Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
13/2 (2020), 65-75. DOI: _http://dx.doi.org/10.21609/jiki.v13i2.818

A COMPARISON OF CLUSTERING BY IMPUTATION AND SPECIAL CLUSTERING ALGORITHMS ON THE REAL INCOMPLETE DATA

Ridho Ananda[1], Atika Ratna Dewi[2], Nurlaili[3]

[1] Faculty of Industrial Engineering and design, Telkom Institut of Technology Purwokerto
[2] Faculty of Informatics, Telkom Institut of Technology Purwokerto
[3] Faculty of Telecommunications and Electrical Engineering, Telkom Institut of Technology Purwokerto
D.I. Panjaitan Street No. 128, Purwokerto, 53147, Indonesia

Email:ridho@ittelkom-pwt.ac.id,atika@ittelkom-pwt.ac.id,nurlaili@ittelkom-pwt.ac.id

Abstract

The existence of missing values will really inhibit process of clustering. To overcome it, some of scientists have found several solutions. Both of them are imputation and special clustering algorithms. This paper compared the results of clustering by using them in incomplete data. K-means algorithms was utilized in the imputation data. The algorithms used were distribution free multiple imputation (DFMI), Gabriel eigen (GE), expectation maximization-singular value decomposition (EM-SVD), biplot imputation (BI), four algorithms of modified fuzzy c-means (FCM), k-means soft constraints (KSC), distance estimation strategy fuzzy c-means (DESFCM), k-means soft constraints imputed-observed (KSC-IO). The data used were the 2018 environmental performance index (EPI) and the simulation data. The optimal clustering on the 2018 EPI data would be chosen based on Silhouette index, where previously, it had been tested its capability in simulation dataset. The results showed that Silhouette index have the good capability to validate the clustering results in the incomplete dataset and the optimal clustering in the 2018 EPI dataset was obtained by k-means using BI where the silhouette index and time complexity were 0.613 and 0.063 respectively. Based on the results, k-means by using BI is suggested processing clustering analysis in the 2018 EPI dataset.

Keywords: *clustering, imputation, missing value, incomplete data*

Abstrak

Adanya nilai data yang hilang tentu akan menghambat proses clustering. Untuk mengatasi hal tersebut, beberapa ilmuwan mengusulkan beberapa metode, diantaranya metode imputasi dan algoritme clustering tertentu. Paper ini membandingkan hasil clustering dari metode tersebut pada data tidak lengkap. Algoritme K-means digunakan pada data imputasi. Algoritme yang digunakan ialah distribution free multiple imputation (DFMI), Gabriel eigen (GE), expectation maximization-singular value decomposition (EM-SVD), biplot imputation (BI), empat algoritme dari fuzzy c-means (FCM) termodifikasi, k-means soft constraints (KSC), distance estimation strategy fuzzy c-means (DESFCM), k-means soft constraints imputed-observed (KSC-IO). Data yang digunakan adalah data indeks kinerja lingkungan (EPI) tahun 2018 dan data simulasi. Hasil clustering optimal pada data EPI 2018 ditentukan berdasarkan index Silhouette,dimana sebelumnya diuji kemampuannya pada data simulasi. Hasil penelitian menunjukkan bahwa index Silhouette memiliki kemampuan yang baik untuk memvalidasi hasil clustering pada data tidak lengkap dan hasil clustering optimal pada data EPI tahun 2018 diperoleh k-means menggunakan data imputasi BI dimana nilai index Silhouette dan kompleksitas waktunya berturut-turut 0.6265 dan 0.063. Berdasarkan hasil tersebut, k-means dengan data imputasi BI direkomendasikan untuk proses clustering pada data EPI tahun 2018.

Kata Kunci: *clustering, metode imputasi, data hilang, data tidak lengkap*

1. Introduction

In the big data, the value of all objects observed is often not obtained completely. It is frequently caused by missing values. Missing values are the lacking information on an object in some of the indicator of measure [1]. Missing values were divided into three categories, namely missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) [2]. MCAR happens when it has no relation among missing values in the same variable or the different variable. In the second type of missing value, if MAR happen so there is a probability where the missing values are depended by available values but there is no relationship among missing values. The last category, NMAR happens when there is not any information about existence of missing values and its cause.

Missing values will really inhibit some analyses that will be used on the dataset, one of them is clustering analysis. To overcome it, some of scientists have found several solutions, namely marginalisation, imputation, and special clustering algorithms. Marginalisation just delete objects or variables that contain missing values [3]. Consequently, the data size significantly decrease if the missing values spread in many objects or variables. The second solution is imputation that is a method to complete missing values with certain values. Finally, the special clustering algorithms are certain clustering algorithms that are applied for handling missing values in the dataset.

On the imputation method, clustering analysis is carried out after the dataset becomes the complete data. Imputation can be executed by filling missing values with their respective variables means, zero values, or other values that are obtained from imputation algorithms. There are two categories for the process of the imputation algorithms that are deterministic and stochastic. The result of the deterministic process in the imputation is consistency different from the stochastic process. Some of the deterministic imputation algorithms that have been found are distribution free multiple imputation (DFMI), Gabriel eigen (GE), expectation-maximization free multiple imputation (EMSVD), and biplot imputation (BI). Those algorithms have been researched in simulation study. Apart from their capability, imputation values will certainly cause bias because they are a guess of a number that is not exact. To know the quality of imputation values, Ananda et.al. have proposed the method to measure the goodness-of-fit of imputation data [4].

For the special clustering algorithms, some of scientists have been proposed them to process clustering in the incomplete dataset. Four algorithms of modified fuzzy c-means (FCM) have been proposed by Hathaway and Bezdek that are whole-data strategy (WDS), partial distance strategy (PDS), optimal completion strategy (OCS), and nearest prototype strategy (NPS) [5]. K-means soft constraints (KSC) has been proposed by Wagstaff [6]. Distance estimation strategy fuzzy c-means (DESFCM) has been proposed by Himmelspach and Conrad [7]. Recently, k-means soft constraints imputed-observed (KSC-IO) have been proposed by Mesquita et.al. [8].

To know quality of clustering needs clustering validity that are categorized into internal clustering validity and external clustering validity. External clustering validity uses the reference clustering to evaluate the clusters obtained. Whereas internal clustering validity only measures quality of clustering obtained that is based on dissimilar measure among objects because there is no reference clustering [9]. In generally, dissimilar measure used on the internal clustering validity is Euclidean distance [10]. The previous researches mostly used the simulation data that contained the reference clustering so the external clustering validity was used in this case. The problem that has not been addressed in previous works is to measure the quality of clustering in the real incomplete dataset that has no the reference clustering. Therefore, this paper will be showed the use of the internal clustering validity in the real incomplete data where in previously it have been tested its capability based on the simulation data. Datasets used in this paper are the 2018 environmental performance index (EPI) and the simulation datasets that are iris, wine, and seeds dataset. Finally, the results of the clustering on the real incomplete data and the simulation data were compared to obtain the optimal clustering.

This paper is arranged as follows. Section II describes material and method used in this research. Section III describes results and discussion. Conclusions and suggestions are on the last section.

2. Material and Method

2.1. Datasets

The datasets used in this paper are the 2018 EPI data and the simulation data. The 2018 EPI data is a project led by Yale University, Columbia University, Samuel Family Foundation, McCall MacBain Foundation, and the World Economic Forum. The data ranks performance of countries where it is based on high-priority environmental issues in two areas, protection of human health (HLT) and protection of

TABLE 1
THE LIST OF VARIABLES IN THE 2018 EPI DATASET

Issues	Symbol	Description
HLT	HAD	Measures the actual outcomes from exposure to indoor air pollution from household use of solid fuels.
	PME	Measures the average annual concentration of $PM_{2.5}$ to which the typical citizen of each country is exposed.
	PMW	Measures the weighted percentage of a countrys population exposed to annual concentrations of $PM_{2.5}$
	UWD	Measures the actual outcomes from lack of access or use of improved sources of drinking water.
	USD	Measures the actual outcomes from lack of access or use of improved sanitation facilities.
	PBD	Measures the actual outcomes from lead exposure
EC0	MPA	Measures the percent of a countrys Economic Exclusion Zone (EEZ) set aside as a marine protected area (MPA).
	TBN	Measures the percent of a countrys biomes in terrestrial protected areas (TPAs), weighted by the prevalence of different biome types within that country.
	TBG	Measures the percent of a countrys biomes in terrestrial protected areas (TPAs) weighted by the prevalence of different biome types around the world.
	SPI	Measures the average area of species distributions in a country under protection, weighted by a country's stewardship for each species.
	PAR	Measures the extent to which a countrys protected areas are ecologically representative.
	SHI	Measures the average loss in suitable habitat for species in a country, weighted by the countrys stewardship for that species.
	TCL	Measures the five-year moving average of percent of forested land lost. Forested land is defined as having $\geq 30\%$ canopy cover.
	FSS	Measures the percentage of a countrys total catch that come from taxa that are classified as either over-exploited or collapsed.
	MTR	Measures the trends in the Regional Marine Trophic Indices of a country, or mean trophic level of the fish catch in each region of the Economic Exclusion Zones.
	DCT	Measures the intensity of CO_2 emissions from the entire economy, as a blend of current-year intensity and a 10-year trend.
	DPT	Measures the intensity of CO_2 emissions per kilowatt-hour of electricity and heat, as a blend of current-year intensity and a 10-year trend.
	DMT	Measures the intensity of methane emissions from the entire economy, as a blend of current-year intensity and a 10-year trend.
	DNT	Measures the intensity of N_2O emissions from the entire economy, as a blend of current-year intensity and a 10-year trend.
	DBT	Measures the intensity of Black Carbon emissions from the entire economy, as a blend of current-year intensity and a 10-year trend.
	DST	Measures the intensity of SO_2 emissions from the entire economy, as a blend of current-year intensity and a 10-year trend.
	DXT	Measures the intensity of NO_x emissions from the entire economy, as a blend of current-year intensity and a 10-year trend.
	WWT	Measures the percentage of wastewater treated, weighted by the connection rate of the population to the wastewater treatment system.
	SNM	Measures the Euclidean distance from an ideal point with optimal nitrogen use efficiency (NUE) and crop yield.

ecosystems (ECO). The 2018 EPI dataset is quantitative data with value in 0 until 100. The data is represented in matrix dataset with ordo 180×24. The data has $237(5.49\%)$ missing value on $89(49.44\%)$ objects and $7(29.17\%)$ variables. Type of missing values in the 2018 EPI data is NMAR because there is not any information about existence of them and its cause. Table 1 shows the list of variables on the 2018 EPI dataset, whereas the list of the objects is showed in Table 2.

TABLE 2
THE LIST OF COUNTRIES IN THE 2018 EPI DATA

Code	Countries	Code	Countries	Code	Countries	Code	Countries	Code	Countries
BLZ	Belize	STP	Safalo Tomafac	ITA	Italy	DJI	Djibouti	BEN	Benin
SEN	Senegal	JPN	Japan	EGY	Egypt	BTN	Bhutan	SLE	Sierra Leone
KWT	Kuwait	ERI	Eritrea	BOL	Bolivia	LKA	Sri Lanka	LVA	Latvia
FJI	Fiji	BWA	Botswana	TJK	Tajikistan	LTU	Lithuania	GMB	Gambia
BFA	Burkina Faso	TZA	Tanzania	LUX	Luxembourg	GEO	Georgia	BDI	Burundi
TLS	Timor-Leste	MKD	Macedonia	GRD	Grenada	KHM	Cambodia	TGO	Togo
MYS	Malaysia	GUY	Guyana	CMR	Cameroon	TON	Tonga	MLT	Malta
HTI	Haiti	CAF	Afrika tengah	UGA	Uganda	MEX	Mexico	IND	India
TCD	Chad	VNM	Viet Nam	MNE	Montenegro	IRN	Iran	COM	Comoros
ZMB	Zambia	MAR	Morocco	IRQ	Iraq	CIV	d'Ivoire	ZWE	Zimbabwe
NLD	Netherlands	JOR	Jordan	COD	Kongo	ALB	Albania	NZL	New Zealand
KAZ	Kazakhstan	DMA	Dominica	ATG	Antigua	NOR	Norway	LBN	Lebanon
ECU	Ecuador	ARM	Armenia	PAN	Panama	LSO	Lesotho	SLV	El Salvador
AUS	Australia	PER	Peru	LBR	Liberia	ETH	Ethiopia	AUT	Austria
POL	Poland	LBY	Libya	GAB	Gabon	BHS	Bahamas	PRT	Portugal
MDV	Maldives	GHA	Ghana	BLR	Belarus	ROU	Romania	MRT	Mauritania
GTM	Guatemala	BEL	Belgium	RUS	Russia	MUS	Mauritius	GIN	Guinea
BRA	Brazil	VCT	Vincent	FSM	Micronesia	GNB	Guinea-Bissau	BRN	Brunei
SVK	Slovakia	MDA	Moldova	HND	Honduras	BGR	Bulgaria	SVN	Slovenia
MMR	Myanmar	IDN	Indonesia	CAN	Canada	KOR	South Korea	OMN	Oman
JAM	Jamaica	CHL	Chile	ESP	Spain	PNG	Papua	KEN	Kenya
CHN	China	SWE	Sweden	QAT	Qatar	KIR	Kiribati	COL	Colombia
CHE	Switzerland	LCA	Saint Lucia	KGZ	Kyrgyzstan	CRI	Costa Rica	TWN	Taiwan
WSM	Samoa	LAO	Laos	HRV	Croatia	THA	Thailand	SAU	Saudi Arabia
MDG	Madagascar	CUB	Cuba	TTO	T- Tobago	SRB	Serbia	MWI	Malawi
CYP	Cyprus	ARE	Arab Saudi	SYC	Seychelles	MLI	Mali	CZE	Czech Republic
GBR	Inggris	SGP	Singapore	MNG	Mongolia	DNK	Denmark	USA	Amerika
SLB	Solomon	MOZ	Mozambique	DOM	Dominican	VEN	Venezuela	ZAF	South Africa
NAM	Namibia	GNQ	Guinea	AFG	Afghanistan	SDN	Sudan	NPL	Nepal
EST	Estonia	DZA	Algeria	SUR	Suriname	NIC	Nicaragua	FIN	Finland
AGO	Angola	SWZ	Swaziland	NER	Niger	FRA	France	ARG	Argentina
TUN	Tunisia	NGA	Nigeria	DEU	Germany	AZE	Azerbaijan	TUR	Turkey
PAK	Pakistan	GRC	Greece	BHR	Bahrain	TKM	Turkmenistan	PRY	Paraguay
HUN	Hungary	BGD	Bangladesh	UKR	Ukraine	PHL	Philippines	ISL	Iceland
BRB	Barbados	URY	Uruguay	COG	Kongo	IRL	Ireland	BIH	Bosnia
UZB	Uzbekistan	RWA	Rwanda	ISR	Israel	CPV	Cabo Verde	VUT	Vanuatu

The simulation dataset used in this paper are Iris, Wine, and Seeds dataset. The Iris dataset is commonly known and consists of 150 objects of Iris plants that are divided into three classes and measured on four variables. The Wine dataset consists of 178 objects that are also divided into three classes and measured on thirteen variables. And also, the Seeds dataset consists of 210 objects that are divided into three classes and measured on seven variables. The incomplete datasets were created from those datasets. The ratio of missing value are $5, 6, 7, 8, 9$, and 10% from all data. The missing value on Iris, Wine, and Seeds dataset were randomly spread in $1(25.00\%), 4(30.77\%)$, and $2(28.57\%)$ variables respectively where the spreads were matched with the spread of missing values in the 2018 EPI data.

2.2. Imputation algorithms

Imputation algorithms used in this paper were distribution free multiple imputation (DFMI) [11], Gabriel eigen (GE) [12], expectation-maximization singular value decomposition (EMSVD) [13], and biplot imputation (BI) [14]. Those algorithms are based on singular value decomposition (SVD) and using multiple regression model. Imputation data obtained is computed the proximity matrix and the covariance matrix to measure the goodness-of-fit of them. Suppose that $\mathbf{X}$ is the proximity matrix of the imputation data and $\mathbf{Y}$ is the proximity matrix of the initial data. The goodness-of-fit of proximity matrix is obtained by using Equation 1.

$$\mathrm{GoF}_p\left(\mathbf{X},\mathbf{Y}\right)=\left(\sum_{i=1}^{r}\sigma_{ii}\right)^2, \qquad (1)$$

Where r and $\sigma_{ii}\left(i=1,2,\cdots,r\right)$ is rank and singular value respectively from $\widetilde{\mathbf{X}}_T'\widetilde{\mathbf{Y}}_T$ atau $\widetilde{\mathbf{Y}}_T'\widetilde{\mathbf{X}}_T$.

$\widetilde{\mathbf{X}}_T$ is $\mathbf{X}$ matrix after the translation-normalization procedure. The measure of $\mathrm{GoF}_p\left(\mathbf{X},\mathbf{Y}\right)$ belong to the interval of $[0,1]$, if $\mathrm{GoF}_p\left(\mathbf{X},\mathbf{Y}\right)\approx 1$ so it means that has a good approximation to represent the dissimilarity measures among objects in the initial data. Conversely, if $\mathrm{GoF}_p\left(\mathbf{X},\mathbf{Y}\right)\approx 0$ so it means that has a bad approximation [4]. Average of the goodness-of-fit of proximity matrix and covariance matrix is decided to be the goodness-of-fit of imputation data in this paper. These that will be processed in clustering analysis are the results of imputation algorithms that the goodness-of-fit of imputation data have more than 0.900. Clustering algorithms used in the imputation data is k-means algorithm.

2.3. K-means clustering algorithm

Clustering algorithms are algorithms that are utilized to put objects into a group based on the similarity measure. Objects in the same group have the high resemblance different from objects in the different group. One of the clustering algorithms that is popular enough is k-means algorithms. K-means is an algorithm that assigns each objects to the cluster having the nearest prototype. The newly researches by using k-means had been done in some areas such as the recommendation system in the selection of specialization course [15], the measure of mangrove areas [16], analysis of education quality in senior high school [17], and mapping the quality of education based on the results of the 2019 national exam in Banyumas Regency [18]. The process of k-means is composed in three steps. The first, partition the objects into k initial clusters arbitrarily or by using certain analysis. The second, assigning an object is to the cluster whose prototype is nearest. Then recalculate the prototype for the cluster receiving the new item and for the cluster losing the item. Finally, repeat the second step until no more reassignments take place.

2.4. Special clustering algorithms for incomplete data

In this paper, we present the seven clustering algorithms for incomplete data. Let any data matrix ${}_n\mathbf{X}_p$ with n objects and p variables has missing values in some variables. Suppose that $[\mathbf{c}_1,\mathbf{c}_2,\mathbf{c}_3,\cdots,\mathbf{c}_k]$ are prototype of the obtained clusters.

2.4.1. Whole-data strategy. Whole-data strategy (WDS) classifies objects that have complete data by using FCM algorithm. Then, objects with missing value are classified into certain cluster and based on the nearest prototype. To know the nearest prototype for ith object that contains missing value, we use Equation 2.

$$\underset{c_j}{\operatorname{argmin}}\, d\left(\mathbf{x}_i,\mathbf{c}_j\right)=\sqrt{\frac{\sum_{k=1}^{p}\left(x_{ik}-c_{jk}\right)^2 w_{ijk}}{\sum_{k=1}^{p} w_{ijk}}}, \quad (2)$$

where $\mathbf{x}_i$ is data of ith object, $\mathbf{c}_j$ is jth prototype, x_{ik} is the value of the kth variable on ith object, and c_{jk} is the value of the kth variable on the jth prototype. w_{ijk} is weight that be 0 if x_{ik} is missing or 1 otherwise.

2.4.2. Partial distance strategy. Partial distance strategy (PDS) classifies objects by using FCM algorithm where there is modification in prototype computation. Prototype is computed by using Equation 3.

$$c_{kj}=\frac{\sum_{i=1}^{n}\left(u_{ik}\right)^2 .x_{ij}.w_{ij}}{\sum_{i=1}^{n}\left(u_{ik}\right)^2 .w_{ij}} \quad (3)$$

c_{kj} is the value of the kth prototype in the jth variable. u_{ik} is the value of membership of ith object in kth prototype. x_{ij} is the value of the ith object in jth variables. w_{ij} is weight that be 0 if x_{ij} is missing or 1 otherwise.

2.4.3. Optimal completion strategy. Optimal completion strategy (OCS) estimates missing values and classifies object into certain cluster simultaneously by optimizing its objective function. Basically, OCS algorithm adopts FCM algorithm on the its process. Imputation process in the OCS algorithm is performed after prototype computation by using Equation 4.

$$x_{ij}^{*}=\frac{\sum_{l=1}^{k}\left(u_{il}\right)^2 .c_{lj}}{\sum_{l=1}^{k}\left(u_{il}\right)^2}, \quad (4)$$

where x_{ij}^{*} is missing value on the ith object in the jth variable. u_{il} is the value of membership of ith object on the lth prototype. c_{lj} is the lth prototype in the jth variables.

2.4.4. Nearest prototype strategy. Nearest prototype strategy (NPS) is similar to the OCS in all steps. Every missing value on the ith object in the jth variable, x_{ij}^{*} is substituted with respective values of the nearest prototype. To know the nearest prototype, we use Equation 2.

2.4.5. K-means soft constraints. K-means soft constraints (KSC) is obtained by the idea where is to define the soft constraints on variables with missing values and to use there as additional information. Suppose that ${}_n\mathbf{X}_p$ is incomplete data, the variables of the data is divided into the dataset of completely observed variables that is ${}_n\widetilde{\mathbf{X}}_q$ and the dataset of variables with missing values that is ${}_n\widehat{\mathbf{X}}_q$ where $p = q + r$. A soft constraint s_{ij} between $\mathbf{x}_i$ and $\mathbf{x}_j$ in ${}_n\widehat{\mathbf{X}}_q$ is computet by using Equation 5 if $i, j \in \{1, 2, \cdots, n\}, \forall k \in \{1, 2, \cdots, r\}, x_{ik} \cup x_{jk} \neq \emptyset$.

$$s_{ij} = -\sqrt{\sum_{k=1}^{r} \left(x_{ik} - x_{jk}\right)^2}, \tag{5}$$

and s_{ij} is zero for otherwise. Henceforth, objects in the ${}_n\widetilde{\mathbf{X}}_q$ dataset are classified by using k-means algorithms where the distance between the ith object and the kth prototype is computed by using Equation 6.

$$d\left(\mathbf{x}_i, \mathbf{c}_k\right) = d_1 + d_2, \tag{6}$$

Where

$$d_1 = w\|\mathbf{x}_i - \mathbf{c}_i\|_2^2$$

and

$$d_2 = (1 - w)\sum_{j=1}^{n} \delta_{ij} s_{ij}^2.$$

$w \in [0, 1]$, δ_{ij} is binary variables that be 1 if $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned to the same cluster and 0 otherwise.

2.4.6. K-means soft constraints imputed-observed. K-means soft constraints imputed-observed (KSC-IO) was presented by Mesquita et.al. to use information from partially complete objects in ${}_n\widehat{\mathbf{X}}_r$ dataset in KSC algorithm [12]. The method developed the KSC algorithm by adding new soft constraints on partially complete objects that were ignored and also on imputed values. For ${}_n\widehat{\mathbf{X}}_r$ dataset we obtain ${}_n\widehat{\mathbf{X}}_r^* = (x_{ik}^*)$ that is imputation data. A soft constraint s_{ij}^* between $\mathbf{x}_i$ and $\mathbf{x}_j$ in ${}_n\widehat{\mathbf{X}}_r$ is computed by using Equation 7, if $i, j \in \{1, 2, \cdots, n\}, \exists k \in \{1, 2, \cdots r\}, x_{ik} \cup x_{ik} = \emptyset$.

$$s_{ij}^* = -\sqrt{\sum_{k=1}^{r} \left(x_{ik}^* - x_{jk}^*\right)^2}, \tag{7}$$

and s_{ij}^* is zero for otherwise. Henceforth, objects in the ${}_n\widetilde{\mathbf{X}}_q$ dataset are classified by using k-means algorithms, where the distance between the ith object and the kth prototype is computed by using Equation 8.

$$d\left(\mathbf{x}_i, \mathbf{c}_k\right) = d_1 + d_2 + d_3. \tag{8}$$

Where

$$d_1 = w_1\|\mathbf{x}_i - \mathbf{c}_i\|_2^2,$$

$$d_2 = w_2\sum_{j=1}^{n} \delta_{ij} s_{ij}^2,$$

and

$$d_2 = w_2\sum_{j=1}^{n} \delta_{ij} \left(s_{ij}\right)^2.$$

$w_1 + w_2 + w_3 = 1, \forall i, w_i \in [0, 1]$ and δ_{ij} is binary variables that be 1 if $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned to the same cluster and 0 otherwise.

2.4.7. Distance estimation strategy fuzzy c-means. Distance estimation strategy fuzzy c-means (DESFCM) was development from FCM by using another variant of the FCM as basis. The variant is based on the membership degrees of data items to the certain prototype. In the first step, ${}_n\mathbf{X}_p$ is divided into the dataset with completely observed object that is ${}_{n_1}\widehat{\mathbf{X}}_p = (\hat{x}_{ij})$ and the dataset with objects contained missing values that is ${}_{n_2}\widetilde{\mathbf{X}}_p$. The initial membership matrix that is ${}_{n_1}\mathbf{U}_k = (u_{ij})$ is initialised randomly. The second step, the cluster prototypes are calculate by using Equation 9.

$$c_{kj} = \frac{\sum_{i=1}^{n_1} \left(u_{ik}\right)^2 \hat{x}_{ij}}{\sum_{i=1}^{n_1} \left(u_{ik}\right)^2}, \tag{9}$$

where c_{kj} is the kth prototype on the jth variable, u_{ik} is the value of membership degree of ith object that is from ${}_{n_1}\widehat{\mathbf{X}}_p$ on the kth prototype, and $\hat{x}_{ij}$ is the ith object that is from ${}_{n_1}\widehat{\mathbf{X}}_p$ on the jth variable. Then ${}_n\mathbf{D}_k = (d_{ik})$ dataset where d_{ik} is distance between the ith object and the kth prototype are calculated by using Equation 10.

$$d_{ik} = \sum_{j=1}^{p} \left(x_{ij} - c_{kj}\right)^2, \tag{10}$$

for all i and j. If $x_{ij} = \emptyset$ then

$$\left(x_{ij} - c_{kj}\right)^2 = \frac{\sum_{l=1}^{n_1} u_{lk} \left(\hat{x}_{lj} - c_{kj}\right)^2}{\sum_{l=1}^{n_1} u_{lk}}$$

where x_{ij} is the ith object on jth variable of dataset, c_{kj} is the kth prototype on the jth variable, u_{lk} is the value of membership degree of lth object that is from ${}_{n_1}\widehat{\mathbf{X}}_p$ on the kth prototype, and $\hat{x}_{ij}$ is the lth object on the jth variable in ${}_{n_1}\widehat{\mathbf{X}}_p$. Then new membership matrix ${}_{n_1}\mathbf{U}_k$ is calculated by using Equation 11.

$$u_{ik} = \left[\sum_{l=1}^{k} \left(\frac{d\left(\hat{\mathbf{x}}_i, \mathbf{c}_k\right)}{d\left(\hat{\mathbf{x}}_i, \mathbf{c}_l\right)}\right)^2\right]^{-1}. \tag{11}$$

The process is iterated to the second step until the residual sum of squares (RSS) between ${}_n\mathbf{D}_k^{(t)}$ and ${}_n\mathbf{D}_k^{(t+1)}$ where they are from tth and $(t + 1)$th iteration respectively is small.

2.5. Clustering validity

Missing values in the 2018 EPI dataset is not ignored because they are categorized into MNAR [2]. Because of that, marginalization is not utilized in this paper. Then, clustering validity used in the data must be using internal clustering validity because there is no reference clustering. Internal clustering validity used is Silhouette index because the validity is better than the other validity [19] [20]. Silhouette index is obtained from the average of Silhouette value each objects that given by using Equation 12.

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{12}$$

where $a(i)$ is average of the distance between ith object and other objects in the same cluster. Whereas, $b(i)$ is average of the distance between ith object and other objects that is in nearest cluster. $s(i)$ is value of the Silhouette index that belong to the interval of $[-1,1]$. If $s(i)\approx 1$, it means that ith object is well matched to its cluster. Conversely, if $s(i)\approx -1$, it means that ith object is not well matched to its cluster [21].

Peladeau et.al. mentioned that process of clustering must be shaped by using dissimilarity measure from the initial data [22]. That statement reinforce motivation to use the dissimilarity measure among objects as the foundation in the internal clustering validity. However, the existence of missing value in data will certainly inhibit computation of dissimilarity measure among objects. The problem is solved if computation of dissimilarity measure uses weighted Euclidean distance that is proposed by Gower [23] and formulated by Equation 13.

$$d\left(\mathbf{x}_i,\mathbf{x}_j\right)=\sqrt{\frac{\sum_{s=1}^{p}\left(x_{is}-x_{js}\right)^2 w_{ijs}}{\sum_{s=1}^{p} w_{ijs}}}, \tag{13}$$

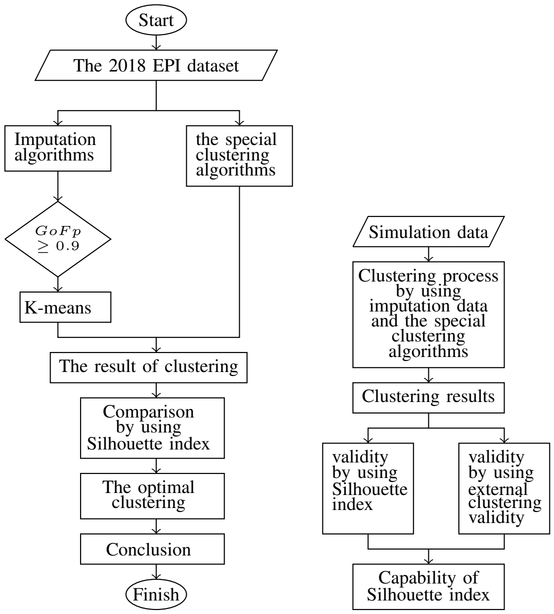
where $d\left(\mathbf{x}_i,\mathbf{x}_j\right)$ is weighted Euclidean distance between ith object and jth object. p is the total number of variables in data. x_{is} is the value of the ith object on sth variable. w_{ijs} is weight that be 0 if x_{is} or x_{js} are missing and 1 otherwise. The computation indirectly uses marginalization method where for all pair of objects that will be computed weight Euclidean distance, it will exactly remove certain variables that at least, one of objects pair has missing value. As a results of that, the distance of every objects pair is not precise if many variables that are removed because of missing value. Exactly, it will also decrease capability of the internal clustering validity because of the dissimilarity measure as its foundation. Therefore, it is very needed method to know capability of the internal clustering validity. In this paper, we use the simulation data that has reference clustering to measure capability of Silhouette index obtained by using weighted Euclidean distance. We validate the results of clustering by using external clustering validity and Silhouette index and then compute their correlation. If the correlation approximate to 1, it means that silhouette index has the good suitability to external clustering validity so it is a reasonably faithful validity to be used. If the correlation approximate to -1, it means that Silhouette index has the converse suitability to external clustering validity where if the external clustering validity shows the highest value for the good clustering so Silhouette index will certainly show the lowest value for it. The second state may be a reasonably faithful validity but be careful to interpret the value of Silhouette index. The unpleasant condition is occurred if the approximate correlation is about 0 where it show that there is not any relationship between Silhouette index and the external clustering validity. Consequently, Silhouette index can not be said a reasonably faithful validity. In this condition, Silhouette index is not suggested. The formula of the correlation uses Equation 14.

$$r(x,y)=\frac{\sum_{i=1}^{m}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\left(\sum_{i=1}^{m}\left(x_i-\bar{x}\right)^2\right)\left(\sum_{i=1}^{m}\left(y_i-\bar{y}\right)^2\right)}} \tag{14}$$

where x and y are vector of Silhouette index and the external clustering validity results respectively where each of element in those vector is validity measure of the results of clustering from each algorithms used in the same order. x_i is ith element of vector of Silhouette index results. $\bar{x}$ is the average of the element of vector of Silhouette index results. y_i is ith element of vector of the external clustering validity results. $\bar{y}$ is the average of the element of vector of the external clustering validity results. m is the total number of the element of the vector.

2.6. Research flow

The research steps used in this paper consists of several stages. Firstly, the clustering results by using the special clustering algorithms and the imputation data by using imputation algorithms will be computed from the 2018 EPI dataset simultaneously. Secondly, it is needed to see the goodness-of-fit of imputation data to know the faithful data imputation to clustering process by using k-means algorithm. Next, the results of clustering are validated by using Silhouette index where in previously Silhouette index has been examined its capability. Then the optimal clustering is determined based on the value of Silhouette index. In the last stage, the result of the optimal clustering is interpreted and given


Start
The 2018 EPI dataset
Imputation algorithms
the special clustering algorithms
$GoFp \geq 0.9$
K-means
The result of clustering
Comparison by using Silhouette index
The optimal clustering
Conclusion
Finish
Simulation data
Clustering process by using imputation data and the special clustering algorithms
Clustering results
validity by using Silhouette index
validity by using external clustering validity
Capability of Silhouette index

(a) (b)

Figure 1. Research flow of (a) the main research steps and (b) the validity of Silhouette index

conclusion. In brief, the research steps are provided on research flow in Figure 1.

3. Result and Discussion

3.1. The missing value distribution on the 2018 EPI dataset

We have known that the 2018 EPI dataset has 237(5.49%) missing value on 89(49.44%) objects and 7(29.17%) variables. Table 3 shows distribution of missing values in objects and Table 4 shows distribution of missing value in variables. From Table 3 we know that 2 countries have missing values in 5(20.83%) variables from 24 measured variables. we also know that 3 missing values is in the most countries and the least missing values is in 25 countries. Furthermore, Table 4 show that variables that have quite a lot of missing value are DPT, MTR, FSS, MPA, and TCL i.e. more than or equal 30 missing values.

3.2. Capability of Silhouette index

This paper uses the correlation between the internal clustering validity and the external clustering validity to measure capability of the internal clustering validity. The internal clustering validity used is Silhouette index and the external clustering validity used are Rand index, Jaccard index, F-measure, and Purity. Figure 2 shows graph of average of the correlation of Silhouette index with the external clustering validity used in the simulation data. From the Figure, we know that the correlation of Silhouette index in dataset used for all ratio of missing value is more than 0.600. Moreover, average of the correlation of Silhouette index is generally about 0.836. It means that Silhouette index has the good suitability to external clustering validity in those simulation dataset so it is a reasonably faithful validity to be used in the incomplete dataset that has the same characteristic with those simulation dataset. Because the spread of missing values in those simulation dataset were matched with the spread of missing values in the 2018 EPI data so Silhouette index is also reasonably used in the 2018 EPI dataset.

TABLE 3
MISSING VALUES ON OBJECTS IN THE 2018 EPI DATASET

Number of countries	Missing values	Percentage
25	1	4.17%
5	2	8.33%
36	3	12.5%
21	4	16.67%
2	5	20.83%

TABLE 4
MISSING VALUES ON VARIABLES IN THE 2018 EPI DATASET

Variables	Missing values	Percentage
SPI	15	8%
SHI	15	8%
TCL	30	17%
MPA	44	24%
FSS	44	24%
MTR	44	24%
DPT	45	25%

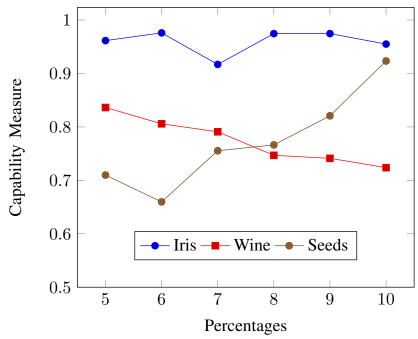
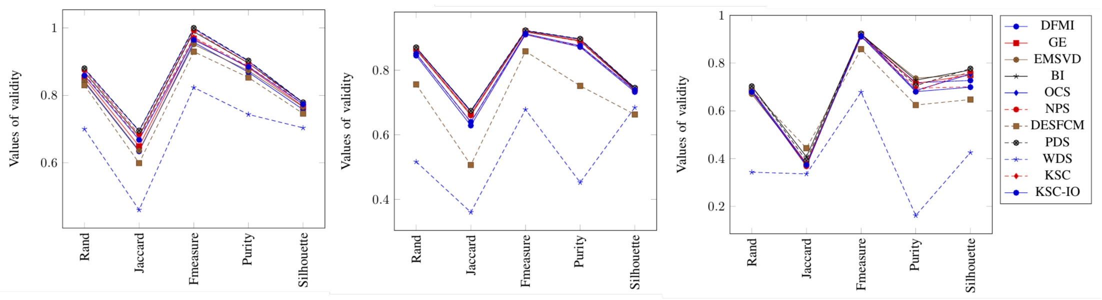


Fig. 2. Graph of the correlation of Silhouette index in the simulation data.



(a) Iris dataset (b) Wine dataset (c) Seeds dataset

Fig. 3. Graph of the quality measure of the clustering results in (a) Iris dataset, (b) Wine dataset, and (c) Seeds dataset.

3.3. The goodness-of-fit of imputation data

Imputation algorithms used in this paper are DFMI, GE, EMSVD, and BI in the 2018 EPI dataset. Suppose that the goodness-of-fit of imputation data is symbolized as $GoFp$, the goodness-of-fit of proximity matrix as **D**, and the goodness-of-fit of covariance matrix as Σ, then Table 5 shows the goodness-of-fit of imputation data for imputation algorithms used. From the table, we know that the goodness-of-fit of every imputation data is more than 0.900. It means that all of imputation data obtained are used in the process of clustering by using k-means algorithm.

TABLE 5
THE GOODNESS-OF-FIT OF IMPUTATION DATA

Algorithm	D	Σ	PA
DFMI	0.9524	0.9913	0.9718
Eigen Gabriel	0.9548	0.9940	0.9744
EMSVD	0.9469	0.9947	0.9708
BI	0.9536	0.9909	0.9722

3.4. Comparison of the clustering results

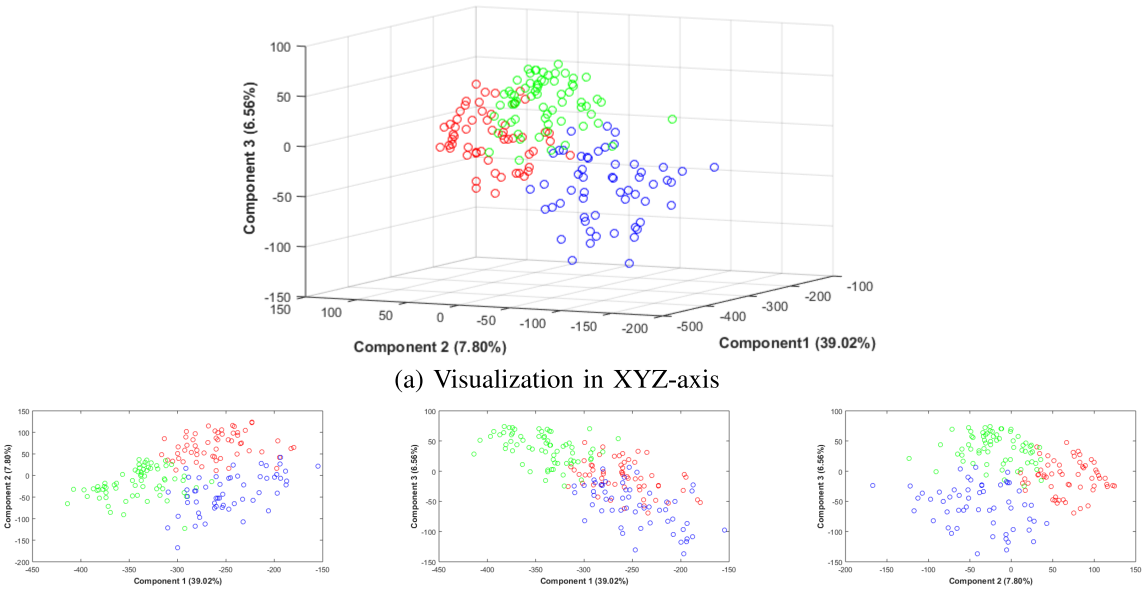
The clustering results from each of clustering algorithms used are compared to know the optimal clustering results. In the simulation data, we use the external validity cluster and Silhouette index to validate the clustering obtained. Whereas in the 2018 EPI dataset, we only use Silhouette index because there is no reference clustering in the dataset. Figure 3 shows average of validation of the clustering results each algorithms in Iris, Wine, and Seeds dataset respectively. In those figures, we know that WDS has the lowest quality of clustering results in generally, whereas the most of clustering results have quality that are quite similar. Furthermore in the 2018 EPI dataset, Table 6 shows average of the clustering quality obtained from repetition 100 times for each algorithms. The table shows that the optimal clustering results is obtained k-means algorithms from imputation data by using biplot imputation (BI). Whereas the lowest clustering quality is obtained by WDS where it is like the results on the simulation dataset. In the simulation data, we also know that quality of k-means from imputation data by using BI has good quality.

TABLE 6
CLUSTERING QUALITY ON THE 2018 EPI DATASET

No	Algorithms	Validity
1	DFMI	0.6256
2	EG	0.6259
3	EMSVD	0.6247
4	BI	0.6265
5	OCS	0.5994
6	NPS	0.5953
7	DESFCM	0.5811
8	PDS	0.5937
9	WDS	0.5809
10	KSC	0.5884
11	KSC-IO	0.603

TABLE 7
TIME COMPLEXITY ON THE 2018 EPI DATASET

No	Algoritme	Time Complexity
1	DFMI	18.21857
2	GE	5.944764
3	EMSVD	0.052468
4	BI	0.062851
5	OCS	1.062831
6	NPS	1.185435
7	DESFCM	20.30097
8	PDS	2.232787
9	WDS	0.829149
10	KSC	1.487316
11	KSC-IO	4.522597



(a) Visualization in XYZ-axis

(b) Visualization in XY-axis

(c) Visualization in XZ-axis

(d) Visualization in YZ-axis

Figure 4. Visualization of countries on the 2018 EPI dataset in (a) XYZ-axis, (b) XY-axis, (c) XZ-axis, and (d) YZ-axis(a) Visuzalization in XYZ-axis.

Table 7 shows average of the time complexity that is also obtained from repetition 100 times. From the table, we know that DESFCM has the highest time complexity, then DFMI is in the second order. Whereas the low time complexity is obtained by EMSVD, as well as BI where their values are quite similar.

Based on the clustering result obtained, if we choose the results of k-means with BI algorithms because of the optimal clustering result in the 2018 EPI dataset and the low time complexity, so we will obtain the visualization of the clustering results in Figure 4. Futhermore, there are three clusters obtained that are the first cluster consists of $58(32.2\%)$ countries, $67(37.2\%)$ in the second cluster, and $55(30.6\%)$ countries in the third cluster.

4. Conclusion

In this paper, we have compared the clustering result by using imputation data and special clustering algorithms on the incomplete dataset, the 2018 EPI dataset and the simulation dataset. The result show that Silhouette index has good ability to validate the clustering results in the real incomplete dataset based on its correlation with the external clustering validity in the simulation dataset. The optimal clustering in the simulation dataset is quite similar on the most clustering results. The optimal clustering in the 2018 EPI dataset is obtained by k-means with BI algorithm and the time complexity of k-means with BI is quite small. Based on the results, k-means with BI algorithm is suggested processing clustering analysis in the 2018 EPI dataset.

Acknowledgement

We thak the support of LPPM IT Telkom Purwokerto through a 2019 internal grant for the financial support of the research for this paper.

References

[1] X. L. Meng, "Missing Data: Dial M for???" *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1325–1330, 2000.

[2] M. Quintero and A. LeBoulluec, "Missing Data Imputation for Ordinal Data," *International Journal of Computer Applications*, vol. 181, no. 5, pp. 10–18, 2018.

[3] C. Cheng and H. Huang, "A Distance-threshold K-NN Method for Imputing Medical Data Missing Values," *Journal of Advances in Computer Networks*, vol. 7, no. 1, pp. 13–17, 2019.

[4] R. Ananda, Siswadi, and T. Bakhtiar, "Goodness-of-fit of Imputation Data in Biplot Analysis," *Far East Journal of Mathematical Sciences*, vol. 103, no. 11, pp. 1839–1849, 2018.

[5] R. Hathaway and J. C. Bezdek, "Fuzzy C-Means Clustering of Incomplete Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 31, no. 5, pp. 735–744, 2001.

[6] K. Wagstaff, "Clustering with Missing Values: No Imputation Required," in *Proceedings of the Meeting of the International Federation of Classication Societies*. Springer, Berlin, Heidelberg., 2004, pp. 649–658.

[7] L. Himmelspach and S. Conrad, "Clustering Approaches for Data with Missing Values: Comparison and Evaluation," in *2010 Fifth International Conference on Digital Information Management (ICDIM)*. IEEE., 2010, pp. 19–28.

[8] D. Mesquita, J. Gomes, and L. Rodrigues, "K-means for Datasets with Missing Attributes Building Soft Constraints with Observed and Imputed Values," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges (Belgium)., 2016, pp. 599–604.

[9] M. Kargar, H. Izadkhah, and A. Isazadeh, "Tarimliq: A New Internal Metric for Software Clustering Analysis," in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. IEEE., 2019, pp. 1879–1883.

[10] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of Internal Clustering Validation Measures," in *2010 IEEE International Conference on Data Mining*. IEEE., 2010, pp. 911–916.

[11] W. Krzanowski, "Cross-validation in Principal Component Analysis," *Biometrics*, vol. 43, no. 3, pp. 575–584, 1987.

[12] K. Gabriel, "Le Biplotoutil Dexploration de Donnees Multidimensionnelles," *Journal de la Societe Francaise de Statistique*, vol. 143, no. 4, pp. 5–55, 2002.

[13] P. Perry, "Cross-validation for Unsupervised Learning," Department of Statistics Stanford University, Tech. Rep., 2019. [Online]. Available: _https://arxiv.org/pdf/0909.3052

[14] W. Yan, "Biplot Analysis of Incomplete Two-way Data," *Crop Science*, vol. 53, no. 1, pp. 48–57, 2013.

[15] R. Ananda, M. Nafan, A. Arifa, and A. Burhanuddin, "Sistem Rekomendasi Pemilihan Peminatan Menggunakan Density Canopy K-Means," *Jurnal RESTI*, vol. 4, no. 1, pp. 172–179, 2020.

[16] T. Cerah, O. Nurhayati, and R. Isnanto, "Perbandingan Metode Segmentasi K-Means Clustering dan Segmentasi Region Growing untuk Pengukuran Luas Wilayah Hutan Mangrove," *Jurnal Teknologi dan Sistem Komputer*, vol. 7, no. 1, pp. 31–37, 2019.

[17] R. Ananda and A. Burhanuddin, "Analisis Mutu Pendidikan Sekolah Menengah Atas Program Ilmu Alam di Jawa Tengah dengan Algoritme K-Means Terorganisir," *Inista*, vol. 2, no. 1, pp. 65–72, 2019.

[18] R. Ananda, "Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency," *Inista*, vol. 2, no. 1, pp. 65–72, 2019.

[19] A. Khairati, A. Adlina, G. Hertono, and B. Handari, "Kajian indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," in *Prosiding Seminar Nasional Matematika (PRISMA)*. Jurusan Matematika FMIPA UNNES., 2019, pp. 161–170.

[20] J. Baarsch and M. Celebi, "Investigation of Internal Validity Measures for K-Means Clustering," in *Proceedings of the International MultiConferenceof Engineers and Computer Scientists*. IMECS 2012, Hongkong., 2012.

[21] L. Vendramin, R. Campello, and E. Hruschka, "On the Comparison of Relative Clustering Validity Criteria," in *Proceedings of the SIAM International Conference on Data Mining*. Sparks, Nevada, USA., 2009.

[22] N. Peladeau, C. Dagenais, and V. Ridde, "Concept Mapping Internal Validity: A Case of Misconceived Mapping?" *Evaluation and Program Planning*, vol. 62, no. 17, pp. 56–63, 2017.

[23] J. Gower, "A General Coefficient of Similarity and Some of its Properties," *Biometrics*, vol. 24, no. 7, pp. 857–871, 1971.