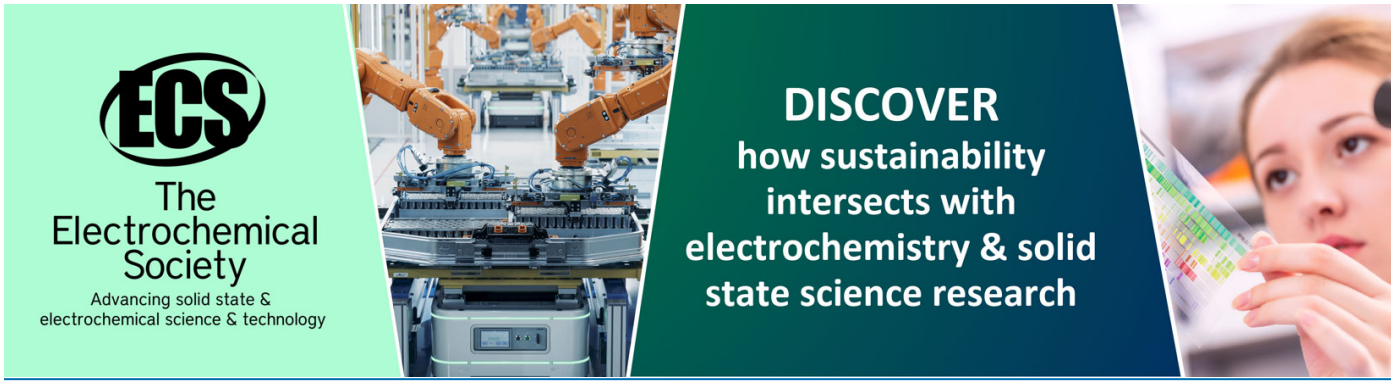
Research on sensing information fusion based on fuzzy theory



Research on sensing information fusion based on fuzzy theory

Jing Yang[1], Yingna Li[1,*], Zhuoheng Wu[2], Zhengang Zhao[1], Chuan Li[1]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China
[2]Guangdong Polytechnic College, Guangdong, China

*Corresponding author e-mail: kmdywln@163.com

Abstract. Aiming at the difficulty of noise estimation and outlier estimation in data fusion, a fuzzy theory data fusion algorithm based on confidence distance is proposed and its application in sensor monitoring is studied. Firstly, according to the characteristics of sensor monitoring values, the characteristics of traditional fuzzy theory data fusion methods are analyzed. Secondly, the confidence distance is used to describe the fuzzy proximity between the data, and the erroneous data is eliminated. Finally, the data is merged by proximity to get the result of the fusion. Experiments show that the difference of the fusion value based on the confidence distance is small, and the fusion result is close to the actual state of the measured object.

1. Introduction

In the data monitoring process, raw data is characterized by incompleteness, ambiguity and redundancy. Fusion technology can improve data quality and improve the performance and accuracy of the mining process, which is one of the important factors affecting data mining efficiency[1]. There are massive sensors in the Internet of Things system[2]. During the sampling process, there is a large amount of noise data and outliers in the monitoring data due to the influence of the external environment or the sensor's dynamic characteristics[3]. Fusion algorithm can improve the reliability of data[4-6].

Common fusion algorithms include generalized likelihood ratio[7], hypothesis testing[8], etc. These methods have difficulties in selecting parameters and may affect the final processing results. The data fusion method based on fuzzy theory is to perform fuzzy clustering on different monitoring data sets, so that no need to select parameters. In the fuzzy theory, the fuzzy closeness between data is generally calculated by means of ratio method, cosine amplitude method, similarity method, maximum-minimum method, average method, absolute value method, and subjective evaluation method. However, fiber optical sensors have a lot of monitoring data. In order to make full use of the sampling values at each time and to improve the fusion accuracy, the confidence distance between data groups is used to describe the degree of nearness[9]. Assigning weights based on the degree of ambiguity of each data allows for a reasonable fusion of data[10]. This algorithm can effectively eliminate outliers and get a more accurate data set.

2. Traditional fuzzy theory data fusion

In the 19th century, G. Cantor, a German mathematician, created the set theory for the first time that established the mathematical foundation for fuzzy theory. American control theory expert L.A.Zadeh proposed a concept of fuzzy sets, which effectively promoted Cantor's set theory that received

attention from people and achieved remarkable results in many areas [11]. Compared with generic set, the fuzzy set has no explicit boundary. In a generic set, the probability that any element belongs to this set is 0 or 1.In a fuzzy set, the degree of membership of its element can be any value in the interval [0,1], each element has a degree that belongs to the collection. Suppose there is a domain of X , A is a mapping of X to interval [0,1].A is called a fuzzy set on X . $A(x)$ is a membership function. The value of $A(x)$ is the degree of membership. If $A(x)=1$, indicating that X completely belongs to the set, the value measured by the sensor supports some assumption. If $A(x)=0$, indicating that X does not belongs to the set, the value measured by the sensor does not support some assumption. If $A(x)=0.5$, indicating that the ambiguity of the senor measured of value is the highest.Traditional fuzzy theory data fusion

Due to the sensor's dynamic characteristics, the collected data set contains noise during the data acquisition process of the sensor. Therefore, it is necessary to pre-process the collected data to obtain a reasonable and correct data to express the status of the monitored object.

Suppose a sensor collects n data in one minute, each observation is x_i, $i=1,2,...,n$, Suppose the status of the monitored object does not change drastically within one minute. According to the fuzzy theory, fuzzy clustering is performed on the data monitored by a sensor in each minute. In the classical fuzzy clustering, similarity is usually described by the following method.

$$r_{ij}=\frac{min\left(x_i,x_j\right)}{max\left(x_i,x_j\right)} \tag{1}$$

In the above formula x_i , x_j are measured values at different times.

When r_{ij} is less than the threshold M , it can be considered that x_i,x_j are not similar. The definition nearness is

$$\alpha_{ij}=\begin{cases}1 & , \ r_{ij}\geq M\\ 0 & , \ r_{ij}<M\end{cases} \tag{2}$$

The nearness matrix of a sensor at all monitoring moments is

$$A=\begin{bmatrix}1 & \alpha_{12} & \cdots & \alpha_{1n}\\ \alpha_{21} & 1 & \cdots & \alpha_{2n}\\ \vdots & \vdots & \vdots & \vdots\\ \alpha_{n1} & \alpha_{n2} & \cdots & 1\end{bmatrix} \tag{3}$$

Normalizing the data, the weight of the measured value at k is

$$w_k=A_k\Big/\sum_{i=1}^{n}A_{ij} \tag{4}$$

Based on the weights calculated in the above formula, the data fusion formula is

$$X=\sum_{i=1}^{n}w_k x_i \quad ,i=1,2,\cdots n \tag{5}$$

When r is less than the threshold M, it can be considered that the monitoring data seriously deviate from the estimated value, and the data at this moment should be removed. Normalize the remaining data according to consistency and calculate the weight. Space fusion in the dimension of time to get one minute accurate monitoring data.

3. Confidence distance-based fuzzy theory data fusion algorithm

The pre-processing method of the fuzzy theory based on the confidence distance is to calculate the confidence distance between the measured values based on the measured values of the sensors at different times. Use a given threshold to find the relational matrix to get the similarity of fuzzy clustering. Finally, determine the weight and fuse the data.

When using a sensor to measure a characteristic parameter, the measured value in one minute obeys the Gauss distribution. The concept of confidence distance is often used to reflect deviations between observations x_i and x_j .Use the probability density function as a sensor characteristic function. Marked as $p_i(x)$ .The confidence distance measure is marked as d_{ij} .

Definition

$$\mathrm{d_{ij}} = 2\int_{\mathrm{x_i}}^{\mathrm{x_j}} \mathrm{p_i(x \mid x_i)d_x} \tag{6}$$

In the expression

$$p_i(x \mid x_i) = \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left\{-0.5\left(\frac{x - x_i}{\sigma_i}\right)^2\right\} \tag{7}$$

Using the error function

$$\mathrm{erf}\left(\mathrm{x}\right) = \frac{2}{\sqrt{\pi}}\int_0^{\mathrm{x}} \mathrm{e}^{-\eta^2}\mathrm{d}_\eta \tag{8}$$

Find

$$\mathrm{d_{ij}} = \mathrm{erf}\left(\frac{\sqrt{\mathrm{x_j - x_i}}}{\sqrt{2}\sigma_\mathrm{i}}\right) \tag{9}$$

If a sensor acquires the same parameter n times in 1 minute, the confidence distance of the measured value in different minutes can constitute the confidence distance matrix D_n

$$D_n = \begin{bmatrix} \mathrm{d_{11}} & \mathrm{d_{12}} & \cdots & \mathrm{d_{1n}} \\ \mathrm{d_{21}} & \mathrm{d_{22}} & \cdots & \mathrm{d_{2n}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{d_{n1}} & \mathrm{d_{n2}} & \cdots & \mathrm{d_{nn}} \end{bmatrix} \tag{10}$$

The smaller the value of d_{ij} , the smaller the deviation of the measured values of the i and j sensors. On the contrary, the deviation of the measured values is even greater. The core of the fuzzy theory clustering algorithm based on confidence distance is to use the confidence distance metric to calculate the data similarity.

The process of the fuzzy theory algorithm based on the confidence distance is shown in Figure 2.The algorithm has 3 core steps. First calculate the confidence matrix, then calculate the degree of nearness, weight, and finally remove and fuse the data.

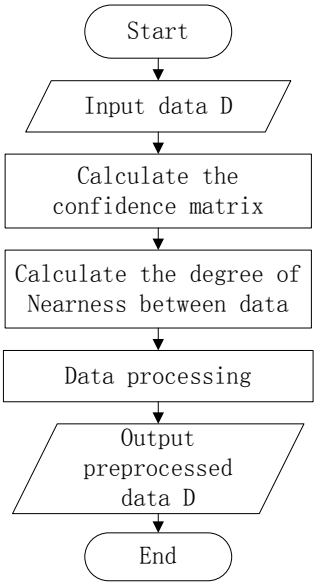


Figure 1. Confidence distance-based fuzzy theory clustering algorithm flow chart

According to the allowable error range of each sensor, the error function can be used to obtain the limit of d_{ij} is M .Use the limit of the error in the following equation to get the degree of support r_{ij} .

$$r_{ij} = \begin{cases} 1, & d_{ij} \le M \\ 0, & d_{ij} \ge M \end{cases} \tag{11}$$

The degree of supporting for each data constitutes a supporting matrix R_{n}

$$R_n = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{12}$$

If $r_{ij} = 0$, then the data collected by the sensors at the time of i and j is considered to exceed the allowable error range, and they do not support each other. If $r_{ij} = 1$, it is considered that the data collected by the sensor at the time of i and the j is within the error range, which is called the data of the sensor at the time i supports the data of the time j .

If the measured data at a certain moment is supported by other multiple moments, the data at that moment is valid. If the measured data at a certain moment is not supported by other multiple moments or is supported by only a few moments, the data collected at this moment is invalid. Such data should be rejected. Define support matrix to represent the support rate for each data. Assume that the support rate threshold is $p_{\min}$, data is available when the support rate is greater than $p_{\min}$, and data is unavailable when the support rate is less than $p_{\min}$ .The support rate p for each data is calculated as follows:

$$p_k = \sum_{i=1}^{n} r_{ik} \Big/ n \tag{13}$$

The counter flag is used to indicate whether the monitoring data at that time is available. If $count_k = 1$ indicates that the data support rate at this moment is greater than the minimum support rate, the data is valid; On the contrary, the data is invalid. $count_k$ is calculated as follows：

$$count_k = \begin{cases} 1, & p_k \ge p_{\min} \\ 0, & p_k < p_{\min} \end{cases} \tag{14}$$

The monitoring data comes from different moments of the same sensor, and the weight is the same at each moment. After eliminating invalid data, calculate weights and data fusion for valid data. The data fusion formula is as follows:

$$X = A_k \bullet count_k \Big/ \sum_{k=1}^{n} count_k \tag{15}$$

4. Fusion algorithm experiments are compared

Based on the sensor monitoring data, an algorithm simulation experiment was described using MATLAB for the above algorithm. On the basis of the experimental data, time efficiency and algorithm effectiveness of the algorithm are compared.

A set of experiments was carried out to test how the algorithm increases the number of algorithms consumed with the number of groups. The amount of data processed by the algorithm in this experiment is n × 6 × 9 (10 ⩽ n ⩽ 6400), which means that there are n sets of data, and each set of data is 6 times every 10 seconds for 9 sensors in one minute. Monitoring data. The experimental data of the algorithm was increased from 10 groups to 6400 groups. The time consumed by the two algorithms is given in Table 1.

Table 1. algorithm time consumption table

Number of data groups	fusion time of fuzzy theory algorithm（s）	fusion time of fuzzy theory algorithm based on confidence distance（s）
10	1.039361	1.12115
50	1.015356	1.215788
100	1.141337	1.472267
150	1.231346	1.487629
200	1.259409	1.56813
250	1.262555	1.600083
300	1.335524	1.814651
400	1.449329	1.910171
800	1.715984	2.198955
1600	2.250309	2.411909
3200	3.150081	3.381059
6400	5.19089	5.521192

From Table 1, we can see that both algorithms increase with the increase of the time consumption of n. In comparison, the fusion time of the fuzzy theory algorithm is less time and the growth rate is relatively slow compared to the confidence fusion algorithm based on the confidence distance. As showed in Fig. 3.4, algorithm 1 is a fuzzy theory fusion algorithm, and algorithm 2 is a confident theory based on confidence distance. Least squares fitting of the above data, algorithm 1 shows a linear change of time s = 0.006n + 1.1173 when the coefficient of determination R2 = 0.9553; algorithm 2 shows a linear change of time when the coefficient of determination R2 = 0.9829 s = 0.006n + 1.4239 . It can be seen from Fig. 3.4 and least-squares fitting that although Algorithm 1 is better than Algorithm 2 in time efficiency, Algorithm 2 and Algorithm 1 show similar linear change in an order of magnitude.

To verify the effect of algorithm fusion and data fusion, the degree of difference was used to measure the processing effectiveness of these two algorithms. Define the degree of difference in the algorithm Di:

$$D_{\mathrm{i}} = \frac{C \times \sum_{i}^{n} \left| x_i - X \right|}{n} \tag{16}$$

Among them, n is the number of data in each group after data fusion; x_i is the i-th data after fusion; X is the fusion data after data fusion; constant C is to improve the accuracy of the different degree. D_i identifies the data difference between pre-processing and fusion data. The larger D_i is, the greater the data difference is, and vice versa. If D_i floats larger, it indicates that the data fusion algorithm have poor performance and stability.

Select 10 groups of data, each group contains 6 data, using two algorithms to preprocess the data to get the fusion result. Then, 6 acquisitions of one sensor from 10 groups of data and their corresponding processed results were used to calculate the difference. According to the formula, D_i is calculated for 10 groups of data, and the accuracy of the algorithm is evaluated by the floating condition of D_i, as showed in Table 2.

$$D_i = \frac{100 \times \sum_{i}^{n} \left| x_i - X \right|}{n} \tag{17}$$

Among these, n ⩽ 6，The number of data to be merged after erroneous data culling will be less than or equal to the original data. Select C=100 to increase the accuracy by 100 times for better observation and comparison. After the data is processed by two groups of algorithms, the degree of difference of the algorithm is obtained according to the formula, as shown in Table 2.The left column of Table 2 is the degree of difference of the pretreatment algorithm of the classical fuzzy theory, and the right column is the degree of difference of the fuzzy theory data fusion algorithm based on the confidence distance. Good fusion algorithms should be close to 0 and make small changes near 0.The following conclusions can be drawn from the comparison of the difference between the average and standard deviation of the two groups data in Figure 5.Algorithm 1 has a high degree of difference, with a low probability of a difference equal to 0, and a large vibration around 0.852596833.The above shows that algorithm 1 has low stability and high degree of difference in data fusion. Compared with Algorithm 1, Algorithm 2 fluctuates around 0.0759, and has a high probability of a difference equal to 0.In summary, the algorithm 2 has high stability, and the degree of difference in data fusion is small.

Table 2. comparison table of algorithm differences

Fuzzy algorithm 1	Fuzzy algorithm 2
0.359483	0.076183
0.336883	0
0.681483	0.14285
0.515767	0
0	0
0.512583	0
0.681483	0.14285
0.26845	0
4.859335	0.02685
0.3105	0.066667

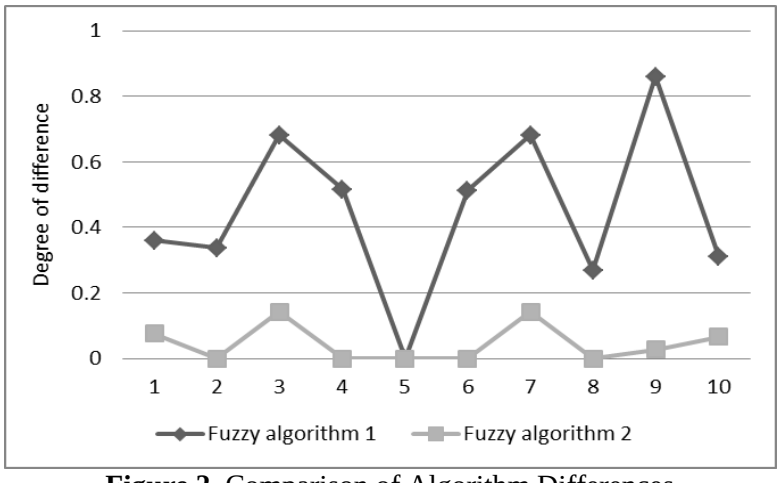


Figure 2. Comparison of Algorithm Differences

The result is that the time efficiency of the classical fuzzy theory fusion algorithm is higher than the fuzzy theory data fusion algorithm based on the confidence distance, but the time consumption of the two algorithms is on a quantitative level, not many differences. Experiment two compares the stability and accuracy of the algorithm for calculating the degree of difference between the two groups of

algorithms. Through analysis, the fuzzy theory data pre-processing algorithm based on confidence distance is high in stability and accuracy. The above two groups of experiments show that the fuzzy theory data pre-processing algorithm based on confidence distance can improve the accuracy and stability of data fusion under the premise of consuming a small amount of time.

5. Conclusion

Confidence distance is introduced into the preconditioning algorithm of fuzzy theory. Fuzzy closeness is calculated by the confidence distance. The paper defines the rate of support between data, obtains the degree of mutual support between measurements at different times, and eliminates the measurement values whose the rate of support is less than the threshold. Determine the weight of each data according to the degree of blurring of different sensors, and finally fuse the data. Experiments show that the fuzzy theory data fusion method based on confidence distance can calculate the data blur degree effectively, determine the outliers quickly and reduce the data redundancy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant no. 51567013.

References

[1] Zhou K, Fu C, Yang S. Big data driven smart energy management: From big data to big insights[J]. Renewable & Sustainable Energy Reviews, 2016, 56:215-225.Reference to a chapter in an edited book:
[2] Ding Zhiming, Gao Xu. A Database Cluster System Framework for Managing Massive Sensor Sampling Data in the Internet of Things[J]. Chinese Journal of Computers, 2012, 35(6):1175-1191.C. D. Smith and E. F. Jones, “Load-cycling in cubic press,” in Shock Compression of Condensed Matter-2001, AIP Conference Proceedings 620, edited by M. D. Furnish et al. American Institute of Physics, Melville, NY, 2002, pp. 651–654.
[3] Thompson W J, Iliadis C. Error analysis for resonant thermonuclear reaction rates[J]. Nuclear Physics A, 2016, 647(3-4):259-279.
[4] García S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining[J]. Knowledge-Based Systems, 2016, 98:1-29.
[5] Liu W, Liu S, Gu Q, et al. Empirical Studies of a Two-Stage Data Preprocessing Approach for Software Fault Prediction[J]. IEEE Transactions on Reliability, 2016, 65(1):38-53.
[6] Shi W, Zhu Y, Huang T, et al. An Integrated Data Preprocessing Framework Based on Apache Spark for Fault Diagnosis of Power Grid Equipment[J]. Journal of Signal Processing Systems, 2017, 86(2-3):221-236.
[7] Jiang W, Zhang C H. Generalized likelihood ratio test for normal mixtures[J]. Statistica Sinica, 2016, 26(3).
[8] Zhang Yuyu, Liang Zhongyao, Lu Wentao. A hypothesis test method with normal distribution assumption for grade assessment of lake nutrient concentration[J]. Acta Scientiae Circumstantiae, 2017, 37(6):2387-2393.
[9] Huang W, Ribeiro A R. Hierarchical Clustering Given Confidence Intervals of Metric Distances[J]. IEEE Transactions on Signal Processing, 2016, PP(99):1-1.
[10] Lu Fang. Data Fusion of Wireless Sensor Networks Based on Fuzzy Theory[J].Laser Journal, 2016, 37(1):145-148.
[11] Kahraman C, Öztayşi B, Onar S Ç. A Comprehensive Literature Review of 50 Years of Fuzzy Set Theory[J]. International Journal of Computational Intelligence Systems, 2016, 9(sup1):3-24.