基于机器学习方法的糖尿病预测分析

目录

摘 要

糖尿病是一组因为胰岛素不正常分泌、生物作用受损引起的终生性代谢性疾病。糖尿病的患者会饱受包括糖尿病肾病、糖尿病神经病变在内的多种并发症，目前还没有根治的治疗方案。我国由于人口多、医疗资源相对欠缺，有着世界上最多糖尿病患者，病人达到1.1亿，每年用于糖尿病的医疗费用占中国公共医疗卫生支出的比例超过13%Error! Reference source not found.，超过3000亿元。流行与防控形势严峻，进行大规模糖尿病筛查难度巨大。而机器学习方法可以通过处理大量的医疗数据获取较好的预测性能，机器学习在未来可广泛应用于医疗领域。运用机器学习算法，构建糖尿病预测模型，实现糖尿病预测，增强我国糖尿病筛查能力，对国民健康状况具有重大意义。

为了构建精确度更高、鲁棒性更好的基于机器学习算法的糖尿病预测模型，本文使用University of California, Irvine（UCI）机器学习数据库中的早期糖尿病风险预测数据集、Kaggle中的皮马印第安人糖尿病数据集和来自天池精准医疗大赛复赛的妊娠期糖尿病（GDM）数据集共三个数据集，并对其进行如下工作：（1）对数据集进行数据描述；（2）对数据使用多种方法进行预处理：使用多种方案数据清理，使用IV

值分析和残差分析两种方法筛选变量；（3）对早期糖尿病风险预测数据集和皮马印第安人数据集使用支持向量机（SVM）、神经网络、XGBoost算法构建预测模型，对天池数据集使用神经网络、XGBoos构建预测模型；（4）使用正确率、精确率、召回率和F1值等评价指标对模型进行评价。

关键词：机器学习；糖尿病；数据挖掘

第 1 章 绪论

1.1 研究背景

糖尿病是由内分泌不足和内分泌利用紊乱引起的以胰岛素不正常分泌、生物作用严重受损症状为主要标志的一组代谢性疾病。随着人们生活水平的提升，糖尿病也从一种不常见的疾病变成了流行病。国际糖尿病联合会发布的数据显示，在2019年共有4.63亿20-79岁的成年人患有糖尿病，且预计这个数字将在2030年达到5.78亿。糖尿病是终生性疾病，具有有超过160种并发症，并发症包括眼部、肾、神经和心血管等组织的慢性病变，对患者的生活造成巨大的影响。糖尿病及其并发症致死率高，每年约有460万人死于这种疾病，平均每7秒就有人因此丧生。

目前糖尿病没有完全治愈的治疗方案，患者需要持续服用降糖药物以促进身体糖代谢，需要花费数额不小的金钱用于糖尿病的治疗和长期服用药物，对糖尿病的预防便显得尤为重要。高危人群是糖尿病发展三个阶段中的第一个阶段，后两个阶段分别为糖尿病前期和糖尿病。糖尿病前期发展为糖尿病的可能性极大，糖尿病前期人群是预防糖尿病的重点人群。目前我国仅糖尿病前期人群就已接近5亿人，在全国14亿人口中占有非常大的比例。对糖尿病高危人群以及糖尿病前期人群进行干预是预防糖尿病的最有力措施。通过控制每日摄入食物总热量，进行高纤维饮食、清淡饮食，进行一定时间和强度运动可以改善胰岛素抵抗，从而预防糖尿病的发展。

在数字化生活逐渐普及的今天，海量的数据正在产生。人们越来越重视数据，政府和企业也开始收集数据，这其中就包括医疗数据。数据驱动的机器学习算法通过学习医疗数据为医疗服务提供快捷和高质量的途径，在医疗领域的应用具有广泛前景。

1.2 研究目的和意义

糖尿病患病人口数量大，2019 年我国糖尿病患者已达 1.164 亿人。国家防控糖尿病工作压力大，因为医疗资源相对缺乏，现阶段国家对潜在的糖尿病人群筛查力度不足，糖尿病的防治形势不容乐观。同时，糖尿病的检测与治疗的工作会需要消耗大量资金，投入大量的人力资源和医疗资源。使用机器学习方法构建糖尿病预测模型对于我国糖尿病防治工作具有重要意义。

本研究通过对多个已有的糖尿病数据库进行数据挖掘，尝试多种数据处理方式，基于机器学习方法构建精准度高、鲁棒性强、速度快的模型，分析出对糖尿病影响大的因素。构建高性能的糖尿病预测模型能够辅助医生进行临床决策，减少检测和治疗成本，减轻糖尿病筛查工作的压力，提高糖尿病筛查的速度和精准度，有利于我国对糖尿病的防治工作。

1.3 国内外研究现状

1.3.1 国外研究现状

糖尿病作为一种终生性疾病严重影响人类身体健康，糖尿病筛查工作至关重要。为了更好确定需要筛查的人群和检测的内容，国外学者已经对糖尿病的影响因素和糖尿病的预测方法进行研究。Jorge Chavarro**Error! Reference source not found.**研究了 8722 名女性中存在的 112 种单核苷酸多态性（SNP）和妊娠期糖尿病（GDM）遗传变异的关系。Mohamed S F**Error! Reference source not found.**使用 2015 年肯尼亚全国 18-69 岁成年人的调查数据，用逻辑回归确定相关因素，指出年龄较大（60-69 岁）和血压升高与糖尿病相关，女性超重或肥胖与糖尿病相关。Mokdad A H**Error! Reference source not found.**的研究成果表明超重和肥胖与糖尿病显著相关，BMI 超过 40 的成年人的糖尿病发生比（Odds Ratio，发生概率与不发生概率的比率）为 7.37。Arianna Dagliati**Error! Reference source not found.**使用随机森林构建了考虑性别、年龄、诊断时间、体重指数（BMI）、糖化血红蛋白（HbA1c）、高血压和吸烟习惯共 7 个变量的糖尿病预测模型。

1.3.2 国内研究现状

我国糖尿病防治形势严峻，但国内学者在糖尿病及其预测方面的研究也取得了一定的进展。茅红艳等人 **Error! Reference source not found.**通过对 10336 例妊娠期糖尿病病例的研究发现存在 8 个遗传多态性与妊娠期糖尿病显著相关。苏健等人 **Error! Reference source not found.**通过多阶段整群随机抽样方法获取数据，指出 BMI、腰围与糖尿病高度相关。洪烨 **Error! Reference source not found.**使用 BP 神经网络、支持向量机（SVM）和 AdaBoost 算法对哈尔滨工业大学 2014 年体检数据进行学习训练构建出糖尿病预测模型。

1.4 研究内容

本文在对糖尿病及其预测的相关文献进行分析的基础上，以 UCI 中的早期糖尿病风险预测数据集、Kaggle 中的皮马印第安人糖尿病数据集和天池精准医疗大赛妊娠期糖尿病数据集，使用机器学习方法建模，针对糖尿病预测问题进行分析研究，具体研究内容如下：

（1）从众多机器学习算法中选择支持向量机、BP 神经网络和 XGBoost 算法进行理论分析，通过不断学习加深对算法的理解，并用以作为后续研究的建模算法。

（2）收集数据集并分别分析，对数据集使用多种处理方案获得多个数据集，用于后续模型的输入。

（3）基于支持向量机、BP 神经网络和 XGBoost 算法对不同的数据集分别构建糖尿病预测模型，对测试数据使用模型进行预测，对不同模型的预测结果计算正确率、精确率、召回率 F1 值作为数据处理方案和模型的评价标准。

论文整体研究路线如图 1-1。

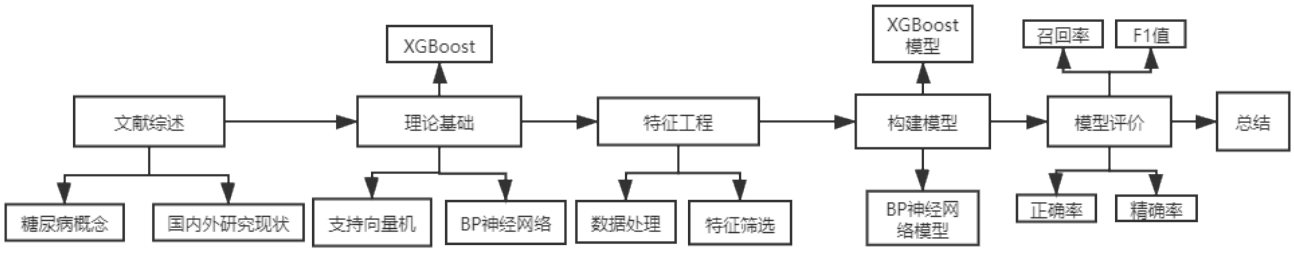

图 1-1 论文整体研究路线

第 2 章 基础算法描述

2.1 支持向量机原理

支持向量机（SVM）是一个广义线性分类器，其思想是寻找一个最好的超平面将样本划分为正类样本和负类样本，如图2-1所示，直观上最好的超平面应该是位于两类样本正中间的粗线代表的超平面。

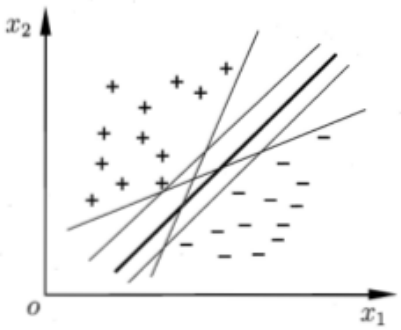

图2-1 区分正负类样本的超平面

设训练样本集为$D=\{(x_1,y_1),...,(x_m,y_m)\}$，$y_i\in\{-1,+1\}$，则超平面可表达为：

$$\omega^T x+b=0 \tag{2.1}$$

其中$\omega=(\omega_1;\omega_2;...;\omega_d)$决定了超平面的方向；$b$为位移项，决定了超平面与原点之间的距离。若将超平面记为(ω,b)，则点x_i及该点到超平面的距离d，有：

$$(\omega^T x+b)*y_i\geq 1 \tag{2.2}$$

$$d=\frac{|\omega^T x+b|}{||\omega||} \tag{2.3}$$

训练集中使得距离d取最小值的一些样本被称为支持向量。取正类样本与负类样本的两个支持向量，计算它们到超平面的距离之和，计算的结果被称为间隔γ，其表达式为：

$$\gamma=\frac{2}{\|\omega\|} \tag{2.4}$$

那么寻找最好的超平面的问题可转化为寻找ω和b使得满足式（2.4）的同时令间隔γ取到最大值，即

$$\begin{aligned}&\max_{\omega,b}\frac{2}{\|\omega\|}\\ &\text{subject to}\quad y_i(\omega^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1,i=1,2,...,m.\end{aligned} \tag{2.5}$$

当γ取最大值，即$\frac{2}{||\omega||}$取最大值时，问题可被转化为找到满足式（2.2）的ω和b使$\frac{1}{2}||\omega||^2$取最小值，即

$$\begin{aligned}&\min_{\omega,b}\frac{1}{2}\|\boldsymbol{w}\|^2\\ &\text{subject to}\quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1,i=1,2,...,m\end{aligned} \tag{2.6}$$

式（2.6）是一个二次规划（quadratic program, QP）问题。目前已有现成的工具包可以解决QP问题，但为了提高算法的性能，我们可以对问题进一步细化，将这个二次规划问题使用拉格朗日乘子法转化为对偶问题（dual problem）。对式（2.6）中的每条约束添加拉格朗日乘子$\alpha_i\geq 0$，可写出该问题的拉格朗日函数：

$$L(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m}\alpha_i(1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)) \quad (2.7)$$

令其对ω和b的偏导为0，得

$$\omega = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \quad (2.8)$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i \quad (2.9)$$

将式（2.8）和式（2.9）代入式（2.7）即可消去ω和b得到式（2.6）的对偶问题

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \quad (2.10)$$

$$\text{subject to } \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\alpha_i \geqslant 0, i = 1,2,...,m$$

解出 α 后，求出 ω 与 b 即可得模型

$$f(\boldsymbol{x}) = \omega^{\mathrm{T}}\boldsymbol{x} + b = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} + b \quad (2.11)$$

为了对偶问题式（2.10）我们可以使用 SMO 算法（Sequential Minimal Optimization）**Error! Reference source not found.**。SMO 的基本思路是在$\sum_{i=1}^{m}\alpha_i y_i = 0$的约束下固定$\alpha_i$之外的所有参数，用其他变量线性表示$\alpha_i$上的极值。SMO 每次选择两个变量$\alpha_i$和$\alpha_j$并固定其他参数求解式（2.10）更新$\alpha_i$和$\alpha_j$的值，并重复这一步骤直至收敛。

以上求解方法是以训练集线性可分为前提，在数据集线性不可分的条件下并不适用。但倘若存在一个高维空间，在此空间内样本经过映射之后线性可分，则前文的求解方法只需稍作修改即可适用。将样本映射于这个空间，找到一个函数$\theta(x)$表示 x 在更高维空间里的特征向量，于是前文关于支持向量机的公式中的 x 都可用$\theta(x)$替换。但由于映射后的空间维数可能非常高，在求解用$\theta(x)$替换后的对偶问题时涉及到$\theta(x)^T\theta(x)$的计算可能变得非常困难，我们尝试设想一个函数

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \theta(\boldsymbol{x}_i)^{\mathrm{T}}\theta(\boldsymbol{x}_j) \quad (2.12)$$

这样我们就不必直接计算高维空间中计算$\theta(x)^T\theta(x)$，直接用$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$对原公式中的$\theta(x)^T\theta(x)$进行替换，得到最终需要求解的对偶问题：

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) \quad (2.13)$$

$$\text{subject to } \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\alpha_i \geqslant 0, i = 1,2,...,m$$

求解后得到模型：

$$f(\boldsymbol{x}) = \omega^{\mathrm{T}}\theta(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i \kappa(x, \boldsymbol{x}_i) + b \qquad (2.14)$$

而$\kappa(\bullet, \bullet)$则被称为核函数。若我们能够知道$\theta(x)$的形式，则也能核函数的具体形式，但这一点我们通常无法做到，故我们通常在构建模型时自行选择核函数。表 2-1 为常见的几种核函数。

表 2-1 常见核函数

核函数	表达式	参数要求
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$	$d \geq 1$
径向基核(高斯核)	$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{\| \boldsymbol{x}_i - \boldsymbol{x}_j \|^2}{2\sigma^2}\right)$	$\sigma > 0$
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\| \boldsymbol{x}_i - \boldsymbol{x}_j \|}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh\left(\beta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \theta\right)$	$\beta > 0, \theta < 0$

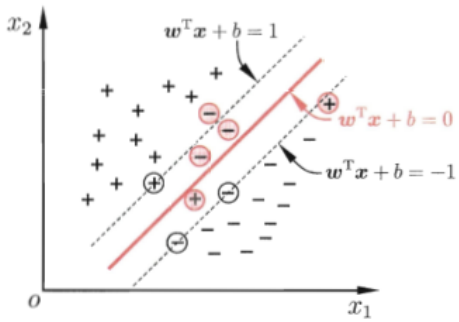

图 2-2 软间隔示意图

因为我们通常无法真正知道适合模型的核函数的具体形式，我们需要允许模型在少量样本上发生错误，因此我们需要设置“软间隔”。软间隔允许部分样本不满足式（2.2），但在使间隔取最大值时也应减少不满足条件的样本。则优化目标可写为：

$$\min_{\boldsymbol{\omega}, b} \frac{1}{2} \| \boldsymbol{\omega} \|^2 + C \sum_{i=1}^{m} \ell(y_i(\boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{x}_i + b) - 1) \qquad (2.15)$$

其中 C 为大于 0 的常数，ℓ为损失函数。下表为常用的三种损失函数。

表 2-2 常用损失函数

损失函数	表达式
hinge 损失	$\ell_{hinge}(z) = max(0, 1-z)$
指数损失	$\ell_{exp}(z) = exp(-z)$
对率损失	$\ell_{log}(z) = log(1 + exp(-z))$

引入松弛变量ξ_i，则可将式（2.15）改写为式（2.16），得到的式子与式（2.6）类似，同样使用拉格朗日乘子法得到对偶问题，并用 SMO 算法求解即可。

$$\min_{\boldsymbol{\omega},b,\xi_i} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{m}\xi_i \quad (2.16)$$
$$\text{subject to} \quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1-\xi_i$$
$$\xi_i \geqslant 0, i = 1,2,\ldots,m$$

2.2 BP 神经网络原理

人工神经网络是一种模仿生物学神经网络结构的模型，其灵感来源于人类脑部中枢神经系统，由大量神经元组成。BP 神经网络（Back Propagation Neural Network）一般指用 BP 算法训练的多层前馈神经网络。

2.2.1 神经元模型

生物神经网络中，一个神经元在激活状态下会改变与其相连的神经元的电位，而当神经元的电位超过了一个阈值时，这个神经元也会处于激活状态。一直沿用至今的“M-P 神经元模型”将这种情况进行抽象化，设有一个神经元与另外 n 个神经元相连，$x_1,x_2,\ldots,x_n$ 为与其相连的神经元对该神经元的输入，$\omega_1, \omega_2, \omega_n$ 对应神经元输入的权重，θ 为该神经元的阈值，则该神经元的输出 y 为

$$y = f(\sum_{i=1}^{n}\omega_i x_i - \theta) \quad (2.17)$$

其中 f(・)又被称为激活函数。“M-P 神经元模型”的原理可由图 2-3 表示，而神经网络则是由多个这样的神经元按一定层次连接起来得到的。

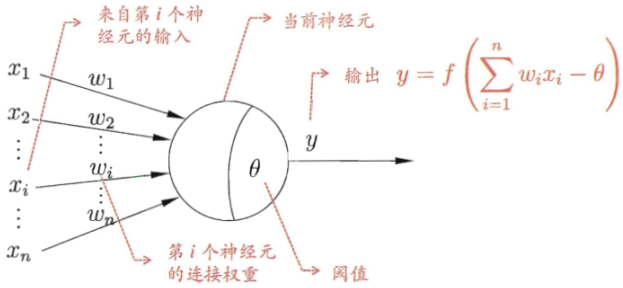

图 2-3 M-P 神经元模型

2.2.2 多层前馈神经网络

多层前馈神经网络指具备输入层、隐含层和输出层的三层以上的神经网络，其中输入层和输出层只有一层，隐含层可有一层或多层。常见的多层前馈神经网络中同层神经元互相不连接，且神经元只与下一层的神经元进行连接。在多层前馈神经网络中，输入层的神经元没有激活函数，输入与输出完全相同；隐含层和输出层的神经元是 M-P 神经元，具有激活函数。对于简单的多层前馈神经网络（如单隐层网络和双隐层网络），影响最终输出结果的因素有输入的数据、神经网络的结构（如网的层数、每一层的神经元个数）、每层神经元激活函数、神经元之间的连接权重、每个 M-P