HASIL CEK_PAK SUGIYARTO_1

by Cek_pak Sugiyarto Pak Sugiyarto

Submission date: 20-Apr-2021 10:39AM (UTC+0700)
Submission ID: 1564316457
File name: FCM_core_and_reduct_dimension_reduction_-_rev.docx (350.34K)
Word count: 5063
Character count: 26782

Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 7, No. 1, August 2021, pp. xx-xx
ISSN: 2338-3070, DOI: 10.26555/jiteki.v7i2.xxxx

Distance Functions Study in Fuzzy C-Means Core and Reduct Clustering

Eliyanto, Joko [1], Surono, Sugiyarto[2]
[1]Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia

ARTICLE INFO

Article history's:

Received
Revised
Accepted

Keywords:

Clustering
Core and Reduct
Dimensional Reduction
Distance Function
Fuzzy C-Means

ABSTRACT

Fuzzy clustering aims to produce clusters that take into account the possible membership of each dataset point in a particular cluster. Fuzzy C-Means Clustering Core and Reduct is a fuzzy clustering method is a Fuzzy C-Means Clustering method that has been optimized using the reduction of Core and Reduct dimensions. The method studied is highly dependent on the distance function used. As a further in-depth study, this study was compiled to see the performance of the Fuzzy C-Means Clustering Core and Reduct using various distance functions. We aim to see how consistent the results of this method are across various distance functions and find the best distance function. The seven distance functions are applied to the same dataset. The seven distances are the Euclidean, Manhattan, Minkowski, Chebisev, Minkowski-Chebyshev, Canbera, and Averages distances. We use UCI Machine Learning datasets for this research. The quality of the clustering results is compared through several measures. Accuracy, Silhouette score and Davies Bouldin Index are used as internal measurements. The results of Fuzzy C-Means Core and Reduct clustering on all distance functions have significantly decreased computational load. Accuracy and purity values can be maintained with values above 80%. There was an increase in the value of the Silhouette Coefficient Score and a decrease in the Davies Bouldin Index after the application of dimension reduction. This means, the quality of the clustering results can be maintained. The distance with the best evaluation result is the Eucliden distance. This method runs consistently across all tested distance functions.

Department of Mathematics, Ahmad Dahlan University, Yogyakarta, Indonesia
Email: sugiyarto@math.uad.ac.id.

1. INTRODUCTION

Current technological developments produce data that is not only large but also continuous. In fact, in recent times, humans have produced more data than all data that has been previously generated [1]. Data at this time is available massively, in large quantities, and in various types [2]. This kind of data is termed big data. This forces us to be able to extract important information from this abundant data.

One of the important information in the data is the data group. Data grouping is very useful for solving various problems in life. This is often applied as in Customer Segmentation, Recommendation, Image Processing, and others [3]. Often the data clusters have not been previously identified. So, supervised learning cannot be applied. One of the things that can be used as the basis for grouping data is similarity. This method of grouping is called clustering [4]–[6]. The next problem is often a data grouped into a group arbitrarily, without considering the possibility to join in other groups. Maybe the computation process will run faster, but its accuracy is questionable. To solve this problem, fuzzy clustering has been proposed as a solution.

The degree of membership is the basis of the fuzzy clustering method [7]. Based on that, say each data point against each exclusion cluster. Fuzzy C-Means clustering is a popular method used in fuzzy clustering[8]. Fuzzy C-Means clustering is a distance-based clustering that applies the concept of fuzzy logic [9]. The clustering process goes hand in hand with the iteration process to minimize the objective function [3], [7], [8]. The objective function is the sum of the multiplication of the distance between the data points to the nearest cluster center with the degree of membership [10]. The more iterations, the decreasing the value of the function should be. The distance function used in this method has a key role[11].

Various studies on the effect of distance in the clustering method have been carried out. Some of the results of previous studies that no distance is more dominant and produce outputs that are not much different. The results of clustering are very dependent on the dataset used[3]. Euclidean and Mahattan / City Block, Chebisev, and Minkowski distances have been identified for their effects in the K-Means Clustering algorithm [11], [12]. The results of both studies indicate that the Manhattan distance has slower computation time than the other distances. In another study, the Euclidean, Mahattan / City Block, Canbera, and Chebisev distances were applied and evaluated on the fuzzy clustering algorithm[13]–[15]. The results of this study concluded that the results of clustering were very dependent on the data used[16]. In our latest research, the combined Minkowski and Chebisev distances can also be used to optimize Fuzzy C-Means clustering[17]. Another form of Euclidean distance, namely Average distance, can also be used in the clustering algorithm and produces better results than Euclidean distance[18].

Another way to optimize the clustering method is to apply the dimension reduction method[19]. Dimension reduction method can reduce data dimension but still maintain data characteristic[20]. One of the dimensional reduction methods is Core and Reduct. The Core and Reduct method from the Rough Set theory is proven to be able to improve the performance of Fuzzy C-Means Clustering at the Euclidean distance function [3], [21]. In this study we are doing an expansion of the research on the last results. We want to know whether the consistent application of Core and Reduct can reduce the computational load on Fuzzy C-Means Clustering with various distance functions. The second objective is that we want to find the best distance for the new method. The data used are also limited to five UCI machine learning data, namely iris data, yeast data, seeds data, sonar data, and hill-valley data [22]. In this study, the method is only implemented on numerical data. The Core and Reduct dimension reduction method used was also developed limited to numerical data only.

2. RESEARCH METHOD

2.1. Fuzzy C-Means Core and Reduct Clustering

Fuzzy C-Means Clustering (FCM) is a clustering method which allows certain data can be induced in two or more clusters [7]. This method was invented by Dunn in 1973 and developed further by Bezdek in 1981. The usual application for this method is for pattern introduction. This method is based on the minimalizing this following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, \quad 1 \le m < \infty \tag{1}$$

Where m is certain real number higher than 1, u_{ij} is the membership degree of x_i, in cluster j, x_i represent the-i data, c_j represent the cluster centroid j, and $\|*\|$ is the norm which state the similarity between data and the cluster centroid.

Fuzzy partition is applied trough the continuous optimization process of the objective function which defined before, with the update of the u_{ij} membership degree matrix and c_j cluster centroid by this following equation:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|}\right)^{\frac{2}{m-1}}} \tag{2}$$

And

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{3}$$

The iteration will be terminated when $\max_{ij}\left\{\left\| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right\|\right\} < \varepsilon$, where ε is the termination criteria valued between 0 and 1, and k is the number of iteration process. This procedure is convergence to minimum local point or saddle point of J_m.

The algorithm of Fuzzy C-Means Clustering consist of this following steps [3]:

1. Initialization of $U = matriks\left[u_{ij}\right], U^{(0)}$.
2. In the steps of –k, measure the cluster centroid vector of $C^{(k)} = \left[c_j\right]$ with $U^{(k)}$.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. Update the value of $U^{(k)}, U^{(k+1)}$.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}$$

4. If $\left\|U^{(k+1)} - U^{(k)}\right\| < \varepsilon$, then terminate the iteration, if not, back to step 2.

In this method, the value of the objective function and membership degree are much related. During the initial iteration process, we assume that each of the data coordinates already have the value of membership degree for each of the existing cluster. Then, this value will be continuously updated trough (2). When (2) is continuously updated, (1) will also keep updated towards its minimum value. One of the challenges of the clustering method is the large computation load. To overcome this, in this method, before entering the clustering process, a dimensional reduction process is carried out using Core and Reduct. The dataset for clustering problems can be viewed as an information table [3]. In the information table , the set of attributes $R \subseteq A$ is called a reduct, if R satisfies the following two conditions:

1) $IND(R) = IND(A)$; (6)

2) $\forall a \in R, IND(R - \{a\}) \neq IND(A)$. (7)

Condition one states that for each object pair that cannot be distinguished by a subset R, it also cannot be distinguished by A, and vice versa. The second condition states that there are object pairs that cannot be distinguished by R - {a} but can be distinguished by A. This means that R is the minimum set of attributes that can maintain the discernibility relationship IND (A). Usually, there is more than one reduction in an information table. The set of all reductions from the information table T is denoted as RED (T).

Then, the cores of the attribute set $R \subseteq A$ are as follows:

$$CORE(R) = \cap RED(R) \quad (8)$$

The following algorithm for Fuzzy C-Means Clustering with Core and Reduct dimensional reduction:

Algorithm 1. ***Fuzzy C-Means Core & Reduct Clustering***

INPUT:
Data input is in the form of variables $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_m\}$ expresses objects and attributes. Data is a $n \times m$ matrix, where n is a lot of data and m is the number of data attributes.

PROCESS BEGIN
1. If the dataset is not numeric data, then encoding data is done, if not then proceed to the next process.
2. Apply the core and reduct method so that the number of variables $Y = \{y_1, y_2, \cdots, y_m\}$ will be a number of new variables $Y = \{y_1, y_2, \cdots, y_p\}$, with $p \leq m$.
3. Applying the fuzzy c-means clustering method so that data clusters are obtained.
4. Cluster evaluation.

PROCESS END

OUTPUT:

The value of the objective function, computational time, purity, Davies Bouldin Index, Silhouette Score and accuracy.

2.2. Distance Function

a. Euclidean Distance

Euclidean distance is known as the most common and applied distance for Fuzzy C-Means clustering process. For x and y coordinates, this distance is defined as:

$$d_{euclidean}(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \quad (9)$$

Where x_k and y_k are the value of x and y on certain dimension of n. This distance become the standard distance for fuzzy c-means clustering method [11], [18].

b. Manhattan Distance

Manhattan distance is defined as the addition of all of the attributes distance. Hence, for two coordinates data of x and y in dimension n, the Manhattan distance for both of the coordinates is defined as [12]:

$$d_{manhatan}(x,y)=\sum_{k=1}^{n}|x_k-y_k| \quad (10)$$

Where x_k and y_k are the value of x and y on certain dimension of n.

c. Chebisev Distance

This distance also known as maximum distance which defined as the maximum value of the existing attributes distance. The distance for two coordinates data of x and y in dimension n is defined as [12]:

$$d_{chebisev}(x,y)=\max_{k=1}^{n}|x_k-y_k| \quad (12)$$

Where x_k and y_k are the value of x and y on certain dimension of n.

d. Minkowski Distance

Minkowski distance is the formulation of the metric distance which defined as [12]:

$$d_{minkowski}(p,x,y)=\sqrt[p]{\sum_{k=1}^{n}|x_k-y_k|^p} \quad (13)$$

p is functioned as the Minkowski parameter. In Euclidean distance, the value of p is equal to 2. For Manhattan distance, the value of p is equal to 1, and for Chebisev distance the value of p is equal to ∞. The metric condition for Minkowski distance is fulfilled as long as $p \geq 1$.

e. Mikowski-Chebisev Distance

Minkowski-Chebisev distance is invented by Rodrigues in 2018 [17]. The combination of Minkowski and Chebisev distance with weight of w_1 and w_2 is defined as (14).

When w_1 is bigger than w_2, this distance is similar to Minkowski distance, and vice versa. The Minkowski-Chebisev distance is defined as [23]:

$$d_{\min-cheb}(w_1, w_2, p, x, y) = w_1 d_{\min kowski}(p, x, y) + w_2 d_{chebisev}(x, y) \quad (14)$$

f. Canberra Distance

Canberra distance is the defined as the addition of the absolute value of fraction difference between two data. The formula of this distance is [16]:

$$d_{canbera}(x, y) = \sum_{k=1}^{n} \frac{|x_k - y_k|}{|x_k| + |y_k|} \quad (15)$$

This distance is very sensitive towards alteration when the value of both of the analyzed coordinates are close to 0. This distance is chosen because of the similarity of its character to Manhattan distance.

g. Average Distance

Average distance is the modification of the Euclidean distance. This modification is applied to improve the clustering result [18]. This distance is defined as:

$$d_{average}(x, y) = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k - y_k)^2} \quad (16)$$

2.3. Cluster Evaluation

Cluster evaluation is applied to determine the clustering algorithm level of accuracy and the availability of cluster label. For this research, four test are applied to evaluate the clustering process, which are purity test , acccuracy test, Davies Bouldin Index(DBI), and Silhouette coefficient score.

1. Purity

Purity is used to calculate the purity of a cluster. Purity calculation for each cluster obtained is done by taking the most objects entered in the C-cluster where $1 < i < C$ and C' are the original h-class with $1 < h < C'$. As for the overall purity of the C cluster, it is done by adding up each purity in the C cluster and dividing it by the number of objects defined as follows:

$$\text{purity}(P, C) = \frac{1}{n} \sum_{i}^{c} \max_{i \le h \le c} , |P_i \cap C_h| \quad (17)$$

where $P = \{P_1, P_2, P_3, \cdots P_c\}$ is the cluster set and $C = \{C_1, C_2, C_3, \cdots, C_c\}$ is the original class set. Poor clustering has a purity value close to 0. This means that there are no cluster results that match the original class. While a good cluster has a value of purity 1. This means that the cluster results are in accordance with the original class.

2. Accuracy

Accuracy is calculated by adding up the number of objects included in the i-cluster, where $1 < i < C$ the exact class is then divided by the number of data objects. Accuracy is defined as follows:

$$r = \frac{\sum_{i=1}^{c} a_1}{n} \quad (18)$$

where:

a_i : number of objects in the i-cluster that correspond to the original class.

n : number of n objects.

Good accuracy results if all clusters match the original class and then divided by the amount of data will produce a value of 1.

3. Davies Bouldin Index

Davies Bouldin Index (DBI) is one of the methods used to measure cluster validity in a clustering method. The purpose of measurement with DBI is to maximize the distance between clusters (inter-cluster) and to minimize the distance between data points (intra-cluster) in the same cluster [17]. The Davies Bouldin Index (DBI) introduced by David L. Davies and Donald W. Bouldin in 1979. Before calculating the DBI, first calculate the variance of each cluster.

$$\text{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

where:

$\bar{x}$: mean of cluster x

n : number of cluster members

Then calculate the Davies-Bouldin Index (DBI) with equations:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{19}$$

with $R_i = \max_{j=1,\cdots,k,\ i\neq j} R_{ij}$ and $R_{ij} = \dfrac{\text{var}\left(C_i\right) + \text{var}\left(C_j\right)}{\left\|c_i - c_j\right\|}$, where

C_i : i-th cluster

c_i : i-th cluster center

4. Silhouette Coefficient Score

Silhouette Coefficient Score (SCS) is an internal metric that measures the cohesiveness and separation of clusters at the same time [24]. SCS calculates the average distance in a cluster and the minimum distance between an object to another cluster as follows:

$$SCS = \frac{1}{n}\sum_{i=1}^{n}\frac{\beta_i - \alpha_i}{\max\left(\alpha_i, \beta_i\right)} \tag{20}$$

α_i is the average distance of objects in a cluster, namely:

$$\alpha_i = \frac{\sum_{j\neq i, x_j \in c_i}\left|x_i - x_j\right|}{\left|c_i\right|}$$

and β_i represents the distance between an object x_i with the center of the cluster w_j, β_i which is calculated as follows:

$$\beta_i = \min\left\{\left|x_i - w_j\right|, j = 1,2,\cdots k, j \neq i\right\}.$$

SCS values can range from -1 to 1 $\left(-1 \leq SCS \leq 1\right)$, where 1 means the grouping solution is "true" and -1 means the grouping solution is "wrong".

2.4. Dataset

The data used in this study is a dataset taken from the UCI Machine Learning website [22]. Table 1 presents a brief description of the dataset used in this study.

Table 1. Dataset

Dataset	Object	Variabels	Class
Iris	150	4	3
Seeds	210	7	3
Yeast	1484	8	10
Sonar	208	60	2
Hill Valley	606	100	2

All of the above datasets are numeric types.

2.5. Research Method

This research is a numerical simulation using the Fuzzy C-Means Core and Reduct Clustering method. The program is structured using Python 3.0. This clustering method is carried out using 7 different distance functions then the results are visualized and analyzed. The research steps are presented in Fig. 1.

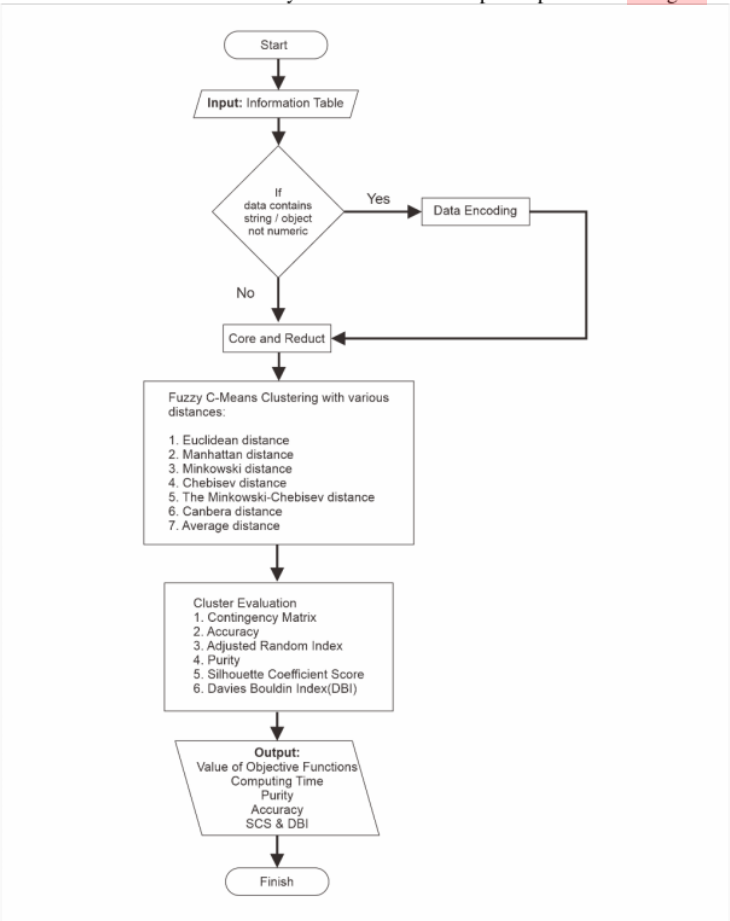

Fig. 1. Research Flow Chart

3. RESULTS AND DISCUSSION

The initial step of this method is to reduce the dimensions of the dataset. The dataset in Table 1 is reduced using the Core and Reduct method. The result of this dimension reduction is a new dataset with fewer variables. The results of the reduction are presented in Table 2.

Table 2. Core and Reduct results

Dataset	Before Reduction	After Reduction
Iris	4	3
Seeds	7	2
Yeast	8	4
Sonar	60	2
Hill Valley	100	2

Based on the result above, Core and Reduct works better for the data with high dimension. As for low dimension dataset, it tends to be difficult to determine the core of the analyzed data. These results are consistent with research in the same field [21]. Data with low dimensions will make the computation load lower and the computation time can be increased significantly.

The main objective of Fuzzy C-Means clustering is to acquire the objective function valued as low as possible. The lower the objective function, the better the result of the fuzzy c-means clustering application. This means that the group of data is more clearly separated. Fig. 2 portrayed the comparison between the values of objective function acquired from Fuzzy-C Means and Fuzzy C-Means Core and Reduct. For 5 simulated dataset, Core and Reduct able to decrease the value of objective function close to 0% remains for the Euclidean distance, Manhattan distance, Canberra distance, and Minkowski distance. These results support previous research which states that Euclidean is one of the best performing distances [12].

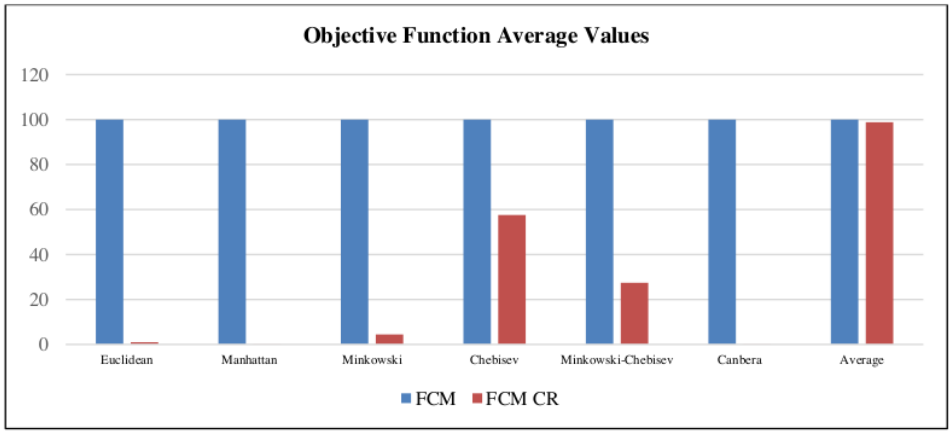

Fig. 2. Objective function average values

The process of computation which works in lower dimension is resulting in the decrease of the computing weight. This phenomenon will affect the computing time and the number of iteration process applied. Fig. 3 depicts the comparison of computing time in these two applied Fuzzy C-Means methods. At all distances, the computation time decreased significantly. The highest drop occurred at the Minkowski distance. However, the Euclidean distance still has the lowest computation time of all the distances used. These results indicate the consistency of the Core and Reduct method which is able to reduce the computational load on the Fuzzy C-Means Clustering method. These results complement previous studies which were limited to the Euclidean and Minkowski-Chebisev distances [3], [17].

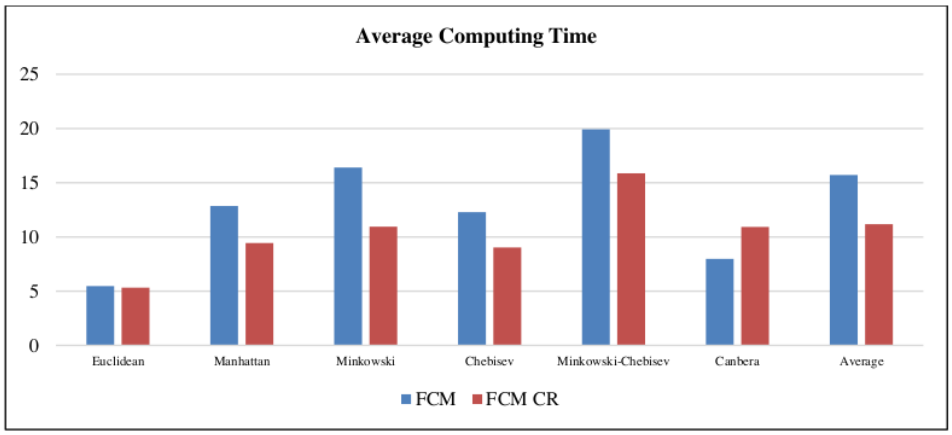

Fig. 3. Average computing time

The lower computation time is usually also due to the iteration time it takes to converge. If the number of iterations is lower, then the fewer steps must be taken. Fig. 4 illustrates the comparison of the average number of iteration until convergent from both of the analyzed methods. The result shows that the Core and Reduct method is relatively good enough to decrease the computing time of Fuzzy C-Means clustering. This behavior caused by the decrease of the number of iteration until convergent in the application of Fuzzy C-Means Core and Reduct. The acquired result is not quite significant because Core and Reduct only decrease the number of attributes, while the data record is stagnant.

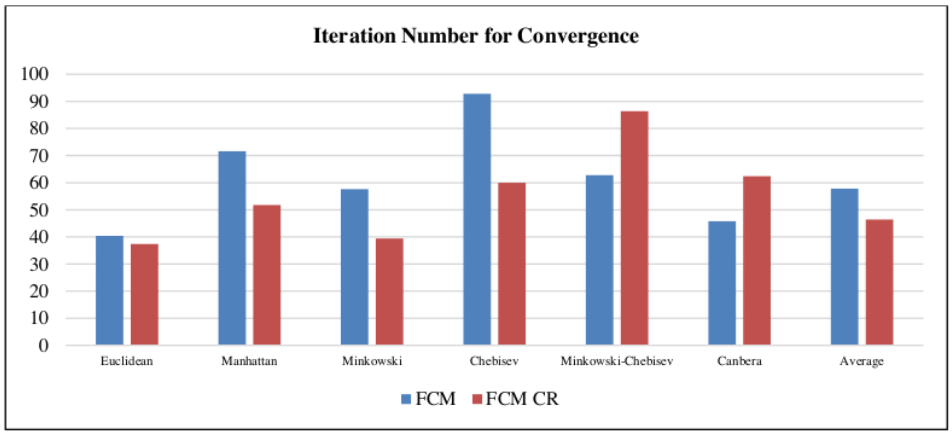

Fig. 4. Number Of Iteration Process Until Convergent

Another expected result from the application of dimension reduction is the good quality of clustering result which has to be similar to the result of common clustering process. The good cluster output should have accuracy and purity valued close to 1. Accuracy describes the accuracy of the working clustering model. Fig. 5 presents the accuracy value for each distance function. The average accuracy for all distances is 0.47. The distance with the highest accuracy value is Euclidean with a value of 0.56. Meanwhile, the distance with the lowest accuracy value is the Minkowski distance with a value of 0.38. However, the impact of the application of the reduction of the Core and Reduct dimensions is to compare the accuracy results before and after the dimensional reduction. It can be seen that Euclidean, Manhattan, Minkowski-Chebisev, Canbera, the average accuracy can be maintained above 80%. At Minkowski's distance, the accuracy drops drastically until only around 60% remains. This result looks like it can be improved by combining it with Chebisev's distance. This supports the research that Minkowski-Chebisev's new distance has a great impact and can be used for optimization of Machine Learning methods.[17], [23].

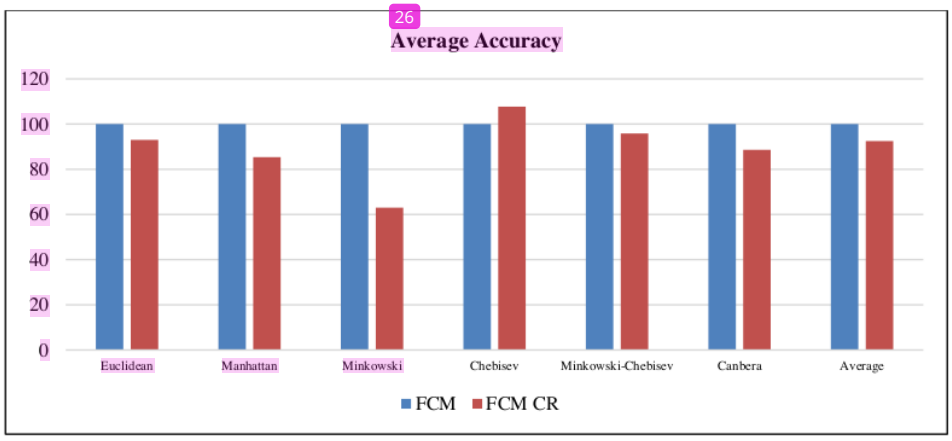

Fig. 5. Average Accuracy

A good cluster result has a purity value close to 1. The result of the average purity value in the Fuzzy C-Means Core and Reduct Clustering method is 0.61. This method succeeded in achieving the highest purity value in the Canbera distance function with an average value of 0.64 and the lowest at the Minkowski-Chebisev distance with an average value of 0.55. This result is quite good, considering the average purity is more than 0.5. When compared with the Fuzzy C-Means method, the purity value can be maintained very well. The purity value on the reduced data is still above 80% of the original result. This means that even though it is reduced, the quality of the clusters can be maintained. The result of this paper expands the previous paper[3]. Fig. 6 presents these results.

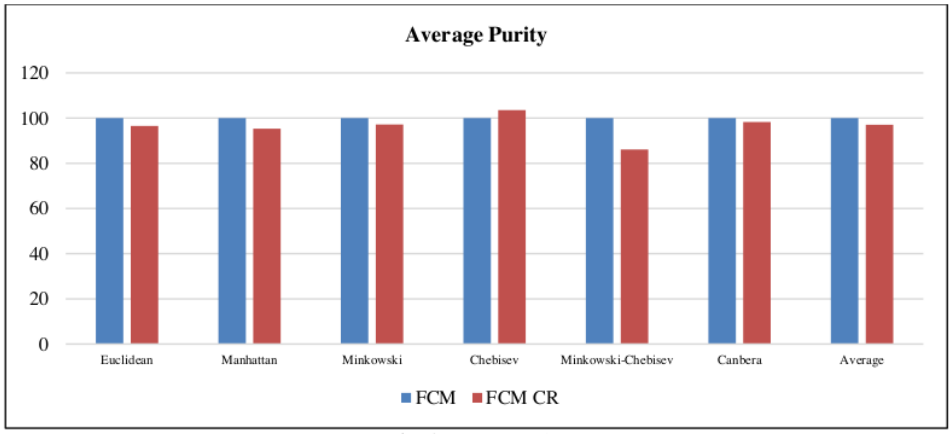

Fig. 6. Average Purity

Another measure of the goodness of a cluster is the Silhouette Coefficient Score. The higher the value of this metric, the more correct the clustering results can be. Interesting research results emerge in this section. In average, Fuzzy C-Means Clustering with Core and Reduct application is able to improve the value of silhouette score from 0.399 to 0.507. This means that the Core and Reduct method consistently across all distance functions can improve the quality of cluster results. Fig. 7 presents the Silhouette Coefficient Score value at each distance for the two methods. Nearly all distance functions have a significant increase for this measure. The Minkowski-Chebisev distance yields the worst score in this case.

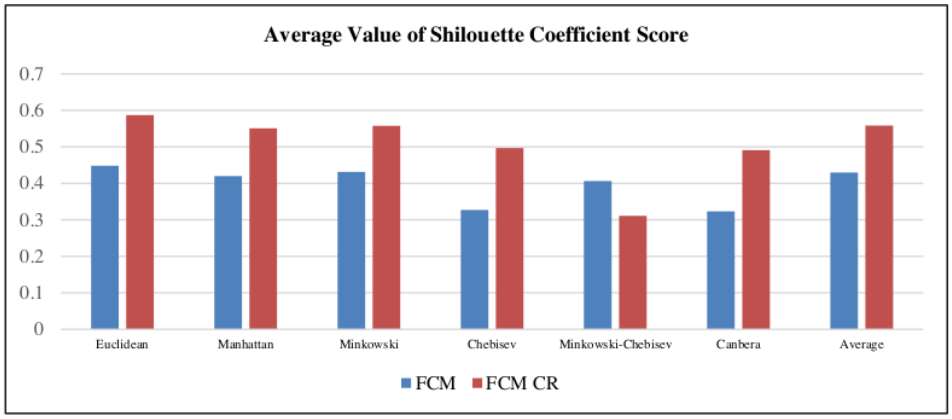

Fig. 7. Average value of Silhouette Coefficient Score

Fig. 8 shows the comparison of the average DBI results between the Fuzzy C-Means Clustering method and Fuzzy C-Means Core and Reduct Clustering. A cluster will be considered to have an optimal clustering scheme if it has a minimal Davies Bouldin Index (DB) (close to 0) [17]. This new method is able to decrease its remaining value to 53%. This means that the reduction of Core and Reduct dimensions increases the results of Fuzzy C-Means clustering. This applies to all distance functions.

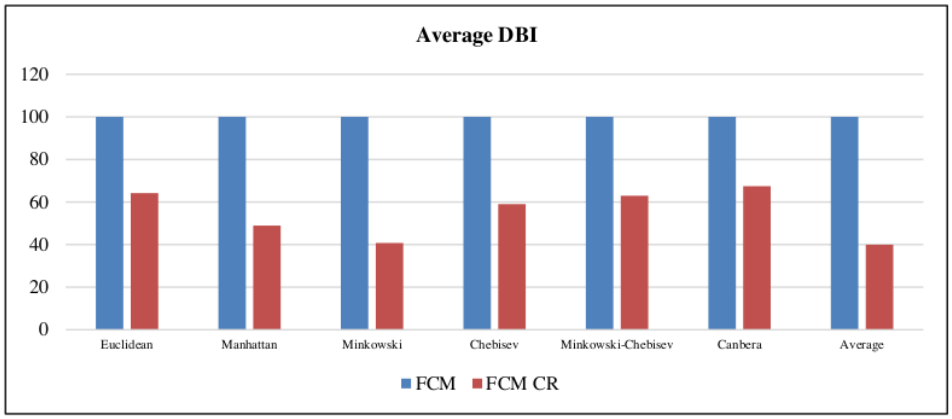

Fig. 8. Average DBI

Few of the aforementioned result are linear with the previous research [25], [26]. R. Zhao, L. Gu, dan X. Zhu also did a research in the same field as this research. Their research resulted in the combination of the C-Means Clustering and Reduct functioned as Rough Set Feature Selection which able to improve the accuracy with value averaged in 1% [21]. The addition of the Core proces which applied in this research guaranteed that the result of the dimension reduction only acquired from the core of dataset only.

The best result of this research is acquired on the certain value of U random to initiate the fuzzy c-means algorithm. In the future, optimization process to determine the initial value for acquiring the best cluster centroid can be developed to further improve the Fuzzy C-Means clustering performance. Table 3 illustrate the valuation of clustering result from the application Fuzzy C-Means Clustering Core and Reduct with 7 different distance function parameter. Based on that result, Euclidean distance still considered as the best distance dunction to be applied.

Table 3. Rank-based Valuation of FCM-CR

Distance function	Number of Iteration	Computing Time	Objective Function	Accuracy	SCS	DBI	Purity	Total
Euclidean	7	7	2	6	7	7	4	45
Manhattan	2	5	1	7	4	5	5	35
Minkowski	3	2	3	5	6	2	6	31
Chebisev	1	4	5	1	1	3	1	17
Minkowski-Chebisev	1	1	4	2	3	4	2	19
Canberra	2	6	7	3	2	6	3	32
Average	1	3	6	4	5	1	7	34

4. CONCLUSION

The Core and Reduct dimension reduction method can reduce the computational burden of the Fuzzy C-Means Clustering method on all distance functions. The value of the objective function can be significantly reduced so that the number of iterations and computation time can also be significantly reduced. These results indicate that the reduction of the Core and Reduct dimensions works consistently on the Fuzzy C-Means clustering method with various distance functions. Even so, the quality of the cluster results from this method can still be maintained. These results are shown in the increase in the Silhouette Coefficient Score, the decrease in DBI, and the accuracy and purity values which are still above 80%. Euclidean distance is the best distance with the result of the number of iterations, computation time, the best Silhouette Coefficient Score. The Fuzzy C-Means Clustering method with the reduction of Core and Reduct dimensions is not recommended for the Minkowski-Chebisev distance function. In the future, research on the development of the Fuzzy C-Means Clustering Core and Reduct method can be applied to image data, video or other data types.

5. REFERENCES

[1] B. Marr, *Big Data In Practice*, 1st ed., vol. 1, no. 1. West Sussex: Wiley, 2016.
[2] Y. Riahi and S. Riahi, "Big Data and Big Data Analytics : Concepts , Types and Technologies Big Data and Big Data Analytics : Concepts , Types and Technologies," *Int. J. Res. Eng.*, vol. 5, no. 9, pp. 524–528, 2018, doi: 10.21276/ijre.2018.5.9.5.
[3] J. Eliyanto, Sugiyarto, Suparman, I. Djakaria, and M. A. H. Ruhama, "Dimension reduction using core and reduct to improve fuzzy C-means clustering performance," *Technol. Reports Kansai Univ.*, vol. 62, no. 6, pp. 2855–2867, 2020, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089822352&partnerID=40&md5=16510c585d8963d28d97b260acec60bd.
[4] W. Purba, S. Tamba, and J. Saragih, "The effect of mining data k-means clustering toward students profile model drop out potential," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, pp. 0–6, 2018, doi: 10.1088/1742-6596/1007/1/012049.
[5] D. P. Ismi, S. Panchoo, and Murinto, "K-means clustering based filter feature selection on high dimensional data," *Int. J. Adv. Intell. Informatics*, vol. 2, no. 1, pp. 38–45, 2016, doi: 10.26555/ijain.v2i1.54.
[6] E. Hardika and S. Atmaja, "Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta," *Int. J. Appl. Sci. Smart Technol.*, vol. 1, no. 1, pp. 33–44, 2019.
[7] K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, "Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, pp. 2760–2770, 2019, doi: 10.11591/ijece.v9i4.pp2760-2770.
[8] A. Gosain and S. Dahiya, "Performance Analysis of Various Fuzzy Clustering Algorithms: A Review," *Procedia Comput. Sci.*, vol. 79, pp. 100–111, 2016, doi: 10.1016/j.procs.2016.03.014.
[9] A. A. hussian Hassan, W. M. Shah, M. F. I. Othman, and H. A. H. Hassan, "Evaluate the performance of K-Means and the fuzzy C-Means algorithms to formation balanced clusters in wireless sensor networks," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 1515–1523, 2020, doi: 10.11591/ijece.v10i2.pp1515-1523.
[10] A. Nurzahputra, M. A. Muslim, and R. Kurniawan, "Online Fuzzy C-Means Clustering for Lecturer Performance Assessment Based on National and International Journal Publication," in *International Conference on Mathematics, Science, and Education*. 2016.
[11] S. Kapil and M. Chawla, "Performance Evaluation ofK-means Clustering Algorithm with Various Distance Metrics," *1st IEEE Int. Conf. Power Electron. Intell. Control Energy Syst.*, pp. 1–4, 2016, doi: 10.5120/19360-0929.
[12] M. K. Arzoo, A. Prof, and K. Rathod, "K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8 . 2," *Int. Res. J. Eng. Technol.*, vol. 4, no. 4, pp. 2363–2368, 2017, [Online]. Available: https://www.irjet.net/archives/V4/i4/IRJET-V4I4484.pdf.
[13] B. Charulatha, P. Rodrigues, T. Chitralekha, and A. Rajaraman, "A Comparative study of different distance

metrics that can be used in Fuzzy Clustering Algorithms," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 2013, pp. 2–5, 2013, [Online]. Available: http://www.ijettcs.org/NCASG-2013/NCASG 38.pdf.
[14] A. Singh, A. Yadav, and A. Rana, "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013, doi: 10.5120/11430-6785.
[15] P. Grabusts, "The choice of metrics for clustering algorithms," *Vide. Tehnol. Resur. - Environ. Technol. Resour.*, vol. 2, no. 1, pp. 70–78, 2011, doi: 10.17770/etr2011vol2.973.
[16] Mahatme and Boyar, "Impact Of Distance Metrics On The Performace Of K-Means And Fuzzy C-means Clustering - An Approach To Assess Student's performance In E-Learning Environtment," vol. 9, no. 1, pp. 888–892, 2018.
[17] S. Surono and R. D. A. Putri, "Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis," *Int. J. Fuzzy Syst.*, vol. 23, no. 1, pp. 139–144, 2020, doi: 10.1007/s40815-020-00997-5.
[18] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Ying Wah, "A Comparison study on similarity and dissimilarity measures in clustering continuous data," *PLoS One*, vol. 10, no. 12, pp. 1–20, 2015, doi: 10.1371/journal.pone.0144059.
[19] B. R. A. Moreira *et al.*, "Classifying Hybrids of Energy Cane for Production of Bioethanol and Cogeneration of Biomass-Based Electricity by Principal Component Analysis-Linked Fuzzy C-Means Clustering Algorithm," *J. Agric. Sci.*, vol. 11, no. 14, p. 246, 2019, doi: 10.5539/jas.v11n14p246.
[20] M. M. Deris, N. Senan, Z. Abdullah, R. Mamat, and B. Handaga, "Dimensional reduction using conditional entropy for incomplete information systems," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11657 LNCS, pp. 263–272, 2019, doi: 10.1007/978-3-030-25636-4_21.
[21] R. Zhao, L. Gu, and X. Zhu, "Combining fuzzy C-means clustering with fuzzy rough feature selection," *Appl. Sci.*, vol. 9, no. 4, 2019, doi: 10.3390/app9040679.
[22] D. Dua and C. Graff, "UCI Machine Learning Repository." University of California, School of Information and Computer Science., Irvine, 2019, [Online]. Available: http://archive.ics.uci.edu/ml.
[23] E. O. Rodrigues, "Combining Minkowski and Chebyshev: New Distances Proposal and Survey of Distances Metrics Using K-Nearest Neighbours Classifier," *Elsevier*, 2018, [Online]. Available: https://doi.org/10.1016/j.patrec.2018.03.021.
[24] F. Wang, H. H. Franco-Penya, J. D. Kelleher, J. Pugh, and R. Ross, "An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10358 LNAI, no. December 2018, pp. 291–305, 2017, doi: 10.1007/978-3-319-62416-7_21.
[25] M. F. Dzulkalnine and R. Sallehuddin, "Missing data imputation with fuzzy feature selection for diabetes dataset," *SN Appl. Sci.*, vol. 1, no. 4, 2019, doi: 10.1007/s42452-019-0383-x.
[26] M. Sammany and T. Medhat, "Dimensionality reduction using rough set approach for two neural networks-based applications," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4585 LNAI, pp. 639–647, 2007, doi: 10.1007/978-3-540-73451-2_67.

BIOGRAPHY OF AUTHORS (10 PT)

About the Author.
Joko Eliyanto
Joko Eliyanto is 24 Instructor & Researcher at the Ahmad Dahlan University Center for Data Science Studies. He received his bachelor's degree in Mathematics and master's degree in the Mathematics Education from Ahmad Dahlan University.
He is interested in Data Science, Machine Learning, and Artificial Intelligence.
He get TensorFlow Developer Certification since 2020.
(*first author*)

About the Author.-
Sugiyarto Surono completed his undergraduate education in Mathematics at Gadjah Mada University and completed his master's degree in Mathematics and Statistics at the same university. He has completed his doctorate degree at Universiti Teknologi Malaysia, majoring in optimization. Currently he is working as a lecturer at the Faculty of Applied Science and Technology at Ahmad Dahlan University while doing some research in fuzzy learning for machine learning and deep learning.

HASIL CEK_PAK SUGIYARTO_1

ORIGINALITY REPORT

20%	9%	16%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.journal.uad.ac.id Internet Source	4%
2	Sugiyarto Surono, Rizki Desia Arindra Putri. "Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis", International Journal of Fuzzy Systems, 2020 Publication	3%
3	Dian Sa'adillah Maylawati, Tedi Priatna, Hamdan Sugilar, Muhammad Ali Ramdhani. "Data science for digital culture improvement in higher education using K-means clustering and text analytics", International Journal of Electrical and Computer Engineering (IJECE), 2020 Publication	1%
4	Submitted to University of Southampton Student Paper	1%
5	eprints.uad.ac.id Internet Source	1%

6	Submitted to Madanapalle Institute of Technology and Science Student Paper	1 %
7	"Rough Sets and Knowledge Technology", Springer Science and Business Media LLC, 2014 Publication	1 %
8	Submitted to The Hong Kong Polytechnic University Student Paper	1 %
9	Submitted to President University Student Paper	1 %
10	Submitted to VIT University Student Paper	<1 %
11	Hamid Karimi, Shahram Jadid, Hedayat Saboori. "Multi-objective bi-level optimisation to design real-time pricing for demand response programs in retail markets", IET Generation, Transmission & Distribution, 2019 Publication	<1 %
12	Annisa Eka Haryati, Sugiyarto. "Clustering with Principal Component Analysis and Fuzzy Subtractive Clustering Using Membership Function Exponential and Hamming Distance", IOP Conference Series: Materials Science and Engineering, 2021 Publication	<1 %

13	Yu Fang, Fan Min. "Cost-sensitive approximate attribute reduction with three-way decisions", International Journal of Approximate Reasoning, 2019 Publication	<1 %
14	Submitted to 54339 Student Paper	<1 %
15	dokumen.pub Internet Source	<1 %
16	journals.plos.org Internet Source	<1 %
17	www.mdpi.com Internet Source	<1 %
18	en.wikipedia.org Internet Source	<1 %
19	Submitted to iGroup Student Paper	<1 %
20	Communications in Computer and Information Science, 2016. Publication	<1 %
21	Submitted to Universiti Putra Malaysia Student Paper	<1 %
22	Submitted to Chandigarh University Student Paper	<1 %

23	V. Torra. "Fuzzy c-means for Fuzzy Hierarchical Clustering", The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ '05., 2005 Publication	<1 %
24	dpj.pitt.edu Internet Source	<1 %
25	Linzi Yin, Zhaohui Jiang. "A Fast Attribute Reduction Algorithm Based on a Positive Region Sort Ascending Decision Table", Symmetry, 2020 Publication	<1 %
26	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1 %
27	Jiancong Fan. "OPE-HCA: an optimal probabilistic estimation approach for hierarchical clustering algorithm", Neural Computing and Applications, 2015 Publication	<1 %
28	docplayer.net Internet Source	<1 %
29	"Intelligent Computing Methodologies", Springer Science and Business Media LLC, 2017 Publication	<1 %

30	Submitted to Bolton Institute of Higher Education Student Paper	<1%
31	Guohua Lv. "ℓmp: A Novel Similarity Measure for Matching Local Image Descriptors", IEEE Access, 2018 Publication	<1%
32	ceur-ws.org Internet Source	<1%
33	"Recent Trends in Information and Communication Technology", Springer Science and Business Media LLC, 2018 Publication	<1%
34	Cong Gao, Yiyu Yao. "Chapter 49 An Addition Strategy for Reduct Construction", Springer Science and Business Media LLC, 2014 Publication	<1%
35	D Juniati, I K Budayasa. "The mathematics anxiety: Do prospective math teachers also experience it?", Journal of Physics: Conference Series, 2020 Publication	<1%
36	Huang, S.C.. "A case study of applying data mining techniques in an outfitter's customer value analysis", Expert Systems With Applications, 200904 Publication	<1%

37	M Faisal, E M Zamzami, Sutarman. "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance", Journal of Physics: Conference Series, 2020 Publication	<1 %
38	Mohamed Lafif Tej, Stefan Holban. "Comparative Study of Clustering Distance Measures to Determine Neural Network Architectures", 2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2018 Publication	<1 %
39	"Global Trends in Information Systems and Software Applications", Springer Science and Business Media LLC, 2012 Publication	<1 %
40	epdf.pub Internet Source	<1 %
41	hdl.handle.net Internet Source	<1 %
42	theses.gla.ac.uk Internet Source	<1 %
43	worldwidescience.org Internet Source	<1 %
44	zdoc.site Internet Source	

<1%

45 Anca Loredana Ion, Liana Stanescu, Dan Burdescu, Stefan Udristoiu. "Improving an image retrieval system by integrating semantic features", 2008 Conference on Human System Interactions, 2008
Publication

<1%

46 Shirkhorshidi, Ali Seyed, Saeed Aghabozorgi, and Teh Ying Wah. "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data", PLoS ONE, 2015.
Publication

<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		