A Study of Double Constrains Handwritten Digit Recognition Based on Adaptive kNN Algorithm

TIAN Sha Sha[1], WANG Hong [1], SHE Wei [2]

1 College of Computer Science
South-Central University for Nationalities, Wuhan, 430074, China
2 *School of Mathematics and Statistic*
South-Central University for Nationalities, Wuhan, 430074, China

Abstract **— In order to improve the accuracy rate of recognition of handwritten numbers, this paper proposes the double constraint recognition method combining structure characteristic and statistical characteristic. This paper designs the recognition method using the row, column and diagonal density of "1" in the sample code as statistical characteristic, and uses the closed area as structure characteristic. The kNN algorithm is used to classify, and dimension reduction is used with dozens of bits eigenvalue instead of 1024 bits in the sample code. The experimental results show that the error rate is only 0.07% higher than no dimension reduction, but calculating speed is increased nearly 10 times. In addition, for calculating the value of k automatically, this paper uses the gradient descent method. The results show that the adaptive method can obtain the k corresponding to the lowest recognition error rate.**

Keywords -- Structure Characteristic, Statistical Characteristic, Dimension Reduction, k-Nearest Neighbor Algorithm

I. INTRODUCTION

As the development of pattern recognition technology and digital image processing technology, handwritten numeral recognition technology has been playing its important role in fields such as zip code recognition, bank check processing and machine automatic input [1]. Identification method is mainly divided into the method based on structural characteristics and the method based on statistical characteristic. Structure features usually include round, endpoint, intersection, outline and statistical features usually include point density, area and so on. Each has advantages and disadvantages [2, 3]. Identification method based on structural features can significantly improve recognition accuracy, and statistical characteristic can reduce the interference of handwritten arbitrary. In order to improve the recognition accuracy, this paper takes double constraints of way of recognition method based on the structural features and statistical characteristics to design a handwritten digit recognition method [4].

Algorithms for implementation of handwritten numeral recognition are mainly k-Nearest Neighbor (kNN) algorithm, artificial neural network and support vector machine (SVM), etc. [5]. Because idea of kNN algorithm is simple, easy to understand and implement, no need to estimate parameters, especially suitable for the multiple classification problems, so this article uses the kNN algorithm to implement the handwritten digit recognition. But as kNN algorithm needs to save all data sets, and calculate the data set distance for each data, its time and space overhead is very large, and while the number of all kinds of different samples are unbalanced, classification error can happen[6, 7]. As a result, this paper will improved kNN algorithm aimed at the handwritten digit recognition by adopting the way of the data collection of dimension reduction to reduce the cost of time and space and use adaptive learning method to calculate of k value.

This article uses the handwritten digital data provided by UCI machine learning repository to conduct experiment and analysis. The data set is handwritten digital image that have been binary processed. Data set includes training set and test set, the training focus have 2000 data, every number has about 200 or so samples, test set has about 900 data, each data sample is 32 * 32 bits.

II. DOUBLE CONSTRAINTS RECOGNITION OF HANDWRITTEN NUMERALS

The key of handwritten numerals recognition is the extraction of characteristic value. If the characteristic value is too little, classification feature may not be obvious, reducing the classification accuracy. If the characteristic value is too much, it will cause mutual interference between characteristic values, increasing of the classification of time overhead at the same time, so can't do real-time identification. For characteristic value, therefore, should not only being selected through the theory analysis, but also from experiment.

Structure features of handwritten numerals are discrete degree, transverse and longitudinal ratio, interval distribution, through number, number, endpoint, line segment and closed area, etc. By structure characteristics analysis experiment of writing digital data sets, we get the following conclusion:

(1) Due to the randomness handwriting, dispersion ratio, transverse and longitudinal ratio, interval distribution, through number, number, endpoint, and line segment

intersection cannot be a good characterization of the features of every number.

(2) When close region is taken as structure characteristic values, for part numbers, it can be a good characterization of its characteristics. For example, "0" is a closed area, and "1" is not closed area. To identify the closed area of 0 and 1, experiments prove its accuracy is up to 99%. But for the identification of closed area "4", the accuracy can reach 73%. Even so, cooperating with statistical characteristic, as a structural feature in recognition process closed area still has a great significance in improving the recognition accuracy.

Handwritten Numbers statistical characteristic mainly are the point density and feature area. Point density is the density of the index "1" in the code. Experiments show that point density is affected by handwritten conceptions. Through verification on the data set, it is found that the characteristics of regional identification accuracy is not high, which is caused by arbitrariness of handwriting. Based on the above analysis, this paper uses the double constraints of "closed area" and "point density" to realize the handwritten digit recognition.

III. DIMENSION REDUCTION OF SAMPLES

DIMENSION REDUCTION METHOD

A digital sample is 32*32 bits, namely 1024-bit binary code. When using kNN algorithm to classify, we need to calculate the Euclidean distance for each test data and the degree of training data to get the familiar degree. Then a classification need do 1024 - bit data operation for 2000 times, which takes a lot of calculation time. If there are tens of thousands of hundreds of millions of a training data, recognition is going to be a long time every time, failing to satisfy the requirement of real-time. Therefore, the dimension reduction is imperative.

Dimension reduction methods can be divided into feature selection and feature extraction. The former is to choose a set of feature subsets from the original attributes or features concentrated according to certain standards, the latter is to transform Meta space into a lower dimension space and map the data to the space. The experimental results show that due to the randomness of handwritten numerals, dimension reduction has reduced recognition rate with extraction methods. The feature selection method can get higher accuracy. This article adopts the way of feature selection for sample dimension reduction.

Feature selection methods of dimension reduction

The purpose of handwritten numerals is to change the 32 * 32 bits of information into sample information of low dimension, less volume. This paper has carried on the analysis and experiment respectively from closed area and density of points.

Through the analysis of the closed area of 0 to 9, and in view of the data set after the experiment, it is found that when take "closed area" as the structural characteristics, identify recognition rate is higher. Therefore, it can use a binary number to represent the characteristics. The first bit indicates whether it is the closed area or not, if it is a double closed area, with a "1", only the figure "8" is accord with the characteristics; if is not in conformity with the characteristics of the double closed area, take it with "0". The second bit indicates whether use the closed area. If use, express in "1", if not expressed in "0". Digits that accord with the characteristics are "0", "6", "8", and "9". It was found that the digit that has no closed area, but easy to cause a closed area number in the handwriting is "4". This is also a factor easy deterrent to any other in experiment, right structure characteristics can be applied in the subsequent study to distinguish.

Every number is made up of 32 * 32 bits of data, a total of 1024 bits; each bit is "0"or "1". Through bits research analysis of each number from 0 to 9, it is found that although the structure of every number is different, the overall structure of the "1" code density difference is not big, which means that it is hard to get ideal recognition rate with the density as characteristic value. Therefore, code line density, density of the column, or diagonal density are analyzed and experimented. According to the implementation results is ideal when take "1" code of density, the density of the column or diagonal density as a statistical characteristic.

Line density of "1" code of the training focus in each row is collected and average density is carried out. "1" density of 32 lines for every number rule is shown in figure 1. From figure 1, it is concluded that line density characteristics of every number's "1" code can be a good characterization of the features of every number.

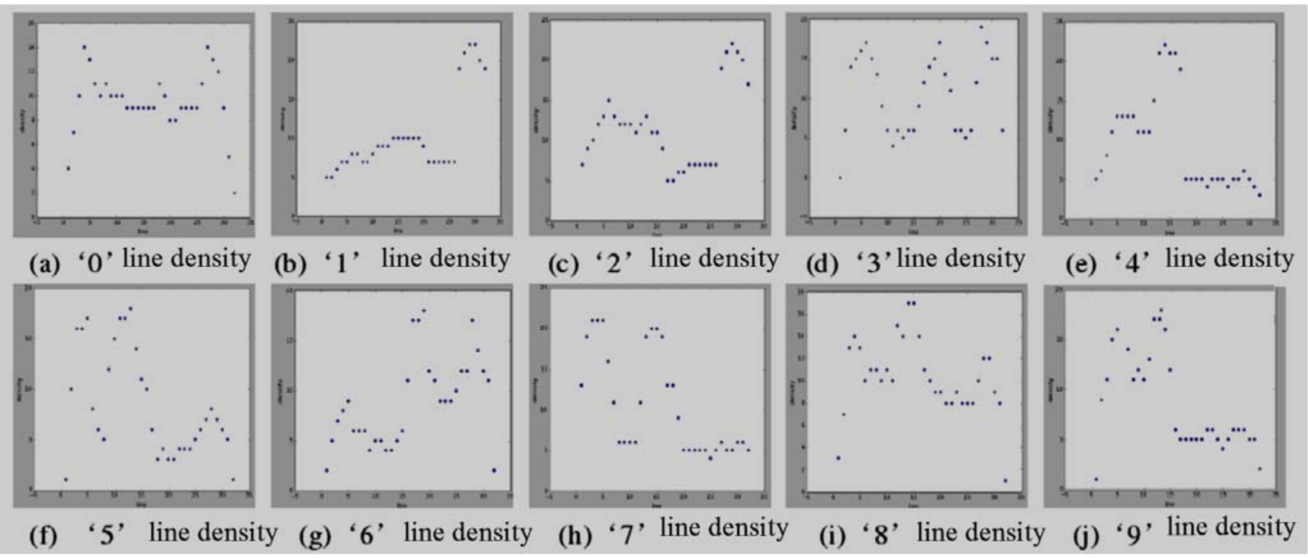

Figure1. The statistical figure of line density of 0 to 9

In the same way, "1" code column density, the density of two diagonal characteristic has carried on the experiment. It is found that they can all be a good characterization of the characteristics of the digital. The results are shown in TABLE I. The experimental results show that using kNN algorithm to do recognition of handwriting numbers, error rate 98 - bit that after the dimension reduction was only 0.07% less than that of than 1024 of the eigenvalue corresponding error rate, but the speed is increased nearly 10 times.

TABLE I ERROR RATE COMPARISON WITH DIFFERENT CHARACTERISTIC VALUE

Error rate	Eigen value	Bits of Eigen value
Line density	14.48%	32
Column density	14.52%	32
Diagonal density	15.96%	32
Course density	5.28%	64
Diagonal density	4.86%	64
Column diagonal density	5.29%	64
The ranks of diagonal density	2.00%	96
Closed area	21%	2
Closed area + diagonal density	1.23%	98
All code in the sample	1.16%	1024

IV. ADAPTIVE K-NN ALGORITHM

For the traditional kNN algorithm, the selection of k value in general is initialized by the designer according to the experience. It is adjusted through the experiment many times, and finally gets the best k value. This method does not possess universality, when the number of data and its training focus changes, k value must be calculated again. In order to improve the generality of classifier, this paper adopts the method of adaptive learning to calculate k values. Set classification results with the results of the test set error rate for $f(k)$, using the gradient descent method obtained by the method of adaptive learning appropriate k value, the specific formula is as follows:

$$k = k - \alpha\nabla_k f(k)$$

Use adaptive k - NN algorithm for handwritten digit recognition, it can be adaptive to calculate k value. Experimental results show that k value gotten by this method is the same as that gotten by repeat experiments. This method can not only get the minimum error rate, but also can effectively shorten the system development time. The experimental data are shown in TABLE II .

TABLE II RECOGNITION ERROR RATE CORRESPONDING TO DIFFERENT K VALUES

k	error rate
2	6.70%
3	1.23%
4	2.10%
5	2.20%
6	2.40%

V. CONCLUSION

This paper has realized a handwritten numeral recognition system of the double constraints of the structural features and statistical characteristics with the adoption of adaptive kNN algorithm. Adoption of the method of sample dimension reduction not only improves the real-time system identification, but also obtains higher identification accuracy. But the sample dimension reduction inevitably brings the loss of sample and reduces the recognition rate, in-depth research is still needed in this area.

ACKNOWLEDGMENT.

This work is supported by the project of Hubei Province Natural Science Foundation(No. BZY14032) and the Fundamental Research Funds for the Central Universities(NO. CZQ14010).

REFERENCES

[1] Shi Enzao. “The KNN quick handwritten digits recognition based on vector projection”. Science and technology, vol. 29, no. 12, pp. 127-129, 2013.

[2] Tan Xueqing Zhoutong, Luoling. “A class average similarity based text categorization algorithm”. Journal of modern information technology, no. 9, pp. 66-73, 2014.

[3] Wang Yimu, Pan Yun, Long Yanchen etc. “Parallel implementation of handwritten numeral recognition based on self-organizing mapping”. Journal of Zhejiang University: engineering science, no. 4, pp. 742-747, 2014.

[4] Xu Qinzhen. “Blended learning model based on improved handwritten Arabic numerals recognition method”. Journal of electronics &information technology, no. 2, pp. 433-438, 2010.

[5] Mejdoub M, Amar C B. “Classification Improvement of Local Feature Vectors over the KNN Algorithm”. Multimedia Tools and Applications, vol. 64, no. 1, pp.197-218, 2013.

[6] Richard J, Chris. “Fuzzy-rough nearest neighbor classification and prediction”. Theoretical Computer Science, vol. 412, no. 42, pp. 5871-5884, 2011.

[7] Pandya D H, Upadhyay S H, Harsha S P. “Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-kNN”. Expert Systems with Applications, vol. 40, no. 10, pp. 4137-4145, 2013.