JOIV : Int. J. Inform. Visualization, 5(1) - March 2021 11-15

INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

A Framework of Mutual Information Kullback-Leibler Divergence based for Clustering Categorical Data

Iwan Tri Riyadi Yanto[a], Ririn Setiyowati[b,*], Nur Azizah[c] , Rasyidah[d]

[a]*Department of Information System, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[b]*Department of Mathematics, Universitas Sebelas Maret, Surakarta, Indonesia*
[c]*Department of Mathematics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[d]*Department of Information Technology, Politeknik Negeri Padang, Indonesia*

*Corresponding author: *ririnsetiyowati@staff.uns.ac.id*

***Abstract*— Clustering is a process of grouping a set of objects into multiple clusters, so that the collection of similar objects will be grouped into the same cluster and dissimilar objects will be grouped into other clusters. Fuzzy k-means Algorithm is one of clustering algorithm by partitioning data into k clusters employing Euclidean distance as a distance function. This research discusses clustering categorical data using Fuzzy k-Means Kullback-Leibler Divergence. In the determination of the distance between data and center of cluster uses mutual information known as Kullback-Leibler Divergence distance between the joint distribution and the product distribution from two marginal distributions. Extensive theoretical analysis was performed to show the effectiveness of the proposed method. Moreover, the proposed method's comparison results with Fuzzy Centroid and Fuzzy k-Partition approaches in terms of response time and clustering accuracy were also performed employing several datasets from UCI Machine Learning. The experiment results show that the proposed Algorithm provides good results both from clustering quality and accuracy for clustering categorical data as compared to Fuzzy Centroid and Fuzzy k-Partition.**

***Keywords*— Kullback-Leibler divergence; mutual information; fuzzy k-means; categorical data; clustering.**

Manuscript received 20 Oct. 2020; revised 20 Nov. 2020; accepted 16 Dec. 2020. Date of publication 31 Mar. 2021.

I. INTRODUCTION

Clustering is a method used in data mining to group objects into several groups or clusters based on information obtained from data that explains the relationships between objects. This clustering aims to make the objects between clusters have a minimum similarity and the objects in one cluster have a maximum level of similarity. Clustering in data mining is useful for finding distribution patterns within a data set that is used for the data analysis process. The similarity of objects is usually derived from the proximity of attribute values that describe objects. In a multidimensional space, objects are usually represented as a point. Clustering is a data segmentation method that has been implemented in various fields such as prediction and business problem analysis of market segmentation, marketing, zoning area to the identification of objects and patterns recognition in the field of computer vision and image processing.

Currently, many algorithms have been developed to cluster the data [2]–[5]. The k-Means Algorithm is one of the most popular among clustering algorithms, and it is still developed today. Researchers still develop this clustering algorithm for grouping large data sets based on their effectiveness and efficiency [6]. One of the k-means clustering algorithm extensions is Fuzzy k-Means clustering proposed by Bezdek [7]. Each object or point in fuzzy clustering has a probability of belonging to each cluster. Unlike in traditional k-means, the probability of each object belongs to only one cluster. The problems where the points are between centers or otherwise ambiguous handled by the fuzzy k-means are done by replacing the distance with probability. In this case, probability can be a function of distance, such as relative probability to the distance inverse.

Although Fuzzy k-means is considered a clustering algorithm with high effectiveness and efficiency, this Algorithm can only be performed on numerical data using distance in determining the center of cluster with each point. Therefore, the distance function should be chosen to determine the center of cluster for categorical data. One of the distance functions used for categorical data is Kullback-Leibler (KL) Divergence [8].

This article suggests a modified Fuzzy k-Means for categorical data clustering based on KL Divergence distance. The distance between the cluster's data and center is determined by using Mutual Information [9], which is KL Divergence distance between the product distribution and the joint distribution from two marginal distributions.

This paper was arranged in the following order. Section II describes Fuzzy k-Means, Entropy, and KL Divergence. Section III explains the proposed method based on KL Divergence to fuzzy k-Means Algorithm. Section IV illustrates the results of the experiment on real world datasets from UCI Machine Learning. Finally, this work is concluded in section V.

II. LITERATURE REVIEW

A. Fuzzy k-Means

The k-Means Algorithm is well known as an efficient algorithm for grouping large data sets [11]. According to Bezdek [7], each pattern in the fuzzy version of the k-Means Algorithm is allowed to have a membership function for all clusters rather than having different memberships on one cluster.

Fuzzy k-Means clustering algorithms group X into k clusters as in the Algorithm [6]. The Algorithm is used to minimize the objective function

$$F(W,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}^{m} d(X_i,Z_l) \qquad (1)$$

by the constrains

$$\sum_{l=1}^{k} w_{li} \text{ for } i=1,\dots,n \text{ with } w_{ki}\in[0,1] \qquad (2)$$

The first step, we formed the Lagrange function L from Eq. (1) and (2). Next, we determine the first derivative of the function L concerning the parameters w_{li}, z_l, λ and equated with 0. So that, we get the following result:

$$w_{li}=\frac{\left(\frac{1}{d(X_i,Z_l)}\right)^{\frac{1}{m-1}}}{\sum_{l=1}^{k}\left(\frac{1}{d(X_i,Z_l)}\right)^{\frac{1}{m-1}}} \qquad (3)$$

$$Z_l=\frac{\sum_{i=1}^{n} w_{li}^{m} X_i}{\sum_{i=1}^{n} w_{li}^{m}} \qquad (4)$$

Thus, we have Fuzzy *k*-Means Algorithm as follows.

Fuzzy *k*-Means Algorithm.

Step 1 : Fix $m\in(1,\infty)$, fix $2\le k\le n$, fix MaxIter and fix $\varepsilon>0$.

Take initials $w_{li}^{(0)}$ and suppose $t=1$.

Step 2 : Calculate $z_l^{(t)}$ with $w_{li}^{(t-1)}$ in equation (3)

Step 3 : Update to $w_{li}^{(t)}$ with $z_l^{(t)}$ in equation (4)

Step 4 : Compute objective function $F(W,Z)^{(t)}$ by (1)

Step 5 : Check the stop condition

IF $\left|w_{li}^{(t)}-w_{li}^{(t-1)}\right|<\varepsilon$, $\left|F(W,Z)^{(t)}-F(W,Z)^{(t-1)}\right|<\varepsilon$ or $t>MaxIter$, THEN Stop.

ELSE $t=t+1$ and return to step 2.

B. Entropy

A single definition is unable to capture an overly broad concept of information fully. However, we can define a quantity from any probability distribution by entropy. Many properties correspond to the intuitive notion of what a size of information derived from entropy. The idea is extended to define reciprocal information, which is a size of the amount of information on one random variable containing another information. Then, entropy becomes self-information from the random variable. Mutual Information is a special case of the more general quantity referred to here as relative entropy. Relative entropy can also be a measure of the distance between two probability distributions [9].

Definition 2.1. Let X be a discrete random variable. Entropy H (X) is defined by

$$H(X)=-\sum_{x\in X} p(x)\log p(x) \qquad (5)$$

Definition 2.2. Pairs of discrete random variables (X,Y) with the joint distribution $p(x,y)$ form the entropy $H\ (x,y)$ which is defined as

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \qquad (6)$$

Relative entropy or Kullback-Leibler Divergence between two probability distributions $p(x)$ and $q(x)$ is defined as

$$D(p||q)=\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \qquad (7)$$

The KL Divergence is a measure of the "distance" between two probability distributions. Since the KL Divergence is asymmetrical and does not follow the triangle's inequality, it is not metric [4].

Suppose that $\{p_1,p_2,\dots,p_n\}$ are sets of discrete probability distributions and $\{\pi_1,\pi_2,\dots,\pi_n\}$ are weights corresponding. Then Jensen-Shannon (JS) Divergence between p_1 and p_2 is written by

$$\begin{aligned} JS_\pi(p_1,p_2) &= \pi_1 D(p_1||\pi_1p_1+\pi_2p_2)+\pi_2 D(p_2||\pi_1p_1+\pi_2p_2) \\ &= H(\pi_1p_1+\pi_2p_2)-\pi_1H(p_1)-\pi_1H(p_1) \end{aligned}$$

with $\pi_1+\pi_2=1$, $\pi_i\ge 0$. It is clear that a measure of the Jensen-Shannon (JS) Divergence is a symmetrical measure in $\{\pi_1,p_1\}$ and $\{\pi_1,p_1\}$ [18]. The distance between a finite number of probability distributions can be measured using the generalization of the JS divergences written in the formula:

$$JS_\pi(\{p_i: 1\le i\le n\})=H\left(\sum_{i=1}^{n}\pi_i p_i\right)-\sum_{i=1}^{n}\pi_i H(p_i) \qquad (8)$$

which is symmetric in the $\{\pi_i,p_i\}$ and $\sum_i \pi_i=1, \pi_i\ge 0$. Thus, based on entropy we can analyze the distance measure ofor categorical data by introducing an important lemma [8] as follows:

Lemma 2.1.

$$\sum_{i=1}^{n}\pi_i D\left(p_i||\sum_{i=1}^{n}\pi_i p_i\right)=H\left(\sum_{i=1}^{n}\pi_i p_i\right)-\sum_{i=1}^{n}\pi_i H(p_i)$$

Definition 2.3. Consider two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass function $p(x)$ and $p(y)$. The Mutual Information $I(X, Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$ [9]:

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \qquad (9)$$
$$= D(p(x,y)||p(x)p(y))$$

C. Kullback-Leibler Divergence

In mathematics, a distance is summarized and abstracted into a metric concept. Kullback-Leibler (KL) Divergence distance is defined for Eq. (7). In most cases, it is easy to see that $D(p||q) \neq D(q||p)$ and $D(p||q) + D(q||r) \geq D(p||r)$, so D is not a metric. Thus, we use definition of mutual information to be presented in the proposition 2.1.

Proposition 2.1. Given data set Q , then Q is partitioned into k clusters. Suppose that random variables X, Y and Z represent the object, the attribute and the cluster, respectively. Suppose that the probability of occurrence of the object x, atrribute y, and cluster z are expressed $p(x), p(y)$ and $p(z)$, respectively. In addition, $n(x, y)$ represents the number of occurences of attribute y in object (x) and $n(x) = \Sigma_y n(x, y)$. Furthermore, we assume that $p(z) = \Sigma_{x\in z} p(x)$. Let $I(X, Y)$ be the mutual information between two random variables X and Y, then

$$I(X,Y) - I(Z - Y) = \sum_{k}\sum\nolimits_{x\in z_k} p(x) D(p(Y|x)||p(Z|z_k))$$

where $p(x) = \frac{n(x)}{\Sigma_x n(x)}$ dan $P(Y|x) = \frac{n(x,y)}{n(x)}$.

We know, $D(p(Y|x)||p(Z|z_k)) = \Sigma_y p(y|x)\log\frac{p(y|x)}{p(Z|z_k)}$

We restate $\Sigma_y p(y|x)\log\frac{p(y|x)}{p(Z|z_k)}$ with D_y to simplify the notation. Furtherrmore, there are four scenarios generated by different combinations of $p(y|x)$ and $p(y|z_k)$ values, namely [8] :

- Scenario 1 : $p(y|x) > 0$ and $p(y|z_k)$. The calculation for D_y is very easy to do. The calculation result is in any real number.
- Scenario 2 : $p(y|x) = 0$ and $p(y|z_k) = 0$. We can simply leave $D_y = 0$ or its equivalent removing this feature..
- Scenario 3 : : $p(y|x) = 0$ and $p(y|z_k) > 0$. In this scenario, "$\log\frac{p(y|x)}{p(Z|z_k)} = \log 0 = -\infty$, which implies that there is an inadequacy in direct computing, but this problem can be solve by applying the L'Hospitals rule, $\log_{x\to 0^+}\log\frac{x}{a} = 0 (a > 0)$. So we can consider $x = p(y|x)$ and $a = p(y|z_k)$ and thus we get $D_y = 0$.
- Scenario 4 : $p(y|x) > 0$ and $p(y|z_k) = 0$. In this scenario, $D_y = +\infty$, which in practise is difficult to handle.

According to Junjie Wu [8], However, the case in scenario 4 is the most difficult case to handle as it is difficult to compute with $+\infty$ in practice. On the other hand, it is clear that the total KL Divergence of $p(Y|x)$ and $p(Y|z_k)$ is infinite if there is some dimension y of scenario 4. This does not work for sparse data because the centroids of such data typically contain many zero-value features. Therefore, assgning instance to centroid is a big challenge for us. This is known as the "zero-value dilemma" [8].

The above problems can be overcome by smoothing sparse data. For example, the entire data set is added with a very small positive value to avoid the zero value of feature [8]. This technique does change the data's scatter property, although this smoothing technique facilitates the calculation of the KL Divergence[8].

III. Proposed method

The Fuzzy k-Means model has been discussed in section II. From the development of Fuzzy k-Means in equations (8) and (9), complex calculations are obtained. Therefore, we propose another model, called Fuzzy k-Means KL Divergence.

Let $\mathbb{Q}$ be a data set. A partition of $\mathbb{Q}$ into k clusters. Suppose that random variables X, Y and Z represent the object, the attribute and the cluster, respectively. Suppose that the probability of occurrence of the object x, atrribute y, and cluster z are expressed $p(x), p(y)$ and $p(z)$. Furthermore, we assume that $p(z) = \Sigma_{x\in z} p(x)$. In addition $n(x, y)$ represents the number of occurrences of attribute y in object x, and $n(x) = \Sigma_y n(x, y)$.

Now, objective function $F_{FKMKL}(W, p(Y|z))$ can be written as follows:

$$F_{FKMKL}(W, p(Y|z)) = \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ki}^{m} p(x_i) D(p(Y|x_i)||p(Y|z_k)) \qquad (10)$$

By the constraint

$$\Sigma_{k=1}^{K} w_{ki} = 1,\ \text{for } i = 1,2,\cdots,n \qquad (11)$$

$$\sum_{y} p(Y|z_k) = 1 \qquad (12)$$

The minimization of the objective function in Eq. (10) is based on Kullback-Leibler Divergence in proposition 2.1. In the case of minimizing $F_{FKMKL}(W, p(Y|z))$, there is a problem with respect to w_{ki} and $p(Y|z_k)$ under constrains of (11) and (12). This problem can be equalized to minimizing.

$$\begin{aligned} &F_{FKMKL}(W, p(Y|z), \lambda_1, \lambda_2) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n} w_{ki}^{m} p(x_i) D(p(Y|x_i)||p(Y|z_k)) \\ &- \lambda_1\left(\sum_{k=1}^{K} w_{ki} - 1\right) - \lambda_2\left(\sum_{y} p(Y|z_k) - 1\right) \end{aligned} \qquad (13)$$

by using the Lagrangian Multiplier concept.

Based on the Lagrange function L_{FKMKL}, the first partial derivatives L_{FKMKL} with respect parameters $w_{ki}, p(Y|z_k), \lambda_1$ and λ_2 are determined and then set equal to 0. The parameters $w_{ki}, p(Y|z_k), \lambda_1$ and λ_2 are determined from the solution of the system of equations $\frac{\partial L_{FKMKL}}{\partial w_{ki}} = 0, \frac{\partial L_{FKMKL}}{\partial p(Y|z_k)} = 0, \frac{\partial L_{FKMKL}}{\partial \lambda_1} = 0, \frac{\partial L_{FKMKL}}{\partial \lambda_2} = 0$ so that it is obtained

$$w_{ki} = \frac{\left(\frac{1}{p(x_i)D(p(Y|x_i)||p(Y|z_k))}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{K}\left(\frac{1}{p(x_i)D(p(Y|x_i)||p(Y|z_k))}\right)^{\frac{1}{m-1}}} \quad (14)$$

$$p(Y|z_k) = \frac{\sum_{i=1}^{n} w_{ki}^{m} p(x_i)p(Y|x_i)}{\sum_{i=1}^{n} w_{ki}^{m} p(x_i)} \quad (15)$$

$$\lambda_2 = -\sum_{i=1}^{n} w_{ki}^{m} p(x_i) \quad (16)$$

$$\lambda_1 = m w_{ki}^{(m-1)} p(x_i)D(p(Y|x_i)||p(Y|z_k)) \quad (17)$$

where

$$p(x_i) = n(x_i)/\sum_{x} n(x_i) \quad (18)$$

$$p(x_i) = n(x_i)/\sum_{x} n(x_i) \quad (19)$$

Fuzzy k-Means KL Divergence Algorithm.

Step 1 : Fix $m \in (1,\infty)$, fix $2 \le k \le n$, fix MaxIter and fix any $\varepsilon > 0$. Take initials $w_{ki}^{(0)}$ and let $t = 1$.

Step 2 : Transformation of data into (19)

Step 3 : Compute $p(x_i)$ by (18)

Step 4 : Compute $p(Y|z_k)^{(t)}$ with $w_{ki}^{(t-1)}$ by (15)

Step 5 : Update to $w_{ki}^{(t)}$ with $p(Y|z_k)^{(t)}$ by (14)

Step 6 : Compute objective function $F_{FKMKL}(W,p(Y|z_k))^{(t)}$ by (10)

Step 7 : Check the stop condition

IF $\left|w_{ki}^{(t)} - w_{ki}^{(t-1)}\right| < \varepsilon$, $\left|F_{FKMKL}(W,p(Y|z_k))^{(t)} - F_{FKMKL}(W,p(Y|z_k))^{(t-1)}\right| < \varepsilon$ or $t > MaxIter$, THEN Stop.
ELSE $t = t + 1$ and return to step 3.

IV. Experiment Results and Discussion

In the experiment, the proposed Fuzzy k-Means KL Divergence was implemented in MATLAB. The clustering results were obtained later in the evaluation of both internal criteria and external criteria. We can compute external criteria that evaluate the clustering quality [12]. To calculate purity, three steps must be taken. In the first step, each cluster was assigned to the most frequent class in the cluster. This task's accuracy was measured by calculating the amount of data set correctly in the second step. The amount of data that had been calculated in the second stage was divided by the number of objects in the third step [12].

$$\text{Purity}(\Omega,\mathbb{C}) = \frac{1}{N}\sum_{k} \max_j |\omega_k \cap c_j| \quad (20)$$

where $\mathbb{C} = \{c_1, c_2, \dots, c_J\}$ is the set of classes and Ω=$\{\omega_1, \omega_{2,}\dots, \omega_K\}$ is the set of clusters. The set of data in ω_k is represented by ω_k and c_j as the set data in c_j (16). Other and the set data in c_j are represented by c_j (16). Random size is another external used to analyze clusters. The adjusted Rand index [5], [13] is corrected for chance version of the rand index that computes how similar the clusters (returned by the clustering algorithm) are. The adjusted rand index is as in (17)

$$R = \frac{\sum_{j=1}^{l}\sum_{k=1}^{K}\binom{n_{jk}}{2} - \binom{n}{2}^{-1}\sum_{j=1}^{l}\binom{n_{j.}}{2}\sum_{k=1}^{K}\binom{n_{.k}}{2}}{\frac{1}{2}\left[\sum_{j=1}^{l}\binom{n_{j.}}{2} + \sum_{k=1}^{K}\binom{n_{.k}}{2}\right] - \binom{n}{2}^{-1}\sum_{j=1}^{l}\binom{n_{j.}}{2}\sum_{k=1}^{K}\binom{n_{.k}}{2}} \quad (21)$$

where n_{jk} represents the number of objects that are in predefined class j and cluster k, $n_{j.}$ indicates the number of objects in a priori class j, $n_{.k}$ indicates the number of objects cluster j, and n is the total number of objects in the data set.

In internal criteria, a clustering result was measured by the clustering accuracy r [6] defined as

$$r = \frac{\sum_{k=1}^{K} a_k}{n} \quad (22)$$

where a_k represented the number of instances occurring in both cluster k and its corresponding class and n represented the number of instances in the data set.

We have real datasets from UCI Machine Learning as follows [12]:

- Zoo data set loads 101 instance and 18 categorical attributes with a total of 7 clusters.
- Soybean small data set loads 47 instances and 35 categorical attributes with a total of 4 clusters.
- Balloon data set loads 20 instances and 4 categorical attributes with a total of 2 clusters.
- Monk data set loads 432 instances and 7 categorical attributes with a total of 2 clusters.

Fuzzy k-Means KL Divergence is run partially given one initial membership function w_{ki} . The matrix initial membership w_{ki} is a random matrix input for Fuzzy k-Means KL Divergence satisfying the constrains (7) and sum probability distributions for cluster center satisfying the constrains (8). From 10 times implementation of Fuzzy k-Means KL Divergence for the Zoo, Soybean small, Balloon, and Monk datasets in varying fuzziness index $m = 2$ with 100 number of iterations, and then the average accuracy, purity, and rand index are calculated. The results are presented as follows.

TABLE I
Comparison Result in Terms of Purity

	KLD	FC	FkP	Improvement (%)
Zoo	0.9403	0.8932	0.8996	5.27
Soybean	0.9167	0.9167	0.9167	0.00
Balloon	0.7917	0.7825	0.8863	13.27
Monk	0.6714	0.53	0.5901	26.68
Average of Improvement				11.30

TABLE II
Comparison Result in Terms of Accuracy

	KLD	FC	FkP	Improvement (%)
Zoo	0.9307	0.8616	0.8568	8.63
Soybean	0.8936	0.9004	0.9066	0.00
Balloon	0.8	0.7985	0.8905	0.00
Monk	0.6713	0.4959	0.6216	35.37
Average of Improvement				11.00

TABLE III
COMPARISON RESULT IN TERMS OF RAND INDEX

	KLD	FC	F*k*P	Improvement (%)
Zoo	0.9451	0.7875	0.7877	20.01
Soybean	0.8982	0.7493	0.7493	19.87
Balloon	0.6632	0.526	0.7134	0.00
Monk	0.5577	0.5	0.5	11.54
Average of Improvement				12.86

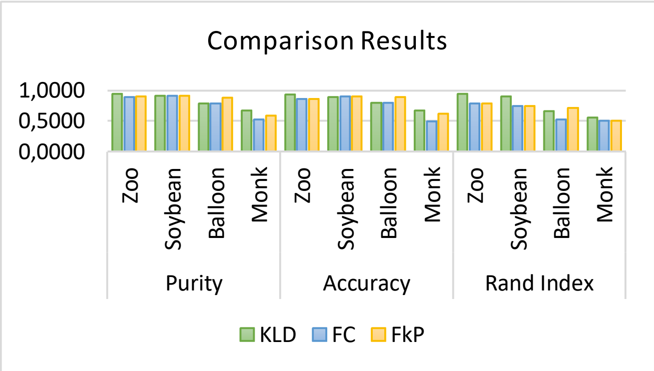

Fig. 1 Validation measure for clustering categorical data

From Table I-III, the overall results show that the KLD achieved an average accuracy of 83% with an average accuracy increase of 11.30%. Likewise, for an average of purity achieved 82.39% with an average purity increase of 11%, and an average of rand index achieved 76.60% with an average rand index increase of 12.86%. In this case, the accuracy level based on the accuracy and quality of clustering based on purity and rand index from Fuzzy *k*-Means KL Divergence give good result for clustering categorical data.

V. CONCLUSION

Based on the discussion results, it can be concluded that the Kullback-Leibler (KL) Divergence can be successfully used for clustering categorical data. The mutual information of KL Divergence between the joint distribution and the product distribution from two marginal distributions is used. The experiment was run using six datasets from UCI Machine Learning to explore the performances. The results are 83%, 82.39%, 76.60% in terms of accuracy, purity, and rand index average, respectively. These experimental results show that the fuzzy k-Means KL Divergence algorithm provides good results both from clustering quality and accuracy for clustering categorical data as compared to Fuzzy Centroid and Fuzzy k-Partition. In future works, we are going to explore the different combination and condition of mutual information of KL Divergence to improve the accuracy.

REFERENCES

[1] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[2] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” J. R. Stat. Soc. Ser. C (Applied Stat., vol. 28, no. 1, pp. 100–108, 1979.

[3] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” Comput. Geosci., vol. 10, no. 2–3, pp. 191–203, 1984.

[4] E. Sutoyo, I. T. R. Yanto, R. R. Saedudin, and T. Herawan, “A soft set-based co-occurrence for clustering web user transactions,” Telkomnika (Telecommunication Comput. Electron. Control., vol. 15, no. 3, 2017.

[5] I. T. R. Yanto, M. A. Ismail, and T. Herawan, “A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering,” Eng. Appl. Artif. Intell., vol. 53, pp. 41–52, 2016.

[6] Z. Huang and M. K. Ng, “A fuzzy k-modes algorithm for clustering categorical data,” IEEE Trans. Fuzzy Syst., vol. 7, no. 4, pp. 446–452, 1999.

[7] J. C. Bezdek, “A convergence theorem for the fuzzy ISODATA clustering algorithms,” IEEE Trans. Pattern Anal. Mach. Intell., no. 1, pp. 1–8, 1980.

[8] *J. Wu, Advances in K-means clustering: a data mining thinking. Springer Science & Business Media, 2012.*

[9] T. M. Cover and J. A. Thomas, Elements of information theory. John Wiley & Sons, 2012.

[10] L.-X. Wang, A course in fuzzy systems. Prentice-Hall press, USA, 1999.

[11] *A. K. Jain and R. C. Dubes, “Algorithms for clustering data,” 1988..*

[12] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to information retrieval, vol. 39. Cambridge University Press, 2008.

[13] L. Hubert and P. Arabie, “Comparing partitions,” J. Classif., vol. 2, no. 1, pp. 193–218, 1985.

[14] D. Dheeru and E. Karra Taniskidou, “{UCI} Machine Learning Repository.” 2017.