Analysis of Handwriting Identification Based on Spectral Clustering

Jian Zhou[1, a], Ning Cai[1, b*] and Xiaokun Liu[2, c]

[1]College of Electrical Engineering, Northwest University for Nationalities, China

[2]school of Foreign Language, Northwest University for Nationalities, China

[a]zhoujianfrank@126.com, [b]caining91@tsinghua.org.cn, [c]lxk0523@163.com

*The corresponding author

Keywords: Spectral clustering; Handwriting identification; Data mining; Module identification

Abstract. This paper discusses the analysis of handwriting based on spectral clustering. It is presented that almost figures and letters can be identified from the approach of spectral clustering. The key idea of our approach is that a novel spectral clustering via local projection distance measure is proposed. With the requisite quantity of figure identification has pay more attention from other areas. According to the existing data were described a similarity affinity matrix or Laplacian matrix, in which computed the eigenvalue and eigenvector of the upon matrix and choose the suitable characteristic vector clustering of different data points.

Introduction

This problem of research in the field of human-computer interaction and the category of pattern recognition is of great importance and special values. As an unsupervised classification technique, clustering has been successfully applied to exploratory data analysis, such as image segmentation [1,2], data mining [3,4,5], signal analysis[6], gene expression analysis [7], sport activities analysis in sport domain [8], and other subjects [9]. During the last decade, a number of clustering algorithms were adequately developed. Nonetheless, many of these algorithms are not effective for data classification when applied on nonconvex data space. Compared with the classical clustering, spectral clustering (SC) [10] has been successfully used to identify irregularly shaped datasets and is supported by linear algebra theory. As we know, data points with high similarity should have uniform density and consistent spatial characteristic. Therefore, the key to estimate whether a pair of data points belong to a specific cluster is how to use the data information between them [11]. In view of the handwritten Numbers and letters signal has a strong randomness and uncertainty and the redundant part will cause letter when writing the endpoint detection is not accurate, the change of characteristic parameters, in turn, affects measure estimation, which reduce the recognition rate. Spectral clustering algorithm is to build a reproduction based on graph theory and have the ability to identify a convex distribution of clustering. We can see the sample points as node figure, data are obtained by a segmentation criterion the best 2 d division. Consequently, we can make a propose that character recognition is the key character feature extraction and to construct weighted undirected graph, see the character of each pixel as a node.

Problem Formulation

According a study shown that many tests show that the correct identification Numbers and letters still has more difficulties. So, this objectively to reduce the amount of data used to train neural network and to the training of the neural network [12, 13]. The same as the data quantity is less, it is difficult to verify the neural network generalization ability and character recognition accuracy.

Problem Analysis. Category of Numbers and letters were 10 and 26 species, categories and less strokes is relatively simple, the identification problem seems to be relatively simple [14].But a lot of tests show that the correct identification Numbers and letters are still has more difficulties such as follows:

(1) Glyph of small amount of information, glyph. Types of Numbers and letters though not much,

the writing has distinct regional characteristics, it is difficult to complete it all the world all kinds of writing commonality character recognition algorithm of high recognition rate.

(2) The order of the words difficult to obtain information. May have a broken pen is due to the character, so it is difficult to get a pen stroke order and character of the characters.

(3) Large deformation characters. Handwritten character due to the factors of the writing, make its likely random in the form of characters, such as the size of the characters, tilt, distortion character and the gray level can affect the effective identification.

For character recognition, neural network is often used as the recognition of tool. Is constructed from the neural network is a kind of based on neural network, and simulates the biology, the basic function of the nervous system, network design is simple and can deal with nonlinear problem, which has been widely used in pattern recognition. Neural network successfully used in character recognition, the premise is to design the structure of the neural network and carries on the appropriate training. If the network structural design is unreasonable, improper training, too little training data can cause neural network generalization ability and lose due.

Model Analysis. Aiming at the existing data, classification algorithm is proposed in this paper are as follows:

Step1: For each of the vector data of the data set for sliding filter. Because of the characters in the process of writing may have broken pen and there may be some noise in the data, so the sliding filter for data preprocessing.

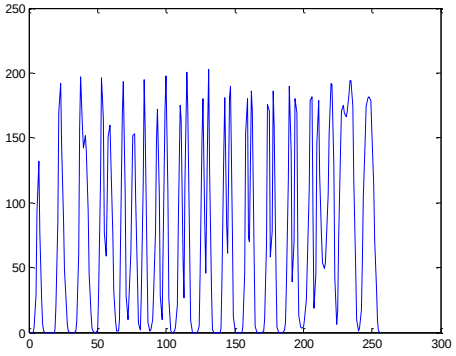

Figure 1. The Original Data of 1

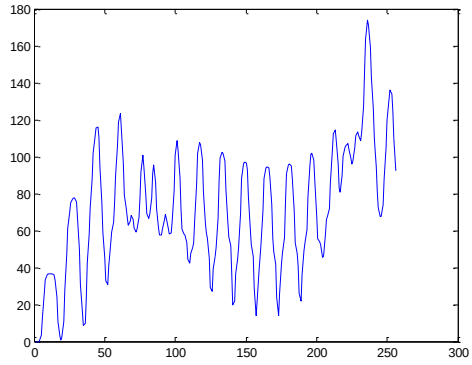

Figure 2. The Data of Sliding Filter of 1

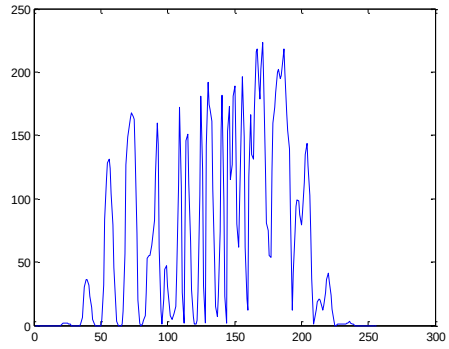

Figure 3. The Original Data of 3

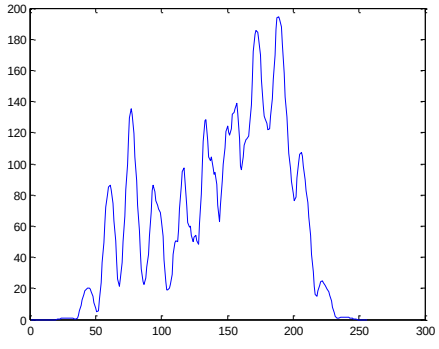

Figure 4. The Data of Sliding Filter of 3

Step2: To filter the data after normalization processing. The sample points as node figure, data are obtained by a segmentation criterion the best 2 d division. As we know, the similarity of the pair of points can be reflected by the distance between the two points. Nevertheless, the pair of points with a longer distance might still belong to the same cluster with a large number of points uniformly

distributed between them. Therefore, the length of the line segment connecting projective image points in LPN can be adjusted by the nonlinear function. According to the spatial structure of the local dataset, a new measure of distance of the pair of points can be obtained through a summation of the values of the length of these line segments as follows:

$$\begin{aligned} \left\| x_{cp} - x_i \right\| &= d(x_i, x_j), \\ \left\| x_{cp} - x_j \right\| &= d(x_i, x_j), \end{aligned} \tag{1}$$

$$\begin{aligned} (x_l - x_{cp1})^2 &\le d^2(x_i - x_j), \\ (x_l - x_{cp2})^2 &\le d^2(x_i - x_j), \end{aligned} \tag{2}$$

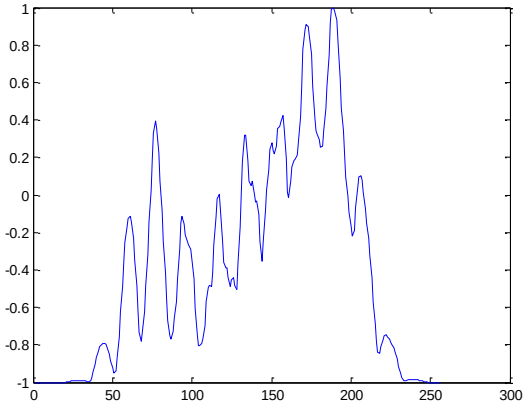

Figure 5. After Sliding Filter Results of Normalization of 3

Step3: To calculate the affinity of vector data of matrix and calculate the similarity of any two points as follows:

$$A_{ij} = \begin{cases} \exp\left\{-d^2(x_i, x_j)/(\sigma_i \sigma_j)\right\}, i \ne j \\ 0, i = j \end{cases} \tag{3}$$

Step4: A normalized processing of affinity matrix as follows:

$$\begin{aligned} L &= D^{-1/2} A D^{-1/2} \\ D_{ij} &= \sum_{j=1}^{n} A_{ij} \end{aligned} \tag{4}$$

Step5: In the new coordinate space with matrix Y every line of figures as coordinates, after the normalization of STEP2 to k-means clustering vector data.

Results Analysis. For the result of objective evaluation classification, classification accuracy indicators used here *Acc*

$$Acc = \max_{map} \sum_{i=1}^{n} \delta(y_i, map(c_i)) / n \tag{5}$$

Table 1 The Classification Results

Character	Classification Accuracy	The Highest Accuracy of Figure	The Lowest Accuracy of Figure
Handwriting	43.9%	3	9

Conclusion

In this paper, the algorithm can write Numbers and letters in a certain degree of recognition. Due to the characteristics of handwritten character data and contest the constraints of the information provided, the algorithm of classification number and the amount of data is less data set can do it better classification, then the classification number and the increase of the amount of data, character recognition difficulty gradually increase, decrease its classification accuracy.

References

[1] A. Rajendran and R. Dhanasekaran, "Enhanced possibilistic fuzzy C-means algorithm for normal and pathological brain tissue segmentation on magnetic resonance brain image," Arabian Journal for Science and Engineering, vol. 38, no. 9, pp. 2375–2388, 2013.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AIMagazine, vol.17, no. 3, pp. 37–53, 1996.

[3] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999.

[4] Ding C,He X,Zha H,et al.Spectral M in Max cut for Graph Partitioning and Data Clustering[C]//Proc. of the IEEE Intl conf .on Data Mining .2001:107-114.

[5] Melia M,Xu L,Multiway cuts and spectral clustering.U.Washington Tech Report.2003.

[6] Witten I H, Frank E. Data Mining: Practical machine learning tools and techniques [M]. Massachusetts: Morgan Kaufmann, 2005: 81-82.

[7] Revor Hastie, Robert Tibshirani, Friedman J J H. The elements of statistical learning [M]. New York: Springer, 2001: 460-514.

[8] Xiang T, Gong S. Spectral clustering with eigenvector selection [J]. Pattern Recognition, 2008, 41(3): 1012-1029.

[9] Shi J,Malik J.Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):887-905.

[10] Chen W, Giger M L. A fuzzy c-means (fcm) based algorithm for intensity inhomogeneity correction and segmentation of MR images[C].From Nano to Macro Marriott Crystal Gateway. Arlington: IEEE Press, 2004:1307-1310.

[11] H. Chang and D.-Y. Yeung, "Robust path-based spectral clustering," Pattern Recognition, vol. 41, no. 1, pp. 191–203, 2008.

[12] X. Zhang, J. Li, and H. Yu, "Local density adaptive similarity measurement for spectral clustering," Pattern Recognition Letters, vol. 32, no. 2, pp. 352–358, 2011.

[13] N. Cai, J. Cao,M. Liu, and H.Ma, "On controllability problems of high-order dynamical

[14] multi-agent systems," Arabian Journal for Science and Engineering, vol. 39, no. 5, pp. 4261–4267, 2014 [14] X. Y. Li and L. J. Guo, "Constructing affinity matrix in spectral clustering based on neighbor propagation," Neurocomputing, vol. 97, pp. 125–130, 2012.

