БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ. BULGARIAN ACADEMY OF SCIENCES

ПРОБЛЕМИ НА ТЕХНИЧЕСКАТА КИБЕРНЕТИКА И РОБОТИКАТА, 53
PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS, 53

София . 2002 . Sofia

Bulgarian Hand-Printed Character Recognition Using Fuzzy C-Means Clustering

Rumiana Krasteva

Central Laboratory of Mechatronics and Instrumentation, 1113 Sofia
E-mail: veso@bgcict.acad.bg

I. Introduction

The theory of fuzzy sets has immediately found its potential application in the fields of pattern recognition and image processing. There are several fuzzy-set models for solving problems in pattern recognition and image processing. Some well known models are: the use of fuzzy membership function, fuzzy clustering, fuzzy rule based systems, fuzzy entropy (measure of fuzziness), fuzzy measure and fuzzy integral.

The recognition accuracy strongly depends on the selected pattern features. There exist various approaches to generate data dependent fuzzy rules. The best known are implementations based on statistics, neural networks, genetic algorithms and fuzzy clustering.

This paper presents an approach for rule base generation for handprinted Bulgarian characters. The general idea of the approach is based on fuzzy C-means clustering algorithm.

MATLAB possibilities are used for the experiments.

Program BitScan is developed for input image format transformation in conformance with MATLAB. The program processes the input character and converts the format into ".dat" file, needed for MATLAB.

The segmentation problem is not discussed in the paper. Experiments are made for each character separately

The basic clustering conception is described in section II. C-means algorithm and results of experiments are discussed in section III.

II. Clustering methods

1. Fuzzy clustering

The goal of a clustering analysis is to divide a given set of data or objects into a cluster, which represents subsets or a group. The partition should have two properties:

- Homogeneity inside clusters – the data, which belongs to one cluster, should be as similar as possible.
- Heterogeneity between the clusters – the data, which belongs to different clusters, should be as different as possible.

The membership functions don't reflect the actual data distribution in the input and the output spaces. They may not be suitable for fuzzy pattern recognition. To build membership functions from the data available, a clustering technique may be used to partition the data, and then produce membership functions from the resulting clustering.

"Clustering" is a process to obtain a partition P of a set E of N objects X_i (i=1, 2,…, N), using the resemblance or disemblance measure, such as a distance measure d. A partition P is a set of disjoint subsets of E and the element P_s of P is called *cluster* and the centers of the clusters are called *centroids* or prototypes.

Many techniques have been developed for clustering data. In this paper c-means clustering is used. It's a simple unsupervised learning method which can be used for data grouping or classification when the number of the clusters is known. It consists of the following steps:

1) Choose the number of clusters – k;

2) Set initial centers of clusters – c_1, c_2,…, c_k;

3) Classify each vector $x_l=[x_{l1}, x_{l2},…, x_{ln}]^T$ into the closest center c_i by Euclidean distance measure:

(1) $$\| x_l - c_i\| = \min \|x_l - c_j\| .$$

(4) Recompute the estimates for the cluster centers c_i. Let $c_i = [c_{i1},c_{i2},…,c_{in}]^T$, c_{im} be computed by:

(2) $$c_{im} = \frac{\sum x_{li} \in \text{cluster}(i^{x_{\text{lim}}})}{N_i} ,$$

where N_i is the number of vectors in the I-th cluster.

5) If none of the cluster centers (c_i, I – 1, 2,…, k) changes in step 4 – stop; otherwise go to step 3.

1. C-means algorithm

The criterion function used for the clustering process is:

(3) $$J(V) = \sum_{k=1}^{n} \sum k_x \in c_i \left| X_k - V_i \right|^2 ,$$

where v_i is the sample mean or the center of samples of cluster I, and $V=\{v_1,v_2,…,v_c\}$.

In the hard clustering process, each data sample is assigned to onty one cluster and all clusters are regarded as disjoint collection of the data set. In practice there are many cases, in which the clusters are not completely disjoint and the data could be classified as belonging to one cluster almost as well to another. Therefore, the separation of the clusters becomes a fuzzy notion, and representation of the data can be more accurately handled by fuzzy clustering methods. It is necessary to describe the data in terms of fuzzy clusters.

The criterion function used for fuzzy C-means clustering is

$$J(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\left|X_k - V_i\right|^2, \quad (4)$$

where:

$x_1,\ldots,x_n$ – n data sample vectors;

$V=\{v_1,v_2,\ldots,v_c\}$ – cluster centers (centroids);

$U=[u_{ik}]$ – cxm matrix, where u_{ik} is the I-th membership value of the k-th input sample x_k, and the membership values satisfy the following conditions:

$$0 \le u_{ik} \le 1 ; \quad i=1,\ldots,c; \; k=1,\ldots,n;$$

$$\sum_{i=1}^{c} u_{ik} = 1 ; \quad k = 1,\ldots, n;$$

$$0 < \sum_{k=1}^{n} u_{ik} < 1; \quad i = 1,\ldots, c;$$

$m \in [1,\infty)$ is an exponent weight factor.

This paper presents an approach for rule base generation, using fuzzy c-means clustering. The algorithm and the experimental results are described in the next section.

III. Rule base generation using fuzzy C-means algorithm

The general idea of the approach is to apply C-means clustering, where the number of the clusters are apriori defined – 5. Rule base generation is divided into 2 phases: Phase 1 – computing cluster centers for the input character; Phase 2 – comparing centroids position. For each of the Bulgarian printed characters, centroid position is calculated in advance. MATLAB possibilities are used in the experiments. For new character processing the sequence of phase1 and phase2 is started. The process algorithm is shortly described in Fig. 1.

Each of the clusters is presented by its centroid position, the calculation being based on membership functions for data, belonging to this cluster. For comparing centroid positions are applied, using a deviation of 10 discrets.

When non-coincidence is detected for three positions, the algorithm starts again for comparison with the next template.

When coincidence exists for more than three positions, the input character is assumed equivalent with template.

For each character a program for "normalization" is executed beforehand. The program is written in TclTK and converts the character according to MATLAB requirements (.dat). The program has a possibility for different file processing (.bmp, .gif,.tif, .jpg,).

Input:

- set of data (A), which need to processed
- set of cendroid positions (B)

Output: Fuzzy rule base generation for A

Method:

- phase 1 – computation cluster positions and centroids;
- deviation – 10 discrets;
- phase 2 – computing and comparing centroid positions.

Fig. 1. Rule base generation algorithm

The number of the clusters – 5, is chosen experimentally and it is enough for rule base generation. Fig. 2 illustrates the experiments for one Bulgarian character recognition using fuzzy C-means.

Fig. 2 (a). Bulgarian handprinted letter

Fig. 2 (a) shows one of the handprinted Bulgarian letters. After processing Fig. 2 (b) shows computed centroid positions. Fig. 2(c) shows membership functions for one of the clusters.

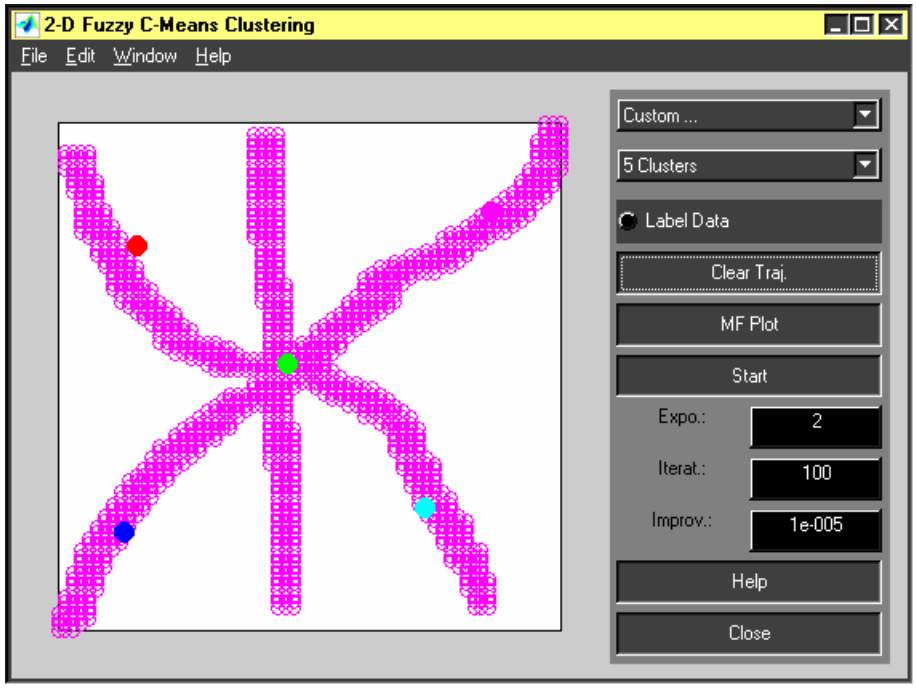

Fig. 2 (b). Determining clusters and **c**entroid positions

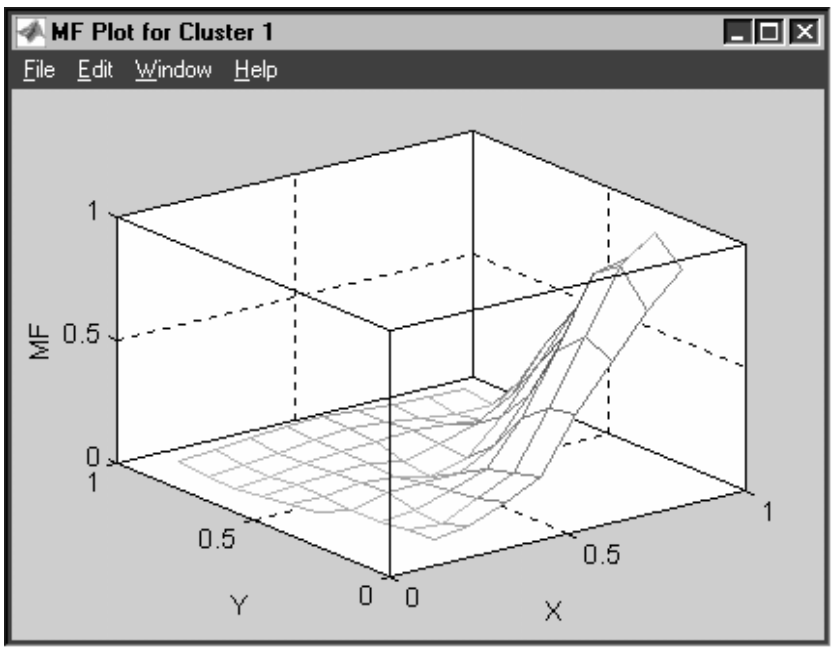

Fig. 2 (c). Membership functions for one of the clusters

In the experiments some errors appeared for some characters depending on various style of writing. In these cases one more iteration is necessary for clusters and centroids determination.

The recognition occurrence depends on the given deviation for centroid positions comparison. A deviation of 10 discrets is chosen for the experiments.

IV. Conclusion

The goal of this experiment is to study and research the possibility for decreasing the recognition criteria. The segmentation problem is not yet discussed. In this experiment the results were tested after using the clustering approach. It is applicable for Bulgarian handprinted characters. Only capital letters have been tested in this experiment.

To improve the recognition results mixed fuzzy set theory for curve categories and clustering approach for each of the curves should be applied.

References

1. Zheru Chi, Hong Yan, Tuan Pham. Fuzzy Algorithms: With Application to Image Processing and Pattern Recognition. World Scientific, 1996.
2. Georgtchev, V., R. Krusteva, A. Boneva, K. Stanischev. Experimentally analysis on old bulgarian text character recognition. – In: MIM 200 IFAC Symposium on Manufacturing, Modeling and Control, University of Patras Rio, Greece (July 12-14,2000), Proceeding ISBN 0 08043554, 124-127.
3. Peters, L, C. Leja, A. Malaviya. A fuzzy statistical rule generation method for handwriting recognition. – Expert Systems, February 1998, Vol. 15, 1998, No 1.
4. Malaviya, A., L. Peters. Fuzzy Feature Description of Handwriting Patterns. – Pattern Recognition, **30**, 1997, No 10, 1591-1604.
5. Wilson, C. L. Evaluation of character recognition systems, neural networks for signal processing III. – In: IEEE, New York, 1992, 485-495.
6. Georgtchev, V., D. Butchvarov, A. Boneva , R. Krasteva, K. Stanischev. Letter characters recognition after information loss. – In: Proceedings “Scientific reports”. Section 3: Mechatronics, ISSN 1310-3946, Sofia, Bulgaria, 1999, 3.39-3.44 (in Bulgarian).

Распознавание болгарских манускриптных букв, используя подход fuzzy C-means

Румяна Крастева

Центральная лаборатория мехатроники и приборостроения, 1113 София

(Резюме)

Обсуждается подход для распознавания манускриптных болгарских букв на основе генерирования правил размитых множеств. Подход основан на алгоритме fuzzy C-means. Эксперименты проводятся в среде MATLAB, используя специальные програмы для трансформации формата. Описаны основная идея подхода и результаты экспериментов.