Computer Science and Application 计算机科学与应用, 2020, 10(5), 914-926
Published Online May 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.105094

Hans汉斯

Prediction of Diabetes Based on Convolutional Neural Network

Ying Wang[1], Heng Zhang[1*], Jian Zuo[1], Kuang-Gi Shu[1,2]

[1]School of Computer and Information Science, South West University, Chongqing
[2]Beijing New Energy Vehicle Technology Innovation Center Co., Ltd., Beijing
Email: *dahaizhangheng@163.com

Received: Apr. 23rd, 2020; accepted: May 6th, 2020; published: May 13th, 2020

Abstract

The remarkable progress of biotechnology and medical science has led to the significant production of biomedical data. For diabetes mellitus (DM), a common chronic disease, a large number of medical data has also been generated in the process of diagnosis and treatment. So the exploration of medical data has become a hotspot. The purpose of this study was to explore the short-term admission probability of discharged patients. According to the probability of re-admission within 30 days, we can judge the effect of this treatment, thus assisting doctors to provide more efficient treatment for patients, so as to improve the quality of life of patients. In this study, the data has been processed, and the improved convolutional neural network algorithm (CNN-EI) was used to perform data mining on the dataset of diabetic case data. The experimental results show that the improved algorithm can well perform high-dimensional and large-sample medical data. Accuracy is 83.7%. The result is compared to the result of other state-of-art methods.

Keywords

Diabetes, Assisted Diagnosis and Treatment, Data Analysis, Machine Learning

基于卷积神经网络的糖尿病预测

王　莹[1]，张　衡[1*]，左　健[1]，徐匡一[1,2]

[1]西南大学计算机与信息科学学院，重庆
[2]北京新能源汽车技术创新中心有限公司，北京
Email: *dahaizhangheng@163.com

收稿日期：2020年4月23日；录用日期：2020年5月6日；发布日期：2020年5月13日

*通讯作者。

文章引用：王莹，张衡，左健，徐匡一．基于卷积神经网络的糖尿病预测[J]．计算机科学与应用, 2020, 10(5): 914-926.
DOI: 10.12677/csa.2020.105094

摘　要

生物技术和医学的显著进步导致了生物医学数据的大量产生。糖尿病(Diabetes mellitus, DM)作为一种常见的慢性病，在诊断和治疗过程中也产生了大量的医学数据。因此，医学数据的探索成为了一个热点。本研究旨在探讨出院病人的短期入院率。根据30天内再次入院的概率，我们可以判断这种治疗的效果。从而协助医生为患者提供更有效的治疗，以便提高患者的生活质量。在本研究中，对数据进行了处理，利用改进的卷积神经网络算法(Convolutional neural network algorithm, CNN-EI)对糖尿病病例数据集进行数据挖掘。实验结果表明，改进后的算法能够很好地处理高维、大样本的医学数据。将该方法的结果与其他先进方法的结果进行比较，其准确率为83.7%。

关键词

糖尿病，辅助诊断与治疗，数据分析，机器学习

1. 引言

糖尿病是一种常见的慢性非传染性疾病。患者的代谢紊乱会导致患者长期高血糖水平。由于该病不易治愈，患者长期高血糖水平给患者的肾脏、心血管和神经系统带来了严重危害，造成了许多并发症，给患者的身心健康带来极大危害。不过，根据世界卫生组织 2017 年发布的《世界卫生统计报告》，2015 年约有 4000 万人死于慢性非传染性疾病。其中，糖尿病是人类第四大死因[1]。报告指出，约 80%的糖尿病患者在中低收入国家[2]，如图 1 所示，这些国家为所有糖尿病患者提供有效治疗的医疗资源有限。因此，开展糖尿病防治工作，帮助患者提高治疗效率，对糖尿病患者尤其是中低收入国家的糖尿病患者是有益的。

随着医疗信息化的发展，各医疗机构在患者诊疗过程中积累了大量的患者医疗电子数据[3]。充分利用大量的医疗数据，探索有价值的治疗规律，帮助医生进行诊断和治疗，从而缓解部分地区医疗资源的不足。

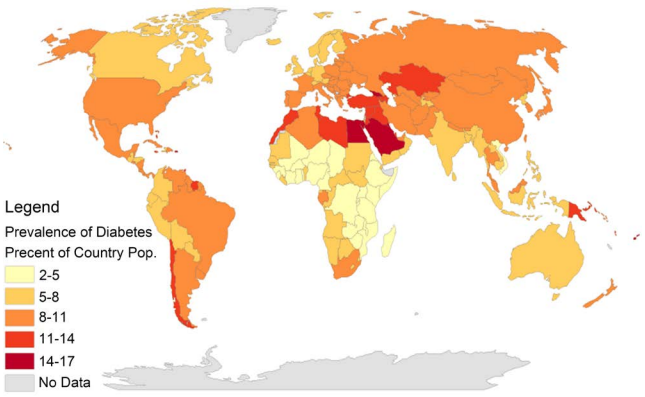

Figure 1. Global distribution map of diabetic patients
图 1. 糖尿病患者全球分布图

机器学习作为多领域交叉学科，近年来已经得到了很大的发展。深度学习作为机器学习的一个分支，由于其强大的特征提取能力，可以用来挖掘数据的深层特征。目前，许多学者利用机器学习算法对糖尿病患者的数据进行研究，以提高糖尿病患者的治疗效果和病情。一般来说，这些研究方法可以分为数理统计方法和机器学习方法[4]。

统计技术可以用来找出数据背后的可能性，在医学领域，统计方法被用来在相对较短的时间内估计复杂疾病的危险因素。加州大学医学院的亚伯拉罕使用统计方法通过监测 16 名糖尿病患者来监测糖尿病的危险因素。研究发现，糖化血红蛋白的糖尿病患者从 9.6%降至 7.2%，可以改善患者的病情，生理状态[5]。根据中医学的特点，李金星等研究人员利用无创设备采集患者的舌、面、舌下、脉搏、气味等信息，并分析各种信息之间的相关性。提出糖尿病的无创诊断策略[6]。

机器学习是处理机器从经验中学习的方法的科学领域[7]。在使用机器学习算法的研究中，研究员 T. Araki 等人提出了一种基于支持向量机(support vector machine, SVM)和主成分分析(principal component analysis, PCA)相结合的冠状动脉形态特征提取系统。该系统用于评估患者的疾病风险[8]。

研究人员 Sneha N.等人采用诸如随机森林、支持向量机(SVM)、K 均值和朴素贝叶斯之类的机器学习方法来选择可用于预测早期糖尿病的属性。结果表明，决策树算法和随机森林算法的准确率分别为 98.20%和 98.00% [9]。叶凌龙等研究员采用混合集成学习方法来分析容易诱发糖尿病患者乳腺癌的危险因素[10]。

作为机器学习的重要技术和研究领域，深度学习可以实现输入信息的逐层提取和筛选。因此，深度学习具有表征学习的能力，可以实现端到端的监督学习和无监督学习[11] [12]。机器学习技术本身可以用来产生良好的特征，这使得机器学习向“自动数据分析”迈进了一步[13]。在糖尿病预防研究领域，研究人员 Zakhriya Alhassan 使用 14,000 名 2 型糖尿病患者的连续数据集，对 LSTM (长–短期记忆，Long Short-Term Memory)和 GRU (门控循环单元，Gated-Recurrent Unit)进行了培训。该模型对 2 型糖尿病具有良好的诊断作用[14]。K. Kogias 提出了利用卷积神经网络(Convolutional Neural Networks 简称 CNNs)创建的两级图像分类方案，目的是将图像中的营养素含量相似的八大食品类别之一分类，然后将其分配给该类别中的特定食品。该算法预测餐后血糖水平的准确率分别为 84.18%和 85.94% [15]。

但是从以上研究结果可以看出，研究人员对糖尿病患者的治疗效果和病情监测主要集中在血糖水平的变化上。很少有研究关注预测住院病人的治疗结果。然而，有重要影响[16] [17]。

因此，本研究将探讨住院患者的病历资料，以追踪住院治疗后出院后短期内再次住院的风险。

本文在对糖尿病患者住院信息进行分析和处理的基础上，使用改进的卷积神经网络预测出院后短期内再次住院的风险。本文的其余部分按以下方式组织。在第二节中，我们将介绍有关数据集的信息；第三节专门阐述拟议方案的框架。第四节讨论了实验和性能比较。第六节总结了论文的结论。

2. 数据集

本节致力于介绍我们工作中使用的数据集。本研究使用的数据集来自 UCI 机器学习库[18]，该库记载了美国 130 家医院 10 年来收治的糖尿病患者的临床护理记录。

根据 UCI 机器学习库的描述，该数据集中的数据满足以下信息[18]：

1) 住院病人的经历。
2) “糖尿病”的遭遇，即在此期间将任何类型的糖尿病输入到系统作为诊断。
3) 住院时间至少 1 天，最多 14 天。
4) 在住院期间进行了实验测试。
5) 在住院期间进行了药物治疗。

根据上述信息，共有 10,766 个数据、54 个属性列和一个满足上述条件的标签变量。标签变量为：出院后 30 天内未再次入院的患者为 0，30 天内再次入院的患者为 1。患者入院记录标签用于判断患者是否有可能在短时间内再次入院，从而判断本次治疗的效果。每个样本包含 54 个属性，包括患者编号、种族、性别、年龄、入院类型、住院时间、检查次数、糖化血红蛋白测试结果、胰岛素释放测试结果、药物数量、糖尿病药物、门诊诊断记录、急诊次数等属性。表 1 显示了一些关键属性和说明。

Table 1. Some key attributes in dataset
表 1. 数据集中的一些关键属性

变量名称	说明
PATIENT_NBR	住院的唯一标识符
RACE	值：白种人，亚洲人，非裔美国人，西班牙裔、其他
GENDER	值：男性、女性、未知/无效
AGE	按 10 年间隔分组：[0, 10), [10, 20), ⋯, [90, 100)
Admission type	对应于 9 个不同值的整数标识符，例如，紧急，突发和不可用
Time in hospital	从入院到出院的整数天数
MEDICAL_SPECIALTY	主治医师专科的整数标识符，对应 84 个不同的值
DIAG_1	初步诊断；848 个不同值
DIAG_2	二级诊断；923 个不同值
DIAG_3	附加辅助诊断；954 个不同值
MAX_GLU_SERUM	指示结果范围或者是否测试。值：“>200”，“>300”，“正常”，若没有测量，则为无
Number of lab	住院期间进行的实验测试次数
A1c test result	指示结果范围或未进行测试
MIGLITOL	稳定/否
READMITTED	住院患者重新入院的天数。若患者在 30 天内重新入院，则值为“<30”，若患者在 30 天后再次入院，则值为“>300”，“无”表示没有再入院的记录。
INSULIN	上升/稳定/下降/否

这项研究对数据集的特征进行了初步分析。在参考医院医生意见的前提下，本研究首先对数据集中的特征进行了分类。在该数据集中，有关患者的年龄，性别和体重的信息已被充分证明对糖尿病有影响。观察数据集发现，住院的患者是不同种族的，考虑到种族可能是糖尿病的潜在病因。因此，将其添加到数据维度。血糖水平可以作为判断糖尿病患者病情变化的重要标准；糖化血红蛋白的量取决于血糖浓度以及血糖和血红蛋白接触的时间，因此可以用作糖尿病监测的重要指标；胰岛素释放实验可以反映出患者胰岛中产生胰岛素的 β 细胞的功能是否正常。 这些生化指标可为医生诊断本病提供重要参考，也对研究目的具有重要价值。在属性的选择上，本研究还选取了医生的诊断记录和用药记录作为关键属性。本研究将这些属性分为四个维度：一是患者的基本生理信息，包括体重，种族，性别等；二是患者的基本入院登记信息，包括患者的住院编号，入院时间等；三是患者病情诊断记录；四是病人的用药信息记录。数据模型如表 1 所示。

根据上述数据集属性的统计结果，该数据集包含了糖尿病诊断和治疗的许多属性。根据糖尿病主治医师的意见，数据集基本上包含了糖尿病诊断和治疗中常用的检测指标和药物种类等信息。用于基于数据集构建预测模型的训练数据非常通用。

2.1. 数据预处理

由于本实验使用的数据集是医院记录的原始数据集，因此该数据集存在数据格式不一致，数据不完整，数据记录错误等问题。由于数据的质量决定了模型分析性能的上限，因此本节重点介绍数据预处理。

数据清理是数据预处理的重要任务。数据清理的第一步是处理原始数据集中的缺失值。病历中数据丢失的主要原因是医务人员记录丢失或未获得患者病历。在这项研究中，数据集计算如下。实验结果如图 2 所示。红色部分代表缺失值的比率。可以看出，该数据集具有良好的完整性，只不过三个属性列的缺失值百分比很高。这些特征包括体重(丢失 96.9%的值)，付款人代码(39.7%)和医疗专科(44%)。研究人员综合了糖尿病医师的意见后，直接删除了付款人码，因为付款人代码的属性与研究结果无关。由于“体重”属性过于稀疏，无法补充，因此本研究将其删除。“医学专科”属性需要进一步分析。

对于缺少值较少的其他属性列，将使用均值方法来完成研究。对于错误数据，本研究使用手动删除和均值替换来删除错误数据。

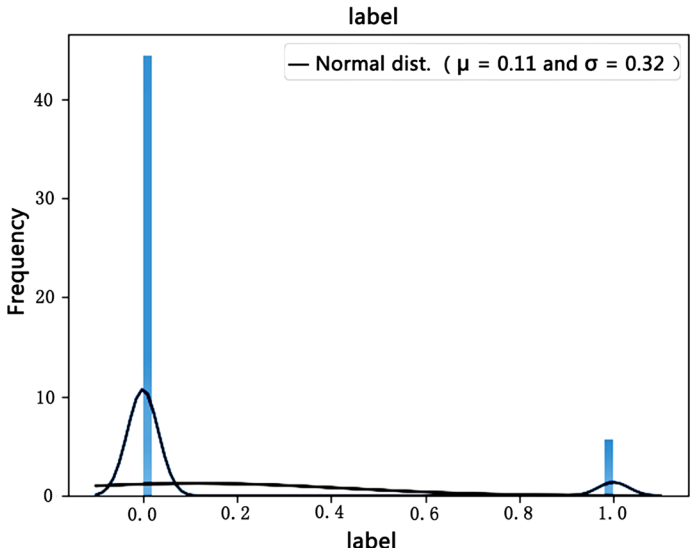

Figure 2. Distribution maps of samples before processing
图 2. 处理前的样本分布图

2.2. 数据平衡

通过对数据的统计分析，两个标签的数据存在严重的数据分布平衡问题。30 天内不复诊与 30 天内复诊的比例为 9:1。

这符合现实中医学数据的分布特征，并且现实生活中的数据大多都是不平衡的。当每个标签类型数据分布不均衡时，以总分类准确率为学习目标的传统分类算法会更多地关注多数类别，从而降低了模型对少数类别样本的分类性能，特别是在疾病诊断和治疗领域，数据不均衡使得实验结果的价值大打折扣，有的甚至变得毫无意义[19]。为了解决这一问题，本研究采用了一种称为多数加权少数过采样技术(Majority Weighted Minority Oversampling Technique, MWMOTE)的新方法，用以有效处理不平衡的学习问题。MWMOTE 首先识别出难以学习的信息丰富的少数类样本，并根据它们与最近的多数类样本的欧氏距离为它们分配权重。 然后利用聚类方法从加权的信息丰富的少数类样本中生成合成样本。这样做的方式是使所有生成的样本都位于少数类群集中[20]。经过实验处理，大多数样品和少量样品的数量基本平衡，如图 3 所示。

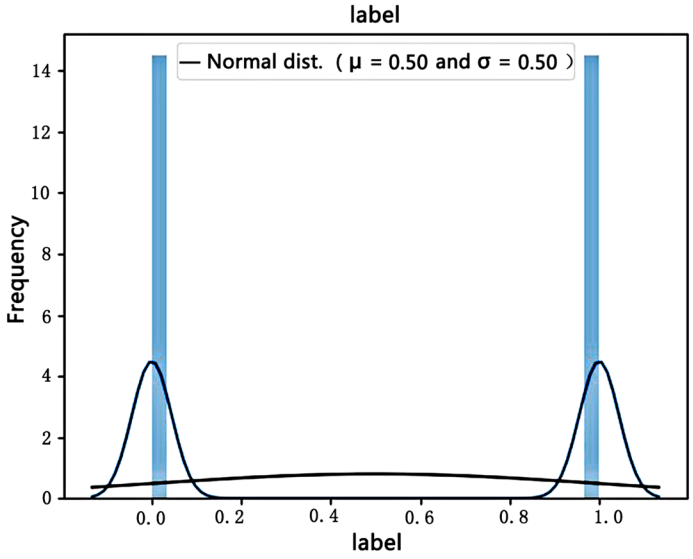

Figure 3. Distribution maps of samples after processing
图 3. 处理后的样本分布图

2.3. 数据特征筛选

本研究将采用深度学习算法进行研究，虽然深度学习算法可以对特征进行深度过滤。但是，鉴于本研究直接使用医院患者的病历资料，有些信息与研究目的无关，并且有些毫无价值的信息甚至可能影响实验目的的实现，因此有必要对特征进行筛选和选择。这项研究主要使用 Xgboost 算法对特征的重要性进行排名[21]。使用 Xgboost 算法计算各特征与预测结果的相关性，即数据特征对预测结果的影响权重。

Xgboost 算法具有精度高、效率高、并发性强等特点，并支持自定义丢失功能。可以用于分类和回归。它可以将特征的重要性输出为随机森林。由于其运算速度快，因此适合作为高维特征选择的重要方法。实验数据通过 10 倍交叉验证法进行划分，采用 Xgboost 算法进行训练，并根据每个特征对实验分类结果的贡献程度进行排序。实验结果如图 4 所示。

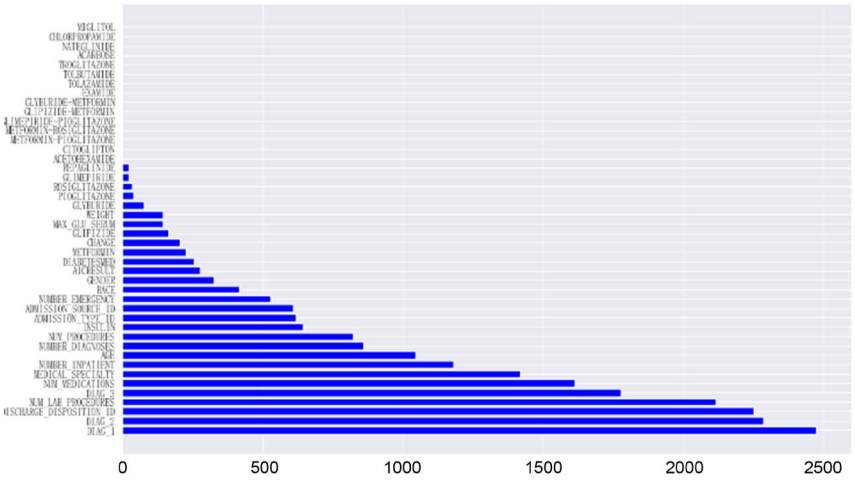

Figure 4. Characteristic weight sorting
图 4. 特征权重排序

实验结果表明，权重超过 500 的特征有 20 个。 同时，权重为 0 的属性共有 12 个，权重为 1 的属性基本上独立，如表 2 所示。

同时，在衡量这些属性的重要性时，本研究还咨询了医院的糖尿病主治医生。这些属性对糖尿病患者的发展影响不大，可以删除。另外考虑到糖尿病专家的意见，将特征值的权值阈值设置为 700，在此步骤的特征之后，将筛选剩余的 36 个重要特征。

Table 2. Characteristics of smaller weight values
表 2. 较小重量值的特征

属性名称	权重	属性名称	权重
LYBURIDE-METFORMIN	3	GLIMEPIRIDE-PIOGLITAZONE	0
METFORMIN-ROSIGLITAZONE	0	GLIMEPIRIDE-PIOGLITAZONE	0
MIGLITOL	0	CHLORPROPAMIDE	0
EXMAID	0	METFORMIN-PIOGLITAZONE	0
ACETOHEXAMIDE	0	TOLAZAMIDE	0
TROGLITAZONE	0	NATEGLINIDE	0

此时，其余特征主要如表 3 所示：

Table 3. Xgboost results of feature filtering
表 3. Xgboost 特征筛选结果

特征类型	特征筛选结果
患者信息	RACE, GENDER
检查报告	A1 CRESULT, INSULIN, MAX_GLU_SERUM. etc
处方	MEDICAL_SPECIALTY, ADMISSION_SOURCE, EMERGENCY. ect
诊断	DIAG_1, DIAG_2, DIAG_3

3. 方法论

本节介绍使用改进的卷积神经网络来预测糖尿病患者的再入院概率。我们首先介绍整个框架，然后详细描述每个过程。

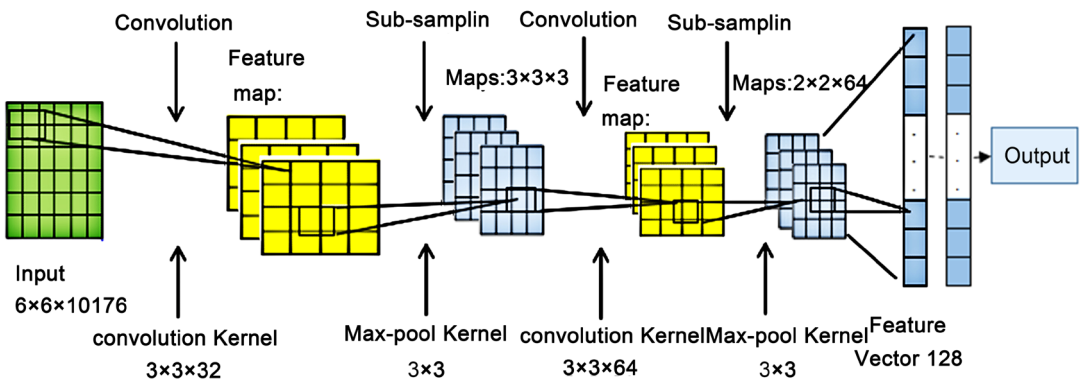

Figure 5. Algorithmic framework
图 5. 算法框架

3.1. 缩写语

图 5 展示了我们使用的方法框架——CNN-EI。主要流程如下：1) 利用滑动窗口机制从数据集中提取每个患者的诊疗记录，保存包括患者生理特征在内的基本信息，然后训练 CNN 作为特征提取器。2) 本研究提出的混合集成学习算法用于对 CNN 提取的特征进行分类。框架的算法执行过程如下：

算法 1 CNN-EI 算法

输入：样本集合 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 x_n 表示 n 个样本，分类标签 $y_n\in\{0,1\}$；T 为 CNN 的迭代次数，层数为 l；基学习算法 ς；集成学习训练轮数 K；混合集成学习者 $H_m(x)$

输出：再入院风险预测结果 $E(x)$

1) 正向传播：For iterator to T

2) For i to n： $a^L=\text{softmax}(2^L)=\text{softmax}(W^La^{L-1}+b^L)$

3) 反向传播： $l=L-1$ to 2

4) 更新 $W^l=W^l-\alpha\sum_{i-1}^{m}\delta^{i,l}*a^{i,l-1}$， $b^l=b^l-\alpha\sum_{i-1}^{m}\sum_{\mu,\upsilon}(\delta^{i,l})_{\mu,\upsilon}$

5) 保存模型 $D=\{(f_1,y_1),(f_2,y_2),\cdots,(f_m,y_m)\}$

6) For 1to m： $H_m(x)$

7) 输出 $E(x)=\text{sign}\sum_{m=1}^{3}w_{H_m}H_m(x)$

3.2. 特征提取器

深度学习作为近年来的热门研究领域，它能通过层次化的网络结构从大量的训练数据中提取出普通分类器无法获得的特征信息，因而得到了广泛的应用。根据本研究中数据量大、数据集特征丰富的特点，我们可以在特征提取中利用深度学习算法的优势。通过比较各种神经网络的特征，发现卷积神经网络可用于处理各种类型的数据，算法中使用的参数较少，当层数较少时，不会出现梯度爆炸和梯度耗散问题，并且不易出现过度拟合和局部最优解的问题。因此，为了发现数据集属性之间潜在的特征关系，本研究使用卷积神经网络提取数据集的特征。

卷积神经网络主要依靠卷积层和池化层进行特征提取和特征选择，使用全连接层进行特征集成，并使用输出层进行特征分类。为了进一步提高卷积神经网络的性能，研究人员近年来进行了一系列研究。一些研究者[22]提出将传统的卷积神经网络用作特征提取的特征提取层。由于本研究使用的特征数量有限且数据集中的数据量很大，因此选择经典的 LeNet-5 网络作为特征提取器[23]。

该网络很小，但包含了卷积神经网络的基本模块。LeNet-5 有 7 层，无输入，每层包含训练参数；每层都有多个特征映射，每个特征映射通过卷积滤波器提取输入的特征，然后每个特征映射有多个神经元。

根据这项研究的特点，调整了 LeNet-5 网络的参数。输入层大小为 6 × 6 × 1，6 × 6 表示数据集中的 36 个特征。在卷积层中，由于输入数据的矩阵较小，因此选择了较小的卷积核。卷积核的数量为 32，并且通过卷积对特征进行升级以获得 6 × 6 × 32 的第二卷积层输入。在下一个池层中，输入是一个 5 × 5 × 64 的特征映射，最终输出 128 个特征。由于数据特征有限，但样本量较大，因此本研究将样本随机分为 5 组小批量，并训练其作为输入以获得丰富的特征集。特征提取器最终获得的特征数为 640 维。

3.3. 混合学习集成算法

受一些研究人员的启发，他们提出用支持向量机或聚类算法代替传统卷积神经网络输出层的 Softmax 函数，以获得性能改进的模型。在本研究中，我们提出用更好的分类器代替传统的卷积神经网络的输出层函数，以进一步优化卷积神经网络的分类能力[24]。

但是，集成学习算法也有一些缺点：与基础学习者相比，其预测速度大大降低。随着基础学习者数量的增加，算法所需的内存也急剧增加。

因此，针对集成学习的不足，考虑到集成算法的有效性，本研究将同态集成算法和异态集成算法相结合，提高了基础学习者的泛化能力，增加了各学习者之间的差异。结合两种集成算法的优点，提出了一种混合集成学习算法。本文选取决策树算法、随机森林算法和朴素贝叶斯算法作为基础学习者。算法的实现过程如下：

算法 2 混合集成学习算法

输入：样本集：$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，样本：$x_n$，标签：$y_n$

基本分类器：$C=\{C_1=DT,C_1=RF,C_1=NB\}$

输出：异构综合分类器 $E(x)=\text{sign}\sum_{m=1}^{3}w_{H_m}H_m(x)$

1) 初始化权值：$w_i=1/N$，$i=1,2,\cdots,N$

2) For $m=1$ to 3：弱分类器 $w_iC_m(x)$

3) 计算弱分类器的错误率(err_m)：$err_m=\dfrac{\sum_{i=1}^{N}w_iI(y_i\neq G_m(x_i))}{\sum_{i=1}^{N}w_i}$

4) 计算 alpha (α_m)：$\alpha_m=\log((1-err_m)/err_m)$

5) 更新权值 $w_i=w_i\exp[\alpha_m\cdot I(y_i\neq G_m(x_i))]$，$i=1,2,\cdots,N$

6) 获得一个集成学习者 $H_m(x)$，A_m

7) 计算同态分类器权重：$A_{H_m}=\dfrac{A_m}{A_1+A_2+A_3}$

8) $E(x)=\text{sign}\sum_{m=1}^{3}A_{H_m}H_m(x)$

为了保证混合集成学习不仅具有良好的分类效果，而且具有良好的分类效率，本研究采用 GridSearch-CV 来调整集成学习者的参数，这些参数对分类精度有很大的影响，如基本学习者的数量和学习率。分析了不同基础学习者数量和学习率对分类准确率的影响，得出了每个基础学习者的最优数和学习率。表 4 列出了基础学习者数量和学习率对分类准确率的影响。

Table 4. Effects of number and learning rate of base learners on classification accuracy
表 4. 基础学习者人数和学习率对分类准确性的影响

基础学习者名称	权重	属性名称	权重
Decision Tree Algorithms	20	0.86	0.82
Random Forest Algorithms	90	0.1	0.86
Naive Bayes	20	0.23	0.795

4. 实验和结果

本节介绍了一系列对比实验以及与其他糖尿病入院风险预测方法的性能比较。

4.1. 评估指标

为了比较和评估我们提出的糖尿病风险预测方法的性能，我们使用准确率(Accuracy)，与召回率(Recall)和精确率(Precision)有关的 F1-Measure [25]和混淆矩阵，如下所示：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

这里，TP (true)是被分类为阳性的阳性病例数。TN、FN 和 FP 分别表示真阴性、假阴性和假阳性数量。召回率代表正确分类的阳性样本的百分比，F1-Measure 是准确率和召回率的加权平均值，代表整体表现。混淆矩阵是一个特定的列联表，允许对临床相关性能进行可视化。ROC 曲线：接收机工作特性。ROC 曲线上的每个点都反映了对同一信号刺激的敏感性。

4.2. 设置和结果

我们对第 2 节中描述的数据集进行了实验。在特征提取层，批量大小和学习率分别设置为 5 和 0.01。d Sklearn 库实现了用于混合集成学习的机器学习算法和其他机器学习分类器。 实验是在配备 NVIDIA Titan x GPU 的计算机上进行的。

1) 无数据平衡预测糖尿病短期再入院风险

为了验证医院实案数据平衡的有效性，本研究通过使用未经数据平衡处理的数据来观察实验结果，对本研究提出的分类预测方法进行了测试。在这项研究中，使用 CNN-EI 预测了 6 × 6 像素的色块。 总共使用了 10176 份病历进行实验。实验结果如图 6 所示，具有数据集且没有数据均衡的实验分类模型的预测准确率为 0.81，召回值为 0.84，特异性为 0.73，精确率为 0.71。

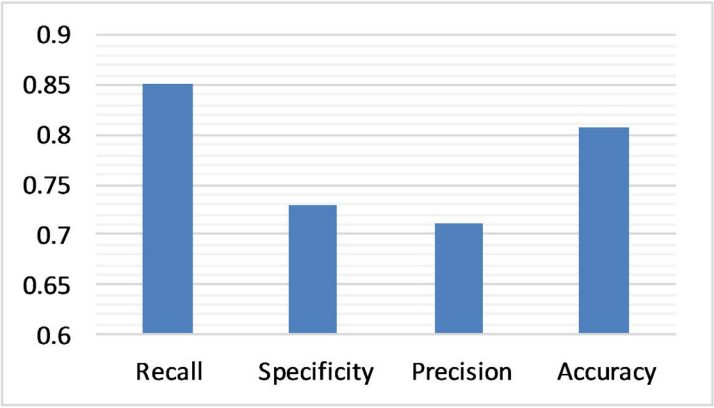

Figure 6. Prediction of short-term re-admission risk of diabetes mellitus without data balancing
图 6. 无数据平衡的糖尿病短期再入院风险的预测

2) 通过数据平衡处理数据集预测糖尿病的短期再入院风险

为了验证数据均衡的必要性，本研究对数据均衡后的数据进行训练。在这项研究中，使用 CNN-EI 预测了 6 * 6 像素的色块，共 180,818 条数据。图 7 显示 CNN-EI 模型的分类精确率为 0.837。可以看出，该模型对糖尿病趋势的预测更为准确，且算法的方差值较小，说明该算法具有良好的稳定性。该算法的召回率为 0.902，模型的准确率为 0.741，模型的特异性为 0.79。综合这些指标可以看出，该模型的性能良好，召回率较高。结果表明，该模型将患者恢复不良误断为恢复良好的概率较低，预测更为准确。

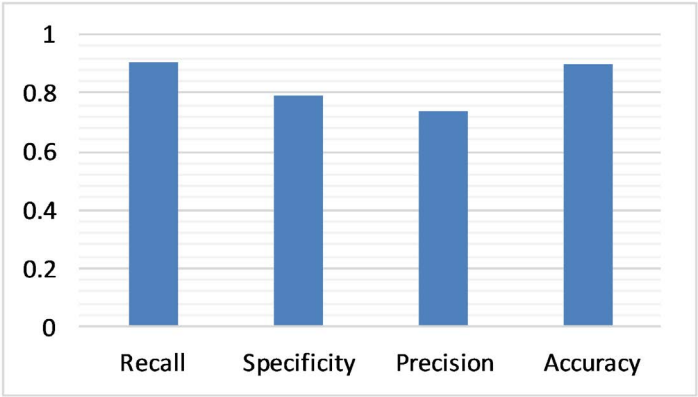

Figure 7. Prediction of short-term re-admission risk of diabetes mellitus through data balance processing data set
图 7. 通过数据平衡处理数据集预测糖尿病的短期再入院风险

在本实验中，研究人员绘制了具有 CNN-EI 算法训练的平衡数据的糖尿病患者的 ROC 曲线，如图 8 所示。

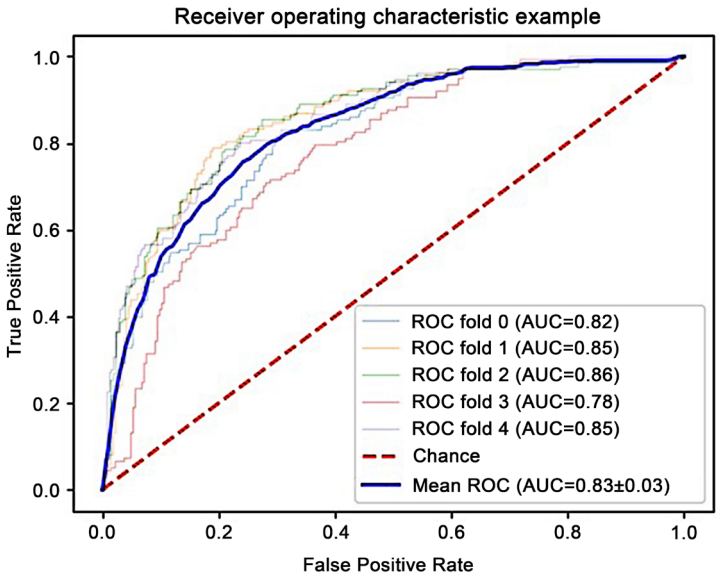

Figure 8. The ROC curves of diabetes mellitus patients with balanced data trained by CNN-EI algorithm
图 8. CNN-EI 算法测试得到的平衡数据的糖尿病患者的 ROC 曲线

从图 8 可以看出，经过五次交叉验证，得到的 ROC 曲线远离对角线，ROC 曲线下面积较大，说明该模型具有较高的可靠性。同时，验证了数据均衡处理的有效性。

3) 对比实验

在验证了数据处理中数据平衡的有效性之后，为了进一步验证本研究的有效性，将数据平衡后的数据集应用于其他三种不同的分类器进行训练：卷积神经网络，决策树，随机森林和朴素贝叶斯，并观察实验结果指标。实验结果如表 5 所示。

其他分类器的实验结果表明，CNN-EI 分类算法具有良好的分类精度和稳定性。根据模型的稳定性分析，CNN-EI 算法的稳定性优于其他算法。在医学研究领域中，模型的预测稳定性与预测精度相比是非常重要的，关系到患者病情的恢复。因此，该模型的良好准确率和算法性能为该算法在实际中的应用提供了可能性。

Table 5. Experimental results on five classifiers of data after balanced treatment of diabetes mellitus cases
表 5. 糖尿病患者平衡治疗后五种数据分类器的实验结果

Models	Accuracy	Recall	Precision	F1-Measur
Decision Tree	0.820 ± 0.02	0.855 ± 0.12	0.721 ± 0.21	0.781
Random Tree	0.823 ± 0.02	0.802 ± 0.12	0.721 ± 0.16	0.771
Naive Bayes	0.795 ± 0.03	0.815 ± 0.09	0.728 ± 0.21	0.776
CNN	0.825 ± 0.06	0.899 ± 0.02	0.722 ± 0.12	0.797
CNN-EI	0.837 ± 0.03	0.902 ± 0.03	0.741 ± 0.16	0.812

5. 论述

本研究对糖尿病患者的原始住院记录进行筛选和平衡，以获得特征简洁、标签分布均衡的数据。然后，根据本研究提出的 CNN-EI 算法，预测患者的短期再入院概率。对比实验表明，所提出的用于实际医学数据的数据平衡过程是必要的，并且可以提高模型的可靠性。同时，通过与其他经典分类算法的实验结果比较，验证了该算法能够提高结果预测的准确性。我们将这些结果与 Beata Strack 等研究中提出的使用患者糖化血红蛋白的值预测短期再入院可能性的可行性和准确性进行了比较[26]。然而，Beata 的结果仅表明糖化血红蛋白可以用作糖尿病患者再入院率的有用预测指标，无法确定特定的相关度。此外，并非每个患者入院时都有 HBA1 值统计信息，因此不可能充分利用有价值的信息来研究患者的再入院率。研究人员 Naveen Kumar Parachur Cotha [27] [28]认为糖尿病和患者人口统计之间的关系是一个多标签问题。这项研究将相同的数据集划分为不同的样本量，并探索了可以通过多个标签预测糖尿病患者的模型。对 1000、10,000、20,000 个样本的数据集进行的实验表明，该研究提出的 BR/JRip 模型具有较高的整体精度，分别为 0.533 (1000 个样本)，0.702 (10,000 个样本)，0.569 (20,000 个样本)。但是，与本研究预测糖尿病的准确性相比，本研究的准确性比 Naveen 的准确性高得多，我们的研究具有更大的实用价值。

6. 结论

针对实际病历数据中存在的数据不平衡问题，本文采用 MWMOTE 算法对数据进行均衡处理，得到具有均衡分类分布的数据。提出了数据的无关特征，得到了 6 × 6 大小的特征矩阵。结合卷积神经网络的特征提取能力和集成学习的出色分类能力，对集成学习算法进行了改进，提出了 CNN-EI 方法来预测处理后数据集中患者的再入院率。实验结果表明，该方法可以预测出 83.7%的再入院率，模型更加可靠。与其他研究相比，本研究综合考虑了患者的各种常见生理指标，可以更准确地预测患者的再次入院概率。

基金项目

本项目受中央高校基本科研业务费项目(XDJK2020B029)、西南大学重大横向项目(SWU41015718、SWU20710953)资助。

参考文献

[1] World Health Organization (2017) World Health Statistics 2017: Monitoring Health for the SDGs.

[2] World Health Organization (2016) Global Report on Diabetes: Executive Summary.

[3] Russell, S. and Norvig, P. (2003) Artificial Intelligence: A Modern Approach. 2nd Edition, Prentice Hall, Upper Saddle River.

[4] Russell, S. and Norvig, P. (2009) Artificial Intelligence: A Modern Approach. 3rd Edition, Prentice Hall, Upper Saddle River.

[5] Sattar, N., Wannamethee, G., Sarwar, N., Tchernova, J., Cherry, L., Wallace, A.M., Danesh, J. and Whincup, P.H. (2006) Adiponectin and Coronary Heart Disease: A Prospective Study and Meta-Analysis. *Circulation*, **114**, 623-629. https://doi.org/10.1161/CIRCULATIONAHA.106.618918

[6] Huang, G.-M., Huang, K.-Y., Lee, T.-Y. and Weng, J. (2015) An Interpretable Rule-Based Diagnostic Classification of Diabetic Nephropathy among Type 2 Diabetes Patients. *BMC Bio-Information*, **16**, S5. https://doi.org/10.1186/1471-2105-16-S1-S5

[7] Li, J.X., *et al*. (2019) Body Surface Feature-Based Multi-Modal Learning for Diabetes Mellitus Detection. *Information Sciences*, **472**, 1-14. https://doi.org/10.1016/j.ins.2018.09.010

[8] Kavakiotis, I., Tsave, O., Salifoglou, A., *et al*. (2017) Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, **15**, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005

[9] Vijayanv, V. and Ravikumar, A. (2014) Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. *International Journal of Computer Applications*, **95**, 12-16. https://doi.org/10.5120/16685-6801

[10] Sneha, N. and Gangil, T. (2019) Analysis of Diabetes Mellitus for Early Prediction Using Optimal Features Selection. *Journal of Big Data*, **6**, 13. https://doi.org/10.1186/s40537-019-0175-6

[11] Ye, L.L., Lee, T.-S. and Chi, R. (2018) A Hybrid Machine Learning Scheme to Analyze the Risk Factors of Breast Cancer Outcome in Patients with Diabetes Mellitus. *Journal of Universal Computer Science*, **24**, 665-681.

[12] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning (Vol. 1). MIT Press, Cambridge.

[13] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436. https://doi.org/10.1038/nature14539

[14] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2015: 114-115.

[15] Alhassan, Z., McGough, A.S., Alshammari, R., Daghstani, T., Budgen, D. and Al Moubayed, N. (2018) Type 2 Diabetes Mellitus Diagnosis from Time Series Clinical Data Using Deep Learning Models. 27*th International Conference on Artificial Neural Networks*, *Rhodes*, Greece, 4-7 October 2018, Vol. 3, 468-478. https://doi.org/10.1007/978-3-030-01424-7_46

[16] Kogias, K., Andreadis, I., Dalakleidi, K. and Nikita, K.S. (2018) A Two-Level Food Classification System for People with Diabetes Mellitus Using Convolutional Neural Networks. 2018 40*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Honolulu, 2603-2606. https://doi.org/10.1109/EMBC.2018.8512839

[17] Umpierrez, G.E., Isaacs, S.D., Bazargan, N., You, X., Thaler, L.M. and Kitabchi, A.E. (2002) Hyperglycemia: An Independent Marker of In-Hospital Mortality in Patients with Undiagnosed Diabetes. *Journal of Clinical Endocrinology and Metabolism*, **87**, 978-982. https://doi.org/10.1210/jcem.87.3.8341

[18] Levetan, C.S., Passaro, M., Jablonski, K., Kass, M. and Ratner, R.E. (1998) Unrecognized Diabetes among Hospitalized Patients. *Diabetes Care*, **21**, 246-249. https://doi.org/10.2337/diacare.21.2.246

[19] Frank, A. and Asuncion, A. (2010) UCI Machine Learning Repository. University of California, School of Information and Computer Science, San Diego.

[20] Maratea, A., Petrosino, A. and Manzo, M. (2014) Adjusted F-Measure and Kernel Scaling for Imbalanced Data Learning. *Information Sciences*, **257**, 331-341. https://doi.org/10.1016/j.ins.2013.04.016

[21] Barua, S., Islam, M.M., Yao, X., *et al*. (2014) MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 405-425. https://doi.org/10.1109/TKDE.2012.232

[22] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785

[23] Lin, K., Lin, Y. and Kong, G. (2018) A XGBoost Algorithm-Based In-Hospital Mortality Prediction Model for Patients with Sepsis in ICU. *Chinese Journal of Health Informatics and Management*, **15**, 536-540+563.

[24] Lecun, Y., Bottou, L., Bengio, Y., *et al*. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[25] Zhou, Z.H., Wu, J. and Tang, W. (2002) Ensembling Neural Networks: Many Could Be Better than All. *Artificial Intelligence*, **137**, 239-263. https://doi.org/10.1016/S0004-3702(02)00190-X

[26] Ujjwal Maulik, S.B. (2002) Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **24**, 1650-1654. https://doi.org/10.1109/TPAMI.2002.1114856

[27] Strack, B., Deshazo, J.P., Gennings, C., *et al*. (2014) Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, **2014**, Article ID: 781670. https://doi.org/10.1155/2014/781670

[28] Cotha, N.K.P. and Sokolova, M. (2015) Multi-Labeled Classification of Demographic Attributes of Patients: A Case Study of Diabetics Patients.