Operations Research and Fuzziology 运筹与模糊学, 2024, 14(3), 384-390
Published Online June 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.143276

Hans汉斯

基于PCA和BP神经网络的糖尿病预测

胡　勇

上海工程技术大学管理学院，上海

收稿日期：2024年1月20日；录用日期：2024年3月19日；发布日期：2024年6月21日

摘　要

糖尿病的发病率正在逐年上升且向低龄化发展，对我国乃至世界的健康安全造成了严重的影响，因此有必要对糖尿病的预测进行研究。本文对皮马印第安人糖尿病数据集进行分类，首先使用主成分分析法将数据从8维降到了3维，接着使用这3维数据建立BP神经网络模型。将基于PCA和BP神经网络的模型与单纯的BP神经网络模型进行对比。结果表明，基于PCA和BP神经网络的模型在精准率、召回率、F值、查准率和Matthews相关系数MCC 5项性能指标上均明显优于BP神经网络，可以作为糖尿病预测的一种有效方法。

关键词

主成分分析，BP神经网络，糖尿病预测

Diabetes Prediction Based on PCA and BP Neural Network

Yong Hu

School of Management, Shanghai University of Engineering Science, Shanghai

Received: Jan. 20th, 2024; accepted: Mar. 19th, 2024; published: Jun. 21st, 2024

Abstract

The incidence of diabetes mellitus is increasing year by year and towards lower age, which has a serious impact on the health and safety of our country and the world, so there is a need to study the prediction of diabetes mellitus. In this paper, the Pima Indians diabetes dataset was categorized by first reducing the data from 8 dimensions to 3 dimensions using Principal Component Analysis (PCA), followed by BP neural network modeling using these 3 dimensions of data. The PCA and BP neural network based model was compared with the BP neural network model alone. The results show that the model based on PCA and BP neural network is significantly better than

文章引用: 胡勇. 基于 PCA 和 BP 神经网络的糖尿病预测[J]. 运筹与模糊学, 2024, 14(3): 384-390.
DOI: 10.12677/orf.2024.143276

BP neural network in five performance indicators: precision, recall, F-value, checking accuracy and Matthews correlation coefficient MCC, and it can be used as an effective method for diabetes prediction.

Keywords

Principal Component Analysis, BP Neural Network, Diabetes Prediction

1. 引言

目前糖尿病已经逐渐发展成为发病率最高的危害人类健康的疾病之一，并且多数患者意识不到糖尿病前期的症状，最终错过了最佳治疗时期。因此，对糖尿病的预测以便及时进行治疗具有十分重要的意义。现有的糖尿病患者识别较多地使用支持向量机、Logistic 回归分类、BP 神经网络、K 近邻算法和决策树算法。邓琳[1]等学者使用支持向量机建立了妊娠期糖尿病的预测模型；李飞[2]等学者使用神经网络模型对检测对象患有糖尿病的可能性进行了评估；Permana B A C [3]等学者使用 C4.5 决策树对糖尿病患者进行了预测；Haohui [4]等学者将 Logistic 回归、支持向量机、Nayeve Bayes、决策树、Random Forest、XGBoost 和人工神经网络 8 种算法作为糖尿病预测模型进行了对比，得出 Random Forest 模型优于其他模型。由以上文献梳理可以看出，大部分学者都是使用其中一种机器学习算法作为糖尿病预测模型，忽略了糖尿病的各特征之间信息存在冗余，各个属性之间存在较强的相关性，直接输入到机器学习算法中，可能会导致模型过拟合。使用主成分分析方法(PCA)可以进行数据精简和降维处理，选取不相关、包含大部分信息的主成分[5]。BP 神经网络作为一种分类和预测方法，通过神经网络建立特征属性与识别对象间的联系来进行分类预测，该方法效果较好，被广泛应用到各个领域。因此，本研究使用 PCA 算法将数据进行降维，然后将降维后的主成分使用神经网络模型来进行糖尿病患者预测。

2. 算法理论

2.1. BP 神经网络

人工神经网络是对生物神经机制研究基础上产生的智能仿生模型。处理单元，或称之为神经元，是神经网络的最基本的组成部分。一个神经网络系统中有许多处理单元，每个处理单元的具体操作都是从其相邻的其他单元中接受输入，然后产生输出送到与其相邻的单元中去。神经网络的处理单元可以分为三种类型：输入单元、输出单元和隐含单元。输入单元是从外界环境接受信息，输出单元则给出神经网络系统对外界环境的作用。隐含单元则处于神经网络之中，它从网络内部接受输入信息，所产生的输出则只作用于神经网络系统中的其它处理单元。隐含单元在神经网络中起着极为重要的作用，BP 神经网络的示意图如下图 1 所示。

假设给定训练集 $C=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_m,y_m\right)\right\}$，隐层共含有 q 个神经元。记隐层第 h 个神经元接受到的输入为 $\alpha_h=\sum_{i=1}^{d}v_{ih}x_i$，输出层第 j 个神经元接受到的输入为 $\beta_j=\sum_{i=1}^{d}w_{hj}b_h$，$b_h$ 为隐层第 h 个神经元的输出。

对训练样本 $\left(x_k,y_k\right)$，记神经网络的输出为 $\hat{y}_k=\left(\hat{y}_1^k,\hat{y}_2^k,\cdots,\hat{y}_l^k\right)$，即

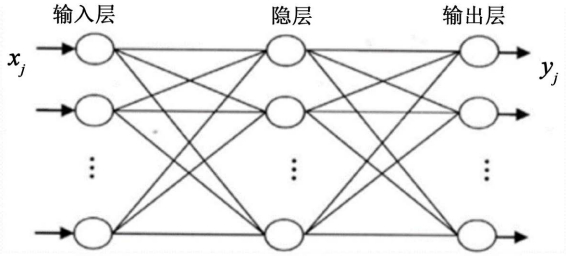

Figure 1. BP neural network and variables
图 1. BP 神经网络及变量

$$\hat{y}_j = f\left(\beta_j - \theta_j\right) \tag{1}$$

均方误差为：

$$E_k = \frac{1}{2}\sum_{j=1}^{l}\left(y_j^k - y_j^k\right)^2 \tag{2}$$

BP 算法基于梯度下降策略，以目标的负梯度方向对参数进行调整，对式(2)的误差给定学习率 ε，有

$$\Delta w_{hj} = -\varepsilon\frac{\partial E_k}{\partial w_{hj}} \tag{3}$$

通过推导，得出 BP 算法的更新公式为：

$$\begin{cases}\Delta w_{hj} = \varepsilon g_i b_h \\ \Delta\theta_j = -\varepsilon g_i \\ \Delta v_{ik} = -\varepsilon e_h x_i \\ \Delta\gamma_h = -\varepsilon e_h\end{cases} \tag{4}$$

其中，$g_i = \hat{y}_j^k\left(1-\hat{y}_j^k\right)\left(y_j^k - \hat{y}_j^k\right)$，$b_h = \dfrac{\partial\hat{y}_j^k}{\partial\beta_j}$，$e_h = b_h\left(1-b_h\right)\sum_{j=1}^{l} w_{hj}g_i$，学习率 ε 控制着算法每一轮迭代中的更新步长，太大容易震荡，太小收敛速度过慢。

BP 神经网络算法步骤为：1) 初始化网络及学习参数，如设置网络初始权矩阵，学习因子 ε 等[2]；2) 提供训练模式，训练网络，直到满足学习要求；3) 前向传播过程：对给定训练模式输入，计算网络的输出模式，并与期望模式比较，若有误差，则执行步骤 4；否则，返回步骤 2；4) 反向传播过程：计算同一层单元的误差，修正权值和阈值，返回步骤 2。

2.2. 主成分分析

主成分分析是一种常用的降维算法，在损失很少信息的前提下把多个指标转化为几个综合指标，通常把转化生成的综合指标称为主成分，每个主成分是原始变量的线性组合，且各个主成分之间互不相关，使得主成分比原始变量具有某些更优越的性能。

设样本有 p 个指标，分别用 $X_1, X_2, \cdots, X_p$ 表示，这 p 个指标构成随机向量 $X = \left(X_1, X_2, \cdots, X_p\right)'$。设随机向量 X 的均值为 μ，协方差矩阵为Σ。对 X 进行线性变换，形成新的综合变量 Y，Y 满足下式：

$$\begin{cases}Y_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\ Y_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\ \vdots \\ Y_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p\end{cases} \tag{5}$$

其中：1) $u_i'u_i=1(i=1,2,\cdots,p)$ 2) Y_i 与 Y_j 相互无关$(i\neq j;i,j=1,2,3,\cdots,p)$ 3) Y_1 是 $X_1,X_2,\cdots,X_p$ 满足原则的所有线性组合中方差最大者，Y_2 是与 Y_1 不相关的 $X_1,X_2,\cdots,X_p$ 的所有线性组合中方差最大者，$Y_3,Y_4,\cdots,Y_p$ 都是 $X_1,X_2,\cdots,X_p$ 的所有线性组合中方差最大者。

满足以上三条原则的综合变量 $Y_1,Y_2,\cdots,Y_p$ 分别称为原始变量的第一、第二…第 p 个主成分，并且各综合变量在总方差中所占的比重依次递减。

2.3. 评价指标

二分类评价指标是基于混淆矩阵(如表 1 所示)得出的，常用的评价指标有精准率、召回率、F 值、查准率和 Matthews 相关系数 MCC。

Matthews 相关系数 MCC：预测分类与观测分类之间的一致性度量指标，取值越大，一致性则越好，分类器效果越好。

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \tag{6}$$

Table 1. Confusion matrix
表 1. 混淆矩阵

真实值 \ 预测值	Yes	No	sum
Yes	TP	FN	P
No	FP	TN	N

精准率：$\mathrm{AR}=(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$

召回率：$\mathrm{recall}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$

查准率：$\mathrm{precision}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$

F 值：$F=2*\mathrm{TP}/(2*\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$

3. 实证分析

3.1. 数据来源

数据来自 UCI 数据集里的皮马印第安人糖尿病数据集，该数据集含有 600 样本，8 个属性变量和 1 个标签变量。其中，标签值为 1 代表患糖尿病，标签值为 0 表示未患糖尿病。这八个属性分别是怀孕次数、血糖值、血压值、皮脂厚度、胰岛素量、BMI、糖尿病遗传性和年龄。

3.2. 数据处理

本实验在 python3 环境下运行，首先利用 StandardScaler 库对数据进行了标准化处理，以消除量纲对模型的影响。然后对属性变量进行 PCA 分析，得出各成分的方差及方差比例见下表 2，并将方差比例进行可视化如图 2 所示。

Table 2. Variance and proportion of variance of each component
表 2. 各成分方差及方差比例

	怀孕次数	血糖值	血压值	皮脂厚度	胰岛素量	BMI	糖尿病遗传性	年龄
方差	2.157	1.740	1.038	0.879	0.752	0.619	0.424	0.404
方差比例	0.269	0.217	0.130	0.110	0.094	0.077	0.053	0.050

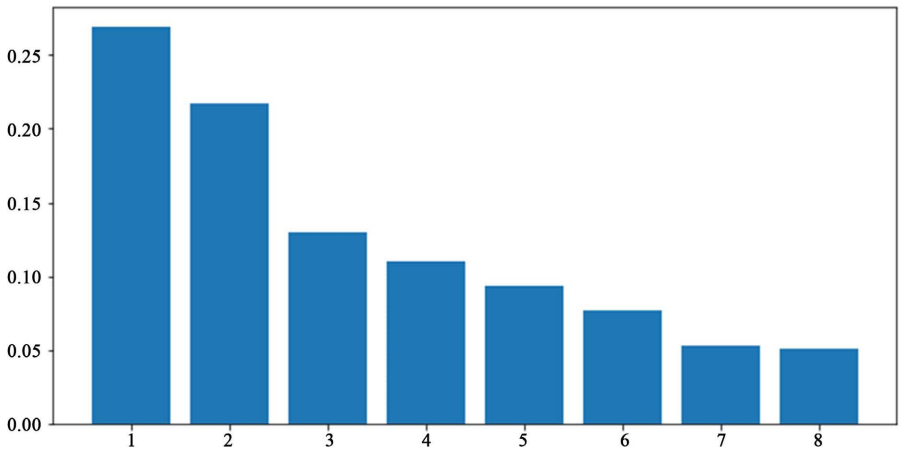

Figure 2. Proportion of variance by attribute
图 2. 各属性方差比例

由图 2 可知，第一主成分、第二主成分和第三主成分提供了大部分信息，因此利用 sklearn 库中的 PCA 将数据降为三维。将降维后的三维数据进行二维可视化如下图 3 所示。

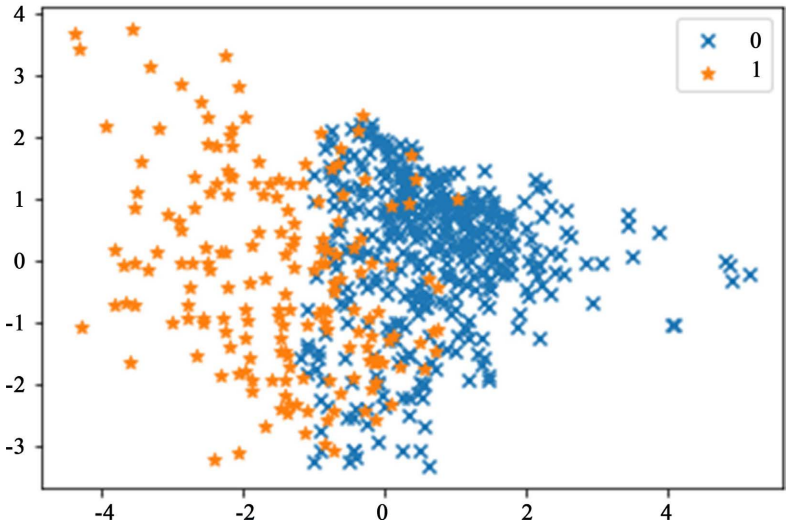

Figure 3. Data visualization after dimensionality reduction
图 3. 降维后的数据可视化

3.3. 模型建立

将数据按照 6:4 的比例划分数据，建立训练集和测试集。神经网络建模利用 sklearn 中的 MLPClassifier 函数，其中参数设置为 solver 为 lbfgs，lbfgs 为 quasi-Newton 方法的优化器，可以更快的收敛并且表现更好；设置 alpha 为 e^{-5}，alpha 为 L2 惩罚参数；hidden_layer_sizes 设置为(5, 2)，表示有两层隐藏层，第一层隐藏层有 5 个神经元，第二层隐藏层有 2 个神经元。模型结果得到训练集准确率为 0.9，测试集准确率为 0.87。

3.4. 模型评估

二分类模型的评价指标有精准率、召回率、F 值、查准率、Matthews 相关系数 MCC 和 ROC (Re-

ceiver Operating Characteristic)曲线，将 PCA + BP 神经网络模型与 BP 神经网络进行对比，得出混淆矩阵如下图 4 和图 5 所示，分类结果比较如下表 3 所示。

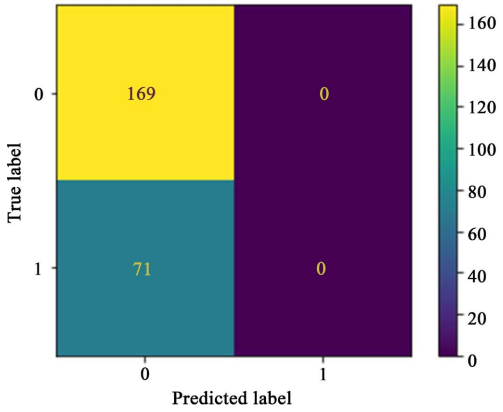

Figure 4. BP neural network confusion matrix
图 4. BP 神经网络混淆矩阵

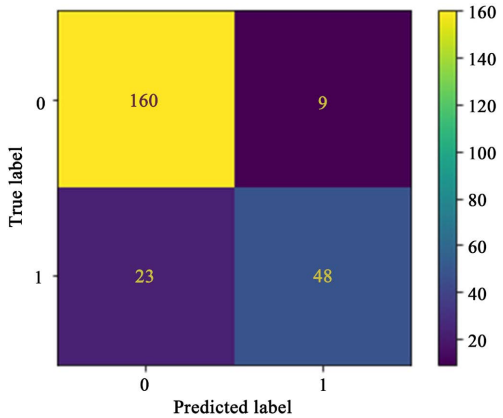

Figure 5. PCA + BP neural network confusion matrix
图 5. PCA + BP 神经网络混淆矩阵

Table 3. Comparison of test set classification results
表 3. 测试集分类结果比较

分类模型	标签	混淆矩阵		精确率	召回率	F 值	查准率	MCC
PCA + BP	1	48	23	0.87	0.68	0.75	0.84	0.67
	0	9	160					
BP	1	0	71	0.71	0	0	0	0
	0	0	169					

ROC 的全称为“受试者工作特征”曲线，是二分类模型优劣的一种评价指标，表示正例排在负例前面的概率，曲线的横坐标表示假阳率，纵坐标表示真阳率。AUC 表示 ROC 曲线下的面积，主要用于衡量模型的泛化能力，AUC 值越大(ROC 曲线越接近左上角)，分类效果越好。文中两模型的 ROC 曲线和 AUC 值如下图 6 所示。

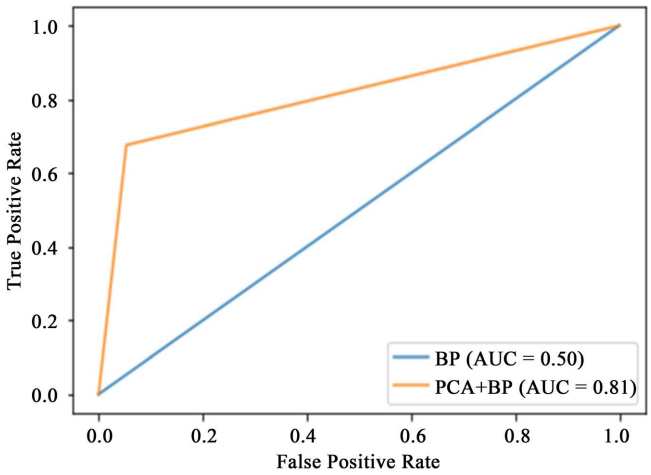

Figure 6. ROC curves and AUC values for BP and PCA + BP models
图 6. BP 和 PCA + BP 模型的 ROC 曲线和 AUC 值

由以上分类评价指标可得，PCA + BP 比 BP 模型的精确率、召回率、F 值、查准率和 MCC 的值都大，且 ROC 曲线下的面积 AUC 也比 BP 大 0.31。因此，从各个指标可以看出，对于糖尿病患者的预测，PCA + BP 神经网络模型的效果比单纯的 BP 神经网络分类模型效果好，在数据集的预测精度表现更佳。

4. 结束语

中国作为糖尿病患病人数最多的国家，利用机器学习算法提高糖尿病预测模型性能及可解释性，对于辅助医生的诊断工作具有重要的现实意义[6]。本文使用 PCA 对皮马印第安人糖尿病数据集进行了降维，良好的处理了数据集中复杂的非线性问题。利用降维后的 3 个主成分使用 BP 神经网络进行分类，预测准确率达到 0.86，比单纯用 BP 神经网络准确率高了 0.16。因此，PCA + BP 神经可以为糖尿病患者预测提供一种有效的方法。有助于及时地对糖尿病作出提早预防和风险控制[7]，进而降低医疗成本，减少误诊率。

参考文献

[1] 郑琳, 倪世伟. 基于支持向量机的妊娠期糖尿病预测模型的构建[J]. 安徽预防医学杂志, 2019, 25(6): 465-468.

[2] 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): 1327-1331.

[3] Permana, B.A.C., Ahmad, R., Bahtiar, H., *et al.* (2021) Classification of Diabetes Disease Using Decision Tree Algorithm (C4.5). *Journal of Physics*: *Conference Series*, **1869**, Article 012082. https://doi.org/10.1088/1742-6596/1869/1/012082

[4] Lu, H., Uddin, S., Hajati, F., *et al.* (2021) A Patient Network-Based Machine Learning Model for Disease Prediction: The Case of Type 2 Diabetes Mellitus. *Applied Intelligence*, **52**, 2411-2422. https://doi.org/10.1007/s10489-021-02533-w

[5] 张志恒, 李超. 基于 PCA-BP 神经网络的审计风险识别研究[J]. 重庆理工大学学报(自然科学), 2021, 35(5): 253-261.

[6] 王鑫, 廖彬, 李敏, 等. 融合 LightGBM 与 SHAP 的糖尿病预测及其特征分析方法[J]. 小型微型计算机系统, 2022, 43(9): 9.

[7] 刘文博, 梁盛楠, 秦喜文, 等. 基于迭代随机森林算法的糖尿病预测[J]. 长春工业大学学报, 2019, 40(6): 604-611.