*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 9 Issue 6 June 2021

Input-Output Clustering Method – Topology Clustering Algorithm

Author: Shin-Jye Lee[1]; Ching-Hsun Tseng[2]

Affiliation: Institute of Management of Technology, National Chiao Tung University, Taiwan[1];

Department of Computer Science, University of Manchester, United Kingdom[2]

E-mail: camhero@gmail.com[1];hank131415go61@gmail.com[2]

DOI: 10.26821/IJSHRE.9.6.2021.9616

ABSTRACT

Developing a new clustering algorithm different from the existing method as well as the strength of the existing method in pattern recognition and machine learning, and effectively presents a unique performance to a variety of linear and non-linear functions and systems as well as the real world relational database system.

Keywords: Cluster Analysis, Input-Output Clustering

1. INTRODUCTION

This clustering method is constructed by integrating the concepts of output constriction, input-output clustering and the technique of 2-part clustering successively. Meanwhile, some of which concepts are innovated by certain concepts of networking, such as networking topology, routing algorithm and the essence of packet routing scheme. Moreover, based on the framework, the rule base of the proposed system is trying to work IF – THEN IF scheme by implementing a main module carrying sub-modules. Further, there are two primary criterions to measure the performance and then contribute the probable criterion working in the entire system, including the measurement of the density of the output constriction, the candidate of index and distance vector routing hop count.

2. IF-THEN IF RULE BASE SCHEME

In the structure of rule base, there is a main module (primary clusters) comprising sub-modules (sub-clusters) in the entire system. In the main module, there are primary clusters, and a primary cluster may carry sub-module(s) (sub-cluster(s)) or not. Meanwhile, based on the relationship of data set distribution, there are three types of distributions in the relational database, including (1) Data set with different outputs; (2) Data set with similar output but non-connective input; (3) Data with similar output and connective input. Therefore, due to smoothly solve the problem of these three types of data set distribution on the rule base, a rule form of "IF - THEN IF" has been schemed in the rule base. The first IF concerns the main module and the primary rule, and the second IF concerns the respective sub-module and the secondary rule set. According to the scheme, a data set will be considered locating at one of primary clusters first, and afterwards decide whether the data set will be considered locating at a sub-module (sub-cluster) of the responding primary cluster or not. If the data set is located at a high certainty region (high population region) in a primary cluster of the main module, then this data set could be recognized locating at a primary cluster. However, if the data set is located at a low certainty region (low population region) in a primary cluster of the main module, then this data set could be

recognized locating at a sub-module of the corresponding primary cluster of the main module.

Hence, a positive scheme of the proposed system can effectively perform to these three types of data set distribution, and which can be concluded as follows:

(1) Data set with different output → Different primary clusters → Different rules
(2) Data set with similar output but non-connective input → The same primary cluster, but different sub-clusters → Different rules
(3) Data set with similar output and connective input → The same primary cluster→ The same rule

3. OUTPUT CONSTRICTION

In principle, the goal of this phase is trying to determine the probable number of output constriction and the proper splitting point between each output constriction. Therefore, the determination of how many output constriction will be made and the definition of the splitting point (the boundary between each output constriction) of the output constriction are the major work in this phase. Basically, it first calculate the density of output space by measuring the distribution of the output variable values of each output constriction, and then determine how many output constrictions has been required or none. If there are lots of low populations data sets distributed at a variety of the location of the geometric space in the relational database, then the determination of more output constriction is essential. After the number of output constriction is being determined, the algorithm start detecting the density of the each output constriction by measuring the distribution of the output variable values of each output constriction and then probably decides the splitting point between each output constriction. Meanwhile, the measurement can also be classified into two types, one is the measurement of the density of the certain output constriction to the whole output constriction, and another one is the measurement of the density between different possible output constriction. Also, the measurement of the density of the output constriction can be expressed as follows:

A. For the measurement of the density of the certain output constriction to the whole output constriction, which can be expressed as follows:

$$\forall Oc = \frac{Oc_i}{Oc_n} * 100\% \quad (1)$$

where Oc_i denotes the population of output variable values of the *i*th output constriction, and Oc_n denotes the whole population of the output variable values, where $1 \le i \le n$.

B. For the measurement of the density between different possible output constriction, which can be expressed as follows:

$$Oc = \frac{Oc_i}{Oc_j} * 100\% \quad (2)$$

where Oc_i denotes the population of output variable values of the *i*th output constriction, and Oc_j the population of output variable values of the *j*th output constriction, where $1 \le i \le n$, $1 \le j \le n$

4. INPUT-OUTPUT CLUSTERING – TOPOLOGY CLUSTERING ALGORITHM

For achieving a good performance, the concept of input-output clustering is being worked in the system, and which also combines certain concepts of networking in particular. Basically, a tuple of the relational database can be regarded as a router located in the subnet or a packet routing in VPN (Virtual Private Network). Therefore, two innovated concepts have been implemented in the proposed input-output

clustering algorithm, including the candidate of index and distance vector routing hop count.

A. The Candidate of Index

Based on the theory of networking, each tuple in the relational database looks like a packet routing in VPN. According to the thinking, the structure of a tuple can be transferred a packet, and which illustrated as follows:

Index(Head)	Input variable1	Input variable2..N	Output variable

Based on the structure, a tuple-packet of SISO can be illustrated as follows:

Index(Input variable X)	Output variable Y

Input variable X is undoubtedly the Index, because there is only one candidate of Index in the packet.

Based on the structure, a tuple-packet of MISO can be illustrated as follows:

Index(Input variable X1)	Input variableX2..Xn	Output variable Y

In MISO, each input variable Xi is one of Index candidates, and the best candidate of Index could be the stable one. The difference between SISO and MISO is MISO possesses more input variables than SISO. In the subsequent simulation examples, the 1st input variable X1 is the primary candidate of Index.

The one goal of Index is re-sorting the data set in ascending order based on the Index input variable values in the relational database system, and then calculate the metrics/geometric distance between each data set in the relational database system.

B. Distance Measure

A global criterion evaluates the geometric distance based on Euclidean distance can be calculated as follows:

$$D_i = \left[\sum_{i=1}^{n} (P_i - Q_i)^2 \right]^{1/2} \qquad (3)$$

where P_i and Q_i presents the coordinate/tuple of the certain dataset, $1 \le i \le n$.

C. Distance Vector Routing Hop Count

There are two types of distance vector routing hop working in the procedure, including minimum distance vector routing hop count and maximum distance vector routing hop count. Basically, the 1st part clustering (the primary cluster of the main module) is probably determined by the maximum distance vector routing hop count, and the 2nd part clustering (the sub-module of primary cluster) is probably determined by the minimum distance vector routing hop count. Meanwhile, the measurement of the distance vector routing hop count is constructed on measuring the "geometric distance" between each data set/cluster/router in the relational database/topology.

(1) Minimum Distance Vector Routing Hop Count

$$D.V._{\min}^{hop} = \frac{\sum_{i=1}^{n-1} |d_{i+1} - d_i|}{n-1} * \theta \qquad (4)$$

where d_i and presents the coordinate of the certain dataset, $1 \le i \le n$, and θpresents the weighted value of the requirement.

(2) Maximum Distance Vector Routing Hop Count

$$D.V._{\max}^{hop} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i} |d_i - d_j|}{n(n+1)/2} * \theta \qquad (5)$$

where d_i and d_j presents the metrics of the certain dataset, $1 \le i \le n$, $1 \le j \le n$, and θ presents the weighted value of the requirement.

D. The Procedure of Topology Clustering Algorithm

The procedure of topology clustering algorithm can be described as follows:

(1) Determine the proper number of output constriction and the probable splitting point between each output constriction.
(2) Define the primary candidate of Index.
(3) Calculate the geometric distance between each data set in the relational database system.
(4) Calculate the minimum distance vector routing hop count and maximum distance vector routing hop count.
(5) Define the primary clusters of the main module (1st part clustering) by adjusting the weight of the value of Maximum Distance Vector Routing Hop Count.
(6) Define the sub-clusters of primary cluster (2nd part clustering) by adjusting the weight of the value of Minimum Distance Vector Routing Hop Count.
(7) 2-part input-output clustering completed, and then generates a fuzzy system by the resultant clusters of 2-part input-output clustering.
(8) Fuzzified and delete useless rules.
(9) Doing the estimation by Method of Least Squares.
(10) Calculate Differential rate, for making the estimated results more closes the original values. (Make the MSE better)

Further, the whole procedure can be briefly illustrated as follows:

Output Constriction → 2-part input-output Clustering → Fuzzified → Method of Least Squares → Differential Rate

E. Additional Issue

If the number of output constriction is being set more than 1.0, then the distribution of data set close around the boundary would be considered carefully. In the procedure of output constriction, the data set and cluster located around the ±20% range of the boundary will be considered. In case the geometric distance between these data set and clusters are lower than the value of minimum distance vector routing hop count, which would be agglomerated into one cluster. Otherwise (the geometric distance between these data set and clusters are higher than the value of minimum distance vector routing hop count), which will still be kept at the original situation.

5. SIMULATION EXAMPLES

A. *Multi-tones Non-Linear Function (MNLF)*

The samples of simulation can be generated by

$$Y=F(X)=100\sin x \qquad (6)$$

where $0 \le x \le 8\pi$, and 100 samples are randomly generated within the definition domain.

Further, the comparison is shown in the Table 1 as follows:

Table 1 – The Robustness Comparison based on MNLF

	MAPE%	MSE	RMSE
0% NOISE	0.064449	38.222009	6.182395
5% NOISE	0.080825	14.456145	3.802124
10% NOISE	0.080825	15.865701	3.983177
50% NOISE	0.080825	29.502337	5.431605

B. *Single-Input-Single-Output (SISO)*

The samples of simulation can be generated by

$$y = 0.6\sin(\pi x) + 0.3\sin(3\pi x) + 0.1\sin(5\pi x) \qquad (7)$$

where $x \in [-1.1]$, and 100 samples are randomly generated within the definition domain.

Further, the comparison is shown in the Table 2 as follows:

Table 2 – The Comparison of SISO

Method	RMSE / Clusters / Parameters		
Pedrycz's Method [1]	0.180/6/18	0.150/8/24	0.147/10/30
	0.144/12/36	0.192/9/27	0.140/12/36
	0.123/15/45	0.114/18/54	0.174/12/36
	0.140/16/48	0.108/20/60	0.100/24/72
	0.149/15/45	0.136/20/60	0.102/25/75
	0.092/30/90	0.141/18/54	0.102/24/72
	0.097/30/90	0.061/36/108	
5%Noise	0.097/6/18		
10%Noise	0.101/6/18		
50%Noise	0.139/6/18		
Proposed Method Topology Algo.	0.016/ 5 main clusters/ 61		
5%Noise	0.007/ 5 main clusters/ 61		
10%Noise	0.008/ 5 main clusters/ 61		
50%Noise	0.01/ 5 main clusters/ 61		

C. Two-Dimensional Non-linear System

The samples of simulation can be generated by

$$y = f(x_1, x_2) = (1 + x_1^{-2} + x_2^{-1.5})^2 \qquad (8)$$

where $x_1 \in [1.5]$, $x_2 \in [1.5]$ and 50 samples are randomly generated within the definition domain.

Further, the comparison is shown in the Table 3 as follows:

Table 3 – The Comparison of 2DNLS

Method	Neurons (Rules)	Parameters	MSE
Sugeno and Yasukawa [2]	6	65	0.079
Wu and Chen[3]	9	19	0.162
Topology Algo.	30	42	0.007

D. Non-Linear Dynamic System

The samples of simulation can be generated by

$$y(k) = g(y(k-1), y(k-2)) + u(k) \qquad (9)$$

where

$$g(y(k-1), y(k-2)) = \frac{y(k-1)y(k-2)[y(k-1)-0.5]}{1+y^2(k-1)y^2(k-2)}$$

$u(k) = \sin\left(\frac{2\pi t}{25}\right)$, $y(0) = y(1) = 0$, $t \in [1.200]$, and 200 samples are randomly generated within the definition domain.

Further, the comparison is shown in the Table 4 as follows:

Table 4 – The Comparison of NLDS

Method	Number of rules	Parameter	MSE
GG-TLS [4]	12	?	3.7E-4
GG-LS [4]	12	?	3.7E-4
EM-TI [4]	12	?	2.4E-4
EM-NI [4]	12	?	3.4E-4
Wang [5]	28	?	3.3E-4

Wang [6]	20	?	6.8E-4
Topology Algo	33	39	1.4E-5

E. A real world relational database of a 2nd-hand company sales system

Total sample: 200 samples

Training sample: 100 samples (randomly picked)

Testing sample: 100 samples (remaining)

Further, the comparison is shown in the Table 5 as follows:

Table 5 – The Comparison of 2nd-Hand Company Sales Data

Training Example	Rules	MAPE%
Chen's method [7]	18	0.14303
The proposed method	20	0.09532
Testing Example	Rules	MAPE%
Chen's method [7]	18	0.19025
The proposed method	20	0.12805

F. Discussion and Analysis

The advantage can be stated as follows:

(a) Good performance on a variety of functions and systems, including linear function, non-linear function, SISO, MISO and the real world relational database system

(b) The reliability of easing the problem of curse of dimensionality

(a) Good robustness on SISO, including linear and non-linear function

Reason: Properly integrate a variety of theories, including 2-part clustering mode, output constriction mode, input-output-oriented clustering mode, the features of networking and database management and so on, and innovate which as a sound method.

The disadvantage can be stated as follows:

(a) General robustness on MISO, including linear and non-linear function

6. CONCLUSION

The method is inspired from astronomy, and which simulates the all data set generated from a variety of functions or in the system as the topology in the universe. Also, the proposed method doesn't only introduce certain classical concepts in pattern recognition and machine learning, such as input-output clustering, output constriction and 2-part clustering, but also a variety of concepts different from pattern recognition and machine learning into achieving a sound performance in the entire procedure, including the concepts of networking and database management. Therefore, it's a totally new clustering algorithm, and which is different from the existing method and the strength of the existing method in pattern recognition and machine learning.

7. REFERENCE

[1] W. Pedrycz, "Linguistic models as a framework of user-centric system modelling," IEEE Transaction on System Man and Cybernetics, Part A, vol. 36, no. 4, pp. 727-745, 2006.

[2] M. Sugeno and T. Yasukawa, "A fuzzy-logic based approach to qualitative modelling," IEEE Transaction on Fuzzy Systems, vol. 1, no.1, pp.7-31, 1993.

[3] Wu, T. P. and Chen, S. M. 1999. A New Method for Constructing Membership Functions and Fuzzy Rules from Training Examples. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 29, No. 1, pp. 25-40.

[4] J. Abonyi, R. Babuska and F. Szeifert, "Modified Gath-Geva fuzzy clustering for identification of Takagi-Sugeno fuzzy models," IEEE Transaction on Systems, Man and Cybernetics, Part B, vol. 32,no.5, pp. 612–621, 1999.

[5] L. Wang and J. Yen, "Extracting fuzzy rules for system modeling using a hybrid of genetic algorithms and Kalman filter," Fuzzy Systems and Sets, vol. 101, no.3 , pp. 353–362, 1999.

[6] J. Yen and L.Wang, "Simplifying fuzzy rule-based models using orthogonal transformation methods," IEEE Transaction on Systems, Man and Cybernetics, Part B, vol. 29, no.1, pp. 13–24, 1999.

[7] Chen, S. M. 2003. Generating Weighted Fuzzy Rules From Relational Database Systems for Estimating Null Values Using Genetic Algorithms. IEEE Transactions on Fuzzy Systems, Vol. 11, No. 4, pp. 495-505.

