文本复制检测报告单(全文标明引文)

№:ADBD2018R_2018053015312720180530154835440174135322 检测时间:2018-05-30 15:48:35

检测文献: 53140607_何凯_计算机科学与技术_深度学习在糖尿病预测研究中的应用

作者: 何凯

检测范围: 中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
大学生论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库
个人比对库

时间范围: 1900-01-01至2018-05-30

检测结果

总文字复制比：2% 跨语言检测结果：0%

去除引用文献复制比：2% 去除本人已发表文献复制比：2%

单篇最大文字复制比：0.7%（血糖预测模型及低血糖预警技术研究）

重复字数：	[491]	总段落数：	[7]
总字数：	[23968]	疑似段落数：	[4]
单篇最大重复字数：	[167]	前部重合字数：	[84]
疑似段落最大重合字数：	[383]	后部重合字数：	[407]
疑似段落最小重合字数：	[29]		

指 标：疑似剽窃观点 疑似剽窃文字表述 疑似自我剽窃 疑似整体剽窃 过度引用

表格： 0 公式： 10 疑似文字的图片： 0 脚注与尾注：0

0%（0）	中英文摘要等（总2724字）
1.8%（29）	第1章绪论（总1603字）
4.3%（383）	第2章研究背景介绍_第1部分（总8913字）
1.2%（50）	第2章研究背景介绍_第2部分（总4073字）
0.8%（29）	第3章深度学习在糖尿病预测的研究（总3705字）
0%（0）	第4章实验结果及分析（总1641字）
0%（0）	第5章总结与展望（总1309字）

（注释：无问题部分 文字复制比部分 引用部分）

1. 中英文摘要等 总字数：2724

相似文献列表 文字复制比：0%(0) 疑似剽窃观点：(0)

原文内容 红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

摘要

深度学习在糖尿病预测研究中的应用

糖尿病是一种非传染性的慢性疾病，糖尿病及其并发症导致的致死率非常高，对患者造成了非常严重的伤害，由其导致

的致死率，仅次于感染、心血管疾病、癌症、创伤等疾病。糖尿病作为一种慢性多发疾病，正逐渐成为全球关注的重点。根据统计，我国是糖尿病患者数量是世界上最多的国家，而且我国糖尿病的患病率这几年也呈现出增长的趋势。

本文的研究主要针对2型糖尿病，在对2型糖尿病血糖预测中，目前比较热门的血糖预测的方法是通过CGMS(continuous glucose monitoring system, 动态血糖监测系统)提供的糖尿病人历史的血糖数据，根据历史的血糖序列数据预测未来某一时间段的血糖值，根据预测可能出现的低血糖或高血糖等不正常的结果，起到对糖尿病人的预警作用。目前对于血糖预测的研究有基于数据驱动的血糖预测研究模型、基于生理模型的血糖预测模型和结合数据驱动和生理模型的预测。

本文的研究并非使用CGMS提供的历史血糖数据，而使用的体检数据以及其他临床指标，数据中包括患有2型糖尿病的人群和正常的人群，通过深度学习的方法根据受检者的信息和血糖值建立模型，通过检查数据，预测血糖值，从而提供给受检者是否具有患糖尿病的可能性。

本文首先采用矩阵补全技术对缺失数据进行填充，预处理；然后，使用深度学习的模型对血糖值进行拟合；最后在测试集上对模型的效果进行分析以及评价。

关键字： 糖尿病， 血糖预测 ， 深度学习 ， 神经网络

Abstract

Application of Deep Learning in Prediction of Diabetes

Diabetes is a non-communicable chronic disease, diabetes and its complications lead to a very high rate of death, the patient caused a very serious injury, caused by the death rate, second only to infection, cardiovascular disease, cancer, trauma and other diseases. Diabetes mellitus, as a chronic and multiple disease, is gradually becoming the focus of global attention. According to statistics, China is the largest number of diabetics in the world, and the prevalence of diabetes in China over the past few years also showed an increase in the trend.

In the study of type 2 diabetes mellitus, the current popular method of blood glucose prediction in type 2 diabetes mellitus is the blood glucose data provided by CGMS (continuous glucose monitoring system, ambulatory blood glucose monitoring), According to the history of blood glucose sequence data to predict the future period of blood sugar value, according to the prediction of possible hypoglycemia or hyperglycemia, such as abnormal results, play a role in the early warning of people with diabetes. At present, the study of blood glucose prediction is based on the data-driven blood glucose prediction model, the physiological model based blood glucose prediction and the combination of data-driven and physiological model prediction.

The study is not based on the historical blood glucose data provided by CGMS, but also on the physical examination data and other clinical indicators, which include people with type 2 diabetes and normal people, and establish models based on the patient's information and blood sugar levels through depth learning, and by examining the data to predict blood sugar values, This provides a possibility for the patient to have diabetes.

In this paper, the matrix complement technology is used to fill the missing data, and then the model is used to fit the blood sugar value, and the result of the model is analyzed and evaluated on the test set.

Keywords： Diabetes， Glucose Prediction ， Deep Learning ， Neural Network

目录

2. 第1章绪论

总字数：1603

相似文献列表　文字复制比：1.8%(29)　疑似剽窃观点：(0)

1	基于时序描述逻辑的UML状态图语义研究 杨海波(导师：李明) -《兰州理工大学博士论文》- 2010-04-15	1.8%（29） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第1章绪论

1.1 课题背景和意义

糖尿病是一种常见的慢性非传染性疾病。据国际糖尿病联盟发布的糖尿病地图第八版指出，2017年全世界共有4.25亿的成年人患有糖尿病，而中国20~79岁的患有糖尿病的人数达1.144亿，占全球患病总人数超过25%。糖尿病伴随着的并发症危害非常大，每年死于糖尿病及其相关疾病达130万人。根据糖尿病的统计资料[1]显示， 2010年我国糖尿病患病率为11.6%，而在这些人群中意识到自己患有糖尿病的人仅为30.1%，只有25.8%的患者进行了治疗，接受治疗的人群中只有39.7%的患者将血糖控制在合理的范围内，而且这几年，我国糖尿病患病率也在逐渐上升，综上可以看出，我国糖尿的诊治率亟待提高[2]。本课题旨在通过糖尿病人的临床数据和体检指标，通过深度学习的技术建模，从而根据受检者的体检信息来预测他们的糖尿病患病程度，帮助提高糖尿病的知情率。

本文采用2型糖尿病人的数据，数据包含正常人群和2型糖尿病人群的受检信息，数据包含他们的临床数据和体检指标，标签值为该样本的血糖值。本文旨在通过深度学习的方法对受检信息和血糖值建模，进一步就可以根据模型预测新的受检人群的血糖值，可以进一步来判断受检者的糖尿病程度，给医生提供参考意见，或者在未来是否有患糖尿病的风险，使得受检者

可以加强预防措施。如果预测结果能指出某一个受检者的预测结果与糖尿病患者相近，而该受检者目前并未患有糖尿病，则说明该受检者受检信息与糖尿病患者接近，受检者可以改变自己的生活习惯等，进一步规避患糖尿病的风险。

1.2 领域研究现状

目前研究的血糖预测[2]多是通过CGMS提供的糖尿病患者的历史血糖数据序列，对其建模并对未来一段时间的血糖变化进行预测，根据预测时间可以分为短时（1-30分钟）、近期（1-30天）等[3]。相对于近期预测，短时预测更有意义，通过预测得到的血糖值，可以对糖尿病患者未来可能出现的低血糖等症状提供预测，从而规避可能出现的危险。从预测的模型上，可以分为基于数据驱动的血糖预测研究方向和基于生理模型的血糖预测研究方向。基于数据便是使用CGMS提供的数据预测；基于模型，便是考虑生理模型在遇到外界条件刺激下，可能做出的反应；还有学者通过CGMS提供的数据做基于数据驱动的预测，并从过外界刺激进行修正，使预测精度更加提高。在数据驱动的预测中，国内外的学者也提出了多种的预测方法，比如自回归模型、极限学习机、支持向量机[5]、神经网络等。

1.3 本文主要工作

本文的工作不同于基于历史数据的序列的血糖预测。首先，预测的数据不同，不同于CGMS提供的糖尿病患者的历史血糖序列数据，本文使用的是受检者的临床数据和体检指标，受检者也包含未患糖尿病的人和患有2型糖尿病的人；其次，预测的目的不同，本文旨在通过预测血糖，发现未患糖尿病的人和2型糖尿病患者的联系，从而进一步为尚未患糖尿病的人提供评估。最后，预测的方法不同，本文使用的深度学习的方法，深度学习在语音识别、图像识别和自然语言处理等领域获得了突破性的进展，本文也尝试将其应用在医学数据上，看看其在医学领域是否也能发挥出巨大的作用。

1.4 本文组织结构

本文的组织结构如下：

第1章：绪论。本章对课题研究的背景及意义做出阐述，还介绍了当前血糖预测领域的研究现状，并且概述了本文所做的工作。

第2章：研究背景和理论基础介绍。本章首先介绍一些关于糖尿病的背景知识，其次对目前血糖预测的研究现状做出阐述，最后介绍算法的背景(如神经网络、卷积神经网络、TensorFlow、评价标准等)。

第3章：深度学习在血糖预测的研究。本章主要介绍数据的来源、数据的预处理方法、实验的流程以及训练的过程。

第4章：实验结果及分析。本章对深度学习预测的结果进行分析。

第5章：总结与展望。整个论文过程进行总结，分析算法的优点与不足。

3. 第2章研究背景介绍_第1部分

总字数：8913

相似文献列表　文字复制比：4.3%(383)　疑似剽窃观点：(0)

1	血糖预测模型及低血糖预警技术研究 申艳蕊(导师：王延年) -《郑州大学博士论文》- 2014-05-01	1.9%（167） 是否引证：否
2	53130504陈振岳计算机科学与技术神经网络的并行化实现 -《大学生论文联合比对库》- 2017-06-01	0.8%（75） 是否引证：否
3	中西医结合治疗糖尿病肾病研究进展 程景;杨冬梅;丁伯平;黄帧桧; -《科技信息》- 2014-02-15	0.6%（55） 是否引证：否
4	2型糖尿病相关的蛋白质组 范菲艳;李铮; -《生命的化学》- 2013-02-15	0.5%（47） 是否引证：否
5	糖尿病的自我防治——专访糖尿病专家张锡明教授 张杨; -《百姓生活》- 2012-01-01	0.5%（45） 是否引证：否
6	糖尿病治疗药物的合理选择 李本明; -《黑龙江医药》- 2014-04-15	0.5%（45） 是否引证：否
7	基于卷积神经网络的图像特征识别研究 杨念聪;任琼;张成喆;周子煜;李倩;邱兰; -《信息与电脑(理论版)》- 2017-07-23	0.5%（45） 是否引证：否
8	中国2型糖尿病防治指南(2010年版) -《中国医学前沿杂志(电子版)》- 2011-12-20	0.4%（36） 是否引证：否
9	糖尿病的诊断和治疗进展(一) 李健; -《基层医学论坛》- 2007-05-01	0.4%（36） 是否引证：否
10	基于廉价传感器的城市大气颗粒污染物监测系统 程云(导师：姜守旭) -《哈尔滨工业大学博士论文》- 2015-06-01	0.4%（32） 是否引证：否
11	11307991009-余露-基于深度卷积网络的图像识别学习框架的研究与搭建 余露 -《大学生论文联合比对库》- 2017-05-22	0.3%（29） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第2章研究背景介绍

2.1 糖尿病背景介绍

糖尿病是一种在世界范围内广泛流传的慢性非传染疾病，严重威胁了人类健康[4]。根据最新数据显示，全世界成年人糖尿病患者约4.25亿，而我国成年人患病人数达1.114亿，是世界上患者最多的国家，而且患病率逐渐上升。2型糖尿病占据了绝大多数的人数，约95%。2型糖尿病不仅面临高血糖的风险，同时还伴随着并发症，如高血压、高血脂、肥胖、高胰岛素血症等[6]。糖尿病及其相关并发疾病给患者带来了严重的危害。据2017版糖尿病防治指南可知[1]，我国未诊断出的2型糖尿病比例达63%，提高糖尿病的知情率很有必要。

糖尿病是是由于胰腺产生的胰岛素不足或者人体无法产生胰岛素时，导致血糖升高，从而无法控制的疾病。胰岛素是调节血糖的激素，长期处于高血糖的状态，会对人体造成非常大的危害[4]。糖尿病根据其病因大概可以分为以下四种：1型糖尿病、2型糖尿病、特殊类型的糖尿病和妊娠糖尿病 [7]。2型糖尿病患者出现高血糖的症状不是因为自身胰腺分泌胰岛素低或者无法分泌，是因为患者出现了胰岛素抵抗(Insulin Resistance, IR)，如果需要降低血糖则需要更多的胰岛素。2型糖尿病的患者多处于中老年，并且可能患有肥胖[3]。1型糖尿病包含有1A型和1B型，1A型病因是由于对胰岛β细胞产生免疫，1B型的病因是胰岛β细胞功能衰退。妊娠糖尿病是指孕妇在孕前糖代谢正常，但是在怀孕后出现了葡萄糖抵抗的症状，孕期中出现持续的高血糖可能导致胎儿畸形，危害非常大。虽然一般情况下，妊娠糖尿病在孕妇怀孕结束之后会恢复正常，但仍有一定的几率转化为永久性的2型糖尿病。特殊类型的糖尿病非常少见。

目前关于糖尿病的诊断国际上通用的方法是根据WHO（Word Health Organization,世界卫生组织）于1999年制定的标准，虽然WHO在2011年也建议采用HbA1c（糖化血红蛋白）作为糖尿病的诊断标准[2]，但是血糖一直是国内糖尿病的判断标准。

根据静脉血浆葡萄糖WHO 1999的判断标准如下：

（1）正常血糖：空腹下测量的血糖不超过6.1mol/L，并且在糖负荷后2h的血糖不超过7.8mol/L。

（2）空腹血糖受损(IFG)，是一种糖尿病前期：空腹血糖介于6.1mol/L和7.0mol/L之间，并且糖负荷后2h血糖低于7.8mol/L。

（3）糖耐量异常(IGT)，是一种糖尿病前期：空腹血糖不超过7.0mol/L，并且糖负荷后2h血糖，介于7.8mol/L和11.1mol/L。

（4）糖尿病空腹血糖不低于7.0mol/L，并且糖负荷后血糖大于等于11.1mol/L。

通过HbA1c的诊断通常以6.3%为切点，大于6.3%即为糖尿病，否则即为正常值。

由于2型糖尿病无法根治，所以其预防显得尤为重要。2型糖尿病的预防[1]根据是否患病等分为如下三个级别：

（1）第一个级别指的是在患病前，通过控制可能导致患病的因素，防止未来患有糖尿病。

（2）第二级是指在患有2型糖尿病之后，应当尽早发现病情，并且针对患病情况进行治疗，控制血糖，发现患有糖尿病。

（3）第三级是指在得知患有2型糖尿病之后，应合理控制糖尿病及其并发症的发展，减少糖尿病带给糖尿病人的损失。

2.2 血糖预测背景介绍

目前的血糖预测都是通过历史的血糖序列数据来推断未来某一时间的血糖值，属于2型糖尿病预防中的第三级，即在得知患者患有糖尿病之后，应合理控制糖尿病及其并发症的发展，从而降低糖尿病的危害。

既然需要历史的血糖序列数据，那么就需要对患者的血糖进行采集。传统的血糖采集方法有自我血糖检测（Self Monitoring of Blood Glucose, SMBG）[3]；另外一种检测方法是通过动态血糖监测系统CGMS[4]来检测。传统的自我血糖检测，是让收集者自己采集自身的血液，并通过血糖仪对其进行分析，这种方式次数有限，无法获取大量的数据来通过训练模型预测未来的血糖值，且过多的采集会使受检者出现抵触情绪。糖化血红蛋白作为另外一种可以衡量受检者是否血糖的信息，具有稳定性，其代表了三个月左右的血糖信息，故也无法用作血糖预测的数据。动态血糖监测系统通过植入皮下的探头来测量受检者体内的血糖信息，相比于传统的血糖采集方法，其具有以下优点：简单便携、不间断等。由于CGMS可以不间断的采集受检者的血糖信息，而且其间隔信息很短，而且受检者不会受影响，所以我们可以通过CGMS采集得到的历史的血糖序列数据来建模预测未来时间的血糖。

目前的血糖预测分为两大方向：基于数据驱动的血糖预测研究方向和基于生理模型的血糖预测研究方向[3]。

（1）基于生理模型的血糖预测：基于生理模型的血糖预测主要将人体考虑为一个系统，当系统的受到外界刺激与影响时，内部的血糖值就会收外部的影响也会相应地变化，其模型复杂但是精度相对高[4]。

（2）基于数据驱动的血糖预测：基于数据驱动的血糖预测模型主要通过动态血糖监测系统提供的患者的历史的血糖变化序列建模，来预测未来一段时间的血糖值。该方法相比于基于生理模型的血糖预测，具有简便的好处，成为了研究的热点。有一些研究以数据驱动为基础，在预测的结果上融入了环境影响的因素，使得的预测的精度进一步提高。

2.3 神经网络和深度学习背景介绍

2.3.1 神经网络背景

神经网络[8]是由若干个神经元相互连接组合而成，最基本的一个神经元模型如图2-1所示。神经元通常具备两种运算，一种是线性运算，通过将权重和输入值取乘积，之后加上偏置（有的地方把偏置也叫做阈值，而且是将乘积之后的结果减去阈值，以模拟生物神经元中的阈值兴奋机制，实际上这两种方法描述的含义是一致的）；另一种是非线性运算，通过非线性的激活函数对线性运算的结果处理。

1. 神经元模型

图2.1 神经元模型

通常神经元包含两个参数：权重W和偏置b。当神经元收到来自上一级的神经元的输出信号作为输入，这些信号通过相应的权重连接到当前神经元，每一个输入信号与一个权重对应；将输入与对应的权重相乘，再将结果加上偏置b。最后将结果通过激活函数处理，处理之后得到的数值即为当前神经元的输出。

单个神经元的运算如下：

其中是神经元的权重， 是输入的属性， 是偏置， 是激活函数， 是经过神经元运算的输出值。

2. 激活函数

激活函数[8]将神经元得到的线性结果处理，使得神经网络具有非线性因素。如果没有激活函数处理，无论多少层的神经元叠加，得到的结果始终是输入的线性组合。当对线性运算的结果加上激活函数的时候，单个神经元就引入了非线性因素，使得模型可以学习非线性的关系。理想的激活函数是图2.2的阶跃函数，阶跃函数刚好对应了神经元的阈值兴奋机制，但阶跃函数由于不连续、不光滑等特点，所以实际中不使用。目前激活函数基本上有Sigmoid、tanh（双曲正切）、ReLU（整流线型单元）等，但都使用的基本上都是ReLU激活函数。

图2.2 阶跃函数

图2.3 Sigmoid激活函数

相比于阶跃函数，Sigmoid具有连续、光滑的特点。但是当通过其训练神经网络时，很容易出现梯度消失（也叫梯度弥散）的现象，究其原因是因为Sigmoid在输入值较大或较小的时候，梯度趋向于0，导致反向传播训练的效率低下。

图2.4 tanh激活函数

相比于Sigmoid激活函数，tanh激活函数关于原点中心对称，并且它的均值为0，使用tanh作为激活函数时通常训练速度比Sigmoid作为激活函数时快，效果比Sigmoid好，所以有时候将Sigmoid激活函数替换为tanh激活函数，会得到性能的提升，但是tanh仍然未解决梯度消失现象。

图2.5 ReLU激活函数

ReLU激活函数在输入大于0时，输出等于输入，在输入小于等于0是输出为0；在输入大于0时，梯度恒定为1，输入小于0时，梯度恒定为0。由于其在大于0的区间时，梯度恒定为1，所以克服了梯度消失现象，提高了训练效率，可以训练更深层的网络，但是其仍具有缺点，当落在负区间时，其梯度为0，无法继续更新，所以出现了死节点。即使当训练层数很深时，右侧的梯度仍然为1，连乘之后依旧是1，所以当网络层数很深的时候，ReLU激活函数的梯度要么为1要么为0，相乘的计算也很方便，所以ReLU激活函数的效果会表现得很好。

3. 神经网络的分类

神经网络的类型有很多种，比如：多层前馈神经网络[8](multi-layer feedforward neural networks，也叫做多层感知机)、ART(Adaptive Resonance Theory,自适应谐振理论)网络、SOM(Self-Organizing Map，自组织映射)网络、级联相关网络、Elman网络和Boltzmann机等。其中使用最广泛的是前馈神经网络。

多层前馈神经网络由输入层、隐含层(可以是多层)和输出层组成。输入层对应了模型的输入，与属性一一对应；隐含层可以为多层，相当于对输入属性做了一个非线性变换。输出层相当于对隐含层的结果做了处理。多层前馈神经网络中，只有相邻层的节点全连接，而且数据是单向传播的，不相邻的层之间的神经元不能连接，且神经元不能出现回路。输出神经元的个数与预测的标签对应。前馈神经网络中隐含层的层数和每一层的神经元节点数目都是由认为给出的，一般通过经验法和试错法确定的。输入结点个数和输出节点个数通过结合具体问题确定，隐层层数、每一层的神经元个数，一般通过经验法和试错法确定。根据万能近似定理[5]可知，一个足够大的前馈神经网络可以学习任何一个函数。

前馈神经网络的参数是权重和偏置，我们迭代地通过数据来训练前馈神经网络，就是训练其神经元的权重和偏置。

4. 反向传播算法

反向传播算法[21](error BackPropagation,简称BP)算法是根据梯度来更新神经网络的参数，一般情况下，我们都是使用反向传播算法来训练神经网络。

反向传播算法共分为两个阶段。第一阶段是正向传播：将属性与输入到对应的输入层中，然后第一个隐含层的每个神经元按照神经元计算输出值得方式，先计算线性结果，再将线性结果通过激活函数处理，当这一层的所有神经元都计算完毕时，再将数据传递给下一层，以此类推，直到最后的输出层，输出层对应了预测的结果。因为到输出的过程，符合网络的方向，没有出现回路，所以叫做正向传播。第二阶段为反向传播，正向传播的结果对应了模型的预测结果，通常情况下，会有一个损失函数来衡量预测值与真实值的差距，当两者的差距越大时，损失函数值越大，差距越小时，损失函数值越小。我们可以通过计算本次的损失值，进一步可以计算出损失函数到每一个神经元节点的梯度，开始时，先计算输出层节点的梯度，当得到某个神经元节点的梯度之后，可以根据单个神经元计算输出值的公式，计算得到输出值关于权重和偏置的导数，再将该节点的导数分别乘上每个参数的导数，即可得每个损失函数关于每个参数的梯度，我们的目标时希望通过反向传播算法来训练神经网络，使其在损失函数上的值尽可能的小，所以我们通过负梯度方向来更新神经网络的参数，给参数减去梯度的倍数，即更新完毕，当本层神经元节点更新完毕，即可更新上一层节点的参数，由于计算梯度的过程只能从后向前计算，所以叫做反向传播。一次正向传播对应了神经网络的一次预测，一次正向传播加上反向传播对应了神经网络的训练，通常情况下，我们训练神经网络是迭代的，可能成千上万次。

如图2-6所示，是一个单隐层的前馈神经网络，由于输出层不包含权重和阈值，所以通常单隐层的前馈神经网络也叫做2层的神经网络，我们以它为例，来介绍反向传播算法。

图2.6 单隐层前馈神经网络

因为输入层不包含权重，所以我们将输入层记作第0层，依次后面一层为第1层，第层后面一层为层。 表示第层的第神经元连接到第层的第个神经元的权重，表示第层的第个神经元的偏置， 表示第层的第个神经元经过输入与权重的乘积之后与阈值的加和，如：

其中， 代表第层第个神经元的输出，对应了第层的输入。 代表第层的第个神经元的输出经过激活函数操作的结果， 其中σ代表激活函数，选择Sigmoid激活函数，将结果压缩到01区间。

经过神经元一层一层的传播，就可以得到输出的预测结果为，真实的标记为，预测结果与真实值的差别通过均方误差来衡量，均方误差公式如下：

网络中每个参数的调整都是在原有基础上调整的，对于任一参数 ,学习的公式如下：

反向传播算法是基于梯度下降的策略，目的是降低损失函数的值，所以我们以负梯度方向更新参数，从而降低损失，使得模型优化。给每个参数对应的梯度再乘上学习率 ，学习率代表学习的步长，通常使用学习率来控制学习的速率，如果学习率设置得过小，则训练速度缓慢。如果学习率设置得太大，则容易发生震荡，可能会导致过拟合。

其中是梯度，即损失函数（现在指的是均方误差）关于第层的第个神经元的权重的梯度。

对于每一个神经元通过输出计算输出值的处理过程，相当于一个关于权重和偏置的函数 ，则预测结果就相当于若干个函数的叠加，梯度的求法也服从链式法则。以隐层第个神经元为例，其参数为和 ，该神经元先计算其线性结果，再将线性结果通过激活函数处理得到隐含层输出值，再影响到了输出层，最后影响到了损失函数。

根据的定义可得

对于Sigmoid函数

则在第2层的第个神经元的梯度为

代表第2层第个神经元的负梯度大小，结合可得该神经元的权重更新大小

同理，我们可以求出该神经元的偏置更新值如下：

第1层神经元结点的梯度值如下：

第1层神经元的权重和偏置更新值如下：

之后再给相应的参数（权重和偏置）减去更新值，一次反向传播的训练值就结束了。通过一次一次地迭代，最终反向传播算法算法会逐渐优化模型再损失函数上的值。当损失函数的值小到一定程度之后，训练就结束了。

2.3.2 卷积神经网络背景

卷积神经网络[9]是为了处理图像而设计的前馈神经网络，它是在前馈神经网络的基础上实现的，它相比于前馈神经网络，多了的部分卷积层和池化层，卷积层的神经元节点不是全连接的，而是根据空间结构部分连接的。科学家通过研究猫的视觉皮层之后发现，每个神经元只会处理部分的视觉图像信息，说明神经元之间的连接是局部的，并非全局的全连接。卷积神经网络的流行是在2012年，Hinton的学生Alex在ImageNet分类图片的比赛ILSVRC(ImageNet Large Scale Visual Recognition Competition)中，通过卷积神经网络AlexNet使得图片分类的准确率大大提高[10]，自此以后，每年的ILSVRC都会有新的进展，卷积神经网络快速发展,深度变得越来越深[11]，目前，已经在计算机视觉方向取得了显著的成果。卷积神经网络的输入层不同于前馈神经网络的输入层，后者的输入数据是一维的没有空间结构，而卷积神经网络的输入层是2维的，包含了空间结构，就像图片分类任务中一样，输入的照片不能丢失其空间结构，卷积神经网络正好利用了这一点。每个卷积层都有若干个卷积核，每个卷积层还包括多个特征映射(feature map)，每个特征映射都是经过卷积操作之后输出得到的结果。卷积层主要提供了两点优势:局部连接(Local Connection)和权值共享(Weight Sharing)，这两点不仅降低了参数量，使得原本复杂到难以计算的模型可以通过普通硬件来计算，而且还降低了过拟合的程度，当参数量很多的时候，神经网络往往会陷入过拟合，从而使模型的泛化性能不够好。池化层主要提供了降采样，相对于卷积的操作，其进一步地降低了参数量，并且使得模型能够拥有轻度的形变。全连接层就类似于前馈神经网络的隐含层，通常从卷积层或者池化层到全连接层之前需要对数据进行扁平化，使得原本的空间结构变成一维的线性结构，之后将一维的数据输入至全连接层，进行处理。输出层与前馈神经网络的输出层一样。通常情况下，卷积层后面紧跟着池化层，构成了卷积池化结构，经过卷积池化结构之后，在连接几个全连接层，将全连接层和输出层相连。

随着卷积神经网络的发展，其深度变得越来越深，从2012年AlexNet的8层深的卷积神经网络，到2014年VGGNet使用的19层深的网络，再到2015年ResNet152层的网络，结构变得越来越深，训练也越来越困难。随着网络的加深，参数的初始化也变得越来越重要，当网络很深时，随机初始化得到的效果不是很好，所以根据不同的激活函数，需要使用不同的初始化方法，当初始化得很好时，网络可以很快地收敛[22]。根据研究[22]，可以通过Xavier方法对深层的卷积神经网络进行训练。通过特殊特初始化方法，可以是权重初始化的不大不小，提升模型的性能，加速学习的过程。

卷积层的运算通常包括一下几点：

（1）按照一定的步长，依次对上层特征映射进行卷积核滤波操作，之后再加上加偏置(bias)。

（2）将（1）中结果，通过激活函数处理，在卷积神经网络中ReLU激活函数最为常用。

（3）对激活函数处理的结果进行池化操作。有时候还会在池化操作之后加上局部相应归一化层，但现在基本不会使用

，因为局部相应归一化层，提高的不多，反而会增加运算。

1. 卷积运算

卷积运算是将卷积核与输入的数据进行计算。通常情况下，输入是一个图片，也就对应了一个2维的矩阵，矩阵中的元素有数值(0-255)，如果是灰度图像的话，只有一个通道；如果是RGB彩色图像的话，则输入就有红绿蓝三个通道。以一个通道且数值大小为0或1为例，如图2.7是一个输入数据的例子。

图2.7 输入数据

卷积核是卷积运算的关键，卷积运算通过卷积核提取出图片的特征，从而应用在全图上。如图2.8是一个3×3的卷积核，通常包括以下几点：

图2.8 卷积核示意图

（1）尺寸。代表了卷积核的大小，卷积核主要是利用了局部数据的相关性，其大小正对应了视觉感受野，尺寸越大，获取的局部信息越多。

（2）深度。对应了卷积核的数量。通常情况下，我们需要提取图片的多个特征。我们通过训练可以使得一个卷积核对应一类特征，比如直线变，斜边等。数量越多，提取的特征就越多。通过一个卷积核滤波，会产生一个结果，也就是每一个特征映射(feature map)，所以深度越大，卷积之后特征映射的数量越多。在图2.8的例子中，我们所举的卷积核的深度为1。

（3）步长。卷积核需要应用在全部数据上，所以它会在原数据上来回扫描，就需要一个参数来控制扫描的快慢。步长越大，跳过的信息越多，所以卷积之后得到的特征映射的尺寸越小。

（4）填充。当我们通过一定的步长对图片进行扫描时，边界部分的数据往往不能像中间的数据与上下左右一起卷积。这时就有了两种策略，第一种是边缘部分用0填充，使得边界数据可以像中间数据一样地卷积，这种方法叫做泛卷积；另一种方法是忽略边界，对无法卷积的部分不进行卷积操作，这种方法叫做严格卷积。

如图2.9，卷积运算是将卷积核依次应用到原数据上，将原数据与卷积核对应的部分的元素相乘，结果加上偏置，之后再应用非线性激活函数ReLU，将激活之后的结果作为特征映射的一个元素值，当对所有的数据卷积完成之后，就得到了一个完整的特征映射。

图2.9 卷积运算结果

2. 池化运算

池化(pooling)，也叫降采样(down sampling)，其目的是提取出原数据中最显著的特征来降低参数量。池化操作主要有两种策略：最大池化，通过选择池化核内最大的数据保留，对其余的数据进行删除操作；平均池化，对特征映射的所有数据取平均作为池化结果。一般都使用最大池化。池化也有尺寸和步长，通常池化操作不会重叠，但是池化操作对数据是一对一的，即有多少个特征映射输入，输出就有多少个特征映射。

3. 卷积神经网络

通过将输入层、卷积层、池化层、全连接层和输出层组织起来，就构成了卷积神经网络[12]。如图2.10是一个包含两层卷积池化和两层全连接层的卷积神经网络。卷积神经网络的输入是一个固定尺寸的图片，输出是4类物体分别对应物体（狗、猫、船和鸟）的概率，概率最大的对应了图片分类的结果。并且真实的标签为对应的one-hot编码，即由4位，分别人代表了不同的物体狗、猫、船和鸟的概率，如果输入图片为对应的物体，则该物体对应的标签处为1，其余的对应的输出值为0。并且训练这种分类算法，一般情况下，损失函数都使用的信息熵（cross entropy）。

第一层卷积深度为3，第二层卷积深度为6，每一层卷积后面都跟有池化层，之后再连接两层全连接层，最后连接到节点为4的输出层。

通过迭代地向卷积神经网络输入图片，以及对应的结果。可以通过反向传播算法对模型进行训练，最终最小化训练误差，训练结束。

图2.10 卷积神经网络

1.1.3 卷软硬件配置

1. 硬件环境

处理器:i5-4200M CPU@2.5GHz

内存:8GB

显卡:NVIDA GeForce GT 755M

2. CUDA介绍

CUDA[13](Compute Unified Device Architecture)，是GPU(Graphics Processing Unit， 图形处理器)厂商英伟达推出的通用并行计算架构，它运行在NVIDA自己的GPU上，开发人员可以通过CUDA并行来加速程序运行效率，以达到高性能目的。目前深度学习的神经网络的训练在传统的CPU（中央处理器）上已经无法承受，通过GPU可以是神经网络的训练提速几十倍。目前卷积层的训练就依赖于英伟达提供的cuDNN[14]闭源实现。

cuDNN是NVIDA推出的针对深度学习专门推出的高度优化实现，目前大多数深度学习框架都是通过cuDNN来驱动GPU计算。

3. TensorFlow介绍

TensorFlow[15]是谷歌公司开源的针对机器学习的库，其内部将计算过程封装成数据流图模式，它的前端支持多种语言，比如Python、C++、Go、Java等，底层使用的C++、CUDA编写。它可以很方便的将程序移植到各种平台以及设备，以及大规模服务器集群。它可以实现多种机器学习算法，是当前非常热门的机器学习库。TensorFlow采用自底向上的编程思想，开发者需要从基本的每一层开始构建深度学习模型。并且TensorFlow内部封装了各种东西，如自动求导、各种类型的优化器、卷积运算、池化运算等，大大加速了我们构建深度模型的效率。

TensorFlow底层可以调用cuDNN，通过GPU来训练神经网络，使得模型的训练速度大大加快。

4. Pandas

Pandas[18]是python科学计算基础包中的提供数据操作的一个工具包，其提供了可以对excel、CSV等表格文件操作的数据结构，并对应的提供了一些列数据处理中可能需要的操作，比如基本的数据填充、表格中值的映射、表格的合并等等。

5. scikit-learn

scikit-learn是一个相对高阶的机器学习库，其内部封装了数据预处理的一些方法（如one-hot编码、标准化归一化）、以及一些机器学习中的模型（如决策树、支持向量机、前馈神经网络等）。其内部还提供了交叉验证、网格搜索以及随机搜索等调参方法的实现。

6. fancyimpute

fancyimpute是在GitHub上开源的一款Python环境下的矩阵填充包，其内部根据矩阵补全方面的论文封装了从简单填充到使用模型填充等各种各样的填充算法。

7. 软件环境

平台:Windows 10

IDE:PyCharm

语言:Python 3.6

主要的包:pandas、scikit-learn、fancyimpute、tensorflow-gpu

2.4 优化方法

优化算法，其目的就是优化某个函数（比如损失函数均方误差）上的取值，通过控制其内部的参数，使得目标函数取得最大值或者最小值的过程。神经网络中，常见的优化算法有梯度下降、随机梯度下降、小批量梯度下降、动量、Nesterov梯度加速法、Adagrad、AdaDelta、Adam等。接下来依次对其进行介绍。

2.4.1 梯度下降

传统的梯度下降算法的每次迭代训练的过程中，需要对所有的样本计算一个梯度，效率十分缓慢。当处理大数据集时可能很费时间。梯度下降和反向传播算法一样，都是通过计算目标神经元的梯度，在对其乘上学习率已完成参数的更新。

2.4.2 随机梯度下降

相比于传统的梯度下降算法，随机梯度下降(Stochastic Gradient Descent,SGD)每次随机抽取一个样本，计算该样本的梯度，用以作为本次迭代中使用的梯度。相比于梯度下降，随机梯度下降具有速度快的好处。

2.4.3 小批量梯度下降

小批量梯度下降(Mini Batch Gradient Descent)是传统的梯度下降算法和随机梯度下降算法的一个折衷，其参数更新的过程与梯度下降算法和随机梯度下降算法相同，唯一的不同点是每次迭代更新所需要的样本数，既不是全部的样本，也不是唯一的样本，而是取全部样本中的子集，对其进行计算梯度，并用计算得到的梯度，来更新参数，优化目标值，所以就叫做小批量

梯度下降算法。其小批量的样本量是一个重要的参数，一般是50-256之间。

2.4.4 动量

动量(Momentum)算法是通过模拟物理中物体运动时的动量而发明的，当前更新的速度，不仅和当前的梯度有关，也和上一时刻的梯度有关。通过一个参数来控制上一时刻的梯度和这一时刻的梯度比值。动量算法可以直接应用在小批量梯度下降上，即计算梯度时，包含了上一时刻的梯度值和这一时刻部分样本的梯度值。

2.4.5 Nesterov梯度加速法

当动量方法到达最小值点时，其仍然具有动量，所以不会立即停止，故有可能导致跨过最值点，所以Nesterov方法就是针对动量方法的这个弊端。

2.4.6 Adagrad、AdaDelta

由于上述方法仍然需要手动地设置学习率，所以自适应学习率的优化器就出现了。Adagrad有一个缺点，就是其计算的学习率总是再逐渐地变小，AdaDelta正好克服了这个缺点。

2.4.7 Adam

Adam(Adaptive Moment Estimation,自适时刻估计方法),是目前效果比较好的自适应学习率算法。其融合了动量和自适应学习率算法，当训练复杂的神经网络或者希望训练时间快的时候，可以考虑使用Adam优化器。

2.5 神经网络的优化方法

由于神经网络中的参数量非常的多，而且层数越来越深，所以深度学习经常面临着过拟合的风险。在我们通过机器学习的算法学习特征和标签之间的联系时，往往会面临三种结果，欠拟合（underfitting）、过拟合（overfitting）和恰好拟合。欠拟合就是我们模拟学习能力不够，在训练集上的性能都没学习好；过拟合指的是在训练集上学习得过了，学习到了一些样本自身的特点，失去了泛化性能，即在训练集上性能很好，在测试集上不够好。我们学习的目的就是为了将其应用在新的样本中，所以泛化性能非常重要。为了控制模型的过拟合，正则化[20]方法就应运而生了。相比于其他的传统的机器学习算法，深度学习的参数非常多，更容易过拟合。常见的控制过拟合的方法有Dropout、L1范数、L2范数等。

2.5.1 Dropout

Dropout[19]是有Hinton等人针对神经网络提出的。其思想在于对于神经网络的某一层，在训练时期，随机以概率p选取其中的部分神经元，丢弃选中的神经元，对该层余下的神经元进行训练；而预测阶段，使用所有的神经元进行预测，并将最终的结果结合概率p对应到真实的预测结果。Dropout策略使得模型，丢弃部分神经元，训练余下的神经元，迫使神经网络学习到更加具有鲁棒性的特征，增强了神经网络的泛化能力。通常情况下，我们将Dropout应用在卷积神经网络的全链接层，大小可以设为0.5，但是可以调整。

2.5.2 L1范数

L1正则化是指给通过L1范数，将参数加入到损失函数中，当训练模型时，也会对参数起到一定的约束。L1范数指矩阵中参数各个分量的绝对值之和。

2.5.3 L2范数

L2正则化又叫做权重衰减（weight decay）[20]，也被称作岭回归（ridge regression），L2范数求解如下

相比于L2范数,L1范数更容易产生稀疏解，即模型更容易解释，但是在神经网络的实际中，L2范数使用的更为广泛，且结果更好。

2.6 评价标准与模型选择

2.6.1 评价标准

MSE(mean squared error，均方误差)，回归任务中，我们使用MSE来衡量预测的结果与真实结果差别的一个评价标准，当MSE越大，代表预测值与实际值差的越大，表示预测效果不好。当MSE为预测值于真实值相同时，均方误差为0。

代表样本数量，代表通过深度学习方法预测得到的第个人的血糖值，代表原始数据中第个人经过体检测量得到的血糖值，最终计算的结果是将每个人预测差值平方后，再除以某个系数。由于数据来自与阿里天池的比赛，官方给出的评价标准是，所以沿用，之所以有是因为求导之后可以抵消系数。

2.6.2 交叉验证

交叉验证(cross validation)是一种评价模型好坏的方法。其思想将数据集均匀的划分为k份，每次取其中k-1份作为训练集，余下1份作为测试集。一次交叉验证是指使用k份数据，每一份都当做测试集测试其在具体指标上的分数，一共训练了k个模型，每个模型都有一个分数，可以取数学期望作为最终的分数。k是交叉验证划分的份数，具体可以结合数据大小设置k值。当k值等于数据集的大小时，就变成了留一法（leave one out），但留一法，运算量太大，所以一般情况下，取多次交叉验证平均作为该模型的分数。

2.6.3 网格搜索

网格搜索（grid search）是机器学习中训练模型时一种常用的调参方法。在机器学习中，模型有自身的超参数，超参数来控制模型学习参数的参数，通常需要人为指定。网格搜索通过穷举所有可能的参数，并通过交叉验证的方式对不同参数进行评估，最终给出所有列表中得分最高的参数。但网格搜索时间消耗太高，如果单个模型的训练时间过长，则网格搜索的时间代价是不可接受的。

所以为了克服网格搜索的缺点，随机搜索（random search）就出现了，随机搜索不是对每个参数进行网格划分，而是在

数据中的每个维度，随机选取某一个点，最后将所有的维度组合起来，就构成了随机的点。

2.6.4 文件格式

在本文中所有的数据集均使用CSV(comma-separated values，逗号分隔值)格式文件。对于CSV文件，其每一列都指代一个属性，属性之间使用逗号分割，属性中不能包含分隔符逗号。且其保存的数据，不会有额外的格式信息，所以可以直接使用记事本等文本阅读软件打开。

在数据预处理程序和模型训练之后输出的程序中，都将结果保存为CSV文件。

2.7 本章小节

首先，本章对介绍了糖尿病方面的背景知识，包括糖尿病的分类、不同糖尿病的区别、血糖预测的研究背景、血糖的评价指标、以及针对2型糖尿病的预防策略。

然后，开始介绍算法方面的背景知识，包括神经网络、卷积神经网络、实验的软硬件环境、反向传播算法、基于梯度下降的优化算法，以及神经网络中常用的正则化方法（如Dropout、L2范数等）。

最后，介绍了预测的评价指标均方误差MSE，以及实验中数据所保存的格式。

5. 第3章深度学习在糖尿病预测的研究

总字数：3705

相似文献列表　文字复制比：0.8%(29)　疑似剽窃观点：(0)

1	基于卷积神经网络和PCA的人脸识别 邢玲;冯倩;穆国旺; -《河北工业大学学报》- 2016-10-15	0.8%（29） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第3章深度学习在糖尿病预测的研究

3.1 数据来源

本次赛题源来自阿里天池大赛——人工智能辅助糖尿病遗传风险预测。本次实验所采取的数据是来自于天池精准医疗大赛初赛的数据。如图3.1所示，官方初赛针对比赛的不同时期，一共提供了3组数据，5个数据文件。

图3-1 数据文件

如表3-1所示，我们使用了初赛公布的数据集train、test a、test b，并将大赛公布的train作为训练集，test a作为验证集，test b作为测试集。

表3-1 文件用途

文件名官方用途条数内容简称

d_train_20180102.csv 训练集 5642 属性、标签 train

d_test_A_20180102.csv A榜测试 1000 属性 test a

d_answer_a_20180128.csv A榜答案 1000 标签 test a

d_test_B_20180128.csv B榜测试 1000 属性 test b

d_answer_b_20180130.csv B榜答案 1000 标签 test b

如图3-2、3-3、3-4，展示了数据的前五列。每个文件都是CSV格式的，其每一行代表一个受检者，第一列是受检者的id，依次对不同的属性都有值与其对应，共有42个属性，其中含有日期、性别等信息，但是部分的属性缺失，最后一列为血糖值，即需要通过模型预测的值。除去id和标签，每个受检者含有体检信息等40项。

这是一个回归任务，其特征为性别、年龄、体检日期等40项，标签是血糖值。我们的目标就是通过深度学习的技术使用40列属性来预测1列标记血糖值。

图3-2 数据展示1

图3-3 数据展示2

图3-4 数据展示3

3.2 矩阵补全技术

矩阵补全(matrix completion)技术旨在通过矩阵中已有的信息来恢复出其丢失的信息。其形式化描述如下：

其中， 表示需要恢复的稀疏信号； 表示矩阵的秩； 表示单个元素； 表示为空， 表示是已有的信息。

因为在集合上的凸包是的“核范数”(nuclear norm):

其中表示的奇异值，核范数指的矩阵的奇异值之和。理论研究表明，在满足一定条件时，若的秩为 ， ，则只需观察到个元素就可能完美恢复出[5]。

虽然核范数方法可以直接求解，但是当矩阵规模较大时，所面临的计算量非常大，常常难以计算，所以当面临大规模矩阵时，通常加上一个软阈值[16]，如下

3.3 数据归一化、标准化

归一化(Normalization)是指通过某种变换将数据映射到小的范围内，其目的是无量纲化，去除数据的单位带来的影响。归一化方法有：

1. Min-Max归一化，通过给所有的数据减去最小值，之后在对所有的数据除以极差，我们可以使原先的数据，不管相差多大，都可以落在01区间内部。其形式化描述如下

2. 对数函数转换

3. 反余切函数转换

标准化(Standardation)是将所有的数据按比例缩放，先求出原始数据中的均值和方差，之后对所有数据减去均值再除以方差，使之落入一个小的特定区间，这个区间通常是(-1,1)，这种方法就是Z-Score，将数据缩放为0均值，1方差

其中是均值， 是方差

3.4 数据预处理

由于数据集样本太少，而且对于受检者不论是低血糖、正常血糖还是高血糖，模型都需要对其进行预测，而且预测前并不知道患者处于哪一类人群，所以除了删除一个性别未知的样本之外，没有做其他的去噪处理。训练集数据中包含一个性别未知的样本，该样本id为580，对预测影响不大，故在预处理之前删除该样本，训练集剩余样本有5641个。数据预处理流程图如图3-5所示，数据预处理由以下步骤组成：

(1) 集合合并。因为数据中有很多空值，我们不希望丢失信息，所以对空值采取算法进行填充，而空值填充时，需要用到数据中蕴含的信息，数据越多填充得越好。而且合并后统一处理，不会出现训练集、验证集和测试集处理不一致的问题。所以对训练集、验证集和测试集进行合并为总集。

(2) 属性编码。属性中包含性别和日期，性别的内容是男或女，为汉字，日期的内容为日期型数据，如10/10/2017。这两者神经网络都不能直接处理，需要对其进行编码。对于性别，将其编码为0或1，“女”编码为0，“男”编码为1。对于日期，选出其中最小的日期，从当天算起，将所有的日期替换为据最小日期的天数。自属性编码之后，所有的值都变成了数值型或空值。

(3) 矩阵补全。经过集合合并之后，对总集进行矩阵补全，由于传统的核范数直接优化的方法，时间复杂度高，在大规模矩阵上不适用等原因，所以矩阵补全算法采用了Soft Impute[7]方法，该方法在大规模数据集上，具有速度快，精度高，稳定等优势。

(4) 归一化。对补全后的数据先按列进行Z-Score标准化，在对其按列进行Min-Max归一化。最终，以确保所有属性都在01区间内。

(5) 集合划分。将归一化后的集合，按照合并时的id划分成训练集、验证集和测试集。集合的划分是根据官方给出的train、test a和test b划分的。

图3-5 数据预处理流程图

数据预处理程序为preprocess.py，经过处理之后输出的数据集如图3-6所示，分别代表训练集、验证集和测试集，其每一行都是包含有43列，前41列为原始属性，最后两列代表标签，分别为血糖和血糖原值，前者代表归一化之后的血糖值，后者代表未归一化的。

图3-6 预处理结果

3.5 深度学习的血糖预测算法介绍

3.5.1 算法思想

深度学习是机器学习的一个分支，随着近几年的发展[7]，深度学习成为了机器学习领域的研究热点之一，本文试图尝试其在生物数据上是否同样适用，故打算深度学习的方法网络对糖尿病人的血糖值进行建模预测。

由于卷积神经网络需要输入2维的数据，而属性共有41个，没有空间结构，所以先移除id属性，因为id对于每一个患者来说都是不同的，所以没有预测意义，相反可能导致模型过拟合，因此还剩下40个属性。对40个属性按照其起始顺序按照5×8进行排列，顺序是从左到右，从上到下。输出为1个节点对应了血糖的预测值。

3.5.2 算法结构

图3-7 卷积结构图

如图3-7卷积结构图所示，卷积神经网络结构采用2层卷积、2层池化和6个全连接层。在本次卷积神经网络中所采用的卷积核的大小为3×3，步长为1×1，激活函数都是ReLU，之后连接上最大池化层，池化层的大小都是2×2，卷积和池化的边界填充都是SAME方式，即泛卷积，边界用0填充以保证卷积后图片的尺寸不会改变。第一层卷积深度为10，第二层卷积深度为15。卷积之后在通过扁平化，之后再连接6个全连接层，隐含层结点都是150，激活函数也都是ReLU。最后输出层只包含1个结点，对应了预测的结果，血糖值。

3.5.3 算法流程

算法运行的流程图如图3-8所示，算法先将预处理之后的数据，转换成具有空间结构的5×8的二维数据。之后在TensorFlow之下，构建卷积神经网络对应的计算图。之后在通过Adam优化器以0.001的学习率对其进行训练。最终，在测试集上测试其在测试集上的预测性能均方误差。

将数据展开成二维的具体方法是，首先去除id列，则数据余下40列数据，便直接对其既有的一维数据进行操作，取其0-8列作为第1行，9-16列作为第二行，等等。对训练集、验证集和测试集使用同样的方法进行操作。

图3-8 程序流程图

3.5.4 算法关键代码

如图3-9，为算法关键代码，这部分代码通过TensorFlow定义了本文中所使用的卷积神经网络的对应的计算图。

图3-9 关键代码

3.6 训练

算法的训练采用了Adam优化器[17]，学习率是0.001，每一次训练使用200个样本，使用全部样本训练结束为1轮，一共训练100轮，时间约为。在训练50轮左右达到最优，之后再训练，验证集的MSE会上升，时间约为20秒。

算法每一次迭代训练，都通过验证集测试其均方误差，并选择出验证集损失最小的迭代次数，对其计算测试集上的损失，并将其保存。

学习曲线如图3-10所示

图3-10 学习曲线

可以看出随着训练迭代的次数的增加，均方误差先降低，后来由于过拟合的原因而增大，则通过验证集选出最佳的迭代次数，在该次数上对测试集进行预测，并获得了测试集上的均方误差，过拟合的愿意也和模型的参数量太大有着密不可分的联系。最佳迭代次数为51，对应的训练集误差为0.8469，测试集误差为0.6546。

3.7 本章小结

本章主要说明了实验的流程以及每个步骤。

首先，说明了数据的来源，来自于阿里云和青梧桐基因联合主办的精准医疗大赛的初赛，并对数据做了一定的介绍，以及分析。

然后，说明了预处理的全过程，包括数据填充，标准化以及归一化等。

最后，对深度学习的血糖预测算法做了介绍，分别从卷积的结构以及训练的各个参数，损失函数的收敛过程。

6. 第4章实验结果及分析

总字数：1641

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第4章实验结果及分析

4.1 实验结果展示

在实验结果中，我们绘制了通过5641个样本训练的卷积神经网络模型，并使用验证集选择除了最佳的迭代次数，并在该迭代次数上，对测试集进行预测，最终将结果保存为CSV文件。

我们最后对实验结果进行绘图展示，横坐标表示样本号，纵坐标表示血糖值，蓝色表示真实的血糖值，橙色表示算法预测的血糖值。

验证集的预测结果如图4-1，可以看出模型偏好于血糖正常的人群，所以大部分的与测试都不会很大或者很小。将验证集按照真实血糖值排序后展示结果如4-2，可以看出，模型在中间人群中的预测值比较接近，但在左侧或者右侧，差别较大，尤其实在，高血糖时，模型虽然能够分辨一些高血糖的人群，但是找出的不够准确。深度学习在验证集MSE为0.8469。

图4-1 验证集预测结果

图4-2 验证集预测结果(排序)

训练集的预测结果如图3-6，将训练集按照真实血糖值排序后展示结果如3-7，根据排序后的测试集图，我们可以看出，模型在低血糖处，错误地将血糖值估计过高，有一个原因是数据量太小，导致验证集和测试集的分布有一定的差异。

最终在测试集上的均方误差MSE为0.6546。

图4-3 测试集预测结果

图4-4 测试集预测结果(排序)

对比验证集和训练集上的结果，我们发现在验证集的排序之后的对比图中，如图4-2中，预测结果为高血糖的样本大部分都集中在中部偏右，中部偏右代表其本身的血糖值就属于相对偏高的人群；而如图4-4，在测试集上，高血糖人群中，整体偏高之外，还有一部分人群，预测的血糖值为高，但其真实值散落在低血糖到高血糖的各个区间，这说明数据集太小，验证集和测试集有一定的差异。

4.2 实验结果分析

从实验结果来看，模型比较偏好血糖值正常的人群。算法对血糖值正确的人的预测相对精确一些，但是对于血糖值偏高或者偏低的非正常人群，预测精度较差。从图4-2中看以看出，当血糖值较高时，预测的血糖值也相应较高，但是模型预测的高的样本中，也有很多真实血糖值不高的样本。模型无法将这部分人群正确地预测为较低的血糖，说明对于模型而言，无法区分他们与高血糖的人群。有两种可能性，其一是，这部分人群和高血糖人群的受检信息一致而导致无法区分，那么进一步就可以说明，这部分人，应当采用2型糖尿病的第一级预防策略，在患病之前，通过合理地人为干预，从而降低未来患2型糖尿病地风险。另一种原因是模型地泛化能力太差，导致预测的结果不够好，可能因为数据量太小，数据中的属性没有采集全面，则可以进一步通过改进预测方法尝试解决这个问题。从回归的评价指标来看，血糖预测的精度仍然具有提高的空间。

本文以血糖值7.0为界限，将大于该值的样本认为其为2型糖尿病患者，作为正例；对小于该值的认为为正常人作为反例；将预测的结果以分类的评价标准对结果进行分析。在验证集和测试集的真实值中，分别包含87、54个正例，其余值均为反

例。在验证集和测试集的预测值中，分别包含了40、35个正例。

对验证集和测试集绘制混淆矩阵图分别为图4-5和图4-6，由图可以计算到验证集和测试集计算查准率（precision）分别为0.45和0.4，计算召回率（recall）分别为0.21和0.26。以传统的糖尿病判断方法对预测结果进行分类得到的查准率和召回率都不高，查准率低代表模型错误地将其中为患有2型糖尿病的患者判断为血糖值高，召回率低说明模型无法找出所有地2型糖尿病患者，这与正负样本不平衡有一部分原因。如果作为分类来评估的话，应该尽可能提高召回率，以确保尽可能多的2型糖尿病患者被预测出来，从而进一步可以发现更多的潜在患者，提高糖尿病的诊治率。

图4-5 验证集混淆矩阵图

图4-6 测试集混淆矩阵图

4.3 本章小结

本章通过对第3章的训练的深度学习模型预测的结果，进行可视化以及基本的分析，从回归和分类两方面进行分析，并指出了目前血糖预测仍存在的问题，并对出现的问题进行讨论。

7. 第5章总结与展望

第5章总结与展望

5.1 工作总结

糖尿病作为三大疾病之一的慢性非传染性疾病，其伴随着的并发症对人体危害非常大。然而，根据调查显示，我国的糖尿病患者知情率却极低。血糖预测可以作为一种临床建议，可以在糖尿病的第一级预防中发挥作用，提供给医生建议，从而帮助提高糖尿病的知情率。

目前有很多的血糖预测研究，都是基于CGMS提供的糖尿病患者的历史糖尿病序列，从而推断未来一段时间内糖尿病患者可能出现的血糖，进而给糖尿病患者起到一定的预警作用。但是这种方法不能够提高糖尿病的知情率，所以本文采取病人的受检信息，旨在通过受检信息来预测病人的血糖值，从而提供给病人一定的意见。进而病人可以采取全面的检查，以确定是否患有糖尿病，如果为患病，则应该加强预防。

本文使用了阿里天池的比赛——精准医疗大赛初赛的数据，该数据包含了2型糖尿病患者和正常人的体检信息以及血糖值。

本文首先对其进行数据预处理，先将日期型数据和性别等信息转换为数字型的，因为数据中有部分体检者没有检查部分体检项，所以采取了矩阵补全技术对其缺失值进行填充，之后对其通过归一化等手段进行缩放到01区间，预处理过程结束。

之后通过深度学习的方法，基于TensorFlow框架构建了一套卷积神经网络模型。之后将预处理得到的数据输入到卷积神经网络中，对其进行迭代地训练，在训练的过程中，使用验证集选取最优的迭代次数，并对测试集进行预测。并将最优的验证集，以及该迭代次数下预测的测试集保存到本地。

最后，对模型预测的血糖值与病人真实的血糖值进行对比。结果发现有一部分人群，深度学习预测其为高血糖，但其真实的体检数据为正常或者低血糖，并对此展开讨论。

5.2 问题与展望

在针对糖尿病的预测的结果中，我们可以看到不论是验证集还是训练集的预测值，在血糖值很高的时候往往难以预测，预测的结果与真实的血糖值相差也很大，不仅如此，还有很多血糖值正常的人的血糖值被预测的很高。

另外，关于那些血糖值实际不高，但预测为高血糖的人，他们目前未患有糖尿病，以后5年内是否会患有糖尿病，仍未可知。如果有跟踪数据，就可以根据跟踪数据判断，这部分人群的预测为高血糖是因为他们的身体指标以及体检信息和糖尿病人的指标和信息相近，所以预测为高血糖。还是，仅仅是模型的问题。

模拟预测的效果不够好。数据集太小，深度学习是一种数据饥饿的算法，通常情况下，在数据量很大的时候其预测效果会很好。所以，需要结合小样本来使用新的深度学习的方法重新建模、预测；下一步还需要考虑数据是否再神经网络上同样适用；并且还要考虑简化卷积神经网络的模型。

本文是以回归的方法对血糖进行建模预测，实验结果以均方误差作为衡量。下一步可以通过分类方法对2型糖尿病患者和正常人进行分类，并通过提高召回率期望模型尽可能地预测出患有糖尿病的人群，并给出假正例（应当为反例，但是模型错误地将其分类为正例）的人群，这部分人群很有研究意义，可能起的身体体质等于糖尿病人类似，但是目前仍未患有糖尿病，应对其进行干预，从而降低或避免其患2型糖尿病的概率。模型预测为反例的人群应该是安全人群，不需要像假正例一样，到人为干预的阶段。

参考文献

[1] Xu Y, Wang L, He J, et al. Prevalence and Control of Diabetes in Chinese Adults[J]. JAMA. 2013;310(9):948–959.
[2] 贾伟平. 中国2型糖尿病防治指南(2017年版)[J]. 中华医学会糖尿病学分会, 2017.
[3] 基于CEEMDAN-ELM的短期血糖预测模型研究[D]. 郑州大学, 2017.

[4] 莫雪. 数据驱动的血糖预测方法研究[D]. 北京化工大学, 2014.郭占丽.
[5] 余丽玲, 陈婷, 金浩宇,等. 基于支持向量机和自回归积分滑动平均模型组合的血糖值预测[J]. 中国医学物理学杂志, 2016, 33(4):381-384.
[6] 廖涌. 中国糖尿病的流行病学现状及展望[J]. 重庆医科大学学报, 2015(7):1042-1045.
[7] 陈倩, 叶俊兵. 糖尿病研究综述[J]. 养生保健指南, 2017(5).
[8] 周志华. 机器学习 : = Machine learning[M]. 清华大学出版社, 2016.
[9] org.cambridge.ebooks.online.book.Author@ea. Deep Learning[M].
[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.
[12] 卢宏涛, 张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(1):1-17.
[13] Sanders J, Kandrot E. CUDA by Example: An Introduction to General-Purpose GPU Programming[M]. Addison-Wesley Professional, 2010.
[14] Chetlur S, Woolley C, Vandermersch P, et al. cuDNN: Efficient Primitives for Deep Learning[J]. Computer Science, 2014.
[15] Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning[J]. 2016.
[16] Mazumder R, Hastie T, Tibshirani R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices[J]. Journal of Machine Learning Research Jmlr, 2009, 11(11):2287.
[17] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
[18] 陶俊杰，陈晓莉.Python科学计算基础教程[M].人民邮电出版社，2017.
[19] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
[20] Jones M, Poggio T. Regularization Theory and Neural Networks Architectures[J]. Neural Comp, 1995, 7(2):219-269.
[21] Fahlman S E. Faster-Learning Variations on Back-Propagation: An Empirical Study[J]. Proceedings of the Connectionist Models Summer School Morgankaufmann, 1988.
[22] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9:249-256.