A Framework of Fuzzy Partition Based on Artificial Bee Colony for Categorical Data Clustering

Iwan Tri Riyadi Yanto
Department of Informatin System
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
yanto.itr@is.uad.ac.id

Dedy Hartama
Departement of Information System
Tunas Bangsa AMIK and STIKOM
Pematangsiantar, Indonesia
dedyhartama@amiktunasbangsa.ac.id

Younes Saadi
Department of Informatios system
University of Malaya
Malaysia
Yunes.saadi@gmail.com

Dewi Pramudi Ismi, Andri Pranolo
Department of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
dewi.ismi@tif.uad.ac.id, andri.pranolo@tif.uad.ac.id

***Abstract*— Fuzzy k-partition (FkP) is an effective clustering technique, which is mathematical model based. Thus, the objective function of FkP is a nonlinear function. Membership random selection is featured by an iterative process, which results in local optima traps easily. It is important to find global optimal consider to nonlinear objective function of the problem. Moreover, Artificial Bee colony (ABC) has ability and efficiently used for multivariable, multinomial function optimization. To this, this paper proposes the hybridization of FkP based on Artificial Bee colony (ABC) a population based algorithm. Some of benchmarks data sets have been elaborated to test the proposed approach. The experiment shows that FkP ABC obtains better results in term of the dun index validity clustering as compared to the baseline algorithm.**

Keywords—data clustering; fuzzy k-partition; artifical bee colony

I. INTRODUCTION

Data clustering is a popular issue commonly appears in various domains of science and technology such as data mining, computer vision and machine learning applications [1], [2]. Clustering divides a given dataset of multiple variables into different groups with common data points located in same cluster. The most well-known clustering algorithm to date is K-means [3]. K-means clustering have been proven to be effective and efficient in the domain of large data clustering due to its ability to deal with high dimensionality compared to other algorithms. However, it is known that K-means has a limitation in terms of dealing with categorical attribute value because of that K-means can only minimize the numerical values of cost function. Therefore, Huang [4] introduced a k-modes approach based that overcomes that issue to support categorical data. Afterwards many improvements of k-modes clustering algorithms are proposed including novel dissimilarity measures implemented in the k-modes clustering [5]–[7] and a fuzzy set based k-means clustering algorithm [8], [9]. In order to improve the output of the fuzzy k-modes approach. Kim, Lee and Lee. [10] introduced fuzzy centroids approach. Based on a different construction on categorical data, a fuzzy approach for documents data clusterization is proposed by Umayahara et al. [11] .

The Fuzzy C-mean (FCM) clustering algorithm [10] and all its variants, such as for clustering numerical data [12], [13], for clustering symbolic data [14], [15] and for clustering categorical data [4], [8], [16], [17] are non-parametric methods via the sum of squared errors inter groups of data to determine cluster where a data point belongs to. Miin-Shen et al. [18] proposed Fuzzy k-partititon (FkP) approach, which is characterized by a multivariate multinomial distributions. The FkP is considered as a clustering based on fuzzy approach used for categorical data.

The FkP works effectively for categorical data, but the membership random selection and the value of probabilities make iterative process resulted in local optimal solution easily since the objective function is nonlinear. To overcome this issue, many optimization algorithms have been introduced such as GA, SA, ant colony, PSO, bee colony. Bee colony is a swarm intelligence based algorithm, in which Bees dance behavior for finding food source is modeled. The dance is used as a way of communication between the bees population. In contrast, in Ant Colony the produced pheromone traces are used as a way to communicate between ants population. The communication between the population individuals is performed locally with a limited information compared to the whole search space. In contrast, when the local outputs are combined, the global output achieves new info more relevant to the whole search space. Artificial Bee colony (ABC) has the ability to escape local optima traps and can be efficiently employed for multivariable and multinomial function optimization approaches [19].

978-1-5090-1721-8/16/$31.00 ©2016 IEEE

II. Related Works

A. Fuzzy K Partition

The Fuzzy k-partition (FkP) model proposed by Yang, Chiang, Chen and Lai. [18] is a technique designed for clustering categorical data set. Multivariate multinomial distribution is foundation of this technique. The technique is applied on a dataset Y composing of I objects ($i = 1,2, \dots, I$) of J discrete variable ($j = 1,2, \dots, J$) every j contains L_j single finite number. FkP technique is using the indicator function $z_1, z_2, \dots, z_k$, where a partition $P = \{P_1, P_2, \dots, P_K\}$ of i into K classes as mutually disjoint sets $P_1, P_2, \dots, P_K$ where $P_1 \cup P_2 \cup \dots \cup P_K = Y$ such that $z_k(y) = 1$ if $y \in P_k$ and otherwise, $z_k(y) = 0$ for every y in Y,$k = 1,2,\dots,K$. This is denoting the data clustering of K classes. Assume the values are represented by Y_{ijl} with a set of L_j binary random variables where y_{ijl} is a realization of Y_{ijl} with $Y_{ijl} = y_{ijl}$, for $i = 1,2, \dots, I$, $j = 1,2, \dots, J$, and $l = 1,2, \dots, L_j$.

Thus, y_{ijl} has a binary value, that is, y_{ijl} has value 0 or 1. Consider Y_i, for $i = 1,2, \dots, I$ is a random type sized by I derived by a distribution $f(y, \lambda)$. Let $P = \{P_1, P_2, \dots, P_K\}$ be a partition of Y. Given a function $Y_1, Y_2, \dots, Y_I$ by using a given partition can be denoted as $\prod_{k=1}^{K} \prod_{y_i \in P_k} f_k(y_i, \lambda_k)$ denoted as Classification Maximum Likelihood (CML) method [20]–[22]. Consider the extension the indicator function $z_{ik} = z_k(y_i)$ to be function $\mu_{ik} = \mu_k(y_i)$ where is μ_{ik} as a fuzzy membership in interval [0,1]. In [12], μ is a Fuzzy k-partition of the dataset Y which is based on fuzzy clustering [23]–[25]. The expansion of maximizing the log CML process is obtained by added the fuzziness to reinforce the membership as presented in [13]. It is represented as in (1).

$$\text{Maximize} \quad J_m(\mu, \lambda) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m \ln f_k(y_i, \lambda_k) \tag{1}$$

$$\text{Subject to} \quad \sum_{k=1}^{K} \mu_{ik} = 1 \,;\; 0 \le \mu_{ik} \le 1; i = 1,2, \dots I$$

Here m is called as fuzziness. The value of m has to be greater than 1. The optimization process $J_m(\mu, \lambda)$ is affected by selecting a fuzzy k-partition and a projection λ to make $J_m(\mu, \lambda)$ maximum.

Consider $f_k(y, \lambda_k)$ as a given distribution with (2).

$$f_k(y; \lambda_k) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{kjl})^{y_{jl}} \quad \text{where} \quad \sum_{l=1}^{l_j} \lambda_{kjl} = 1 \; \forall k, j, \tag{2}$$

where λ_{kjl} is a probability of value l for the j th parameter by element i with the kth extreme profil that is $P(Y_{ijl} = 1 | Y_i \text{ in } k \text{ class}) = \lambda_{kjl}$. By replacing $f_k(y, \lambda_k)$ by considering the multivariate multinomial distribution, the mathematical model is represented by (3).

$$J_m(\mu, \lambda) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m \ln \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\lambda_{kjl})^{y_{ijl}}$$
$$= \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m \sum_{j=1}^{J} \sum_{l=1}^{l_j} \ln (\lambda_{kjl})^{y_{ijl}} \tag{3}$$

The maximization of the FkP objective function $J_m(\mu, \lambda)$ can be obtained by updating the equation as in (4) and (5).

$$\lambda_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^m \cdot y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^m} \tag{4}$$

$$\mu_{ik} = \left[\sum_{s=1}^{K} \left(\frac{\sum_{j=1}^{J} \sum_{l=1}^{L_j} \ln(\lambda_{kjl})^{y_{ij}}}{\sum_{j=1}^{J} \sum_{l=1}^{L_j} \ln(\lambda_{sjl})^{y_{ij}}} \right)^{\frac{1}{m-1}} \right]^{-1}. \tag{5}$$

B. Bee Colony

The Bee Colony or artificial Bee Colony has been developed by Karaboga and Ozturk. It is a metaheuristic approach that adopt the honeybee swarm behavior on looking for the nectar. The sources of food are considered as a solution candidate. Moreover, the food amount of a nectar equals to the achieved solution value of objective function. To this, it can be known that the bee's population number is equivalent to the food sources number (SN). The optimization process starts with the initialization of the population. It is assumed that every solution represented by D parameters where $X_i^t = \{x_{i1}^t, x_{i2}^t, \dots, x_{iD}^t\}$ describes the i[th] solution generated in time t by D parameters. The new solution from the previous one is defined using a differential expression. Equation (6) is defined the employed process.

$$v_{ij}^t = x_{ij}^t + \varphi_{ij}^t (x_{ij}^t - x_{pj}^t) \tag{6}$$

where $p \neq i, p \in \{1,2, \dots, SN\}$ is choose randomly. φ_{ij}^t is a scaling factor generated randoml $i = 1,2, \dots, SN, J = 1,2, \dots, D$. Regarding to the minimization problem, the solution in the employed process is define by (7).

$$x_{ij}^{t+1} = \begin{cases} v_{ij}^t & f(v_{ij}^t) < f(x_{ij}^t) \\ x_{ij}^t & \text{otherwise} \end{cases} \tag{7}$$

The employed results will be update by the onlooker process. It is to find the better solution based on the highest of probability value p_i which is chosen by (8).

$$p_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \tag{8}$$

Here fit_i is the fitness value of the solution i. To calculate a particular optimization problem, the process is denoted by (9).

$$\mathrm{fit_i} = \begin{cases} \frac{1}{1+\mathrm{f}(\mathrm{X_i^t})} & \text{if } \mathrm{f}(\mathrm{X_i^t}) \geq 0 \\ 1 + |\mathrm{f}(\mathrm{X_i^t})| & \text{if } \mathrm{f}(\mathrm{X_i^t}) < 0 \end{cases} \quad (9)$$

$\mathrm{f}(\mathrm{X_i^t})$ is the cost value associated to $\mathrm{X_i^t}$. Then, the onlooker process is defined by (10)

$$\mathrm{on_{ij}^t} = \mathrm{x_{ij}^t} + \varphi_{\mathrm{ij}}^{\mathrm{t}}(\mathrm{x_{ij}^t} - \mathrm{x_{kj}^t}) \quad (10)$$

where $\mathrm{k} = \mathrm{argmax}(\mathrm{p_i}|\mathrm{i} = 1,2,\dots,\mathrm{SN})$. $\varphi_{\mathrm{ij}}^{\mathrm{t}}$ is scaling factor generated randomly. The solution in the onlooker process is the new solution of the problem $\mathrm{x_{ij}^{t+1}}$ and is define by (11).

$$\mathrm{x_{ij}^{t+1}} = \begin{cases} \mathrm{on_{ij}^t} & f(\mathrm{on_{ij}^t}) < f(\mathrm{x_{ij}^t}) \\ \mathrm{x_{ij}^t} & \text{otherwise} \end{cases} \quad (11)$$

Equation (12) will be used to be new solution by changing the rejected solution if there is no significant improvement of the solution after a determined number given.

$$x_{ij} = x_{min_j} + rand(0,1)(x_{max_j} - x_{min_j}) \quad (12)$$

where x_{min_j} is a lower bound and x_{max_j} is the upper bounds on the value of the *j*th parameter.

C. Proposed Method

In this section, the proposed approach is presented. It is called Fuzzy K partition based on Artificial Bee Colony (FkPABC), which we refer to ABC algorithm applied to find the membership function of FkP objective function.

In FkPABC, the position of a particle is denoted by X. It describes the fuzzy relation of a set of data objects, $= \{o_1, o_2, \dots, o_n\}$, to set of cluster centers, $C = \{c_1, c_2, \dots, c_k\}$. X is represented by (13).

$$X = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1k} \\ \vdots & \ddots & \vdots \\ \mu_{n1} & \cdots & \mu_{nk} \end{bmatrix} \quad (13)$$

Where μ_{ij} represents the membership function. We use (6) as employed process and (9) as onlooker process in order to update the positions of the bees by using the matrix operation. It may violate the constrains stated in (1). Therefore, it is necessary to adopt the position matrix and make it normalized. First, most of the negative values are changed into zero, in case where the values in a row of the position matrix are equal to zero, we apply (12) to re-evaluate considering the range [0,1] and then the matrix undergoes the transformation as (14) considering the constrains. The pseudo code is described in Fig. 1.

$$X = \begin{bmatrix} \mu_{11}/\sum_{j=1}^{k}\mu_{1j} & \cdots & \mu_{1k}/\sum_{j=1}^{k}\mu_{1j} \\ \vdots & \ddots & \vdots \\ \mu_{n1}/\sum_{j=1}^{k}\mu_{nj} & \cdots & \mu_{nk/\sum_{j=1}^{k}\mu_{nj}} \end{bmatrix} \quad (14)$$

FkPABC
Input: Categorical data set
Output: Clusters

Begin
1. Initialize the parameters including population size P, max iteration
2. Create a swarm with P particles
3. Calculate the Update λ applying (4)
4. Update the position for each particle using (6) as employed process.
5. Update the position for each particle using (9) as onlooker process.
6. If terminating condition is not met, go to step 3.

End

Fig. 1. The psedo code of FkPABC

The termination conditions in the introduced approach is based on the maximum number of iteration or μ_{ik} estimates to be stabil $\| \mu_{ik}^{it} - \mu_{ik}^{it-1} \| < \varepsilon$) or no improvement of the objective function ($| J_{m_m}^{it}(\mu,\lambda) - J_m^{it-1}{}_m(\mu,\lambda)| < \varepsilon$.

D. Computation and Discussion

The main focus of the experiment consists of evaluating performance of the proposed method. The Dunn Index [26] attempts to measure the cluster sets that are near to each other and those clusters are well disjointed. The Dunn's validation index can be computed as in (15).

$$Dn = \min_{1\leq k\leq K}\left(\min_{k+1\leq m\leq K}\left(\frac{d(c_k,c_m)}{\max_{1\leq n\leq k} d'(n)}\right)\right) \quad (15)$$

where $d(c_i, c_j)$ denotes the inter-clusters distance between cluster k and cluster m. The distance measures all the data points of the dataset. In order to validate of the proposed approach, three methods are developed via the UCI datasets, such as: Balloon dataset which is containing 20 objects and 4 categorical variables, Monk with 432 observations and Car dataset with 1728 instances.

In the experiment, all distance obtaining Dunn index in (14) are calculated using Hamming distance. It is known that internal criterion require clusters to be with a high intra-cluster outputs and a low inter-cluster outputs in terms of similarity. As a results, the algorithms that produce with very high Dunn Index are considered successful. Furthermore, the Dun index is shown in Fig. 2.

Based on the Fig 2. Fuzzy partition based on ABC has highest Dunn Index than FkP and Fuzzy centroid. It means that FkP ABC is more desirable to perform clustering the categorical data.

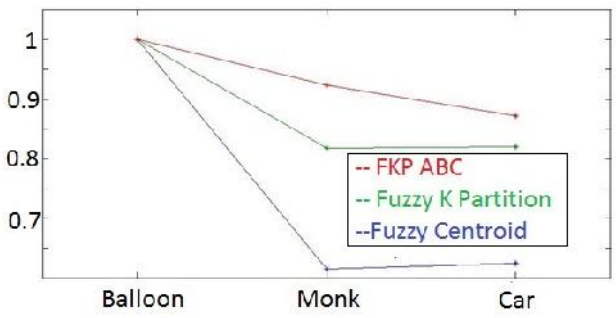

Fig. 2. Dunn Index

III. CONCLUSION

We have discussed clustering technique with emphasize on Artificial Bee Colony (ABC) technique collaborated with Fuzzy partition. We have presented an alternative approach to find the solution the problem of data categorical clustering. The approach is basically using ABC optimization. The proposed algorithm has been successfully applied to some different benchmark data sets. Simulation results have shown that the proposed approach obtains better results in term of the dun index validity. In order to improve the performance and as a future works, we plan to combine the ABC to other algorithms and also to apply the algorithm to other Big data sets

REFERENCES

[1] H. Xu and Z. Tian, "An optimal spectral clustering approach based on Cauchy-Schwarz divergence," *Chinese J. Electron.*, 2009.

[2] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, 1999.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. fifth Berkeley Symp.*, 1967.

[4] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, 1998.

[5] Z. He, S. Deng, and X. Xu, "Improving k-modes algorithm considering frequencies of attribute values in mode," *Int. Conf. Comput.*, 2005.

[6] M. Ng, M. Li, J. Huang, and Z. He, "On the impact of dissimilarity measure in k-modes clustering algorithm," *IEEE Trans. Pattern*, 2007.

[7] O. San, V. Huynh, and Y. Nakamori, "An alternative extension of the k-means algorithm for clustering categorical data," *Int. J. Appl.*, 2004.

[8] Z. Huang and M. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Trans. Fuzzy Syst.*, 1999.

[9] W. Meng, X. Han, Z. Chen, and H. Zhang, "Multi-Agent Reinforcement Learning Based on Bidding," *2009 First Int.*, 2009.

[10] D. Kim, K. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognit. Lett.*, 2004.

[11] K. Umayahara, S. Miyamoto, and Y. Nakamori, "Formulations of fuzzy clustering for categorical data," *Int. J. Innov.*, 2005.

[12] M. Khalilia, J. Bezdek, M. Popescu, and J. Keller, "Improvements to the relational fuzzy c-means clustering algorithm," *Pattern Recognit.*, 2014.

[13] J. M. Leski, *Fuzzy c-ordered-means clustering*, no. 0. 2004.

[14] F. de Carvalho, "Fuzzy c-means clustering methods for symbolic interval data," *Pattern Recognit. Lett.*, 2007.

[15] K. Dobosz and W. Duch, "Understanding neurodynamical systems via fuzzy symbolic dynamics," *Neural Networks*, 2010.

[16] D. Parmar, T. Wu, and J. Blackhurst, "MMR: an algorithm for clustering categorical data using rough set theory," *Data Knowl. Eng.*, 2007.

[17] M. Yang, Y. Chiang, C. Chen, and C. Lai, "A fuzzy k-partitions model for categorical data and its comparison to the GoM model," *Fuzzy Sets Syst.*, 2008.

[18] M. S. Yang, Y. H. Chiang, C. C. Chen, and C. Y. Lai, "A fuzzy k-partitions model for categorical data and its comparison to the GoM model," *Fuzzy Sets Syst.*, vol. 159, no. 4, pp. 390–405, 2008.

[19] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm," *Appl. Soft Comput.*, 2011.

[20] P. BRYANT and J. A. WILLIAMSON, "Asymptotic behaviour of classification maximum likelihood estimates," *Biometrika* , vol. 65 , no. 2 , pp. 273–281, Aug. 1978.

[21] A. J. Scott and M. J. Symons, "Clustering Methods Based on Likelihood Ratio Criteria," *Biometrics*, vol. 27, no. 2, pp. 387–397, Jun. 1971.

[22] M. J. Symons, "Clustering Criteria and Multivariate Normal Mixtures," *Biometrics*, vol. 37, no. 1, pp. 35–43, Mar. 1981.

[23] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[24] K.-L. Wu and M.-S. Yang, "Alternative c-means clustering algorithms," *Pattern Recognit.*, vol. 35, no. 10, pp. 2267–2278, Oct. 2002.

[25] M.-S. Yang, "A survey of fuzzy clustering," *Math. Comput. Model.*, vol. 18, no. 11, pp. 1–16, Dec. 1993.

[26] J. C. Dunn†, "Well-Separated Clusters and Optimal Fuzzy Partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974.