文章编号:1005-1228(2024)01-0007-03

基于 KNN 算法在糖尿病预测中的应用

梅　俊，陈建敏

（黄山职业技术学院 工业与财贸系，安徽 黄山　245000）

摘　要：人工智能技术在海量医疗数据中，通过技术手段实现疾病预测，为辅助治疗提供重要依据。文章分析了机器分类算法 KNN 算法的流程，以及在糖尿病数据中的具体实例。通过划分糖尿病数据集，计算 KNN 算法中的 K 值，确定选取最佳 k 值，达到最优准确率。通过实验验证 KNN 算法在糖尿病数据集上，该模型进行糖尿病预测有效。

关键词：KNN 算法；糖尿病预测；人工智能；数据集；k 值

中图分类号：TP311.13　　**文献标志码：**A

DOI:10.19414/j.cnki.1005-1228.2024.01.017

Application of KNN Algorithm in Diabetes Prediction

MEI Jun, CHEN Jian-min

（Huangshan Vocational and Technical College, Department of Industry and Finance and Trade, Huangshan 245000, China）

Abstract: Artificial intelligence technology achieves disease prediction through technological means in massive medical data, providing important basis for auxiliary treatment. This paper analyzes the machine classification algorithm KNN algorithm process, as well as specific examples in diabetes data. By dividing the diabetes data set, calculate the K value in the KNN algorithm, and determine the best k value to achieve the optimal accuracy. The experiment verifies that KNN algorithm is effective in predicting diabetes on diabetes dataset.

Key words: KNN algorithm; diabetes prediction; artificial intelligence; dataSet; k value

随着科学的不断进步，人们越来越关注健康问题。人工智能技术日新月异，对于海量医疗健康数据来说，通过技术手段预测疾病、研究疾病发展趋势和影响因素，可以为人们的健康提供重要支撑。糖尿病[1]是一种常见的慢性病，近年来，随着人们生活水平的提高，我国成年人糖尿病患病率不断攀升，已高于全球平均水平，而预防和治疗方法都十分有限。机器学习是人工智能的一个重要分支，在医疗领域中，机器学习算法可以从海量的医疗数据中发现规律和趋势，并快速有效地制定相应的诊断和治疗计划。针对糖尿病等慢性疾病，机器学习算法可以通过分析患者的生理指标和医疗数据，为辅助治疗提供有力支持。KNN 算法[2]是一种常见的分类算法，本文通过研究机器学习 KNN 算法在糖尿病数据集中的应用[3-4]，实现疾病预测。

1　数据预处理

本文研究数据集选用印第安人糖尿病数据集，数

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPress	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
2	10	125	70	26	115	31.1	0.205	41	1
3	1	97	66	15	140	23.2	0.487	22	0
4	0	137	40	35	168	43.1	2.288	33	1
5	13	145	82	19	110	22.2	0.245	57	0
6	3	158	76	36	245	31.6	0.851	28	1
7	3	88	58	11	54	24.8	0.267	22	0
8	4	103	60	33	192	24	0.966	33	0
9	4	111	72	47	207	37.1	1.39	56	1
10	3	180	64	25	70	34	0.271	26	0
11	9	171	110	24	240	45.4	0.721	54	1
12	1	103	80	11	82	19.4	0.491	22	0
13	1	101	50	15	36	24.2	0.526	26	0
14	1	89	66	23	94	28.1	0.167	21	0
15	3	78	50	32	88	31	0.248	26	1
16	2	197	70	45	543	30.5	0.158	53	1
17	1	189	60	23	846	30.1	0.398	59	1
18	5	166	72	19	175	25.8	0.587	51	1
19	0	118	84	47	230	45.8	0.551	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	11	143	94	33	146	36.6	0.254	51	1
24	5	88	66	21	23	24.4	0.342	30	0
25	8	176	90	34	300	33.7	0.467	58	1
26	7	150	66	42	342	34.7	0.718	42	0
27	7	187	68	39	304	37.7	0.254	41	1

图 1　印第安人糖尿病数据集片段

收稿日期：2022-11-14

基金项目：安徽高校自然科学研究一般项目（项目编号：KJ2020H02）；安徽高校自然科学研究重点项目（项目编号：2023AH053101）

作者简介：梅俊（1983—），安徽池州人，女，讲师，硕士，主要研究方向：数据挖掘，云计算，机器学习；陈建敏（1985—），安徽黄山人，男，副教授，硕士，主要研究方向：大数据、数据分析。

据显示基于给定的医疗措施预测皮马印第安人 5 年内糖尿病的发病情况。数据集中每个字段值数量并不均衡，数据结果显示患病或者没有患病。该数据集包含 768 条数据，其中每条数据含有 8 个输入变量和 1 个输出变量。数据集片段如图 1 所示。

输入字段的名称含义见表 1。

表 1　数据集字段含义

列表名	含义
Pregnancies	怀孕次数
Glucose	血糖
BloodPressure	血压
SkinThickness	皮脂厚度
Insulin	胰岛素
BMI	体重指数
DiabetesPedigreeFunction	糖尿病遗传函数
Age	年龄
Outcome	结果（0 表示未患糖尿病，1 表示患有糖尿病）

对该数据集进行数据分析，结果显示没有重复数据，但第 2 ~ 5 列属性值（血糖、血压、皮脂厚度、胰岛素、体重指数）根据常识不可能为 0，血糖等指标为 0 值在医学领域是无意义的。为了使数据更合理，需对数据进行清洗。

本次实验对数据集进行如下的数据处理：先对原有数据集第 2 ~ 5 列的零值用缺失值来取代，然后计算第 2 ~ 5 列的每列平均值，用每列的均值填充该列的缺失值，确保数据集数据误差与实际值最小，从而最大保证测试数据的准确性。

根据上述分析对数据进行数据清洗后，数据集预处理后信息如图 2 所示。此时的数据显示较为合理。

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigri	Age	Outcome
2	6	148	72	35	80	33.6	0.627	50	1
3	1	85	66	29	80	26.6	0.351	31	0
4	8	183	64	21	80	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	21	80	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	69	21	80	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	21	80	32	0.232	54	1
12	4	110	92	21	80	37.6	0.191	30	0
13	10	168	74	21	80	38	0.537	34	1
14	10	139	80	21	80	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	69	21	80	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	21	80	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	21	80	35.4	0.388	50	0

图 2　数据预处理后信息

2　KNN 算法

KNN 算法又称 K– 近邻算法，是通过计算该测试点到若干个已知点的各字段之间的距离，找出最近 k 个已知点，统计其类别，数量最多的类别即为测试点类别。

2.1　KNN 算法思想及流程

KNN 算法思想总结如下：在训练集中数据和标签已知的情况下，输入测试数据，将测试数据的特征与训练集中对应的特征进行相互比较，找到训练集中与之最为相似的前 k 个数据，则该测试数据对应的类别就是 k 个数据中出现次数最多的分类。算法流程：首先计算测试数据与各个训练数据之间的距离；其次按照距离的递增关系进行排序，选取距离最小的 k 个点，确定前 k 个点所在类别的出现频率，即计算投票数；最后返回前 k 个点中出现频率最高的类别作为测试数据的预测分类。

2.2　KNN 算法的优缺点

KNN 算法优点：KNN 模型 [5–6] 的使用无需事先进行训练，模型结构简单易懂，降低了学习的估计错误，只需要一个 k 值作为超参数，其准确率和灵敏度都很高，适用于大规模的自动分类。

KNN 算法缺点：存在着样本不平衡的问题；KNN 算法必须事先设置 K 值，选取的 K 值大小对 KNN 算法的分类效果影响很大。当 K 的数值较小，且对邻近点附近实例的敏感性较高，易受噪声点影响，从而产生过拟合问题。当 K 的数值太大，则等于在更大的邻域中使用一个训练例子进行预测，意义不大。KNN 算法需要对全部试验数据与训练数据进行计算，因此计算工作量大、耗时长、计算复杂。

2.3　KNN 算法在数据集中分析

按照 KNN 算法的思想，以图 1 数据集中前十条作为训练集，第十一条数据作为测试数据分析 KNN 算法在该数据集中的应用实例。首先计算该测试数据与已知训练数据之间的距离，按照欧氏距离公式：$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$，结果见表 2。

表 2　数据集第十一条数据与前十条数据欧式距离结果

数据	距离值
11->1	49.6
11->2	38.7
11->3	79.6
11->4	38.7
11->5	106.9
11->6	22.5
11->7	55.1
11->8	24.4
11->9	472.8
11->10	29.4

对于上表值，按照计算的距离值进行排序，结果见表 3。

表 3　距离排序表

排序	数据序号	距离值	结果值
1	6	22.5	0
2	8	24.4	0
3	10	29.4	1
4	2	38.7	0
5	4	38.7	0
6	1	49.6	1
7	7	55.1	1
8	3	79.6	1
9	5	106.9	1
10	9	472.8	1

设 k 值为取 k 个邻近点，当 k=1 时，测试数据距离第 6 条训练数据距离最近，第 6 条数据结果为 0，因此测试数据预测结果判断为 0 表示未患糖尿病；当 k=3 时，测试数据距离第 6、8、10 训练数据最近，结果分别为 0、0、1，0 的投票数为 2，大于 1 的投票数为 1，因此该测试数据预测结果判断为 0 是未患糖尿病；当 k=5 时，测试数据距离第 6、8、10、2、4 最近，0 投票数为 4，大于 1 的投票数 1，而当 k=7 时，该测试数据通过计算投票数得出预测结果判断为还是 0；当 k=9 时，该测试数据 1 的投票数为 5，大于 0 的投票数为 4，这时该测试数据预测结果为 1。从实例分析中可以看出 k 值的不同影响预测结果。因此，KNN 算法需要确定一个最佳的 k 值，使预测结果准确率更高。

3　K 折交叉验证

K 折交叉验证是将训练集划分成 K 个大小相等的子数据集，遍历每个子数据集作为一次验证集，剩下的 K-1 个子数据集作为建模集，最终会得到 K 个子数据集的评估结果，求其均值作为最终精度结果，一般 K 取 5 或 10。

4　实验步骤

根据上述实例分析，对整个数据集进行实验，步骤如下：首先将整个数据集划分为训练集和测试集。本次数据集规模比较小，为了避免单次划分时数据划分不平衡，本次试验使用 K 折交叉 [7] 验证，分别统计 10 次、5 次划分数据集，对多次评估的结果取平均，避免模型过拟合现象，训练数据和测试数据比例为 7 ∶ 3。其次分别计算训练数据、测试数据准确率。使用 python 语言中 Sklearn 库进行模型搭建，计算出准确率，分析准确率。最后根据准确率确定最佳 k 值。将 KNN 算法分析思想应用在本数据集中，分别计算 k 值在 1 ~ 20 范围内训练集和测试集的预测准确度，分别计算不同 k 值时，KNN 算法的准确率。

当 K 折交叉验证为 10 次划分时，训练集和测试集的预测准确度见表 4。

表 4　10 次交叉验证划分数据集准确率

不同准确率	k=15	k=13	k=11	k=9	k=7	k=5
训练集准确率	0.703 6	0.698 0	0.707 2	0.725 9	0.709 2	0.699 9
测试集准确率	0.710 3	0.727 5	0.731 9	0.705 6	0.705 8	0.710 3

当 K 折交叉验证为 5 次划分时，训练集和测试集的预测准确度见表 5。

表 5　5 次交叉验证划分数据集准确率

不同准确率	k=15	k=13	k=11	k=9	k=7	k=5
训练集准确率	0.709 3	0.701 9	0.685 2	0.696 2	0.700 0	0.688 9
测试集准确率	0.709 7	0.714 0	0.722 5	0.709 6	0.735 6	0.722 6

从表中可以看出，测试集准确度由低向高在 k 值的最佳选择为 k=11，当大于 11 后呈下降趋势，因此确定 k 值为 11。

5　结束语

本文通过对糖尿病数据集中各特征值对每条病人数据进行是否患有糖尿病的 KNN 算法分类实验，实验结果表明，使用 K 折交叉验证方法选择参数 k 时，准确率较高。不足之处是应用于小数据集效率较高，但当数据规模变大时，KNN 算法计算过大，效率变低，后续将进一步研究 KNN 改进算法在疾病数据中的应用。

参考文献：

[1] 中国老年 2 型糖尿病防治临床指南编写组 . 中国老年 2 型糖尿病防治临床指南 [J]. 中国糖尿病杂志 ,2022,30(1):2-51.

[2] SUN J, DU W, SHI N. A survey of KNN algorithm[J]. Information Engineering and Applied Computing, 2018(1):10.

[3] 吴兴惠 , 周玉萍 , 邢海花 , 等 . 机器学习分类算法在糖尿病诊断中的应用研究 [J]. 电脑知识与技术 ,2018,14(35):177-178.

[4] 李芳君 . 基于机器学习的医学数据分类算法研究 [D]. 济南 : 山东大学 ,2020.

[5] 郭躬德 , 黄杰 , 陈黎飞 . 基于 KNN 模型的增量学习算法 [J]. 模式识别与人工智能 ,2010,23(5):701-707.

[6] 张迪 . 基于 KNN 和神经网络算法的数据挖掘与预测模型研究 [J]. 太原师范学院学报 (自然科学版),2023,22(2):29-34.

[7] BENGIO Y, GRANDVALET Y. No unbiased estimator of the variance of K-Fold cross-validation[J]. The Journal of Machine Learning Research,2004(5):1089-1105.