

基于GCC-KNN算法的病人糖尿病遗传分类预测方法

吴守政¹, 周鹏举²

¹阳光融和医院, 山东 潍坊

²四川大学网络空间安全学院, 四川 成都

收稿日期: 2024年9月12日; 录用日期: 2024年10月16日; 发布日期: 2024年10月28日

摘要

糖尿病是世界上常见的慢性病, 及时查验和治疗非常有必要。在目前数字化背景下, 将机器学习与医疗安全相结合具有重要意义。针对医院检查人群中潜在糖尿病病人遗传分类识别进行研究。由于糖尿病病人分类识别的特征值较多且密集, 且本文主要的需求是保证其识别的准确率, 故在KNN算法的基础上进行改进, 使用了改进的GCC-KNN模型来对其进行分类识别, 通过网格搜索优化算法确定了K值的最优参数, 以及将不同距离进行对比, 选择了该模型的最优距离切比雪夫距离, 实现了医院检查人群中潜在糖尿病病人遗传的初步划分。经过实验对比, GCC-KNN模型准确率在潜在糖尿病病人遗传分类识别中均优于其他对比模型。

关键词

糖尿病, 机器学习, 网格搜索与交叉验证, 切比雪夫距离, GCC-KNN

Genetic Classification and Prediction Method for Patients with Diabetes Based on GCC-KNN Algorithm

Shouzheng Wu¹, Pengju Zhou²

¹Sunshine Union Hospital, Weifang Shandong

²School of Cyber Science and Engineering, Sichuan University, Chengdu Sichuan

Received: Sep. 12th, 2024; accepted: Oct. 16th, 2024; published: Oct. 28th, 2024

Abstract

Diabetes is a common chronic disease in the world, so it is necessary to check and treat it in time. In

the current digital context, combining machine learning with medical security is of great significance. The genetic classification and identification of potential diabetes patients in the hospital inspection population were studied. Because the feature values for classification and recognition of diabetes patients are more and more intensive, and the main demand of this paper is to ensure the accuracy of their recognition, it is improved on the basis of the KNN algorithm, using the improved GCC-KNN model to classify and recognize them, determining the optimal parameters of K value through the grid search optimization algorithm, and comparing different distances, selecting the optimal distance Chebyshev distance of this model, realizing the preliminary division of the genetics of potential diabetes patients in the hospital inspection population. Through experimental comparison, the accuracy of GCC-KNN model is superior to other comparison models in genetic classification and recognition of potential diabetes patients.

Keywords

Diabetes, Machine Learning, GridSearchCV, Chebyshev Distance, GCC-KNN

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

糖尿病[1]等各种慢性病的患病率现阶段逐年提高,单一的医疗卫生服务体系难以有效满足人民群众的健康需求。将治病和防病结合起来,重点关注早期居民的糖尿病问题,有利于尽快治疗,高效治疗,最大程度地减少居民慢性病影响健康问题。近些年来,我国糖尿病患病人数高速增长,糖尿病防治成为多地开展数字化医疗防控的重点。国内外许多专家关于如何有效及时预防和及时治疗糖尿病展开了研究,主要考虑聚焦于如何根据去医院检查人员在相关的各项指标情况下融合数字化背景以及人工智能有效地发现潜在的糖尿病患者,从而更好地及时进行相关治疗,减少糖尿病对于人群的影响。而发现潜在的糖尿病患者的关键的一个方面是对糖尿病患者的遗传风险的预测。通过这一方面的考虑,我们可以通过对于进行各项检查的人群进行糖尿病的分析 and 预测,将未确认糖尿病但指标不正常有糖尿病趋势的人群进行潜在的糖尿病居民的划分,分为潜在糖尿病人群和非糖尿病人群。

目前在大数据技术和人工智能技术飞速发展下,越来越多的机器学习方法被运用到人群的糖尿病预测和分类研究中。在糖尿病分类预测研究中,刘巧红[2]提出了基于 XGBoost 算法的糖尿病分类预测模型及应用,胡月[3]提出了基于 Stacking 集成学习的糖尿病风险预测研究,凌雄娟[4]提出了基于机器学习算法如 XGBoost 和逻辑回归算法的糖尿病预测,吴纵凌使用基于卷积神经网络对糖尿病进行了预测,薛博元[5]通过改进 CatBoost 算法对糖尿病回归和分类预测进行模型研究。

此次论文研究以糖尿病遗传风险预测数据集(天池精准医疗大赛)为研究对象,以根据病人的基本情况指标、乙肝指标、血常规指标、肝功能指标、肾功能指标等构建出病人潜在糖尿病遗传风险预测的评价特征。根据评价特征构建了 GCC-KNN 模型(GridSearch CV and Chebyshev Distance, K-Nearest Neighbor),使用了 KNN 算法作为基础,网格搜索及交叉验证方法[6]寻找参数的最优值,确定 K 值的最优值。根据不同的距离公式[7]进行各项评价指标对比,决定了此论文中使用的距离公式确定为切比雪夫距离。构建好 GCC-KNN 模型后对糖尿病遗传风险预测进行相应的分类预测,从而将糖尿病病人遗传风险分类分为潜在糖尿病人群和非糖尿病人群。各项评价指标表明该模型在糖尿病遗传风险预测方面具有较好效果,有效为潜在糖尿病人群提供预警,辅助医生进行早期干预及诊断,从而降低糖尿病发病风险。

2. 相关技术基础

2.1. KNN 算法

KNN 算法主要用于机器学习方面, 是属于比较基础的算法。它可以用于分类、还可以用于预测。它是通过根据不同的距离来进行特征值的分类。KNN 方法如下: 对于每一个输入的向量, 就是空间的每一个点, 输出值为其对应的预测标签。KNN 不需要普通的学习过程, 它在训练数据完成后, 对其进行相应的划分, 划分后的标签作为最终的预测结果。后面输入无标签数据, 无标签数据的特征与之前的测试集数据的特征进行相似度测试, 再获取测试集中特征最接近的数据的标签。KNN 中有一个 K 值参数, 即选择训练集中 k 个特征, k 一般来说是小于等于 20。

2.2. GridSearchCV 算法(网格搜索和交叉验证算法)

网格搜索属于一种遍历方法, 从而寻找使用的模型的最合适的参数。交叉验证将论文中要用到的数据集进行划分, 如分为 n 份, 在这 n 份中选择其中的一份用作测试, 剩下的 n-1 份用来训练, 因为我们使用的模型中包含了一些超参数, 这些超参数对于算法的结果影响很大, 因此我们要组成不同的参数, 从而找到最合适的组合参数来构建好相应的模型。它的原理是只需要将若干参数传递给网格搜索对象, 它自动帮我们完成不同超参数的组合、模型训练、模型评估, 最终返回一组最优的超参数。交叉验证解决模型的数据输入问题(数据集划分)得到更可靠的模型, 网格搜索解决超参数的组合, 构成一个模型参数调优的解决方案。

2.3. 多种距离公式

2.3.1. 欧几里得距离(Euclidean Distance)

点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 欧几里得距离为公式(1)所示。

$$d_{12} = \sqrt{\sum_{k=1}^n (x_{1k} - x_{2k})^2} \quad (1)$$

2.3.2. 曼哈顿距离(Manhattan Distance)

点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 曼哈顿距离为公式(2)所示。

$$d_{12} = \sum_{k=1}^n |x_{1k} - x_{2k}| \quad (2)$$

2.3.3. 切比雪夫距离(Chebyshev Distance)

点 $a(x_{11}, x_{12}, \dots, x_{1n})$ 和 $b(x_{21}, x_{22}, \dots, x_{2n})$ 切比雪夫距离为公式(3)所示。

$$d_{12} = \max(|x_{1i} - x_{2i}|) \quad (3)$$

2.3.4. 汉明距离(Hamming Distance)

采用两个字符串 s1 和 s2, 这两个字符串长度相同, s1 和 s2 的距离: 将其中一个变为另外一个所需的最小字符替换次数。

3. 数据特征及分析

3.1. 数据集

此次论文的研究数据集以糖尿病遗传风险预测数据集(天池精准医疗大赛)为研究对象。我们使用的这个数据集有 5642 份数据, 每一个病人存在不同的 41 个特征, 特征主要包含了病人的各种基本身体情

况，比如基本情况指标、肝功能指标、生化指标、乙肝五项指标、血常规指标以及血糖目标变量等主要的指标。我们对这份数据集进行了一定的预处理，然后根据需求进行了各种不同的算法的试验来进行比较和对比。表 1 详细展示了数据集病人信息的特征变量类别说明。

Table 1. Dataset patient information

表 1. 数据集病人信息

特征类型	特征名称
基本情况指标	Id、性别、年龄
肝功能指标	天门冬氨酸氨基转换酶、碱性磷酸酶、r-谷氨酰基转换酶、丙氨酸氨基转换酶、总蛋白、球蛋白、白蛋白、白球比例
生化指标	甘油三酯、高密度脂蛋白胆固醇、低密度脂蛋白胆固醇、总胆固醇、肌酐、尿酸、尿素
乙肝五项指标	乙肝表面抗原、乙肝表面抗体、乙肝 e 抗原、乙肝 e 抗体、乙肝核心抗体
血常规指标	白细胞计数、红细胞计数、血红蛋白、红细胞平均血红蛋白量、红细胞压积、红细胞体积分布宽度、红细胞平均体积、血小板体积分布宽度、红细胞平均血红蛋白浓度、血小板计数、血小板平均体积、血小板比积、嗜酸细胞%、淋巴细胞%、中性粒细胞%、单核细胞%、嗜碱细胞%

3.2. 数据平衡性分析

本数据集中男女性别比例基本一致，所以不存在性别对数据平衡性的影响(图 1)。

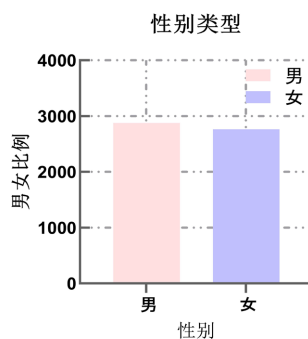


Figure 1. Sex type

图 1. 性别类型

4. 面向病人糖尿病遗传分类预测模型——GCC-KNN 模型

4.1. 模型框架

针对病人特征指标信息较多，且需要保证病人糖尿病遗传的预测准确率的情况下。本章节结合基于切比雪夫距离和网格搜索和交叉验证算法的 GCC-KNN 算法，用于进行潜在病人糖尿病遗传分类预测(图 2)。

4.2. 基于网格搜索和交叉验证算法对最优参数 K 值的确定

使用了相应的优化算法对此次论文中的 k 值参数进行参数的最优值的寻找，根据目的需要，设置了 K 值参数选择的范围是 1~20 之间的整数，经过 GridSearchCV 对 K 参数寻找后，本篇文章中选择的 K 值参数最合适值确定为 18。

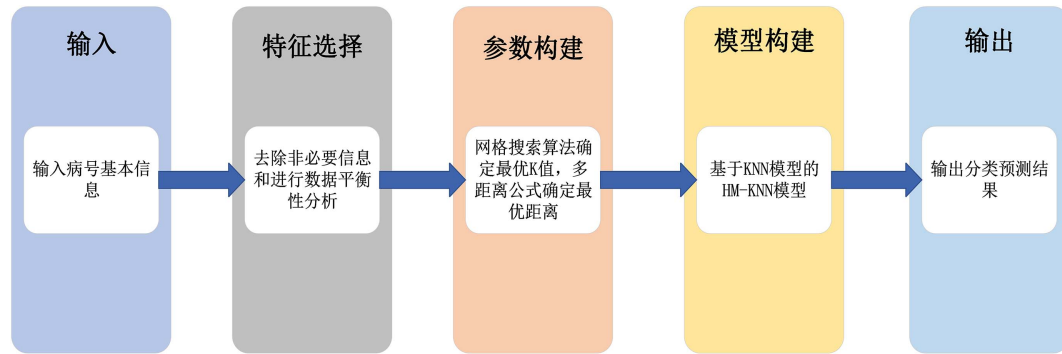


Figure 2. GCC-KNN algorithm model diagram

图 2. GCC-KNN 算法模型图

4.3. 基于多种距离公式确定最优距离公式

使用了各种不一样的距离放在 KNN 算法中，运行后再通过检验的标准来进行选择。此篇文章中需求是为了对糖尿病人群遗传风险进行预测和评估，在对准确率、精准率、召回率，F1 值等进行计算的过程中，着重对准确率这项指标进行了选择。使用欧几里得距离、曼哈顿距离、汉明距离、切比雪夫距离这四个不同的距离进行了糖尿病人群遗传风险的评估，得出的准确性判断结果中，切比雪夫距离最适合本篇文章的需要。

5. 面向病人糖尿病遗传分类预测模型——GCC-KNN 模型

5.1. 实验对象和具体设计

我们的实验对象是糖尿病遗传风险预测数据集(天池精准医疗大赛)，选用 70%用作我们的训练，选用 30%用作我们的测试。以 KNN (欧几里得距离)、C-KNN (切比雪夫距离)、M-KNN (曼哈顿距离)、H-KNN (汉明距离)、G-KNN (网格搜索及交叉验证与欧几里得距离)、GM-KNN (网格搜索及交叉验证与曼哈顿距离)、GH-KNN (网格搜索及交叉验证与汉明距离)、GCC-KNN (网格搜索及交叉验证与切比雪夫距离)用在糖尿病遗传人群分类预测中去，具体步骤如下采用，对 Accuracy (准确率)和 Precision (精准率)、Recall (召回率)、F1-Score (F1 值)排序来综合考量，以及主要考量潜在糖尿病人群遗传风险评估的准确率。

5.2. 实验结果分析

Table 2. KNN Algorithm and Accuracy

表 2. KNN 算法及准确率

算法	准确率	精准率	召回率	F1 分值
基于欧几里得距离的 KNN 算法	0.832	0.343	0.067	0.112
基于切比雪夫距离的 KNN 算法	0.833	0.333	0.056	0.096
基于曼哈顿距离的 KNN 算法	0.834	0.333	0.073	0.119
基于汉明距离的 KNN 算法	0.829	0.316	0.067	0.111
基于网格搜索算法和欧氏距离的 KNN 算法	0.841	0.400	0.011	0.022
基于网格搜索算法和切比雪夫距离的 KNN 算法	0.845	0.800	0.022	0.044
基于网格搜索算法和曼哈顿距离的 KNN 算法	0.844	1.000	0.017	0.033
基于网格搜索算法和汉明距离的 KNN 算法	0.843	1.000	0.011	0.022

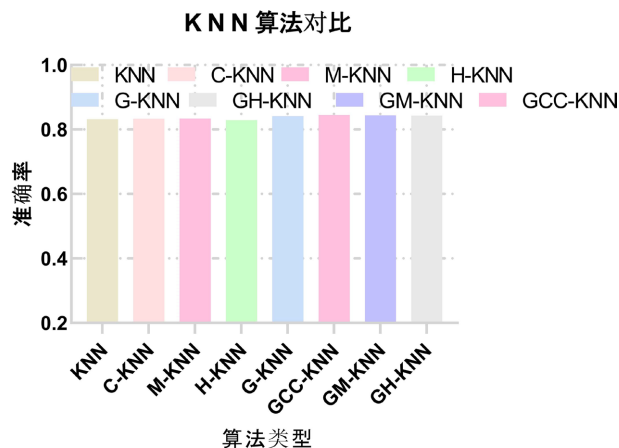


Figure 3. KNN algorithm comparison

图 3. KNN 算法对比

从表 2 的数据比较可以得到, GCC-KNN 算法与其他算法的比较中, 各方面效果更佳, 在此文章中最需要的准确率效果比较中, GCC-KNN 也是效果更好的。从图 3 可以明显的看出, GCC-KNN 与 KNN 比较, 准确率高出 1.3%、GCC-KNN 与 M-KNN 比较, 准确率高出 1.2%、GCC-KNN 与 C-KNN 比较, 准确率高出 0.9%、GCC-KNN 与 H-KNN 比较, 准确率高出 1.6%、GCC-KNN 与 G-KNN 比较, 准确率高出 0.4%、GCC-KNN 与 GM-KNN 比较, 准确率高出 0.1%, GCC-KNN 与 GH-KNN 比较, 准确率高出 0.2%。

6. 结论

糖尿病分类预测方法可以对人群中潜在糖尿病人群遗传进行分类预防, 本篇文章首先查看了糖尿病病人性别数量的情况, 病人的性别比例是很平衡的, 不用进行优化。基本情况指标、肝功能指标、生化指标、乙肝五项指标、血常规指标都与糖尿病认证信息相关, 根据对于糖尿病人群遗传分类预测的需要, 提出了 GCC-KNN 算法, 该论文算法需要确定恰当参数的过程中, 运用了 GridSearchCV 方法找到了最适合的参数, 该论文中基础的距离性能非最高的问题下, 选择对于提高糖尿病人群遗传概率测算更针对性的距离公式, 即 Chebyshev Distance 改进。我们通过使用不同的评价方法进行评价之后, 数据显示现在的糖尿病人群中遗传几率偏高, 从糖尿病人群的遗传入手, 从而减少糖尿病人群的比例和发病率是很重要的。其测试的效果表明 GCC-KNN 算法能够在糖尿病人群的遗传分类预测领域具体使用中效果更合适, 展现了与以前医院、诊所等对于糖尿病人群遗传测试不一样的, 机器学习与医学实践分析相结合的新的思路方法。

参考文献

- [1] 心怡, 张永泽, 阳成虎, 等. 数字健康背景下糖尿病基层医防融合服务的需求研究[J]. 中国全科医学, 2024, 27(31): 3958-3965.
- [2] 刘巧红, 马雨生, 蔡雨晨. 基于 XGBoost 算法的糖尿病分类预测模型及应用[J]. 现代仪器与医疗, 2023, 29(4): 1-6+11.
- [3] 胡月. 基于 Stacking 集成学习的糖尿病风险预测研究[D]: [硕士学位论文]. 哈尔滨: 黑龙江大学, 2024.
- [4] 凌雄娟, 王俊杰. 基于机器学习算法的糖尿病预测[J]. 现代信息科技, 2024, 8(14): 59-63+68.
- [5] 薛博元. 基于改进 CatBoost 算法的糖尿病回归和分类预测模型研究[D]: [硕士学位论文]. 银川: 北方民族大学, 2024.

- [6] Nalini, M., Yamini, B., Fernandez, F.M.H. and Uma Priyadarsini, P.S. (2024) Enhancing Anomaly Detection Efficiency: Introducing Grid Searchbased Multi-Population Particle Swarm Optimization Algorithm Based Optimized Regional Based Convolutional Neural Network for Robust and Scalable Solutions in High-Dimensional Data. *Biomedical Signal Processing and Control*, **96**, Article ID: 106651. <https://doi.org/10.1016/j.bspc.2024.106651>
- [7] 王紫薇, 徐凯, 侯益明. 基于不同距离公式的 KNN 算法对鸢尾花的分类[J]. 无线互联科技, 2021, 18(13): 105-106.