

Based On K-Means and Nearest Neighbor Algorithm for Fuzzy System Used for Data Fitting

Yi Hu, Jixia Han and Songyi Dian*

College of Electrical Engineering and Information technology, Sichuan University, Chengdu, China

*scudiansy(at)mail.scu.edu.cn

Abstract. This paper proposes using classical clustering algorithm in fuzzy system to fit experimental data. The main purpose is to reduce the number of fuzzy rules by clustering. K-means algorithm and nearest neighbor algorithm clusters classify data into classes so that each cluster corresponds to a fuzzy rule, the number of fuzzy rules is also determined by the number of clusters. We compare and discuss the results of data fitting between K-means algorithm and nearest neighbor algorithm. It find that two algorithm both can achieve better data fitting performance.

1. Introduction

Nowadays, many things cannot be simply divided into yes or no. The fuzzy system provides new ideas and tools for people to analyze and solve this problem. The theory of fuzzy systems was founded by Zadeh in 1965. It is a system based on knowledge or rule. With the continuous development of fuzzy theory, fuzzy systems began to be combined with various algorithms. In paper [1] improved the T-S fuzzy system on stabilization. The author [2] did research on a on smooth compactly supported radial fuzzy system. In paper [3] used participatory search algorithm in fuzzy systems modeling. In [4] proposed a rainfall forecasting using fuzzy system based on genetic algorithm.

Clustering is an unsupervised identification and classification process that divides the data into different groups. Clustering has been widely used in many fields, including signal processing, pattern recognition, bio-engineering, and image segmentation. Several clustering methods have been proposed. Classical clustering algorithms include K-means algorithm and nearest neighbor algorithm. In order to improve the classification speed and effectiveness, people have made the following efforts. The paper [5-7] proposed a K-means clustering algorithm based on feature weight. In paper [8-9] enhanced K-means by using PSO algorithm. The research [10] used nearest neighbor algorithm in human activity recognition.

This paper applies the K-means algorithm and nearest neighbor algorithm to the fuzzy system. The basic idea is to reduce the fuzzy rules by clustering, and one set corresponds to a fuzzy rule. Experiments show that the fitting error of data pairs can be reduced.

2. Fuzzy system structure

The fuzzy system is composed of four parts: fuzzy rule base, fuzzy inference machine, fuzzifier, and defuzzifier. We often use product inference engine, single-valued fuzzifiers and center average defuzzifier.

The fuzzy rule base is composed of the following fuzzy “If-Then” rules:



$$R_u^{(l)} : \text{if } x_1 \text{ is } A_1^l \text{ and } \dots \text{and } x_m \text{ is } A_m^l, \text{ then } y \text{ is } B^l \quad (1)$$

Where A_i^l and B^l are fuzzy set in $U^i \subset R$ and $V \subset R$, $x = (x_1, x_2, \dots, x_n)^T \in U$ and $y \in V$ are respectively input variables and output variables of fuzzy system.

It is given a finite set of input-output data couples $(x_0^l; y_0^l) (l = 1, 2, \dots, N)$. Fuzzy system consists of N rules like (1), it described as follows [11]:

$$f(x) = \sum_{l=1}^N y_0^l \exp\left(-\frac{|x - x_0^l|^2}{\sigma^2}\right) / \sum_{l=1}^N \exp\left(-\frac{|x - x_0^l|^2}{\sigma^2}\right) \quad (2)$$

Where y_0^l is equal to the center of B^l and the fuzzy set $\mu_{A_i^l}$ uses a Gaussian membership function given by:

$$\mu_{A_i^l}(x_i) = \exp\left(-\frac{|x_i - x_{0i}^l|}{\sigma}\right) \quad (3)$$

The appropriate value σ can make the fuzzy system to fit all input-output data couple within any error. We have $|f(x_0^l) - y_0^l| < \varepsilon$ ($l = 1, 2, \dots, N$) when every $\varepsilon > 0$.

3. Clustering algorithm

3.1. K-means algorithm

The K-means algorithm proposed by MacQueen is a classic algorithm to solve the cluster analysis problem [12]. The main principle of k-means is to divide a group of data into k clusters based on the distance between the data and have a higher similarity within the cluster and a lower degree of similarity between clusters. Algorithm works by minimizing the cost function using an iterative optimization technique as follows:

$$J = \sum_{i=1}^K J_i = \sum_{i=1}^K \sum_{j=1}^N w_{ji} \|X_j - C_i\|^2 \quad (4)$$

Where K and N are the number of clusters and samples in data set, respectively. $\|X_j - C_i\|$ is the Euclidean distance between the i th cluster center C_i and j th data X_j . w_{ji} presents the weight of the data X_j in the i th cluster. It can be calculated by following equation:

$$w_{ji} = \begin{cases} 1, & \text{if } \|X_j - C_i\| \leq \|X_j - C_m\|, \forall m \neq j \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$$s.t. \sum_i^K w_{ji} = 1, \forall j = 1, 2, \dots, N; \sum_i^K \sum_j^N w_{ji} = N$$

The following data clustering techniques using the K-means algorithm as follows.

Step1: Initialize algorithm parameters, including initial cluster random centers, cluster number, iteration number and termination criterion;

Step2: Calculate the euclidian distance of the remaining data points from the means of all the clusters. If the data point X_j is determined to belong to the i th cluster, the weight value $w_{ji} = 1$ otherwise it is 0.

Step3: Calculate the cost function. If $\|J^{k+1} - J^k\| < \varepsilon$ then stop; else continue.

Step4: Update the cluster centers using the following formula(6) and go to step2.

$$C_i = \sum_{j=1}^N w_{ji} X_j / \sum_{j=1}^N w_{ji} \quad (6)$$

3.2. Nearest neighbor algorithm

Nearest neighbor algorithm is highly popular and effective in the field of pattern recognition. It is different from K-means algorithm that it does not need to specify the cluster number. Similarly, it divides a group of data into k clusters based on the distance.

The algorithm of nearest neighbor is as follows:

Step1: Set up radius r and select the first data as initial cluster centers.

Step2: Calculate the euclidian distance of the remaining data points from the centers of all clusters.

$\|X_j - C_i\|$ is the shortest Euclidean distance between the i th cluster center C_i and j th data X_j .

(a) If $\|X_j - C_i\| > r$, X_j becomes a new cluster centers.

(b) If $\|X_j - C_i\| \leq r$, X_j is assigned to C_i .

Step3: Until all point is assigned to the nearest cluster based on its distance from the center of each cluster and stop.

4. Experiment and discussion

It is given N input-output data couples. We separately use K-means algorithm and nearest neighbor algorithm in fuzzy system.

Step1: Using K-means to cluster data couples and until the cost function stabilize or using nearest neighbor to cluster data couples and until all point is assigned to the nearest cluster.

Step2: If the cluster C_i ($i = 1, 2, \dots, K$) has m data couples, $A^i = y_i^1 + y_i^2 + \dots + y_i^m$, and $B^i = m$.

Step3: The fuzzy system is as follows:

$$f(x) = \sum_{i=1}^K A^i \exp\left(-\frac{|x - C_i|^2}{\sigma}\right) / \sum_{i=1}^K B^i \exp\left(-\frac{|x - C_i|^2}{\sigma}\right) \quad (7)$$

For K-meas algorithm, we initialize initial random cluster centers, cluster number, iteration number and termination criterion. The result is shown below.

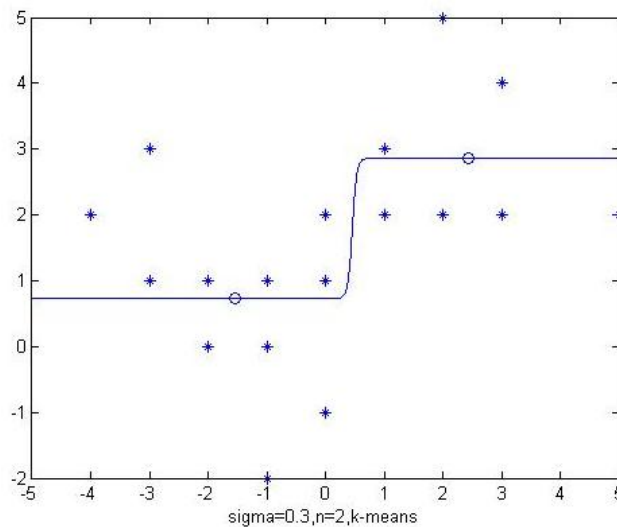


Figure 1. Data fitting of K-means algorithm using 2 clusters

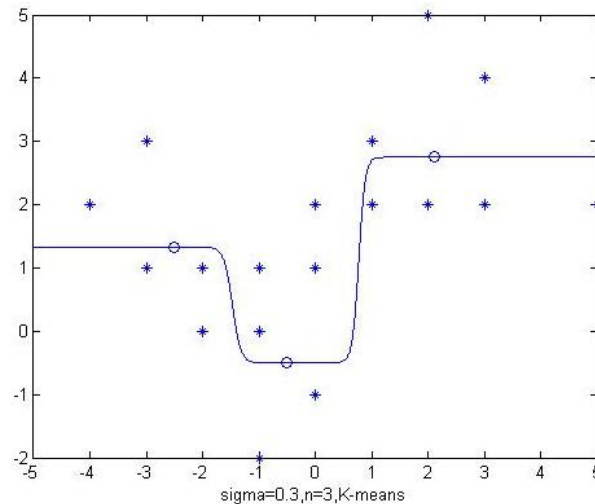


Figure 2. Data fitting of K-means algorithm using 3 clusters

From the above figure 1 and figure 2, we can find that it have a poor performance on fitting data when cluster number is 2. It shows that the fitting effect is related to the number of clusters. We can have better data fitting effects by appropriately increasing the number of clusters.

For nearest neighbor algorithm, we set up radius and initial cluster centers. The result is shown below.

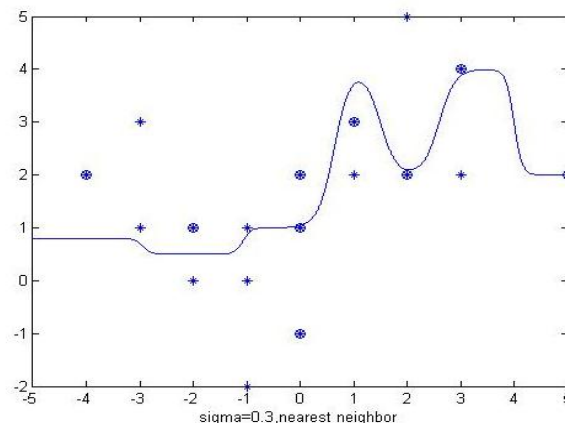


Figure 3. Data fitting of nearest neighbor algorithm

Nearest neighbor algorithm does not need to specify the cluster number. From the above figure 3, we find that it is over-fitting because of having too much cluster. When there are too many data couples, it is easy to over fitting.

Mean Squared Error(MSE) can evaluate the effect of data fitting. The smaller the value of MSE, the better the accuracy of the prediction model to describe the experimental data.

$$MSE = 1/n * \sum_{i=1}^n (X_n - X_n^*)^2 \quad (8)$$

Where X_n and X_n^* is the Experimental data and Model data, n is the number of data.

We use MSE to measure the data fitting results and calculate MSE values for K-means algorithms and nearest neighbor algorithm separately. As shown in Table 1 below.

Table 1. The MSE of K-means and nearest neighbor algorithm.

	K-means algorithm (cluster number=2)	K-means algorithm (cluster number=3)	nearest neighbor algorithm
MSE	1.6133	1.2684	2.0301

From Table 1, we find that the MSE of K-means algorithm (cluster number=3) is the smallest. It has a better behavior on data fitting. But using nearest neighbor algorithm can't fit data well. If we increase the cluster number of K-means algorithm, we can get better data fitting results.

5. Conclusion

This paper reduces the number of fuzzy rules by combining clustering algorithms and fuzzy systems. Classic clustering algorithms include K-means and nearest neighbor. The performance on fitting data of the fuzzy system is related to the number of clusters. The disadvantage of the K-means algorithm is that the number of clusters needs to be set in advance, but the nearest neighbor algorithm does not need to. In the case of a large number of data, the nearest neighbor algorithm tends to generate too many clusters leading to over fitting. We can improve the nearest neighbor algorithm that if the membership of a cluster has not changed for a long time, cancel this cluster.

Acknowledgment

This work is supported by Fundamental Research Special Funds for Central Universities (20826041A4133).

6. Reference

- [1] Lian, Z., He, Y., Zhang, C., & Wu, M. (2017). Improved stabilization conditions for T-S fuzzy systems with interval time-varying delay. Chinese Control Conference (pp.384-389).
- [2] David Coufal M.(2017). On Smooth Compactly Supported Radial Fuzzy System (pp.39-43). IEEE
- [3] Liu, Y. L., & Gomide, F. (2017). Fuzzy systems modeling with participatory search algorithm. Fuzzy Systems Association and, International Conference on Soft Computing and Intelligent Systems(pp.1-6). IEEE.
- [4] Nhita, F., & Adiwijaya. (2013). A rainfall forecasting using fuzzy system based on genetic algorithm. Information and Communication Technology(pp.111-115). IEEE.
- [5] Wang, S., Fan, Y., Zhang, C., Xu, H. X., Hao, X., & Hu, Y. (2008). Subspace Clustering of High Dimensional Data Streams. Ieee/acis International Conference on Computer and Information Science (pp.165-170). IEEE Computer Society.
- [6] Ren, J. T., Shi, X. X., Sun, J. H., Huang, H. Y., Yin, J., M.(2006). An improved k-means clustering algorithm based on feature weighting. computer science,33(7), 186-187.
- [7] Zeng, B., Zhao, W., Luo, C., & Chen, B. (2010). The Optimization Arithmetic of K-means Clustering Based on Indirect Feature Weight Learning. International Conference on Computer and Communication Technologies in Agriculture Engineering(cctae 2010) (volume (Vol.2, pp.243-246).
- [8] Gupta, A., Pattanaik, V., & Singh, M. (2017). Enhancing K means by unsupervised learning using PSO algorithm. International Conference on Computing, Communication and Automation (pp.228-233).
- [9] Wang, X., & Sun, Q. (2017). The Study of K-Means Based on Hybrid SA-PSO Algorithm. International Symposium on Computational Intelligence and Design (pp.211-214). IEEE.
- [10] Zul, M. I., Muslim, I., & Hakim, L. (2017). Human Activity Recognition by Using Nearest Neighbor Algorithm from Digital Image. International Conference on Soft Computing, Intelligent System and Information Technology (pp.58-61). IEEE Computer Society.
- [11] Wang, L. X. (1996). A course in fuzzy systems and control. Prentice-Hall, Inc.
- [12] Tou, J. T., & Gonzalez, R. C. (1974). Pattern recognition principles., 17(10), 274-282.