Application of KNN Algorithm in Diabetes Prediction

Jun Mei[1], Jianmin Chen[1,*]

[1]Department of Industry and Finance and Trade, Huangshan Vocational and Technical College, Anhui 245000, China

[*]Corresponding Author

Abstract

Artificial intelligence technology realizes disease prediction through technical means in massive medical data, providing an important basis for auxiliary treatment. The article analyzes the process of the machine classification algorithm KNN algorithm, 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. Through experiments, it is verified that the KNN algorithm is effective in predicting diabetes on the diabetes data set.

Keywords

KNN algorithm, diabetes prediction, artificial intelligence, dataSet, k value.

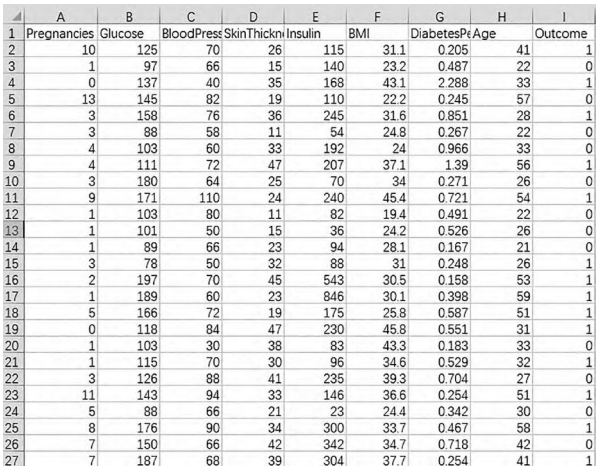
1. Introduction

With the continuous advancement of science, people are paying more and more attention to health issues. Artificial intelligence technology is advancing rapidly. For massive medical and health data, predicting diseases and studying disease development trends and influencing factors through technical means can provide important support for people's health. Diabetes[1] is a common chronic disease. In recent years, with the improvement of people's living standards, the prevalence of diabetes among adults in my country has been rising and is higher than the global average. Prevention and treatments are very limited. Machine learning is an important branch of artificial intelligence. In the medical field, machine learning algorithms can discover patterns and trends from massive medical data and quickly and effectively formulate corresponding diagnosis and treatment plans. For chronic diseases such as

diabetes, machine learning algorithms can provide strong support for auxiliary treatment by analyzing patients' physiological indicators and medical data. KNN algorithm[2] is a common classification algorithm. This article studies the application of machine learning KNN algorithm in diabetes data set[3,4] to achieve disease prediction.

2. Data preprocessing

The research data set in this article uses the Indian diabetes data set.



	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

Figure 1: Snippet of Indian Diabetes Dataset

Shown to predict 5-year incidence of diabetes among Pima Indians based on given medical measures. The number of values in each field in the data set is not balanced, and the data results show disease or no disease. This data set contains 768 pieces of data, each piece of data contains 8 input variables and 1 output variable. A fragment of the data set is shown in Figure 1. The meaning of the input field names is shown in Table 1.

Table 1: Data set field

List name	
Pregnancies	(Number of pregnancies)

Glucose
Blood Pressure
Skin Thickness
Insulin
BMI
Diabetes Pedigree Function
Age
Outcome (0 means not suffering from diabetes, 1 means indicates having diabetes)

Data analysis of this data set shows that there is no duplicate data, but the attribute values in columns 2 to 5 (blood sugar, blood pressure, sebum thickness, insulin, body mass index) cannot be 0 according to common sense. Blood sugar and other indicators have a value of 0 in medicine. Fields are meaningless. In order to make the data more reasonable, the data needs to be cleaned. In this experiment, the data set was processed as follows: first, the zero values in columns 2 to 5 of the original data set were replaced with missing values, and then calculated. Calculate the average value of each column from columns 2 to 5, and use the average value of each column to fill in the missing values in that column to ensure that the data set error and the actual value are minimized, thereby maximizing the accuracy of the test data.

After data cleaning based on the above analysis, the preprocessed information of the data set is shown in Figure 2. The data display at this time is more reasonable.

1	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

Figure 2: Information after data preprocessing

3. KNN algorithm

3.1. KNN algorithm idea and process

The idea of the KNN algorithm is summarized as follows: when the data and labels in the training set are known, input the test data, compare the characteristics of the test data with the corresponding features in the training set, and find the top k data in the training set that are most similar to them. Then the category corresponding to the test data is the category that appears most frequently in the k data. Algorithm process: First, calculate the distance between the test data and each training data; secondly, sort according to the increasing relationship of the distance, select the k points with the smallest distance, determine the frequency of occurrence of the category of the first k points, that is, calculate the number of votes; finally return to the top k The category with the highest frequency among the points is used as the predicted classification of the test data.

3.2. Advantages and disadvantages of KNN algorithm

Advantages of the KNN algorithm: The use of the KNN model[5,6]does not require prior training. The model structure is simple and easy to understand, which reduces the estimation error of learning. It only requires a k value as a hyperparameter. The accuracy and sensitivity are both high and suitable for large-scale automatic classification.

Disadvantages of the KNN algorithm: There is the problem of sample imbalance; the KNN algorithm must set the K value in advance, and the selected K value has a great

impact on the classification effect of the KNN algorithm. When the value of K is small and the sensitivity to instances near adjacent points is high, it is easily affected by noise points, resulting in over-fitting problems. When the value of K is too large, it is equivalent to using a training example to predict in a larger neighborhood, which is of little significance. The KNN algorithm needs to calculate all test data and training data, so the calculation workload is large, time-consuming, and complex.

3.3. KNN algorithm analyzes in data set

According to the idea of the KNN algorithm, the first ten items in the data set in Figure 1 are used as the training set, and the eleventh item of data is used as the test data to analyze the application examples of the KNN algorithm in this data set. First, calculate the distance between the test data and the known training data, the results are shown in Table 2.

Table 2: Euclidean distance results between the eleventh piece of data and the first ten pieces of data in the data set

Data	Distance value
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

For the values in the above table, they are sorted according to the calculated distance value, and the results are shown in Table 3.

Table 3: Distance sorting table

sort	Data serial number	distance value	result value
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

Let the k value be k neighboring points. When k=1, the test data is the closest to the 6th training data, and the 6th data result is 0.

Therefore, the test data prediction result is judged to be 0, which means that there is no diabetes; When k=3, the test data is closest to the 6th, 8th, and 10th training data, and the results are 0, 0, and 1 respectively. The number of votes for 0 is 2, and the

number of votes greater than 1 is 1. Therefore, the prediction result of the test data is judged as 0 means no diabetes; when k=5, the test data is 6, 8, 10, 2,4.Most recently, 0 votes were 4, votes greater than 1 were 1, and when k=7, the test data calculates the number of votes to determine that the prediction result is still 0; when k=9, the number of votes for the test data 1 is 5, and the number of votes greater than 0 is 4. At this time, the prediction result of the test data is 1. It can be seen from the example analysis that different k values affect the prediction results. Therefore, the KNN algorithm needs to determine an optimal k value to make the prediction results more accurate.

4. Conclusion

This paper conducts a KNN algorithm classification experiment on whether each patient data has diabetes by using each feature value in the diabetes data set. The experimental results show that when using the K-fold cross-validation method to select parameter k, the accuracy is higher. The disadvantage is that it is more efficient when applied to small data sets. However, when the data size becomes larger, the KNN algorithm calculation is too large and the efficiency becomes low. The application of the KNN improved algorithm in disease data will be further studied in the future.

References

[10] Writing Group of Clinical Guidelines for the Prevention and Treatment of Type 2 Diabetes in the Elderly in China. (2022). Clinical Guidelines for the Prevention and Treatment of Type 2 Diabetes in the Elderly in China. Chinese Journal of Diabetes, 30(1), 2-51.

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

[12] Wu, X., Zhou, Y., Xing, H., et al. (2018). Research on the application of machine learning classification algorithm in diabetes diagnosis. Computer Knowledge and Technology, 14(35), 177-178.

[13] Li, F. (2020). Research on medical data classification algorithm based on machine learning [Doctoral dissertation, Shandong University].

[14] Guo, G., Huang, J., & Chen, L. (2010). Incremental learning algorithm based on KNN model. Pattern Recognition and Artificial Intelligence, 23(5), 701–707.

[15] Zhang, D. (2023). Research on data mining and prediction models based on KNN and neural network algorithms. Journal of Taiyuan Normal University (Natural Science Edition), 22(2), 29–34.