ON A CLUSTERING OPTIMIZATION WITH GENETIC ALGORITHM OF FUZZY C-MEANS AND FUZZY GUSTAFSON-KESSEL (CASE STUDY: FISHER'S IRIS)

ORIGINALITY REPORT

14%	13%	12%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

7%

★ "Fuzzy Systems in Medicine", Springer Science and Business Media LLC, 2000

Publication

Exclude quotes	Off	Exclude matches	< 2%
Exclude bibliography	On		

ON A CLUSTERING OPTIMIZATION WITH GENETIC ALGORITHM OF FUZZY C-MEANS AND FUZZY GUSTAFSON-KESSEL (CASE STUDY: FISHER'S IRIS)

by Lasman Parulian Purba

Submission date: 30-Jun-2021 01:35PM (UTC+0700)
Submission ID: 1614056621
File name: lassification_Rate,_Objective_function_Lasman_Parulian_Purba.doc (102K)
Word count: 2269
Character count: 12249

ON A CLUSTERING OPTIMIZATION WITH GENETIC ALGORITHM OF FUZZY C-MEANS AND FUZZY GUSTAFSON-KESSEL (CASE STUDY: FISHER'S IRIS)

Kiki Indah Novitasari[1], Lasman P. Purba[2], dan Wahyu S. J. Saputra[3]

[2]*Program Studi Teknik Industri, Fakultas Teknologi Industri, Universitas Pelita Harapan Surabaya,*
[1, 3]*Program Studi Teknik Informatika, Fakultas Teknologi Industri,Universitas Pembangunan Nasional "Veteran" Jawa Timur*
Jl. Raya Rungkut Madya, Gunung Anyar Surabaya, 60294. Phone / Fax. : +62-31-8706369
Corespondence email: 1asmanparulianpurba@gmail.com

ABSTRAK

A Fuzzy Gustafson-Kessel (FGK) is one of clustering method using adaptive norm-distance to detect the shape of each data cluster. This algorithm is a development of Fuzzy C-means (FCM) that the result remains local minimum solution, thus the genetic algorithm (GA) approach is used to solve that problem. Then the clustering process uses MATLAB R2012a using FGK Algorithm with GA optimization. This optimization process from FGK clustering using GA is started by inputting the tested data, Fisher's Iris. Resulting the matrix of cluster center from FGK process, then the evolution must be done using GA to make matrix of cluster center more optimal. Based on the test, it can be summarized that the optimization from FGK clustering of Fisher's Iris data set using GA will be better by minimizing objective function. Thus, the objective function value of FGK-GA resulted is smaller than FGK in all tested cluster values. Based on the FGK-GA classification rate 90.31% is more than the average value of FGK classification rate 90%. The test showed that the best cluster is 3 and this value is similar to Fisher's Iris data set classified in 3 classes.

Kata kunci: Clustering, Fuzzy Gustafson-Kessel, Genetic Algorithm, Fisher's Iris, Classification Rate, Objective function

1. INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) (http://kangedi.lecturer.pens.ac.id 2015). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics (https://en.wikipedia.org). Machine learning is a subfield of computer science (www.britannica.com) that evolved from the study of pattern recognition and computational learning theory in artificial intelligence (www.britannica.com). Machine learning explores the construction and study of algorithms that can learn from and make predictions on data (Kohavi *et al*., 1998). Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions (Bishop, 2006) rather than following strictly static program instructions. In clustering, the similarity of the object is usually obtained from the proximity of attribute values that describe each data object. While each data object is usually represented as a point in a multidimensional space.

Fuzzy Gustafson - Kessel (FGK) is a method for clustering data as a development of Fuzzy C - Means (FCM) method. Some researchers applied the FCM algorithm to a variety of cases, including: data clustering of teaching faculty performance (Luthfi, 2007), the analysis of poor families (Wardani, 2010), and the determination of the end value of the lecture (Khoiruddin, 2007). Random initiation sensitivity on FCM, causing clustering results obtained is stuck in a local minimum.

The minimization of the FCM functional represents a nonlinear optimization problem that can be solved by using a variety of methods, including iterative minimization, simulated annealing or genetic algorithms (Babuska, 2009, Gustafson *et al*., 1978). The most popular method is a simple Picard iteration through the first-order conditions for stationary points of fuzzy c-means functional, known as the fuzzy *c*-means (FCM) algorithm (Babuska, 2009). In

this paper, it was the same initiation of both FCM method and FGK method, as known also from (Babuska, 2009, Bezdek, 1980, Mauliyadi *et al*., 2013), that's why it is a need an approach to optimize local minimum problems that may occur on FGK method too. Researcher (Widyastuti *et al*., 2007, Febriawan, 2011) used the GA algorithm to eliminate the local minimum problem regarding the clustering. Then, it is necessary for an accurate analysis of the FGK algorithm with genetic algorithm approach in order to know the best cluster optimization results. Because the genetic algorithm approach can usually eliminate such problems (Babuska, 2009, Gustafson *et al*., 1978, Kusumadewi, 2003, http://entin.lecturer.pens.ac.id).

In this paper, study about clustering optimization of the data used will be focus by using a combination of FGK and GA. The data used in this study is the Fisher's Iris dataset. This dataset consists of 150 samples of data, divided into 3 classes where each class consists of 50 samples of data that is Iris - Setosa , Iris - versicolor and Iris - virginica. Each class of the Iris using four features, namely the petal width, petal length, crown width, and length of the crown.

Parameter of the FGK method was the level of vagueness (fuzziness) is 2, the iterations limit is 100, and the accuracy was 0.00001. 2. For the GA, number of individual initial population 100, the number of generations 50, Crossover probability (Pc) = 0.9, the probability of mutation (Pm) 0.2. These parameters will be tested on several cluster that is 3, 4, and 5.

2. FGK ALGORITHM

Gustafson and Kessel (Babuska, 2009, Gustafson *et al*., 1978, Bezdek, 1980) extended the standard FCM algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set. In this method, each iteration update the clustering norm inducing matrix Ai, so that each cluster can adjust the shape of its group with the data. Here is the algorithm of FGK (Babuska, 2009):

$$J(Z;U,V,\{A_i\}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m D_{ikAi}^2 \tag{1}$$

Each cluster has its own norm-inducing matrix A_i, which yields the following inner-product norm:

$$D_{ikAi}^2 = (z_k - v_i)^T A_i (z_k - v_i) \tag{2}$$

By using the Lagrange multiplier method, the following expression for A_i is obtained:

$$A_i = [\rho_i \det(F_i)]^{1/n} F_i^{-1} \tag{3}$$

where Fi is the *fuzzy covariance matrix* of the *i*th cluster, Z is the data to be grouped, U is a matrix of initial partitions, V is the matrix center group.

The matrices A_i are used as optimization variables in the *c*-means functional, thus allowing each cluster to adapt the distance norm to the local topological structure of the data.

3. GENETIC ALGORITHMS

Genetic algorithms are evolutionary algorithms in the field of artificial intelligence inspired by evolutionary biology such as inheritance, selection, mutation, where the most powerful individual is a winner in the competition environment. GA consists of eight components: encoding scheme, the value of fitness, selection, parents, crossover, mutation, elitism (for AG manifold generational replacement), the replacement of the population, and termination criteria (Kusumadewi, 2003).

Encoding scheme

To be able to be processed using a GA, a problem must first be converted into individual shapes which represented by one or more chromosomes with a specific code. GA represents the gene (artificial), in general, as a real number, decimal or binary.

The value of fitness

On the issue of optimization, if a solution sought to maximize a function h (known as the maximization problem), then the fitness value used is the value of the function h, i.e. f = h (where f is the fitness value). But if the problem is to minimize the function h (minimization problem), then the function h can not be used directly. This is due to the GA using a rule that individuals with higher fitness values will have the ability to survive higher than the low fitness values. Therefore, the value of fitness for a minimization problem is f = 1 / h, which means that the smaller the value h the greater the value f. But this function can be problematic if h is 0, the result could be worth infinity f. To overcome these problems, h needs to be coupled with a number which is considered very small, so the formula of its fitness function becomes:

$$f = 1/(h+a)' \tag{4}$$

Selection

The process of selecting two individuals as parents usually done proportionally based on the values of its fitness. One commonly used method of selection is the roulette - wheel. As the name, this method implies mimicking the roulette - wheel which each individual occupies a reduction shall circle on the roulette wheel, in proportion to its fitness value.

Crossover

Crossover is the operator of genetic algorithm involving two mains for forming a new chromosome. Crossovers produce a new point in the search space that is ready to be tested. This operation is not always carried out on all existing individual. Individuals were randomized to do with Pc crossing between 0.6 - 0.95. If the crossing is not done, then the value of the parent will be lowered to the descent (Febriawan, I. 2011).

Mutation

Mutation of floating numeral will be the non-uniform mutation or known as dynamic mutation. This mutation was designed to be well with the aim of achieving a high level of accuracy.

Elitism

Because the selection is done randomly, then there is no guarantee that the highest fitness of an individual will always be selected. Even if the individual is worth the highest fitness elected, that individual may be damaged because the process of crossing over. To maintain the highest fitness worth individuals are not lost during evolution, need to be made one or two of copies. This procedure is known as elitism.

Replacement population

Individual removal procedure is like the elimination of the oldest individual or individuals who have the highest fitness value. Elimination of individuals can be performed on a parent only or on all individuals in the population.

4. VALIDITY CLUSTERING

Calculation of the validity of clustering needed to determine whether the results obtained from the grouping by FGK algorithm has included the best results. The clustering validity used is Partition Coefficient (PC), Entropy Classification (CE) and Partition Index (SC). The partition coefficient is used to measure the closeness of all prototype samples of the selected input. CE is entropy classification which translates as the degree of uncertainty of the object. SC is a function of the validity of compactness and separation, which is the ratio between the average distances of the sample with the selected prototype with a minimum distance between the prototypes. Validity of clustering of each of these processes on the FGK and the FGK - GA will be calculated.

5. DISCUSSION

FGK piloted in several clusters, i. e. 2, 3, 4, and 5. Tests on some of these clusters are used to determine the optimal number of groups that can be processed on GA. After testing against several clusters including second, 3rd, 4th, and 5th then the value of the objective function, Jm, the validity of clustering FGK, and the validity of clustering FGK - GA for each cluster were obtained.

Table 1. Values of Jm FGK and FGK – GA

Sum of cluster (c)	Jm values on FGK	Jm values on FGK-GA
2	105.0524	73.5494
3	94.5802	62.3988
4	91.0258	19.5770
5	90.4040	11.9416

Table 2. Values of clustering validity of FGK

c	Partition coefficient	Classification entropy	Separation index
2	0.7404	0.4061	0.042
3	0.7278	0.4662	0.0049
4	0.6201	0.6979	0.062
5	0.5351	0.8952	0.0066

Table 3. Values of clustering validity of FGK-GA

c	Partition Coefficient	Classification Entropy	Separation Index
2	0.7060	0.4500	0.0279
3	0.6377	0.6146	0.0063
4	0.3972	1.0948	0.0191
5	0.4420	1.0313	0.0062

Table 4. Classification rate of c=3

Experiment number-	FGK (%)	FGK-GA (%)
1st	90%	92%
2nd	90%	90%
3rd	90%	90.67%
4th	90%	88.67%
Average	90%	90.31%

Table 4 showed that the value Jm of FGK-GA is mostly smaller than the value Jm of FGK in all grades of tested cluster. FGK-GA classification rate is 90.31% was greater than the average value FGK classification rate which is 90%. Through the tests performed as seen on Table 4, the best values were obtained in cluster number 3. This value corresponds to the Fisher 's Iris dataset that has been classified into three classes.

6. CONCLUSIONS

The fuzzy clustering by using Fuzzy Gustafson-Kessel (FGK) algorithm and Fuzzy Gustafson-Kessel with Genetics Algorithm (FGK-GA) was better than FGK standard. It was found that the objective function, Jm of the FGK is 94.5802, then Jm of the FGK-GA is 62.3988. From 4 times of experiment, the value of classification rate of FGK-GA was found about 90.31% and 90% for FGK.

REFERENCES

Babuska, R. (2009). *Fuzzy and Neural Control. Delft University of Technology*. Delft University of Technology, Germany.

Bezdek, J. C. (1980). A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, NO. 1, PAMI-2, 1-8.

Bishop, C. M. (2006). "Pattern Recognition and Machine Learning". *Springer*, ISBN 0-387-31073-8.

Febriawan, I. (2011). "*Optimasi Hasil Clustering Fuzzy C-Means Menggunakan Algoritma Genetika (Studi Kasus : Clustering Data Bunga Iris)*". Skripsi thesis. Universitas Brawijaya, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Brawijaya, Malang.

Gustafson, E. E. and Kessel, W. C. (1978). Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proc. IEEE,* 761-766.

http://entin.lecturer.pens.ac.id,2015

http://kangedi.lecturer.pens.ac.id, 2015.

https://en.wikipedia.org

Khoiruddin, A. A. (2007). "Menentukan Nilai Akhir Kuliah dengan Fuzzy C-Means". *Seminar Nasional Sistem dan Informatika 2007*, Bali.

Kohavi, R. and Foster P. (1998). "Glossary of terms". *Machine Learning,* Vol. 30, 271–274.

Kusumadewi, S. (2003). *"Artificial Intelligence"*. Graha Ilmu, Yogyakarta.

Luthfi, E. (2007). *Fuzzy C-Means untuk Clustering Data (Studi Kasus : Data Performance Mengajar Dosen)*. STMIK AMIKOM, Yogyakarta.

Mauliyadi, A., Sofyan, H., and Subiyanto, M. (2013). Perbandingan Metode Fuzzy C-Means (FCM) dan Fuzzy Gustafson-Kessel (FGK) Menggunakan Data Citra Satelit Quickbird (Studi Kasus Desa Lubuk Batee, Aceh Besar). *Jurnal Matematika,* Vol. 00, 01-05.

Wardani, I. (2010). *Analisa Keluarga Miskin dengan Menggunakan Metode Fuzzy C-Means Clustering*. Politeknik Elektronika Negeri Surabaya, Surabaya.

Widyastuti, N., and Hamzah, A. (2007). Penggunaan Algoritma Genetika dalam Peningkatan Kinerja Fuzzy Clustering untuk Pengenalan Pola. *Berkala MIPA*, Vol. 17 (2).

www.britannica.com