计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):182−188 [doi: 10.15888/j.cnki.csa.007363]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合批归一化的多层感知机糖尿病预测诊断模型①

胡清礼[1], 胡建强[1], 余小燕[2]

[1](厦门理工学院 计算机与信息工程学院, 厦门 361024)
[2](上海工程技术大学 电子电气工程学院, 上海 201620)
通讯作者: 胡清礼, E-mail: huqingli2014@outlook.com

摘　要: 糖尿病的早期发现, 对成功控制、预防并发症, 降低患病率具有重要意义. 现有基于机器学习建立的糖尿病诊断模型, 由于泛化能力不足而导致精度较低. 为此, 本文提出结合批归一化的多层感知机模型, 保证模型中数据分布的一致性. 基于 PIMA 数据集进行训练评估, 实验结果表明该模型用于糖尿病早期识别泛化能力好、收敛速度快且有较高的准确率.

关键词: 糖尿病; 机器学习; 批归一化; 泛化能力

引用格式: 胡清礼,胡建强,余小燕.结合批归一化的多层感知机糖尿病预测诊断模型.计算机系统应用,2020,29(5):182–188. http://www.c-s-a.org.cn/1003-3254/7363.html

Multi-Layer Perceptron Diabetes Prediction Model Combined with Batch Normalization

HU Qing-Li[1], HU Jian-Qiang[1], YU Xiao-Yan[2]

[1](School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China)
[2](School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: The early detection of diabetes is of great significance for successful control of diabetes, prevention of complications, and reduction of prevalence. Existing diabetes diagnosis models based on machine learning have weak precision due to insufficient generalization ability. Therefore, this study proposes a multi-layer perceptron model combined with batch normalization to ensure the consistency of data distribution in the model. The proposed model is based on the PIMA training set for training evaluation. The experimental results show that the model has sound generalization ability in early recognition of diabetes, fast convergence, and high accuracy.

Key words: diabetes; machine learning; batch normalization; generalization ability

1　引言

糖尿病是一个日益严重的全球问题, 对患者的日常生活和身体健康影响巨大, 甚至威胁患者的生命. 近年来, 糖尿病的患病率愈来愈高, 根据国际糖尿病联盟最新数据指出, 2017 年全球约 4.25 亿成人患糖尿病, 若按照当前趋势持续下去, 预计到 2045 年, 糖尿病患病人数将达到 6.29 亿[1]. 早期发现糖尿病, 尽早进行强化治疗, 可以使接近 50% 的患者病情缓解, 早期发现便可通过运动饮食将血糖控制在正常范围.

糖尿病诱发的病因病理复杂, 医学界尚未建立统一的预测模型, 可以完全科学的早期发现、诊断糖尿病. 国内外许多研究者围绕糖尿病的早期预测展开许多有益的工作. 具体包括:

文献[2]采用朴素贝叶斯、支向量机和决策树等

① 基金项目: 福建省自然科学基金 (2019J01856); 赛尔网络下一代互联网创新项目 (NGII20160708)
Foundation item: Natural Science Foundation of Fujian Province (2019J01856); CERNET Innovation Program for Next Generation of Internet (NGII20160708)
收稿时间: 2019-09-20; 修改时间: 2019-10-15; 采用时间: 2019-10-22; csa 在线出版时间: 2020-05-07

算法, 在 PIMA 测试集上获得 76.3% 的准确率. 文献[3]提出基于支持向量机的糖尿病预测模型, 采用径向基作为核函数, 通过三折交叉验证和网格搜索确定最优惩罚参数和核参数, 在 PIMA 测试集上获得 78.39% 的准确率. 文献[4]提出基于奇异值 (Singular Value Decomposition, SVD) 和 Boosting 级联学习的糖尿病风险分级模型, 提高预测的智能性和准确率. 文献[5]提出结合遗传算法和朴素贝叶斯算法的糖尿病预测模型, 即利用遗传算法进行特征选择, 朴素贝叶斯进行分类决策, 在 PIMA 数据集上获得了 78.7% 的准确率. 上述模型由于精度较低, 很难满足实际的需求.

文献[6]提出概率人工神经网络方法 (PNN) 诊断 2 型糖尿病, 即一个隐藏层, 神经元数量与训练数据集数量相等, 该方法的预测精度略有提高, 但泛化能力低, 影响模型精度提升. 文献[7]使用多层感知机 (MLP) 模型, 由于层数或神经元的个数过多反而更容易产生过拟合现象, 泛化能力不足, 在 PIMA 数据集上测试精度不高. 过拟合使得模型对未知数据拟合效果差, 导致泛化能力下降, 影响整个模型的精度. 文献[8]采用 Dropout 方法解决多层感知机过拟合问题, 该方法根据设置的失活率使神经元随机失活, 减少神经元参加训练, 以此来降低过拟合问题, 该方法在一定程度上抑制了过拟合, 提高了模型的精度. 为了保证模型具有较高的精度, 需要能够学习到更多的特征, 势必会增加网络的层数和神经元个数, 导致整个网络训练速度变慢, 即收敛速度慢.

多层感知机的隐藏层能够分析并学习数据的特征, 通常其层数越多或层上神经元个数越多, 该网络表征数据特征的能力就越强, 但事实上并非如此, 其层数或神经元的个数过多反而更容易产生过拟合现象, 使得模型对未知数据拟合效果非常差. 因此, 在保证能够充分学习到数据特征的前提下, 解决多层感知机的过拟合现象是一个重要问题. 此外, 随着多层感知机的层数增加, 每一层的参数更新会导致数据分布在输入和输出时发生变化, 经过多个隐藏层后, 最后面的隐藏层的输入数据分布会发生剧烈的变化, 这就使得靠后的隐藏层需要不断去重新适应靠前的隐藏层的参数更新, 这会严重影响模型的收敛速度, 甚至由于数据分布不一致从而导致模型泛化能力不足.

针对上述问题, 本文将批归一化算法与多层感知机结合, 构建新的多层感知机深度神经网络模型, 该模型在每个隐藏层后面增加一个批归一化层, 该批归一化层对前一层的输出进行归一化处理并学习其分布特性, 保证模型中数据分布的一致性. 经过实验验证, 结合批归一化算法的多层感知机深度神经网络模型不仅具有较好的泛化能力, 而且收敛速度快, 同时其准确率较上述方法有明显提高.

2 结合批归一化算法的多层感知机

为解决多层感知机的过拟合问题和数据分布不一致带来的负面影响, 将批归一化算法与多层感知机结合, 解决多层感知机在训练过程中的过拟合问题, 同时减小数据分布不一致带来的不利影响, 提高模型在训练过程中的收敛速度. 批归一化算法最初由 Ioffe 等[9]提出, 该算法能够对输入数据的分布状况进行学习, 保证数据分布的一致性, 减小数据分布改变对训练的影响, 加快训练速度[10]. 此外, 该算法能够在一定程度上消除使用 Dropout 的情况, 即降低过拟合, 提高模型的泛化能力[11].

结合批归一化算法的多层感知机模型如图 1 所示. 对数据集进行预处理, 使模型学习效果更好. 采用 z-score 方法将数据转换为均值为 0, 方差为 1 的正态分布; 然后将预处理后的数据送至网络的输入层, 依次通过 3 个隐藏层, 对数据的特征进行提取, 每个隐藏层后面都有一个批归一化层, 对每个隐藏层输出的分布进行学习和调整, 最后在输出层进行分类决策得出结果.

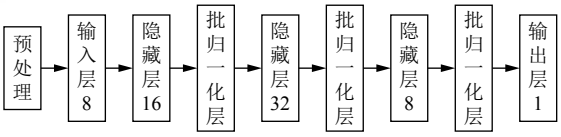


图 1 结合批归一化算法的深度神经网络模型

批归一化算法单独地对每一个神经元的特征进行批归一化处理, 并且每次只在一个训练用的 mini-batch 上计算均值和方差, 对数据归一化的处理就是让数据的均值为 0, 方差为 1, 计算公式如式 (1) 所示.

$$\hat{x}^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\sqrt{Var\left[x^{(k)}\right]}} \tag{1}$$

其中, $x = \left(x^{(1)}, \cdots, x^{(d)}\right)$是输入数据, $\hat{x}^{(k)}$是对应的$x^{(k)}$批

归一化处理后的数据. 输出在经过归一化处理后会改变学习到的特征分布, 批归一化算法引入的可学习的重构参数γ、β, 通过重构变换恢复出原始学到的特征, 其表达式如 (2) 所示.

$$y^{(k)}=\gamma^{(k)}\hat{x}^{(k)}+\beta^{(k)} \tag{2}$$

结合批归一化算法的多层感知机网络结构在每个隐藏层后有一个批归一化层, 对隐藏层的输出进行处理, 可将其与隐藏层看成一个结合层. 其结构如图 2 所示.

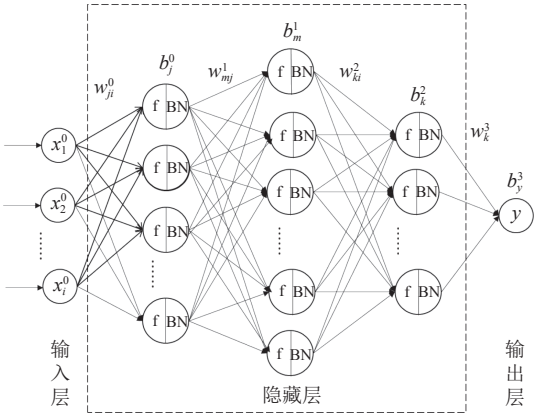


图 2 结合批归一化算法的多层感知机网络结构

假设一个 mini-batch 有m个数据, 第l层经过一个 mini-batch 后, 该层的第i个神经元有m个输出, 则对应的均值、方差以及归一化处理如式 (3)、式 (4) 以及式 (5) 所示.

$$\mu_{Bi}^{l}=\frac{1}{m}\sum_{k=1}^{m}y_{ik}^{l} \tag{3}$$

$$\sigma_{B}^{l2}=\frac{1}{m}\sum_{k=1}^{m}\left(y_{ik}^{l}-\mu_{Bi}^{l}\right)^{2} \tag{4}$$

$$\hat{y}_{ik}^{l}=\frac{y_{ik}^{l}-\mu_{Bi}^{l}}{\sqrt{{\sigma_{B}^{l}}^{2}+\varepsilon}} \tag{5}$$

其中, $\hat{y}_{ik}^{l}$为第l层第i个神经元的第k个数据经归一化处理的结果. 将输出的数据进行归一化处理后, 必然会改变其原有学习到的特征分布, 因此需要对归一化的结果进一步处理, 使之恢复到原有学习的特征分布, 在批归一化算法中引入两个重构参数γ、β, 对归一化的结果进行线性变换.

$$Y_{i}^{l}=\gamma_{i}^{l}\hat{y}_{ik}^{l}+\beta_{i}^{l} \tag{6}$$

其中, Y_{i}^{l}为第l层第i个神经元经过批归一化层后的输出, γ_{i}^{l}和β_{i}^{l}分别为第l层第i个神经元的两个重构参数, 每一个神经元都有一对这样的参数, 且这些参数可以根据前一层输出的分布特点进行学习, 通过这两个参数, 使得输出经过归一化处理后, 依然能够恢复原先的分布.

隐藏层与批归一化层形成的结合层输出为:

$$Y_{i}^{l+1}=\gamma_{i}^{l+1}\frac{f\left(z_{i}^{l+1}\right)-\mu_{Bi}^{l+1}}{\sqrt{\left(\sigma_{B}^{l+1}\right)^{2}+\varepsilon}}+\beta_{i}^{l+1} \tag{7}$$

$$z_{i}^{l+1}=\sum_{j=1}^{N}W_{ji}^{l}Y_{j}^{l}+b_{i}^{l} \tag{8}$$

最后一层输出层 y 为:

$$y=g\left(\sum_{k=1}^{N}W_{k}^{3}Y_{k}^{3}+b_{y}^{3}\right) \tag{9}$$

式中, g 为输出层的激活函数, 用于决策最终的类别.

3 糖尿病诊断模型

使用结合批归一化算法的多层感知机建立糖尿病诊断模型, 该模型搭建及训练所用电脑的 CPU 配置为 Intel Core i7-8750H@2.20 GHz, 16 GB 内存, 模型采用 Keras[12], sklearn[13]进行实现, 并且使用 Tensorflow GPU 加速训练. 整个模型建立流程图如图 3 所示.

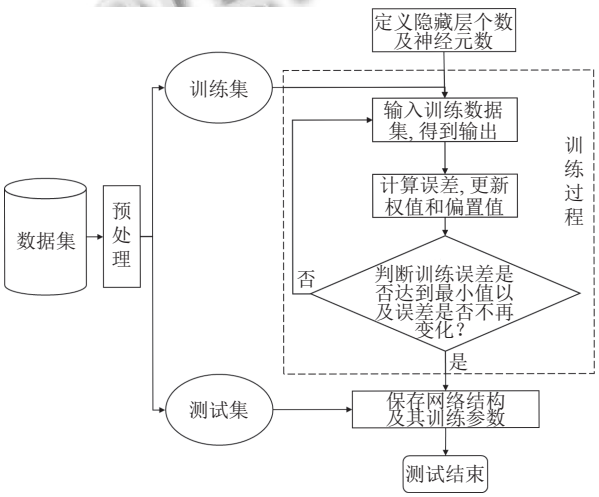


图 3 糖尿病诊断模型建立流程图

糖尿病诊断模型建立流程:

(1) 定义网络隐藏层个数及神经元个数. 给定网络

的初始隐层层数和神经元个数, 通过后续的训练和测试效果, 对隐层层数和神经元个数做加减, 使得模型效果好.

(2) 开始训练过程, 反复训练, 直到符合要求. 将前期划分出来的训练集输入网络中, 并在训练集中划分 0.1 作为验证集微调模型的超参数, 直到训练误差达到最小值或误差不再变化, 训练结束.

(3) 将符合要求的网络及训练参数保存下来, 用于测试集. 对第 (2) 步中训练好的网络, 保存下来, 使用测试集评估网络的性能.

(4) 重复 (1) 到 (3) 寻找最优参数组合.

4 实验

4.1 训练数据集

本文建立的模型采用 PIMA 印第安糖尿病数据集 (PIDD)[14]进行训练. 该数据集中共有 768 条数据, 9 个属性, 数据集属性如表 1 所示. 该数据集中的第 9 个属性是分类标签, 其值为 0 和 1, 0 表示没有糖尿病, 1 表示患有糖尿病. 标签为 0 的有 500 个, 为 1 的有 268 个. 前 8 个属性是每个样本的特征.

表 1 PIDD 属性

属性序号	属性名	属性描述
1	Pregnancies	怀孕次数
2	Glucose	血浆葡萄糖浓度
3	BloodPressure	血压值
4	SkinThickness	皮脂厚度
5	Insulin	胰岛素含量
6	BMI	身体质量指数
7	DiabetesPedigreeFunction	遗传指数
8	Age	年龄
9	Outcome	分类标签

4.2 数据预处理

为了提升模型的训练效果, 对该数据集进行一些预处理. 通过 pandas 中的 describe 函数对整个数据集进行描述, 如表 2 所示. 可以看出数据集中各属性的取值范围较大. 而且该数据集属性量纲不统一, 对模型的训练影响较大. 因此, 本文采用 sklean 框架中 StandardScaler 标准化方法, 其通过删除均值和缩放到单位方差来标准化特征, 原理公式为:

$$z = \frac{x-u}{s} \tag{10}$$

其中, z 为标准化后的值, s 为方差, x 为样本特征值, u 为平均值.

表 2 PIDD 的描述信息

属性序号	平均值	标准差	最小值	25%	50%	75%	最大值
1	3.85	3.37	0	1	3	6	17
2	120.90	31.97	0	99	117	140.25	199
3	69.11	19.36	0	62	72	80	122
4	20.54	15.95	0	0	23	32	99
5	79.80	115.24	0	0	30.5	127.25	846
6	31.99	7.88	0	27.3	32	36.6	67.1
7	0.47	0.33	0.078	0.24	0.37	0.63	2.42
8	33.24	11.76	21	24	29	41	81

箱型图[15]是目前最受欢迎的数据分析图之一, 如图 4 所示, 该图能够直观表示数据离散分布状况, 上四分位数和下四分位数的间距越小, 说明分布越集中, 否则越分散. 箱型图的一个优点就是基本不受异常值的影响, 因此常用来判断数据集中是否存在异常值.

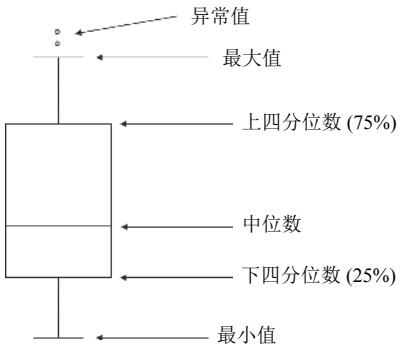


图 4 箱型图

对 PIDD 应用箱型图, 其结果如图 5 所示, 可以直观的看出数据的分布状况, 虽然箱型图对数据分析比较客观, 但是同样需要结合实际情况对箱型图给出的异常值进行判断, 比如怀孕次数中的 3 个异常值是符合实际的, 还有皮脂厚度、遗传系数以及年龄等, 考虑到实际情况, 这些数据不予处理. 综合箱型图和实际情况, 对数据进行处理, 因为 PID 样本量比较小, 若将包含异常样本的数据删除, 将会丢失其他特征信息. 因此, 在本文中, 对异常值按照缺失值处理, 即使用平均值来替换异常值. 除了异常值, 该数据集中同样包含许多缺失值, 对这些缺失值用该列的平均值进行替换.

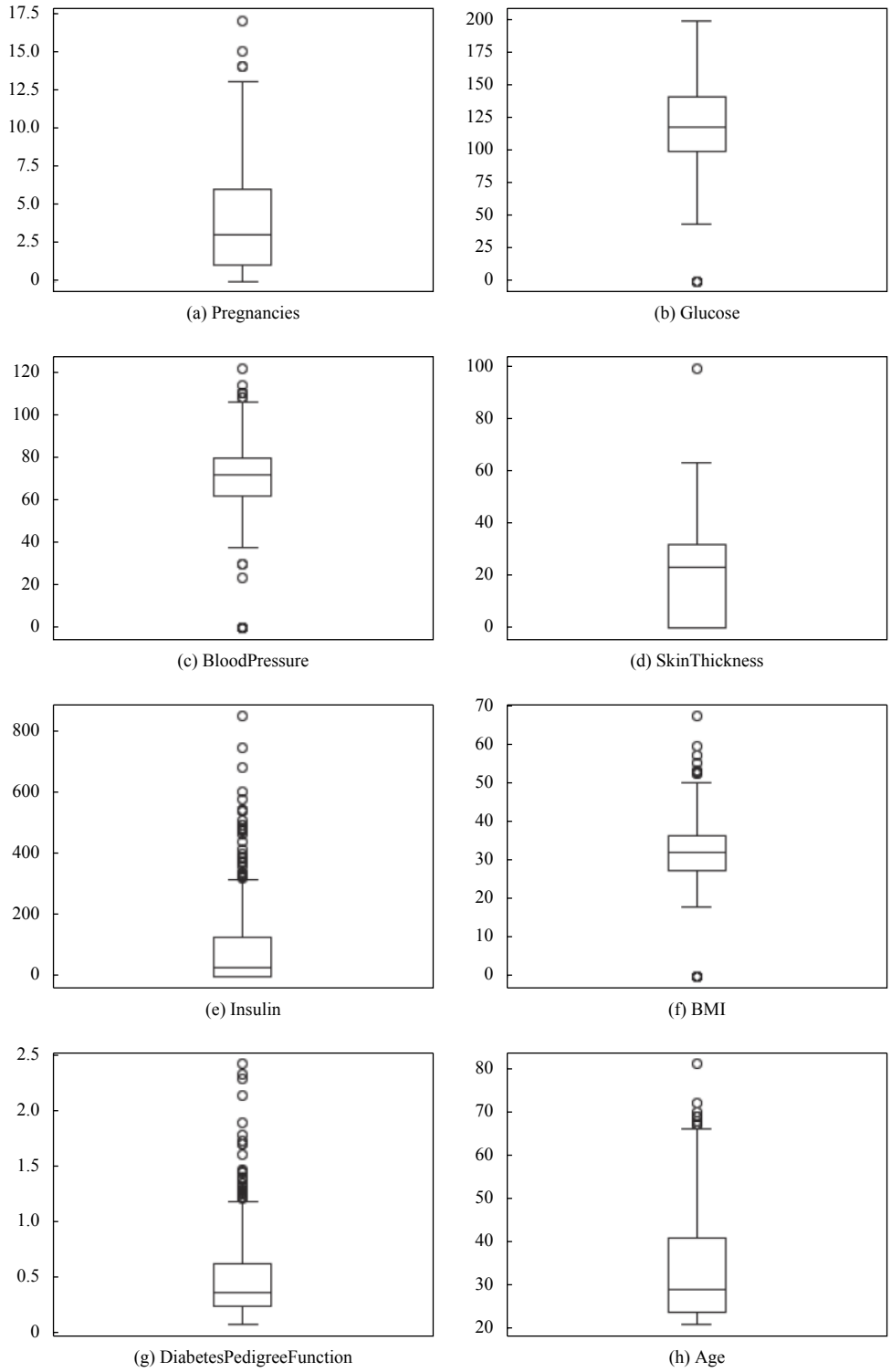


图 5 PIDD 中前 8 个属性的箱型图

4.3 模型训练

将处理后的数据按照 9:1 比例划分为训练集和测试集, 即训练集有 691 个样本, 测试集有 77 个样本. 将训练集送至本文搭建好的神经网络中进行训练, 训练过程中划分 0.1 的训练集做验证集, 调整网络超参数. 经过多次实验调参, 选择最优模型, 即 3 个隐藏层, 各

层神经元个数分别是 16、32、8, 输出层仅有 1 个神经元. 隐藏层的激活函数选用 Relu 激活函数, 该激活函数能够使得输出的均值更接近于 0, 输出层的激活函数选用 Softplus 激活函数, 该激活函数更接近生物学的模型且具有一定的稀疏能力, 可进一步优化网络性能[16]. 采用均方误差作为损失函数, 优化器采用 Adadelta 优化算法.

4.4 实验验证及结果分析

本文共做了 3 组实验, 分别是 (a)、(b) 和 (c). (a) 组实验即为本文建立的糖尿病诊断模型; (b) 组实验采用文献[8]中提出的模型; (c) 组实验将 (a) 组中所有批归一化层去除, 其他参数不变, 作为一个 MLP. 3 组实验的数据集以及预处理都相同.

3 组实验训练过程中训练和验证的准确率与损失值的变化如图 6 所示. 图 6(a) 中训练和验证的损失值在 175 轮训练时趋于稳定, 且两者基本重合, 网络拟合效果好; 图 6(b) 中训练和验证的损失值在 300~350 轮之间已经稳定, 较 (a) 组收敛稍慢; 图 6(c) 中很明显的可以看出, 其训练的收敛速度远远慢于 (a) 组, 说明结合批归一化算法的多层感知机建立的糖尿病诊断模型收敛速度显著提高.

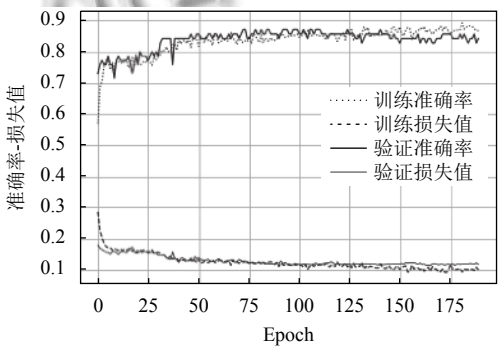
由图 6 中的点虚线和粗实线 (即训练准确率和验证准确率) 可以看出, (c) 组实验训练集准确率远远高于验证集, 即模型对训练集过度拟合, 泛化能力最差; (a) 组和 (b) 组两者之间误差较小, 泛化能力好, 说明批归一化算法能够在一定程度上消除使用 Dropout 的情况, 提高模型的泛化能力.

3 组实验的实验结果如表 3 所示, 分别记录各组实验的训练准确率、测试准确率以及 AUC 值. 由表中数据可知, (a) 组实验效果最好.

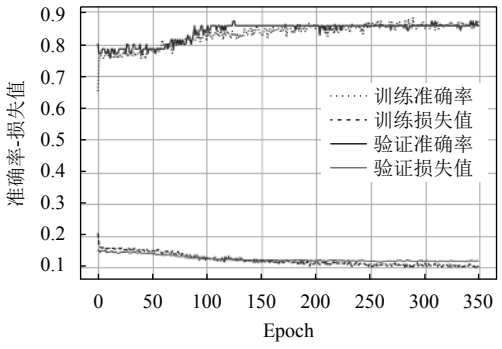
综上所述, 本文建立的糖尿病诊断模型的使用批归一化算法, 提高了模型的收敛速度, 同时其泛化性能有明显提高.

5 结束语

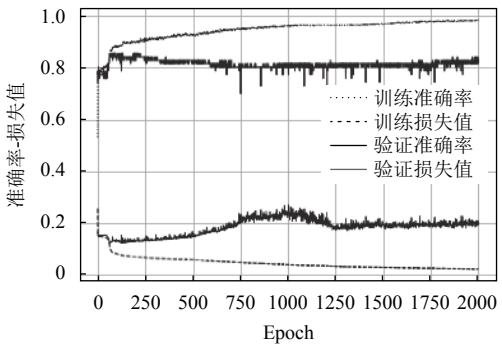
本文主要研究多层感知机在构建糖尿病诊断模型中出现的过拟合导致的泛化能力不足以及收敛速度慢等问题, 提出将批归一化算法运用到多层感知机中, 利用批归一化算法能够学习数据分布, 提高训练速度等优点, 重新构建新的多层感知机深度神经网络模型. 批归一化算法通过对隐藏层的输出数据进行批归一化处理, 并能够学习数据的分布, 保证了网络每次输出的分布一致, 加快收敛速度, 并在一定程度上能够缓解过拟合问题. 本文通过对比实验, 在 PIMA 数据集上验证, 提出的神经网络模型在训练集和测试集分别达到了 88.28% 和 92.21% 的准确率, 泛化能力好且收敛速度比其他两组实验快, 可见基于批归一化算法改进的多层感知机神经网络不仅能够有效降低神经网络的过拟合问题, 提高模型的泛化能力. 同时, 改进的多层感知机神经网络模型相比较其他方法在精度上有所提高. 下一步工作将利用本文建立的糖尿病诊断模型与实际应用结合起来, 对模型的适用性作进一步探究.



(a) MLP+BN



(b) MLP+Dropout



(c) MLP

图 6 3 组实验训练过程中训练和验证的准确率和损失值的变化

表 3 3 组实验的实验结果记录

实验序号	方法	训练准确率 (%)	测试准确率 (%)	AUC
(a)	MLP+BN	88.28	92.21	0.97
(b)	MLP+Dropout	88.42	89.61	0.96
(c)	MLP	97.69	85.71	0.93

参考文献

1 国际糖尿病联盟. 全球糖尿病概览. 8 版. 2017. 43.

2 Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Computer Science, 2018, 132: 1578–1585. [doi: 10.1016/j.procs.2018.05.122]

3 刘阳, 孙华东, 张艳荣, 等. 基于支持向量机的糖尿病预测模型研究. 哈尔滨商业大学学报 (自然科学版), 2018, 34(1): 61–65, 74.

4 胡建强. 一种基于云雾辅助的移动健康监护系统设计. 厦门大学学报 (自然科学版), 2019, 58(4): 608–613.

5 Choubey DK, Paul S, Kumar S, *et al*. Classification of pima Indian diabetes dataset using naive Bayes with genetic algorithm as an attribute selection. Proceedings of International Conference on Communication and Computing Systems. Nanjing, China. 2017. 451–455.

6 Mohapatra SK, Swain JK, Mohanty MN. Detection of diabetes using multilayer perceptron. In Bhaskar MA, Dash SS, Das S, et al, eds. International Conference on Intelligent Computing and Applications. Singapore: Springer. 2019. 109–116.

7 Soltani Z, Jafarian A. A new artificial neural networks approach for diagnosing diabetes disease type II. International Journal of Advanced Computer Science and Applications, 2016, 7(6): 89–94.

8 Ashiquzzaman A, Tushar AK, Islam MR, *et al*. Reduction of overfitting in diabetes prediction using deep learning neural network. In: Kim KJ, Kim H, Baek N, eds. IT Convergence and Security 2017. Singapore: Springer, 2018. 35–43.

9 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 448–456.

10 朱威, 屈景怡, 吴仁彪. 结合批归一化的直通卷积神经网络图像分类算法. 计算机辅助设计与图形学学报, 2017, 29(9): 1650–1657. [doi: 10.3969/j.issn.1003-9775.2017.09.008]

11 张德园, 杨柳, 李照奎, 等. BN-cluster: 基于批归一化的集成算法实例分析. 沈阳航空航天大学学报, 2018, 35(3): 72–80. [doi: 10.3969/j.issn.2095-1248.2018.03.010]

12 Galea A, Capelo L. Applied Deep Learning with Python: Use Scikit-learn, TensorFlow, and Keras to Create Intelligent Systems and Machine Learning Solutions. Packt Publishing, 2018.

13 Hackeling G. Mastering Machine Learning with Scikit-learn. 2nd ed. Packt Publishing, 2017.

14 UCI. UCI Machine Learning. https://www.kaggle.com/uciml/pima-indians-diabetes-database. 2016.

15 Wickham H, Stryjewski L. 40 years of boxplots. The American Statistician, 2011. [doi: 10.1016/j.apradiso.2013.11.121]

16 孙艳丰, 杨新东, 胡永利, 等. 基于 Softplus 激活函数和改进 Fisher 判别的 ELM 算法. 北京工业大学学报, 2015, 41(9): 1341–1348.