IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 6, Ser. I (Nov. – Dec. 2022), PP 01-03
www.iosrjournals.org

Research on K-means clustering classification technology

Chu Fang
College of Economics and Management,Zhaoqing University, Zhaoqing City, Guangdong, China

Abstract: *Clustering is a data analysis technique that can form different subsets of similar objects through clustering methods, and the objects in the same subset have some similar properties. Common methods include the number of adjacent points in the same space, and the The shortest spatial distance in the coordinate axis, etc., the application fields include machine learning, data mining, pattern recognition, image analysis and bioinformatics. This research is mainly divided into two parts, the first is the application of K-means clustering to food image cutting; the second is the application of K-means clustering to general numerical data. Both parts are modified to facilitate image segmentation. In the first part, this study will demonstrate that the application of food grade analysis can cut a better number of groups, and the second part is the K - The application of the mean method clustering is more expanded and can be applied to general values, and the part of modifying the K-means method clustering is to improve the error of the initial cluster center point in the selection of random numbers, and the selection of the number of subgroups , hope to select a better number of groups. Finally, it can be verified by experiments that the modified K-means clustering proposed in this study is helpful to improve the effectiveness of the original K-means clustering.*

Key Word*: Image Segmentation;C-Ablation; K-Means Clustering; Extended K-Means Clustering*

Date of Submission: 22-10-2022 Date of Acceptance: 05-11-2022

I. Introduction

Clustering is a technique for data analysis that is widely used in many fields, including machine learning, information mining, pattern recognition, image analysis, and bioinformatics. Clustering is to combine similar objects into many subsets through the method of classification, so that the objects in the same subset have similar properties. Common ones include the number of adjacent points in the same space and the shortest space in the coordinate axis. distance etc. The objects of this research are image analysis and general numerical analysis. As for the image analysis, because an image can contain a lot of information, for example, an image can basically represent the information in the image. People, things, times, places, objects, but images are formed by a variety of features, including color, brightness, texture, corners, edges, and so on. Therefore, in order to understand the information inside the image and the meaning it wants to convey, the image must be analyzed. However, the image needs to be processed before analyzing the image. The method of this study is to use the classification technology. Therefore, this research is a method of image analysis using classification technology. However, the image is analyzed after obtaining the image, and the method of using classification method to analyze the image includes K-means clustering (K-means clustering)[1] [2].There are various methods such as fuzzy clustering (Fuzzy c-means), robust algorithm (Robust algorithm), and the method selected in this study is K-means clustering. better, and more powerful than the robust algorithm. Therefore, the technique adopted in this research is K-means clustering, and the original K-means clustering selects the initial The cluster center point is selected using random numbers, and the selected cluster center point may be too close or the same phenomenon, and such an initial cluster center point may cause a reduction in the number of groups or selection errors, while the original K- The mean method clusteringalso cannot automatically determine the number of groups, which must be determined by the user.Judging the number of groups, but whether the selection of the number of groups is appropriate will affect the effect of cutting. Therefore, it is worthwhile to design a selection mechanism for the center point of the initial group and the selection mechanism for the optimal number of groups[3] [4].

For the selection of the number of groups, the first step of the general K-means clustering algorithm is to input the number of groups, but this research design is based on the K-means clustering algorithm that does not set the number of groups. The use mechanism is to set a threshold value (ε). In the hypothesis, this study sets the background color group of the group number as the benchmark group and compares it. The termination condition of the algorithm is that when the benchmark group shows a convergence state , in general, when using the K-means clustering method to cut before imaging, it is worth discussing why the number of groups to be grouped is determined. For example, the analyzed images are edible beef, and the intuitive classification will be divided into background, fat and lean parts, but in the experiment, the images are simply divided into three types

of images. The segmentation will be incomplete, and the image segmentation cannot be supported when it is divided into four categories[28]. Therefore, a mechanism is provided to select a better number of groups. The relevant variables are defined as follows:

- Vm: Standard tuple.
- C(Vm): The cluster center point of the reference group.
- t: $t \geq 2$ is the number of groups. Algorithm: Extended K-Means Clustering

Input: threshold value (ε), image X = {x1, x 2,…, xn}, xi is the grayscale value of each point. Output: Optimal number of groups and images after grouping.

II. Research model and hypothesis

This chapter mainly describes the experiment and discussion on the image cutting ability of K-means clustering when applied to food quality analysis, as well as the experimental presentation of extended K-means clustering for general numerical data grouping, including the selection of the initial cluster center point, The selection of the number of clusters by extending the K-means method.

Canadian Beef Image: In Figure 3.1 Canadian Beef Prime grades, the Prime grade beef of the four high quality grades (A, AA, AAA, Prime) of Canadian beef grades is represented, and its characteristics must be at least 4 mm Thick, firm and white in back fat, or slightly reddish or amber in color, and more oily on the inside, if these same standards of beef are present in the image, it is called Canadian Prime Prime, and the image shown in Figure 3.2 Beef grade A of the four high-quality grades graded for Canadian beef, characterized by a minimum thickness of 4 mm, firm and white or slightly reddish or amber-colored back fat, and little to trace oil inside Flowers,

III. Empirical Research

Extending the selection of the initial cluster center point of K-means clustering, it can be known from the experimental results in Section 3.1 that the K-means clustering designed in this study can effectively classify and cut food images, while for other types of The data analysis did not have a better effect and application, so this study improved its method and designed another extended K-means method to cluster general numerical data. This section will introduce the comparison and discussion of the selection of the extended K-means method clustering on the initial cluster center point and the original K-means method clustering on the initial cluster center point.

The data used in this section is generated using the mvnrnd (Multivariate normal random numbers) system command in MATLAB, and the analysis data includes 50 groups, which are divided into c3_1~10, c4_1~10, c5_1~10, c6_1~10, c7_1 ~10, c3 represents 3 groups, c4 represents 4 groups, and so on, and the number of points in each group is 25 points, and the center point is generated using random numbers from (0 ,0)~(100,100), and the density of the group is the covariance generated between (0,0)~(10,10) using random numbers. Taking Table 3.3 and Figure 3 as an example, c3_5 generates In the data of , c3 indicates that there are three center points and three covariances, and also indicates that there are three groups of data on the plane in the two-dimensional image.

IV. Result

In the experiment of this research, it can be known that the use of classification technology can effectively cut the image, and the cut image can be analyzed. In this research, the K-means method clustering is modified, so that the image cutting can be selected. A better initial cluster center point will not cause segmentation errors or slow execution efficiency due to the same or close initial center points, and the advantage of not setting the number of clusters can avoid human perception errors such as excessive cutting Or lack of groups, find a better number of groups to facilitate image cutting and subsequent image analysis, and use the classified and cut images to analyze the images to learn the meaning and information contained in the images. In the second part, we can know that the extended K-means clustering in the study is not only effective for image cutting, but also for multi-dimensional data analysis. Compared with the original K-means clustering, the mean method clustering has a good improvement rate, but the selection of the number of subgroups improves the shortcomings of the original K-means method clustering that cannot select the number of groups, and makes the extension The K-means method of clustering can select the best number of clusters.

At present, this research is only conducted on the two food images of apple and beef. In the future work, we hope to conduct experiments on other images, such as portrait images, satellite images, etc., and expect more diverse applications in classification technology research. For example, adding the function of self-learning and strengthening the function of selecting the optimal number of groups can correct the shortcomings of the original classification technology.

References

[1]. S.C. Ahalt, A.K. Krishnamurty, P. Chen, and D.E. Melton (1990), "Competitive Learning Algorithms for Vector Quantization," Neural Networks, Vol. 3, pp. 277-291.

[2]. H. Akaike (1973), "Information Theory and Extensions of the Maximum Likelihood Principle", Second Int'l Symp. Information Theory, p. 267-281.

[3]. H. Akaike (1974), "A New Look at the Statistical Model Identfication", IEEE Trans. Automatic Control 716-723.

[4]. J. Bezdek (1981), "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.

[5]. H. Bozdogan (1987), "Model Selection and Akaike's Information Criterion the General Theory and its Analytical Extensions", Psyhometrika 52, pp. 345-370.

[6]. J. Breckenridge (1989), "Replication Cluster Analysis: Methods, Consistency and Validity", Multivariate Behavioral Research.

[7]. J. Brendon Woodford (2008), "Evolving Neural Computational Systems for Horticultural Applications," Applied Soft Computing, Vol. 8, p. 8. 564-578.

[8]. Y.M. Cheug (2005), "On Competitor Penalty Controlled Competition Learning and Automatic Cluster Number Selection for Clustering", IEEE Trans. Knowledge Data, Vol. 17, p. 17. 1583-1588.

[9]. K. Chen and Ch. Qin (2008), "Beef Marbling Segmentation Based on Visual Thresholding", Computers and Electronics in Agriculture, Vol. 62, No. 2, p. 2. 223-230.

[10]. J. Friedley and S. Dudoit (2001), "Application of Resampling Methods to Estimating the Number of Clusters and Improving the Accuracy of Clustering Methods", Technical Report 600, UC Berkeley Statistics Division, September.

[11]. B. Fritzke (1994), "Growing Cell Structure - Self-Organizing Networks for Unsupervised and Supervised Learning", Neural Networks, vol. 7. No. 9, pp. 1441-1460.

[12]. R.M. Gray (1984), "Vector Quantization," IEEE ASSP Magazine, vol.1, pp. 4-29.

[13]. Guo, C.L. Philip Chen and M.R. Lye (2002), "Cluster number selection for a small set of samples using a Bayesian yin-yang model", IEEE Trans. Neural Networks, Volume 13, no. 3, pp. 757-763.Meyers-Levy, J.; Sternthal, B. Gender differences in the use of message cues and judgments. *J. Mark. Res.* **1991**, *28*, 84–96.

Chu Fang. "Research on K-means clustering classification technology." *IOSR Journal of Computer Engineering (IOSR-JCE),* 24(6), 2022, pp. 01-03.