KNN report

kai ZHANG, Mengyu Pan

September 2020

Abstract

This experiment evaluated the performance of KNN algorithm in solving binary classification problems. The evaluation results confirmed that KNN algorithm could give relatively good prediction results. The effect of K value was also analyzed and it indicated that the K value should be well set when using KNN algorithm.

1 Introduction

k-nearest neighbor algorithm(KNN) is proposed by Thomas Cover in 1969 [1] . They proposed that we can classify the new data by its nearest neighbors. This algorithm has a significant advantage that it doesn't rely on complex statistic theory while it brings much calculation.

2 Method

In our method, there are four steps:

(1)Filter Dataset: before we apply KNN algorithm, we check our dataset. If there is invalid data in the dataset, we replace it with the average value. For example, we eliminate the '?' in the Breast Cancer Wisconsin (Diagnostic) Data Set.

(2)Distance Calculation: we divide our dataset randomly into two parts: train data and test data, where test data takes up to 20 percent in the dataset. As you can see in the figure 1, for each test data in the dataset, we calculate the distances between it and other train data.

$$distance = \sqrt{\sum (AttributeOfTestData - AttributeOfTrainData)^2} \quad (1)$$

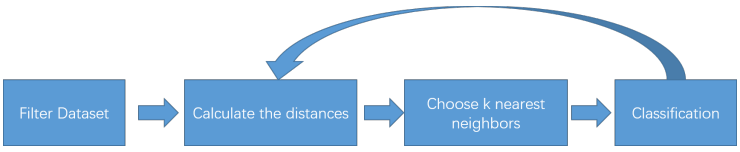

Figure 1: schema of KNN method

		True condition	
		True	False
Predicted condition	Positive	True Positive	False Positive
	Negative	True Negative	False Negative

Figure 2: Confusion matrix

(3)Classification: then we choose its nearest neighbors and classify it. If most of its neighbors belong to one class, it's more likely to belong to the same class. Taking k=5 as an example, after the calculation, we choose 5 nearest neighbors and decide its class by calculating the average value of its neighbors' class.

(4)Iteration: at last, we iterate this step until the end of the test data.

3 Experiments

3.1 Data set

Two datasets were utilized in our experiments to train our method and evaluate accuracy.

(1) Breast Cancer Wisconsin (Diagnostic) Data Set [2] is a classic dataset for binary classification, in which the features are computed from a digitized image of a fine needle aspirate(FNA) of a breast class. The total number of instances is 699 (458 for Benign, 241 for Malignant) and for each instance, it consists of 10 real-valued features and the corresponding class attribute.

(2) Haberman's Survival Data Set [2] contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. There are total 306 instances in this dataset, and 4 attributes (3 features and 1 class) for each instance. In our experiment, the ratio between the number of training instances and the testing instance is 0.8:0.2.

3.2 Metric

To evaluate the performance of our experiment in the two datasets, we adopt the confusion matrix and F1 score to measure the results. As shown in table **??**, the confusion matrix includes the TP, FP, TN, and FN, which describe the number of instances in each classification case (wrongly classified, correctly classified).

F1 score measures the accuracy of the classification results. It is the ratio of overlap area to total area, as shown in

$$F1score = 2 \times \frac{areaofoverlap}{totalarea} = 2 \times \frac{A \cap B}{A + B} \tag{2}$$

Dataset	K	TP	FP	TN	FN	F1 score
Breast Cancer Wisconsin Data Set	1	102	0	1	34	0.8571
	3	101	1	1	34	0.8523
	5	102	0	1	34	0.8571
	7	102	0	0	35	0.8536
	9	102	0	0	35	0.8535
Haberman's Survival Data Set	1	35	11	7	9	0.7955
	3	42	4	13	3	0.8317
	5	43	3	13	3	0.8431
	7	44	2	13	3	0.8544
	9	43	3	13	3	0.8431

Figure 3: Quantitative evaluation results

3.3 Evaluation results and discussion

In the evaluation experiments, we analyzed the performance of our KNN algorithm in the above two datasets. Additionally, the impact of different K was also investigated. The results were represented in **??**. By observing the F1 score in both datasets, we can see that the value of K affects the final accuracy of the classification results. For example, in the Breast Cancer Wisconsin Data Set, the most appropriate value is 1 or 5, but in the Haberman's Survival Data Set, k=7 achieves the best performance. Therefore, in order to obtain the best classification results by KNN algorithms, the value of K should be searched and evaluated.

4 Conclusion

In this experiment, we use the KNN algorithm to solve classification problems. To judge the performance of the KNN algorithm, it was evaluated on two datasets. The final results verify that the KNN algorithm could produce relatively good classification results but different K could affect the final performance. In conclusion, when using the KNN algorithm, the K value should be seriously searched and set.

References

[1] Thomas Cover and P. E. Hart. "Nearest Neighbor Pattern Classification". In: *IEEE Transactions on Information Theory* 13.21 - 27 (1967).

[2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.