Paper 37

by The Jin Ai

A Particle Swarm Optimization-based Clustering for Non-Metric Data

Ririn Diar Astanti†, The Jin Ai
Department of Industrial Engineering
Universitas Atma Jaya Yogyakarta, Yogyakarta, Indonesia
Tel: (+62) 274-487711, Email: ririn@mail.uajy.ac.id, jinai@mail.uajy.ac.id

Voratas Kachitvichyanukul
Department of Industrial and Manufacturing Engineering
Asian Institute of Technology, Pathumtani, Thailand
Tel: (+66) 2-5245683, Email: voratas@ait.asia

Abstract. Advance development of information technology enables an organization to get their transaction data easily or it is called as a big data era. The data they have are meaningless unless there is an effort to mine those data to be valuable information that can be used for the managerial level to make a decision. One of the methods to mine the data is clustering technique. To the best of author knowledge there are many researches have been found dealing with clustering metrics for metric data however there is limited research dealing with clustering technique for non-metric data. This paper proposes Particle Swarm Optimization (PSO)-based clustering for non-metric data.

Keywords: big data, clustering, non-metric, particle swarm optimization

1. INTRODUCTION

Advanced development of information technology has change the company run its business. For example a technology such as Point of Sale (POS) terminal is nowadays implemented in many modern retails. Those technologies enable the company to record their transaction data in example, using POS enable a retail to get detail data about customer purchase. Another example is that using optical scanning and bar code enable the company to record the inventory product easily. Therefore, nowadays industry is facing of what it is called as Big Data era.

The ability of each organization to gain many data is meaningless if the data cannot be processed to become information that is useful for the managerial to make decision. Therefore the challenge in this big data era is how we can retrieve, process and analyze in a large volume of data or it called as data mining technique (DMT) (Liao et al., 2012; Weiss & Indurkhya, 1998). Turban et al. (2007) defines data mining as "*the process that uses statistical, mathematical, artificial intelligence and machine-learning technique to extract and identify useful information and subsequently gain knowledge from large databases*" Other researcher such as Liao et al. (2012) also stated that "*data mining have formed a branch of applied artificial intelligence (AI)*". Similar definition regarding data mining have been proposed by several researchers such as Berson et al. (2000), Lejeune (2001), Ahmed (2004) and Berry and Linoff (2004).

According to Liao et al. (2012), they are several major kinds of data mining methods. One of them is clustering. According to Dong and Qi (2009) and Jain et al. (1999), clustering is an exploratory data technique that is very useful such as for data mining and pattern classification. Different with discriminant analysis, clustering technique is a grouping technique for unlabeled data.

Clustering technique can be classified into 2 techniques. They are hierarchical and non-hierarchical technique (Jain et al. (1999). Particle Swarm Optimization is proposed by Kennedy and Eberhart (1995). Since from the earliest development of Particle Swarm Optimization, one of the common application of this optimization algorithm is on data clustering, especially clustering for metric data, i.e. Van Der Merwe & Engelbrecht (2003), Chen & Ye (2004). After that, enormous researches are conducted on the topic of PSO on clustering. These researches recently have been reviewed by Rana et al. (2011), Alam et al. (2014), and Esmin et al. (2015).

Based on the review papers, there are three important issues related to the application of PSO on data clustering. First, it is well known that many variants of PSO are exists in the literature and most of them have been applied on data clustering, for example: cooperative PSO (Zhang et al., 2016), niching PSO (Ma et al., 2015), Second, various

† :Corresponding Author

clustering applications are exists being solved by PSO, for example: feature selection (Lane et al., 2013), data streams clustering (Fong et al., 2016), text document clustering (Abualigah et al., 2017), medical images processing (Vishnuvarthanan, 2017). Third, the PSO is commonly being combined or hybridized with other techniques when it is applied on data clustering. Several techniques that are commonly used are K-means (Niu et al., 2017), Nelder Mead (López García et al., 2014), and GRASP (Marinakis et al., 2008).

As data can be divided into metric and non-metric data, therefore a proper method for data mining dedicated to each data type is needed. In the particular application of PSO on data clustering, however, majority of focus is on the metric data. Therefore, there is a room for exploring in the area of PSO application on non-metric data.

2. HOMOGENEITY AND HETEROGENEITY DEFINITIONS

Let consider a clustering problem with all non-metric data. Number of objects considered in the clustering is N with V classifier variables. The number of cluster created by the algorithm is K. The objective is to create K number of clusters, in which the homogeneity of each cluster is maximize and the heterogeneity among cluster is also maximize. In order to convert this problem into optimization problem, we need to define homogeneity and heterogeneity into quantitative terms.

Let us consider cluster k, which consists of N_k objects. Homogeneity among objects in this cluster, can be identified by the similarity of each classifier variable across objects. In here, we define the similarity of variable v in this cluster as

$$S_{k,v} = \begin{cases} 1, & \text{if} \quad x_{k,1,v} = x_{k,2,v} = \cdots = x_{k,N_k,v} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where

$x_{k,i,v}$: the value of classifier variable v in the object i of cluster k

After the similarity of all variables are obtained, the homogeneity of cluster k can be defined as:

$$G_k = \frac{1}{V}\sum_{v=1}^{V} S_{k,v} \tag{2}$$

For whole clusters, the homogeneity measurement can be defined as:

$$\bar{G} = \frac{1}{K}\sum_{k=1}^{K} G_k \tag{3}$$

In order to express heterogeneity into quantitative term, we define the difference of variable v between cluster j and cluster k as

$$D_{(j,k),v} = \begin{cases} 0, & \text{if} \quad S_{j,v} = S_{k,v} = 1 \text{ and } x_{j,1,v} = x_{k,1,v} \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

After the difference of all variables are obtained, the heterogeneity between cluster j and k can be defined as:

$$H_{(j,k)} = \frac{1}{V}\sum_{v=1}^{V} D_{(j,k),v} \tag{5}$$

For whole clusters, the heterogeneity measurement can be defined as:

$$\bar{H} = \left(\sum_{j=1}^{K} \sum_{k=j+1}^{K} H_{j,k} \right) \Big/ \sum_{k=1}^{K-1} k \tag{6}$$

Therefore, the clustering problem can be stated as optimization problem of maximizing weighted total homogeneity and heterogeneity measurement:

$$Z = w_g \bar{G} + w_h \bar{H} \tag{7}$$

where

w_g : weight of homogeneity measurement

w_h : weight of heterogeneity measurement

It is noted that $w_g + w_h = 1$.

3. DECODING METHOD

The most important issues before implementing PSO to any optimization problem are defining the solution representation, i.e. how the problem is being represented in the PSO, and decoding method, i.e. the relationship between solution in the domain of the PSO and the solution of the optimization problem. In this paper, a solution representation of non-metric clustering problem with N objects to form K clusters is particle with N dimension, in which each particle dimension is encoded as a real number within range [0, K].

It is noted that each particle dimension is representing each object, i.e. dimension d is representing object d. The position value of dimension d is indicating the cluster number that includes object d. The conversion of position value into cluster number is following this equation

$$X_h = \lceil \theta_{lh} \rceil \tag{8}$$

where

X_h : cluster number of object h

θ_{lh} : position value of a particle l in the dimension h

Following illustration is provided to give example how the particle position is being decoded into clusters. Given the particle position is [0.67; 0.42; 1.76; 0.87; 1.45], conversion using equation (13) is lead to cluster number [1; 1; 2; 1; 2]. Therefore, the cluster 1 consists of object 1, 2, 4 and cluster 2 consists of object 3,4.

4. THE PROPOSED PSO ALGORITHM

A PSO variant called GLNPSO (Pongchairerks and Kachitvichyanukul, 2005) is applied here, in order to solve the clustering problem with all non-metric data. The algorithm is presented below. Two problem specific steps are inserted into the GLNPSO Algorithm, which became step 2 and step 3 of the algorithm.

In order to give overall information about the algorithm, the algorithm is rewritten here although this algorithm is similar to the application to other problem, i.e. vehicle routing problem (Ai and Kachitvichyanukul, 2009). Only step 2 and step 3 of the algorithm are different.

Particle's position is converted to cluster in the step 2 (See Section 3) and the performance of cluster, which is weighted total homogeneity and heterogeneity measurement (See Section 2), is calculated in the step 3. In this framework, the particles are initialized in step 1. The iteration of particles movement is described by steps 2-8, in which the particles' fitness value are evaluated in steps 2-3, their cognitive and social information are updated in steps 4-7, and their positions are updated in step 8. Step 9 is the controlling step to repeat or stop the iteration.

Notation

τ	:	Iteration index; $\tau = 1 \ldots T$
l	:	Particle index, $l = 1 \ldots L$
h	:	Dimension index, $h = 1 \ldots H$
u	:	Uniform random number in the interval $[0,1]$
$w(\tau)$	:	Inertia weight in the τ^{th} iteration
$\omega_{lh}(\tau)$	:	Velocity of the l^{th} particle at the h^{th} dimension in the τ^{th} iteration
$\theta_{lh}(\tau)$	:	Position of the l^{th} particle at the h^{th} dimension in the τ^{th} iteration
ψ_{lh}	:	Personal best position (pbest) of the l^{th} particle at the h^{th} dimension
ψ_{gh}	:	Global best position (gbest) at the h^{th} dimension
ψ^{L}_{lh}	:	Local best position (lbest) of the l^{th} particle at the h^{th} dimension
ψ^{N}_{lh}	:	Near neighbor best position (nbest) of the l^{th} particle at the h^{th} dimension
c_p	:	Personal best position acceleration constant
c_g	:	Global best position acceleration constant
c_l	:	Local best position acceleration constant
c_n	:	Near neighbor best position acceleration constant
$\theta^{\max}$	:	Maximum position value
$\theta^{\min}$	:	Minimum position value
Θ_l	:	Vector position of the l^{th} particle, $[\theta_{l1} \; \theta_{l2} \; \cdots \; \theta_{lH}]$
Ω_l	:	Vector velocity of the l^{th} particle, $[\omega_{l1} \; \omega_{l2} \; \cdots \; \omega_{lH}]$
Ψ_l	:	Vector personal best position of the l^{th} particle, $[\psi_{l1} \; \psi_{l2} \; \cdots \; \psi_{lH}]$
Ψ_g	:	Vector global best position, $[\psi_{g1} \; \psi_{g2} \; \cdots \; \psi_{gH}]$
Ψ^{L}_l	:	Vector local best position of the l^{th} particle, $[\psi^{L}_{l1} \; \psi^{L}_{l2} \; \cdots \; \psi^{L}_{lD}]$
R_l		The l^{th} set of vehicle route
$Z(\Theta_l)$	:	Fitness value of Θ_l
FDR	:	Fitness-distance-ratio

PSO Algorithm

1. Initialize L particles as a swarm, generate the l^{th} particle with random position Θ_l in the range $[\theta^{\min}, \theta^{\max}]$, velocity $\Omega_l = 0$ and personal best $\Psi_l = \Theta_l$ for $l = 1 \ldots L$. Set iteration $\tau = 1$.
2. For $l = 1 \ldots L$, decode $\Theta_l(\tau)$ to a set of clusters R_l.
3. For $l = 1 \ldots L$, compute the performance measurement of R_l, and set this as the fitness value of Θ_l, represented by $Z(\Theta_l)$.
4. Update pbest: For $l = 1 \ldots L$, update $\Psi_l = \Theta_l$, if $Z(\Theta_l) < Z(\Psi_l)$.
5. Update gbest: For $l = 1 \ldots L$, update $\Psi_g = \Psi_l$, if $Z(\Psi_l) < Z(\Psi_g)$.
6. Update lbest: For $l = 1 \ldots L$, among all pbest from K neighbors of the l^{th} particle, set the personal best which obtains the least fitness value to be Ψ^{L}_l.

7. Generate nbest: For $l=1\ldots L$, and $h=1\ldots H$, set $\psi_{lh}=\psi_{oh}$ that maximizing fitness-distance-ratio (FDR) for $o=1\ldots H$. Where FDR is defined as

$$FDR=\frac{Z(\Theta_l)-Z(\Psi_o)}{|\theta_{lh}-\psi_{oh}|} \quad \text{which } l\neq o \tag{9}$$

8. Update the velocity and the position of each l^{th} particle:

$$w(\tau)=w(T)+\frac{\tau-T}{1-T}\left[w(1)-w(T)\right] \tag{10}$$

$$\begin{aligned}\omega_{lh}(\tau+1)=\;&c_p u\left(\psi_{lh}-\theta_{lh}(\tau)\right)+c_g u\left(\psi_{gh}-\theta_{lh}(\tau)\right)\\&+c_l u\left(\psi^L_{lh}-\theta_{lh}(\tau)\right)+c_n u\left(\psi^N_{lh}-\theta_{lh}(\tau)\right)\\&+w(\tau)\omega_{lh}(\tau)\end{aligned} \tag{11}$$

$$\theta_{lh}(\tau+1)=\theta_{lh}(\tau)+\omega_{lh}(\tau+1) \tag{12}$$

If $\theta_{lh}(\tau+1)>\theta^{\max}$, then

$$\theta_{lh}(\tau+1)=\theta_{lh}(\tau+1)-\theta^{\max} \tag{13}$$

$$\omega_{lh}(\tau+1)=0 \tag{14}$$

If $\theta_{lh}(\tau+1)<\theta^{\min}$, then

$$\theta_{lh}(\tau+1)=\theta^{\min}+\left[\theta^{\min}-\theta_{lh}(\tau+1)\right]$$

$$\omega_{lh}(\tau+1)=0$$

9. If the stopping criterion is met, i.e. $\tau=T$, stop. Otherwise, $\tau=\tau+1$ and return to step 2.

CONCLUDING REMARKS

Two important elements of the PSO implementation for solving clustering problem with non-metric data are proposed in this paper, which are the solution representation and the decoding method. In addition, the performance of formed clusters is defined in term of homogeneity and heterogeneity measurements. Using these proposed definitions, the PSO algorithm is ready to be implemented as a computer program to solve the intended non-metric clustering problem.

REFERENCES

Ai, T. J., & Kachitvichyanukul, V. (2009). Particle swarm optimization and two solution representations for solving the capacitated vehicle routing problem. *Computers & Industrial Engineering*, *56*(1), 380-387.

Ahmed, S.R. (2004). Applications of data mining in retail business. Information Technology: Coding and Computing, 2, 455 – 459.

Alam, S., Dobbie, G., Koh, Y.S., Riddle, P., Ur Rehman, S. (2014). Research on particle swarm optimization based clustering: A systematic review of literature and techniques. Swarm and Evolutionary Computation, 17, 1-13.

Berry, M.J.A., and Linoff, G.S. (2004). Data mining techniques second edition – for marketing, sales, and customer relationship management. Wiley.

Berson, A., Smith, S., and Thearling, K. (2000). Building data mining applications for CRM. McGraw-Hill.

Chen, C.-Y., & Ye, F. (2004). Particle swarm optimization algorithm and its application to clustering analysis. Conference Proceeding - IEEE International Conference on Networking, Sensing and Control, 2, 789-794.

Dong, J., and Qi, M. (2009). A new algorithm for clustering based onparticle swarm optimization and K-means. International Conference on Artificial Intelligence and Computational Intelligence.

Esmin, A.A.A., Coelho, R.A., & Matwin, S. (2015). A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 44 (1), 23-45.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: a review. ACM Computing, Surveys.31 (3), 264-323.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, 4, 1942-1948.

Krajewski, L.J., Ritzman, L.P. (1987). Operations Management, Strategy and Analysis, Reading Addision-Wesley Publishing, MA

Lane, M.C., Xue, B., Liu, I., Zhang, M. (2013). Particle swarm optimisation and statistical clustering for feature selection. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8272 LNAI, pp. 214-220.

Lejeune, M.A.P.M (2001). Measuring the impact of data mining on churn management. Internet research: Electronic Networking Applications and Policy, 11, 375 – 387.

Liao, S-H., Chu, P-H and Hsiao, P-Y. (2012). Data mining techniques and applications – A decade review from 2011 to 2011 Expert Systems with Applications, 39, 11303-11311.

Ma, D., Ma, J., Xu, P. (2015). An adaptive clustering protocol using niching particle swarm optimization for wireless sensor networks. Asian Journal of Control, 17 (4), pp. 1435-1443.

Rana, S., Jasola, S., & Kumar, R. (2011). A review on particle swarm optimization algorithms and their applications to data clustering. Artificial Intelligence Review, 35(3), 211-222.

Turban, E., Aronson, J.E., Liang, T.P., and Sharda, R. (2007). Decision support and business intelligence systems Eight edition. Pearson Education.

Van Der Merwe, D.W., & Engelbrecht, A.P. (2003). Data clustering using particle swarm optimization. 2003 Congress on Evolutionary Computation, CEC 2003 - Proceedings, 1, art. no. 1299577, 215-220.

Weiss, S.H., and Indusrkhya, N. (1998). Predictive Data Mining: A Practical Guide. San Fransisco, Ca: Morhan Kaufmann Publishers.

Zhang, Y., Xia, C.-H., Gong, D.-W., Rong, M. (2016). Streaming data clustering using cooperative particle swarm optimization. Kongzhi yu Juece/Control and Decision, 31 (10), pp. 1879-1883.

Paper 37

ORIGINALITY REPORT

27%	23%	23%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	mini-chip.eu Internet Source	7%
2	Ai, T.J.. "Particle swarm optimization and two solution representations for solving the capacitated vehicle routing problem", Computers & Industrial Engineering, 200902 Publication	5%
3	Jingneng Ni, Jiuping Xu, Mengxiang Zhang. "Constructed wetland planning-based bi-level optimization to balance the watershed ecosystem and economic development: A case study at the Chaohu Lake watershed, China", Ecological Engineering, 2016 Publication	3%
4	Atiwat Boonmee, Kanchana Sethanan. "A GLNPSO for multi-level capacitated lot-sizing and scheduling problem in the poultry industry", European Journal of Operational Research, 2016 Publication	1%

5	docshare.tips Internet Source	1%
6	Submitted to Universitas Atma Jaya Yogyakarta Student Paper	1%
7	Submitted to University of Sunderland Student Paper	1%
8	documents.mx Internet Source	1%
9	www.vpaa.ait.ac.th Internet Source	1%
10	E.W.T. Ngai, Li Xiu, D.C.K. Chau. "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, 2009 Publication	1%
11	orbilu.uni.lu Internet Source	1%
12	nrid.nii.ac.jp Internet Source	1%
13	Submitted to University of Birmingham Student Paper	<1%
14	Submitted to Yuan Ze University Student Paper	<1%

15 Ai, The Jin, Jeffry Setyawan Pribadi, and Vincensius Ariyono. "Solving the Team Orienteering Problem with Particle Swarm Optimization", Industrial Engineering and Management Systems, 2013.
Publication
<1%

16 www.encyclopedias.biz
Internet Source
<1%

17 Gunjan Soni, Vipul Jain, Felix T.S. Chan, Ben Niu, Surya Prakash. "Swarm intelligence approaches in supply chain management: potentials, challenges and future research directions", Supply Chain Management: An International Journal, 2019
Publication
<1%

18 www.cs.uoi.gr
Internet Source
<1%

19 Jiao Zhao, Lixin Tang. "A particle swarm optimization for the quay crane scheduling problem with non-interference constraints", 2009 IEEE International Conference on Automation and Logistics, 2009
Publication
<1%

20 www.aut.upt.ro
Internet Source
<1%

21 Studies in Computational Intelligence, 2014.

Publication

<1%

22 Submitted to University of Northumbria at Newcastle
Student Paper

<1%

23 www.emeraldinsight.com
Internet Source

<1%

24 "Intelligent Computing Methodologies", Springer Science and Business Media LLC, 2018
Publication

<1%

25 Jiuping Xu. "Bi-Random Multiple Objective Decision Making", Lecture Notes in Economics and Mathematical Systems, 2011
Publication

<1%

26 Ririn Diar Astanti, The Jin Ai, Hunyh Trung Luong, Hui Ming Wee. "Two techniques for solving nonlinear decreasing demand inventory system with shortage backorders", International Journal of Operational Research, 2018
Publication

<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		