Tumor imaging diagnosis analysis based on improved KNN algorithm

Cailing Wang1b, LeiChao Li*1a, SuQiang He1 and Jing Zhang1

1Computer technology, Xi'an Shiyou University, Xi'an, Shaanxi Province, 710065, China

a*1162401679@qq.com
b*cailingw@xsyu.edu.cn

Abstract. As a simple, effective and non-parameter analysis method, knn is widely used in text classification, image recognition, etc. [1]. However, this method requires a lot of calculations in practical applications, and the uneven distribution of training samples will directly lead to a decrease in the accuracy of tumor image classification. To solve this problem, we propose a method based on dynamic weighted KNN to improve the accuracy of classification, which is used to solve the problem of automatic prediction and classification of medical tumor images based on image features and automatic abnormality detection. According to the classification of tumor image characteristics, it can be divided into two categories: benign and malignant. This method can assist doctors in making medical diagnosis and analysis more accurately. The experimental results show that this method has certain advantages compared with the traditional KNN algorithm.

1. Introduction

Tumors are one of the hazards to human life today. If they can be detected and treated at an early stage, the survival rate of cancer patients can be greatly improved. With the development of society and changes in the environment, the incidence of malignant tumors has been on the rise. Computer tomography (CT) medical imaging has been widely used in clinical diagnosis. CT is used as a screening tool for tumor diagnosis, helping radiologists to detect and locate pathological changes more accurately, and is helpful for early tumor diagnosis. Early detection and treatment is the key to improving the effect of tumor treatment and patient survival rate. However, analyzing tumor CT images requires a lot of professional experience from doctors, and some features of tumor images are not obvious. It is difficult for doctors to make accurate judgments about the benign and malignant tumors due to the influence of experience factors when analyzing some small local features. Therefore, it is a very valuable and meaningful technology to help doctors quickly and accurately judge the benign and malignant tumors and realize the automatic detection of tumor images.

The k nearest neighbor (k-nn) algorithm is a simple algorithm in computer machine learning. It is a typical simple, effective, non-parametric classification method[2]. The algorithm can provide good classification accuracy in the field of image classification. We added a dynamic weighting function to the traditional KNN algorithm, and proposed a dynamic weighted KNN method, which can effectively predict and classify tumor images.
2. Algorithm description

2.1. Overview of KNN algorithm
The core idea of the KNN algorithm is that the features of a sample are most similar to the features of the k samples in the data set. If most of the k closest samples in the feature space belong to a certain category, the sample also belongs to that category, and have the characteristics of samples in this category. When determining the classification decision, this method only determines the category of the sample to be divided according to the category of the closest K samples, and does not need to perform operations such as discriminating the class domain. Therefore, the KNN algorithm works better for class domains with cross-overlapping conditions than other classification algorithms.

2.2. KNN algorithm introduction
KNN is classified by measuring the distance between the characteristic values of unclassified points and the feature values of classified points. The idea is: If a certain category of the k samples with the closest distances of unclassified points in the feature space has the highest frequency, then the unclassified points belong to this category. K is usually a very small integer.

In the KNN algorithm, the selected neighbors are all correctly classified objects. As shown in Figure 1, how to assign a category to the green square, is it a red five-pointed star or a blue triangle? When K=5, there are 2 five-pointed stars and 3 triangles in the dashed circle. At this time, the category of the green square is judged to be the blue triangle. When K=10, there are 6 five-pointed stars and 4 triangles in the solid circle. At this time, it is judged that the category of the green square is a red five-pointed star.

![Figure 1. KNN principle.](image-url)

This also shows that the result of the KNN algorithm largely depends on the choice of K. In KNN, the distance between data feature points is calculated as the index of dissimilarity between each feature point, avoiding the matching problem between feature points, and Euclidean distance is usually used here.

\[
d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}
\]

(1)

2.3. Algorithm flow
For each feature point of an unknown category attribute in the dataset, perform the following operations in sequence:
(1) Calculate the distance between the feature point in the known category data set and the feature point currently tested;
(2) Sort in ascending order of distance;
(3) Select k points with the smallest distance from the current feature point;
(4) Determine the frequency of occurrence of the category of the first k feature points;
Return the type with the highest frequency of the top k feature points as the predicted classification of the current feature point.

2.4. Dynamically weighted K-NN

The main disadvantage of the KNN algorithm in classification is, for example, when the sample density is unbalanced, the sample size of one category is large, while the sample size of other categories is small. It may cause the large-capacity class of samples to account for the majority of the K neighbors of the current feature point during data input. The algorithm only calculates the nearest K neighbors. If the number of samples in a certain category is large, no matter whether the samples of the category are not close to the target samples, or the samples of this category are very close to the target samples, the classification will be biased. In order to avoid this, a dynamic weighted KNN algorithm can be used. The idea is that the neighbors with a small distance from the sample have a large weight.

Algorithm flow:
1. Calculate the distance between the feature point in the known category data set and the feature point currently tested;
2. Sort in ascending order of distance;
3. Select k points with the smallest distance from the current feature point;
4. Calculate the corresponding weights of the first K points according to the distance;
5. Determine the frequency of occurrence of the category of the first k feature points; determine the category of the first K points in turn, if the current category already exists, add the corresponding weight to the frequency of the current category; if not, then the frequency of this feature point The weight is the frequency of the new category;

Return the type with the highest frequency of the top k feature points as the predicted classification of the current feature point.

2.5. Gaussian function weighting

The Gaussian function is the density function of the normal distribution. The Gaussian function is used to optimize the weight of samples at different distances. When the distance between the training sample and the test sample is larger, the weight of the distance value is smaller. Closer neighbors get more weight, while farther neighbors get less weight, and a weighted average is used. The formula is:

\[ f(x) = a \cdot e^{-\frac{(x-b)^2}{2\delta^2}} \]  

(2)

Among them, a is the height of the peak of the curve, b is the coordinates of the center of the peak, and \( \delta \) is called the standard deviation.

Figure 2. Gaussian function image.
2.6. Sigmoid function weighting
Sigmoid function is a common sigmoid function in biology, also known as sigmoid growth curve. In information science, because of its simple increment and inverse function, sigmoid function is often widely used as an activation function of neural network to map variables between 0 and 1. The formula is:

$$S(x) = \frac{1}{1 + e^{-ax}}$$

(3)

3. Experimental results and analysis
The experimental data is real data, from the benign and malignant tumor data set of kaggle, each tuple data has symmetry, smoothness, compactness, texture, radius, perimeter, area, concave point, concavity, fractal dimension, etc. Multiple characteristics.

3.1. The impact of test set proportion on accuracy
In Figure 4, the x-axis represents the percentage of the test set to the total number of samples, and the y-axis represents the percentage of misclassification rate. It can be seen from Figure 3 that as the proportion of the test set increases, the misclassification rate also shows an overall upward trend. Among them, when the test set accounted for 10%, the error rate was the lowest.
3.2. The impact of k value on accuracy
It can be seen from Figure 2 that when the test set proportion is 10%, the error rate is the lowest, so we continue to carry out the experiment of the influence of the value of K on the accuracy when the test set proportion is 10%, and test them separately. The accuracy of traditional KNN, GaussianKNN, and sigmoidKNN when K takes different values. In Figure 5, the X-axis represents the value of K, the range is 1-9, and the Y-axis represents the percentage of misclassification rate. It can be seen from Figure 5 that when the same sample is tested, when K is 1, the error rate of the three is the same. In traditional KNN, when K is 2-4, the error rate is the highest, and when K>=5, the error rate tends to be stable. In GaussianKNN, when K is 1-3, the error rate is the lowest, and when K>=8, the error rate is the highest and tends to be stable. In sigmoidKNN, when K is 2-4, the error rate is the highest, and when K>=7, the error rate is the lowest and tends to be stable.

Overall, the performance of GaussianKNN and sigmoidKNN are better than traditional KNN, and the performance of sigmoidKNN is better than GaussianKNN, with the lowest misclassification rate of about 4.35%.

3.3. The influence of the parameter value of Gaussian function and sigmoid function on accuracy
The variable parameter of the Gaussian function is δ, and the variable parameter of the sigmoid function is α. The values of δ and α will affect the size of the weight, thereby affecting the frequency of each category, resulting in changes in accuracy. It can be seen from Figure 6 that when δ is 9-10, the accuracy of the Gaussian function is the highest, and the misclassification rate is about 5.80%. When the variable parameter α>=1 of the sigmoid function, the accuracy is the highest and tends to be stable, and the misclassification rate is about 4.35%. Therefore, sigmoidKNN is better than GaussianKNN in the predictive classification of benign and malignant tumor image data sets.

Figure 6. The influence of the parameter values of Gaussian function and sigmoid function on accuracy.

4. Conclusion
In order to solve the problem of automatic prediction and classification of medical tumor images based on image features and automatic abnormality detection, this paper proposes an improved dynamic weighted KNN algorithm, which can effectively reduce the misclassification rate in the case of uneven sample distribution. Through the test on the benign and malignant tumor image data set, more satisfactory results have been obtained.

References
[1] Lu, L. R., & Fa, H. Y. (2004). A density-based method for reducing the amount of training data in knn text classification. Journal of Computer Research and Development, 41(4), 539-545.
[2] Liu, Y. D., & Niu, H. M. (2011). Knn classification algorithm based on k-nearest neighbor graph for small sample. Computer Engineering.
[3] Wang, Z. L., Teng, T., Wang, Q., Huang, Z. Y. (2021). Improved KNN Classification Algorithm Based on K-nearest Neighbor Graph for Remote Sensing Images. Journal of Geospatial Information, 19(02), 33-36.

[4] Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. IEEE Trans Systems Man & Cybernetics, 6(4), 325-327.

[5] Han, E. H. S., Karypis, G., & Kumar, V. (2001). Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. Springer, Berlin, Heidelberg.

[6] Hu, Y., Peng, G., Wang, Z., Cui, Y., & Qin, H. (2020). Partition Selection for Large-Scale Data Management Using KNN Join Processing. Mathematical Problems in Engineering, 2020.

[7] Ramteke, R. J., & Monali, K. Y. (2012). Automatic medical image classification and abnormality detection using k-nearest neighbour. International Journal of Advanced Computer Research, 2(4), 190.

[8] Raikwal, J. S., & Saxena, K. (2012). Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. International Journal of Computer Applications, 50(14).