA A, ZAHADI E D, et al. Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus[M]//SNASEL V, PLATOS J, EI-QAWASMEH E. International Conference on Digital Information Processing and Communications. Berlin/Heidelberg:Springer-Verlag , 2011: 537.

[11]江燕,帅仁俊,张姝,等. 基于 KPCA-LSSVM 的健康档案空腹血糖水平预测研究[J]. 计算机工程与应用,2018,54(13):241.

[12]李飞,王贻坤,朱灵,等. 基于神经网络模式识别的糖尿病无创风险评估方法研究[J]. 光谱学与光谱分析,2014,34(5):1327.

[13]RAMESH S, BALAJI H, IYENGAR N C S N, et al. Optimal predictive analytics of pima diabetics using deep learning[J]. International Journal of Database Theory and Application, 2017, 10(9): 47.

[14] ASHIQUZZAMAN A, TUSHAR A K, ISLAM M R, et al. Reduction of overfitting in diabetes prediction using deep learning neural network[M]//KIM K, KIM H, BAEK N. IT Convergence and Security 2017. Lecture Notes in Electrical Engineering. Singapore:Springer, 2017,449: 35.