Fuzzy c-shells clustering algorithm

N B I Pratiwi and D R S Saputro
Universitas Sebelas Maret, Department of Mathematics, Faculty of Mathematics and Natural Sciences.
Jl. Ir. Sutami 36A, Surakarta, 57126, Central Java, Indonesia.
E-mail: nafisaberlianaindahpratiwi@gmail.com

Abstract. Clustering is a process to classify data into some clusters or classes so that data in the same cluster have maximum similarity and data between clusters have minimum similarity. Generally, clustering is separated into two methods, hierarchical and non-hierarchical methods. Non-hierarchical methods are separated as soft clustering and hard clustering. Non-hierarchical methods that mostly used is K-Means algorithm. K-Means grouping object into certain member and not included in different clusters. Another approaches in clustering technique which is based on fuzzy sets theory is known as fuzzy clustering. Fuzzy clustering is one kind of soft clustering that also widely used because of many advantages than the hard clustering. Several varieties of fuzzy clustering are Fuzzy C-Means (FCM) and Fuzzy C-Shells (FCS). FCM is characterized by centre of cluster and FCS uses radius as additional parameter. This paper concerns FCS clustering algorithm using review of several references. Hence, FCS is characterized by centre of cluster and radius.

1. Introduction
Data mining is a logic process which is used in big data to find many useful information. The purpose of data mining process is to obtain patterns that previously unknown. Data patterns that have been known can be used as the base in next analysis, decision making, and to discover many kinds of object and the phenomenon. Learning sets are included in data mining are supervised learning and unsupervised learning [1]. Unsupervised model is kind of important process in data mining which used in research by grouping data into some categories or clusters based on statistical properties or well known as clustering analysis.

Clustering analysis is multivariate technique aims to grouping objects based on themselves characteristics. Clustering is a process grouping object such that every object which have closest characteristics with the other objects will be placed in one common cluster. Cluster that is formed has high internal homogeneities and high external homogeneities [2]. There are three types of clustering technique, hierarchical clustering, non-hierarchical (partitional) clustering, and hybrid clustering (composite between hierarchical and non-hierarchical clustering). Hierarchical clustering technique is including single linkage, complete linkage, average linkage, median, and ward. Non-hierarchical clustering or also well known as partitioning clustering includes K-Means, Adaptive K-Means, K-Medoids, and fuzzy clustering [3]. This research discusses the non-hierarchical (partitional) clustering, Fuzzy C-shells (FCS). This method is suitable for big data and have continuous variable.

In every clustering process either hierarchical or non-hierarchical, the cluster is formed such that every object is placed exactly in one cluster. There is an exception, case as mentioned previously.
cannot be proceed to place an object exactly in a cluster because that object is located between two or more another clusters so that it is necessary to use fuzzy clustering technique which is considering the membership degree fuzzy set as basic weighting [4]. Clustering with fuzzy is keep on developing because generally, data cannot be separated strictly into a cluster, but has tendency as membership degree that has value in interval 0 and 1 [5,6].

The classification of clustering algorithm is divided into two types, hierarchical and non-hierarchical clustering as mentioned before. Hierarchical clustering forms tree structure called dendrogram that represents data relation. Tree structure consists of number of clusters that include all object and for each two clusters are independent [7]. Otherwise, non-hierarchical clustering or partitional clustering is characterized with numbers of clusters K which is an input parameter [8]. That method relocates data points from one cluster to others cluster that started by initial cluster configuration. Classic clustering algorithm or well known as hard / crisp clustering has characteristic each data placed in one cluster, but possibly uncertainties or fuzziness hence there is a chance that datum placed in one cluster or more in similar period then hard/crisp clustering has developed and the resulting the approximation based on fuzzy named fuzzy clustering.

Clustering with Fuzzy C-Means (FCM) is an example of clustering technique which is development of non-hierarchical / partitional clustering with fuzzy concept. FCM is first recognized by Dunn on 1973 then FCM is used in pattern recognition field [9]. FCM method is often used in clustering because this method gives smooth results and sufficiently effective to increase homogeneity in each cluster result. Matters that occurs in the use of FCM are selecting the number of clusters and reducing sensitivity to the initial value [10]. The next development of Fuzzy clustering algorithm is Fuzzy C-Shells (FCS). Both methods, FCM also FCS, has difference in the algorithm. FCM is characterized by the centre of cluster, otherwise FC is characterized by centre of cluster and its radius. Fuzzy clustering is beneficial when clusters aren’t well separated and the boundary is ambiguous [8,10,12]. Hence, in this article concerns the FCS algorithm. This article discusses the FCS algorithm which observes radius and centre of clusters. The FCS algorithm is the development of FCM algorithm which only note the centre of cluster. The first paragraph after a heading is not indented (Bodytext style).

2. Research Method
This research is a theoretical research conducted by analysing several related literatures such as clustering technique, the development of clustering technique, fuzzy clustering, fuzzy shells clustering and fuzzy c-shells clustering. The results of this research were linked and ordered to disclose the relations among the theories. The procedures that applicable on this research are: 1) conduct theoretical study of some related materials, 2) scrutinize the results of theoretical studies and analysing the relation among fuzzy clustering and fuzzy c-shells clustering by observing its behaviour, and the effectiveness in the algorithm, 3) conduct the conclusion.

3. Research Method
3.1. Clustering
Clustering is unsupervised learning because data that is used is not labelled. Clustering technique is often applied in feature space [13]. Clustering is process to cluster data based on information that was found on data which is describe an object [14]. Clustering algorithm is used to recognize compact and not compact surface in object space then find the entirety pattern distribution also the correlation between attributes [15]. Clustering algorithm aims to split dataset so then object which is settle in same cluster has maximum characteristic similarity and object that settles in another cluster and different cluster has minimum characteristic similarity.

Clustering is achieved by three purpose that are description classification, data simplify and relation identification. Homogeneity concept is the most basic thing in clustering. There are three methods that used to determine the homogeneity such that, correlation measurement, distance measurement,
association measurement. Distance measurement is measurement which is most used that is applied to data with metric scale, as Euclidean distance. Substantial distance shows slightly characteristic, on the contrary, concise distance shows that the object similar with another object. A cluster is identified based on centre cluster or centroid [16].

Distance measurement that commonly used is Euclidean distance which is written as

$$d(z_u, z_w) = \sqrt{\sum_{j=1}^{Nd} (z_{u,j} - z_{w,j})^2} = \| z_u - z_w \|$$

Where $z_{u,j}$ is $u^{th}$ data on variable $j$, $z_{w,j}$ is $w^{th}$ data on variable $j$.

The results of the $d(z_u, z_w)$ represents the similarity of each attribute, when the distance becomes smaller then the similarity is greater [17].

Clustering algorithm generally divided based on two technique that are most used, hierarchical clustering and non-hierarchical clustering or well known as partitional clustering. Hierarchical clustering resulting cluster tree (dendrogram) using heuristic separation or merging technique [16]. Cluster tree defined as tree which shows sequence from clustering process with every cluster becomes a partition in dataset. Clustering algorithm uses separation technique to results cluster tree called as divisive. Popular clustering algorithm based on merging clustering to results cluster tree is well known as agglomerative. Non-hierarchical clustering or partitional clustering divides the dataset into specific cluster. Partitional clustering divides $n$ objects into $c$ cluster [18]. Each cluster are represented by centroid which is description wrap up from all objects that includes in a cluster [19]. A centroid is vectors contain a number to every variable with every numbers are mean of a variable to observe in a cluster. Centroid can be recognized as mean of multi-dimensional of a cluster. If it was obtained to have large numbers of clusters then the centroid could be clustered and resulting hierarchical clustering on dataset.

Non-hierarchical clustering or partitional clustering is branching into two types, hard clustering and soft clustering [20]. Classic clustering or generally called as hard/crisp clustering is clustering algorithm that every data point originated from only one cluster. This case needs lots of previous information related to the cluster. Hard clustering grouping data with partition series from a cluster to the other cluster includes all members [8]. Hard clustering algorithm involves agglomerative and divisive clustering. Both methods organize data based on nearby distance matrix. Classic clustering is not suitable to be used on data that has not countable and limited circumscription between clusters [18]. The deficiency of using hard / crisp clustering is the outcome cluster sensitive with initial cluster centroid and possibly convergent to local optima. Soft clustering enables data to join a cluster or more than one cluster. On soft clustering case, fuzzy sets is used for clustering data hence data will be grouped based on corresponding membership value [21].

3.2. Fuzzy Clustering

A datum might contain uncertainties or fuzziness, that state will make obstacles in determining cluster if uses the hard/crisp clustering that has characterized each data belongs to exactly one cluster. Fuzzy clustering analysis let the gradual membership from data to cluster measured as degree in $[0,1]$ [22]. That circumstances make flexibility data included into one cluster or many clusters. Fuzzy membership function has proved based on definition and interpretation of fuzzy set. Given sets of objects $O_1, O_2, ..., O_n$ on fuzzy clustering with $k$-fuzzy cluster $C_1, C_2, C_3, ..., C_k$ represented by partition matrix written as

$$M = [w_{ij}](1 \leq i \leq n, 1 \leq j \leq k)$$
where \( w_{ij} \) is membership degree from \( 0_i \) in fuzzy cluster \( C_j \). Fuzzy clustering is either methods to determine optimum cluster in vector space based on normal Euclidean form for a distance between vectors [23] that aims to cluster \( n \) object that presented as vector into \( c \) cluster based on similarity with centre cluster that measured through distance function. The advantages fuzzy clustering rather than hard/crisp is having lesser tendency to get trapped in local minimum. Membership function asserts ambiguous or not ambiguous a datum when it settles on a cluster. Because of the membership on clusters tend to be fuzzy so then there is no label indicates data will be placed on certain cluster. Data points will be placed on most similar cluster. Fuzzy clustering prevailing data points will be settled in a cluster based on membership degree which depends on shortest distance to the centre of cluster. The development of fuzzy clustering algorithm based on different principles, as fuzzy relation, objective function, etc. Fuzzy clustering algorithm widely used is based on objective function.

Reference [24], fuzzy clustering algorithm purposes to minimization objective function weighted distance from data to cluster. Objective function has mentioned about is written as

\[
J(X, U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m d^2(v_i, x_k)
\]

with the term

\[
\sum_{k=1}^{n} u_{ik} > 0 \text{ for all } i \in \{1, 2, ..., c\}
\]

\[
\sum_{i=1}^{c} u_{ik} = 1 \text{ for all } k \in \{1, 2, ..., n\}
\]

fuzzy c-shells functional is defined by \( X = \{x_1, ..., x_n\} \subseteq \mathbb{R}^p \) is a dataset, and numbers of fuzzy clusters is \( c \), membership degree of data \( x_k \) to cluster \( i \). A cluster denoted by \( u_{ik} \in [0,1] \). \( v_i \) is a prototype or vector parameter to cluster \( i \), and \( d(v_i, x_k) \) is distance between prototype \( v_i \) to datum \( x_k \) [25]. Parameter \( m > 1 \) determined as fuzziness index. According to Pal [26] best value of \( m \) which can be chosen on the interval \([1.5-2.5]\) then based on the interval give results the mean and the midpoint which is equal to 2. for \( m = 2 \) is value that generally used. For value of \( m \to 1 \), cluster tend to crisp. The lack of hard/crisp clustering cannot declare the difference between data points located in centre cluster and data points on the boundary of the cluster [27].

3.3. Fuzzy C-Shells Clustering

Fuzzy shells clustering algorithm (FSC) has many similarities with Fuzzy C-Means algorithm (FCM), both algorithms have same \( v_i \) in some cases [25]. FSC adds new parameter to the algorithm and becomes the difference between FCM. The behaviour of FSC is each data point, including data points which indicated as noise data points must collected in functional minimization. Fuzzy c-shells algorithm behaves differently, the algorithm does not notice the noisy data points on the process, this technique simply aims to cluster data points. Fuzzy c-shells clustering algorithm is clustering algorithm that utilize prototypes in circle and ellipse shape. Fuzzy c-shells algorithm (FCS) mentioned as fuzzy clustering completed with geometry attribute. FCS algorithm is the first clustering algorithm which uses concept of nonlinear prototypes In FCM algorithm [28]. FCS detects structure of different clusters with empty interior or annulus on data in 2 dimensions, “shells” is a curved surface that used instead of common prototype such as point, line, plane and hyperplane on traditional fuzzy algorithm. Figures which represents shells shown in Figure 1, the source of Figure 1 is [9].
FCS is the expansion of fuzzy c-means algorithm (FCM) with additional parameter, radius. FCS is characterized with centre of cluster and radius. FCS algorithm has basic form hyper-spherical shells with $p$ dimensions which the characteristic is centre and radius. By using norm that resulting symmetric positive matrix and finite. On FCS algorithm was obtained shells with ellipsoidal structure as cluster prototype as shown on Figure 2.

**Figure 2.** Shells distance geometrically where $D_{ik}$ is distance between prototype shells boundaries and datum $x_k$.

Given $\mathfrak{R}$ as sets of real numbers dan $\mathfrak{R}^p$ is sets of $p$-tuples resulting basic form hyper-spherical shells on center $v_i, v \in \mathfrak{R}^p$ and radius $r$. Let $R$ is $c$-tuple and $R = \{r_1, r_2, ..., r_n\} \in \mathbb{R}^+$ is obtained a set of basic form of hyper-spherical shells written as

$$\{x \in \mathbb{R}^p : (x - v)^T I (x - v) = r^2\}$$

with $I$ is $p \times p$ identity matrix. Distance between point $x_k \in \mathbb{R}^p$ with hyper-spherical shells prototype is

$$(D_{ik})^2 = (\| x_k - v_i \| - r_i)^2$$

where

$$r_i = \frac{\sum_{k=1}^{n}(u_{ik})^m \| x_k - v_i \|}{\sum_{k=1}^{n}(u_{ik})^m}$$

and

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}.$$

Let $U \in M_{fc}$ is fuzzy $c$-partition of $X$ ($M_{fc}$ is fuzzy $c$-partition set of $X$) and yields the function on fuzzy $c$-shells.
\[ J: \mathbb{M}_{fc} \times \mathbb{R}^c \times \mathbb{R}^c \rightarrow \mathbb{R}^+ \]

Hence, objective function on FCS is written

\[ J(U,V,R) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m (D_{ik})^2 \]

Where \( m \) is a fuzziness index with \( m \in [1, \infty) \), \( u_{ik} \) is the membership degree and \( d^2(v_i, x_k) \) is norm of Euclidean distance.

Fuzzy c-shells algorithm is explained as follows:

4.1 Define numbers of clusters, \( 2 \leq c < n \), with \( n \) is numbers of data points.
4.2 Define the exponent (weight) = \( m \), \( 1 < m < \infty \).
4.3 Set the iteration counter \( j = 0 \), Initializing the fuzzy c-partition \( U^0 \), where \( U \) is a matrix

\[ U = \begin{pmatrix} \mu_{11}(x_1) & \ldots & \mu_{1n}(x_n) \\ \vdots & \ddots & \vdots \\ \mu_{c1}(x_1) & \ldots & \mu_{cn}(x_n) \end{pmatrix} \]

Where initial matrix partition is chosen randomly with \( \mu \) is value based on row and a column, also applicable to \( x \).
4.4 Calculate the centre of cluster \( (v_i) \) and the radius of cluster \( (r_i) \) with equation below

\[ v_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m}, \quad r_i = \frac{\sum_{k=1}^{n} (u_{ik})^m \| x_k - v_i \|}{\sum_{k=1}^{n} (u_{ik})^m} \]

4.5 Calculate distance \( (D_{ik})^2 \). \( D_{ik} \) is distance between \( k \)-feature vector, \( x_k \) and the \( i \)-prototype defined as

\[ (D_{ik})^2 = (\| x_k - v_t \| - r_t)^2 \]

Where \( \| . \| \) is Euclidean distance, centre \( v_t \) and radius \( r_t \) from basic form shell cluster.

Update membership iteration at \( j \), \( U \) by the equation

\[ u_{ik} = \frac{1}{\sum_{j=1}^{n} \left[ \frac{D_{ik}^2}{D_{jk}^2 \| \|} \right]^{1/m-1}} \]

4.6 Comparing the \( U^j \) and \( U^{j-1} \). If \( |U^j - U^{j-1}| \leq \varepsilon \), then the process stops.

Otherwise \( |U^j - U^{j-1}| > \varepsilon \), set \( j = j + 1 \) and turn into step 3.

4. Conclusion

This paper has discussed the clustering algorithm with approximation based on fuzzy concept which grant simplicity in process of construct clusters on datasets. Fuzzy clustering algorithm based on objective function well known as fuzzy c-means clustering is fuzzy clustering algorithm most widely used then developed to utilize prototypes with geometry structure such as curved surface (circles, ellipses, etc), Fuzzy C-Shells (FCS). FCS is equipped with radius and centre of cluster. FCS algorithm
is the first algorithm which is utilizing the non-linear prototype from FCM algorithm and as reference to develop new algorithm with another varieties of prototypes. FCS algorithm is often used in detects nonlinear space on image segmentations then can be used to clusters dataset with better accuracy because of the additional parameter.

Acknowledgements
We would like to acknowledge Universitas Sebelas Maret which provide authors chance to present this paper and thanks a lot to all committee of AD INTERCOMME 2019.

References
[1] Chitra K and Maheswari D 2017 A Comparative Study of Various Clustering Algorithms in Data Mining. International Journal of Computer Science and Mobile Computing 6(8) pp. 109-115
[2] Sofyan H, Said M A, Affan M and Bawahidi K 2005 The Application of Fuzzy Clustering to Satellite Images Data. Proc. of the 6th WSEAS Int. Conf. on Fuzzy Systems pp. 100-103
[3] Oyelade O J, Oladipupo O O, and Obagbuwa I C 2010 Application of k-means clustering algorithm for prediction of Students’ Academic performance. International Journal of Computer Science and Information Security 7(1) pp. 292-95
[4] Abonyi J and Szeifert 2002 Supervised Fuzzy Clustering for the Identification of Fuzzy Classifiers Journal Elsevier 24 pp. 2195-207
[5] Hoppen F and Klawonn F 2004 Learning Fuzzy Systems – An Objective Function-Approach, Mathware & Soft Computing pp. 11143-62
[6] Dunn J 1973 A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact well-Separated Cluster Jurnal of Cybernetic 3 32-7
[7] Hartigan J A 1937 Clustering Algorithm (USA : John Wiley and Sons, Inc., Publication)
[8] Tagarelli A 2012 XML Data Mining : Model, Methods and Applications. (USA : Information Science Reference)
[9] Bezdek J C 1981 Pattern Recognition with Fuzzy Objective Function Algorithms (New York: Plenum Press)
[10] Duan L, Yu F, and Zhan L. An Improved Fuzzy C-means Clustering Algorithm 2016. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery
[11] Xu R and Donald C W 2009 Clustering (USA : John Wiley and Sons, Inc.)
[12] Hansen P, Jaumard B 1997 Cluster Analysis and Mathematical Programming Mathematical Programming 79(1) pp. 191-21
[13] Dave R N 1992 Generalized Fuzzy C-Shells Clustering and Detection of Circular and Elliptical Boundaries. Journal Pergamon Pattern Recognition 25(7) pp. 713-21
[14] Tan P N, Steinbach M and Kumar V 2006 Introduction to Data Mining (Boston:Pearson Education)
[15] Bharati M R 2006 Data Mining Techniques and Applications. Modern Institute of Information Technology and Research 1(4) pp. 301 – 05
[16] Omran, Mahamed G H, Engelbrecht P A and Ayed S 2007 An Overview of Clustering Methods. Intelligent Data Analysis 11 (2007) pp. 583-605
[17] Arora J, Khatter K and Tushir M. 2019. Fuzzy c-Means Clustering Strategies: A Review of Distance Measures. Software Engineering. Advances in Intelligent Systems and Computing 731 pp. 153-162
[18] Jafar, Mohamed O A and Sivakumar R 2013 A Comparative Study of Hard and Fuzzy Data Clustering Algorithms With Cluster Validity Indices. ERCICA
[19] Rai P and Subha S 2010 A Survey of Clustering Techniques. International Journal of Computer Applications (0975 – 8887) 7(12)
[20] Grover N 2014 A study of various Fuzzy Clustering Algorithms. International Journal of 
Engineering Research 3(3) pp. 177-81
[21] Bora J D, Gupta and Anil Kumar 2014 A Comparative Study Between Fuzzy Clustering 
Algorithm and Hard Clustering Algorithm IJCTT
[22] Abonyi, Benjamin and Mark 2007 Advance in Fuzzy Clustering and Its Application (USA : 
John Wiley and Sons, Inc., Publication)
[23] Ravi V, Srinivas E R and Kasabov NK 2007 On-Line Evolving Fuzzy Clustering. IEEE, 
International Conference on Computational Intelligence and Multimedia Application 347-51
[24] Klawonn F and Annette K 1999 Fuzzy Clustering Based on Modified Distance Measures 
International Symposium on Intelligent Data Analysis IDA Advances in Intelligent Data Analysis pp. 291-301
[25] Dave R N 1990 Fuzzy Shells Clustering and Applications to Circle Detection in Digital Images. 
International Journal of General Systems. 16(4) pp. 343-55
[26] Pal N R and Bezdek J C 1995 On Cluster Validity for the Fuzzy c-Means Model. IEEE 
Transactions On Fuzzy Systems 3(3)
[27] Kruse R, Christian D and Marie J L 2007 Advances in Fuzzy Clustering and its Applications 
(England : A John Wiley and Sons, Inc., Publication)
[28] Dave R N 1992 Boundary Detection through Fuzzy Clustering Proceedings FUZZ-IEEE 
International Conference on Fuzzy System pp. 127-34