MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 6 | No. 1 | Halaman 16 - 30
DOI : https://doi.org/10.26760/mindjournal.v6i1.16 Juni 2021

Sequence Clustering in Process Mining for Business Process Analysis Using K-Means

NUR FITRIANTI FAHRUDIN

Program Studi Sistem Informasi, Institut Teknologi Nasional Bandung, Indonesia
Email : nurfitrianti@itenas.ac.id

Received 8 Maret 2021 **|** *Revised* 22 April 2021 **|** *Accepted* 12 Juni 2021

ABSTRAK

Proses Discovery merupakan teknik utama dalam proses mining yang bertujuan untuk menghasilkan sebuah model dari event log. Namun dalam implementasinya ditemukan masalah, karena banyak varian proses yang terdapat pada event log. Hal ini membuat hasil proses discovery sulit untuk dipahami. Penelitian ini di awali dengan mengelompokan event log menggunakan metode K-Means sebagai tahap pre-processing. Hasil dari tahap pre-processing ini kemudian di modelkan menggunakan teknik proses mining. Namun, pada saat metode K-Means ini di terapkan penentuan jumlah cluster yang optimal sangatlah penting. Kesalahan dalam menentukan nilai K dapat menurunkan nilai fitness dan precision dari model yang dihasilkan. Berdasarkan hasil pengujian pada data set issue tracking dengan jumlah case 1091 dan jumlah event 7924 yang terbagi ke dalam empat cluster nilai precision meningkat dari 0,49 menjadi 1 dan nilai fitness meningkat dari 0,34 menjadi kisaran 0,61-1 pada cluster 2, 3 dan 4.

***Kata kunci**: K-Means, proses mining, event log, clustering, sequence clustering*

ABSTRACT

Process Discovery as the main technique in the mining process aims to produce a model of an event log. However, in the implementation, there is a problem found, for a lot of process variants contained in the event log. This makes the results of the discovery process difficult to understand. This research begins by grouping event logs using the K-Means method as a pre-processing stage. The results of this pre-processing stage are then modeled using the process mining technique. However, determining the optimal number of clusters is crucial. Mistakes in determining the K value can reduce the fitness value and precision of the resulting model. Based on the test results on the issue tracking data set with the number of cases 1091 and the number of events 7924 which is divided into four clusters the precision value increased from 0.49 to 1 and the fitness value increased from 0.34 to 0.61-1 in clusters 2, 3 and 4.

***Keywords**: K-Means, process mining, event log, clustering, sequence clustering*

1. INTRODUCTION

In the era of information technology, data are processed using an information system that is capable of recording every activity stored in the database. These data can be used as analytical material, for companies to make strategic decisions better. Companies can use the data, as an evaluation material both to improve business process performance or can be used to monitor business process performance. Before making improvements to business processes, several things need to be done, including choosing a business process that will be improved. The next step is to understand the business process in detail. The process of business understanding is commonly referred to as process discovery **(Dumas, La Rosa, Mendling, & Reijers, 2013)**.

In recent years a method has been developed to automate process discovery, namely ABPD (Automated Business Process Discovery). ABPD is also known as process mining. This study proposes an analytical model with a process mining technique approach. Process mining techniques are used to analyze and evaluate the companies business processes **(Bolt, de Leoni, & van der Aalst, 2018; Omar AlShathry, 2016; Tax, Lu, Sidorova, Fahland, & van der Aalst, 2018).** Process mining is relatively new research that combines learning machines and data mining on one hand and process modeling and analysis on the other **(Lu, Fahland, & van der Aalst, 2015)**. The process mining aims to extract knowledge or information from a file or historical data contained in the Information System **(van der Aalst, 2016)**. As has been known, this information system automates many steps in the business process that were previously done manually.

However, the implementation of process mining in the real world still faces obstacles **(Liu, Alshangiti, Ding, & Yu, 2018)**. Many process mining techniques work well when the data used has a good structure, on the other hand when data have a high noise level, it is difficult to find useful information. The model generated from this kind of log becomes difficult to analyze **(Liu et al., 2018; Veiga & Ferreira, 2010).** This will create a model that looks like spaghetti **(Veiga & Ferreira, 2010)**. One of the ways to overcome the problem is by reducing the number of cases to be analyzed **(Rebuge & Ferreira, 2012)**.

Clustering is a data mining method that performs unsupervised modeling. Jung and Kang compared two clustering algorithms; K-Means and Expectation-Maximization (EM) **(Jung, Kang, & Heo, 2014)**. EM is a method that aims to group data based on probability **(Gyu, Soo, & Heo, 2014)**. There are still several methods of data grouping, one of which is K-Means. K-Means is similar to EM **(Jung et al., 2014)**, if the EM data is mapped based on the probability value then the K-Means data is mapped based on distance. According to Jung and Kang, the K-means method has a performance comparable to the EM method **(Jung et al., 2014)**. They explained that the K-means method is more accurate than EM but the K-means method is more time consuming than EM **(Jung et al., 2014)**.

Several researchers have conducted studies related to sequence clustering. Rebudge et al present the Markov-chain and Expectation-Maximization algorithms for classifying event-logs **(Rebuge & Ferreira, 2012)**. They map each "sequence" into the cluster that has the highest probability. Richard and Hyejin use sequence clustering for credit risk analysis. They use a sequence matrix to group similar company transition behaviors, and firms that show momentum in their transition will be part of the same cluster **(Le, Ku, & Jun, 2021)**. Their study uses the K-means clustering method, and the results show that the clustering model can predict better.

Thus, by reflecting on existing theories and studies, the topic of sequence clustering is raised as the core of this research, which specifically examines the effect of the number of clusters on the K-Means method for modeling business processes with process mining.

This research developed a method for sequences clustering into several groups based on their probability values. The purpose of this research is to create a simpler process model so that the model is easier to analyze. Rebudge et al. use an Expectation-Maximization algorithm to clustering the event log. Therefore in this study, K-means is proposed as a method for conducting sequence clustering and the final results are compared with EM. It aims to see which model is better. This paper also discusses methods for determining the optimal number of clusters into which data can be grouped. Three methods were compared, they are DBI, Silhouette Coefficient, and Elbow.

This paper is organized as follows: Section I introduces, Section II provides the method of sequence clustering, Section III presents the results and discussion of sequence clustering applications in the mining process preprocessing stage, finally, Section IV conclusion.

2. RESEARCH METHOD

Sequence clustering has been introduced previously in addition to the methodology for conducting BPA (Business Process Analysis) **(Rebuge & Ferreira, 2012)**. Sequence clustering aims to reduce the amount of data contained in the event log and can be used as a preprocessing step. The application of sequence clustering can produce a simpler model for each cluster. The methodology consists of stages are the preparation of an event log, log inspection, sequence clustering analysis, control-flow analysis, performance analysis, organizational analysis, transfer of results.

As mentioned previously in the introduction, Sequence clustering is a sub-methodology that includes techniques for grouping event logs and as a pre-analyze stage process **(Rebuge & Ferreira, 2012)**. In this paper, this technique is adopted as one part of the process in methodology. To conduct business process analysis, not all stages of the BPA process used by previous researchers were applied. Organizational aspects are not analyzed, this research only focuses on aspects of control flow or process discovery and performance analysis. The following in figure 1 is the stage for conducting control flow analysis.

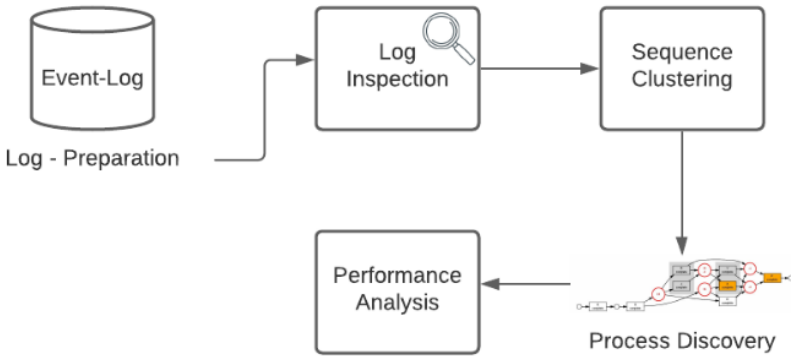

Figure 1. Bussines Process Methodology

The first stage in log preparation is identifying the primary processes, where the final results of this stage are used to determine the attributes in the event log. Inspection logs provide initial analysis such as the number of events involved, number of cases, where this information is useful to support the next stage. Next, the event log is partitioned into smaller groups in the sequence clustering stages. Process Discovery aims to find the pattern of business processes contained in the event log. The patterns of business processes are then evaluated to assess the performance of the results of the modeling that has been done.

2.1. Preprocessing

In this section, the main focus is on preparing event logs before grouping them. Some of the steps taken before sequence grouping were log preparation and log inspection. The log inspection stage is important for this stage because at this stage an analysis of the things contained in the event log is carried out. This log inspection usually gives first impressions about event logs, such as a number of cases, a number of events, start events, end events, etc **(Fukui, Okada, Satoh, & Numao, 2019)**.

To make an event log can be clustered, manipulation is needed to make the event log meet the following characteristics.

1. Event logs are converted into a group of sequences where a caseId represents one sequence.
2. Each sequence is represented with orderly discrete symbols
3. Each symbol represents each activity or event Id
4. Transform the format of an event log, which initially had a vertical format to horizontal

After the event log was manipulated, several steps need to be done to calculate the probability value of each sequence. The probability value of each of these sequences is represented by a first-order Markov chain. The steps used in finding the probability value of the sequence are as follows:

1. Calculation of single activity appearance frequency
   To find the frequency of 1-sequence, the first thing to do is to register all the items in the event log. Each item or discrete symbol will be calculated for the number of occurrences or frequency.
2. Calculation of two consecutive activities appearance frequency
   To find a 2-sequence frequency, the first thing to do is the same as step 1, which is to look for the 2-sequence combination found in the event log. Next is to calculate the number of occurrences or the frequency of each 2-sequence combination.
3. Calculation of Transition Matrix
   The next step is to calculate the probability of 2-sequence impressions in Equation 1.

$$P(E) = \frac{X}{N} \tag{1}$$

   N is the frequency of the previous event and X is the two consecutive activities' appearance frequency. The result for this step is the NxN matrix which represents the first-order Markov chain.
4. Calculation of two consecutive activities probability
   The next step is to calculate the sequence probability for each case impression in Equation 2.

$$P(X) = \prod_{i=2}^{L} P(x_i|X_{i-1}) \quad (2)$$

2.2. Sequence Clustering

Event logs consist of a set of sequences, where each case has a sequence **(Ferreira, Zacarias, Malheiros, & Ferreira, n.d.)**. Clustering aims to separate or segment data into several clusters, based on certain characteristics where the label of each data is unknown **(Putu et al., 2014)**. Sequence clustering is a bioinformatics technique that aims to classify sequence data sets that have similarities **(Le et al., 2021)**. A set of sequences represents an event log, which is also called as a trace **(Song, Günther, & Van Der Aalst, 2009)**. Therefore, the grouping sequence is also called trace clustering or sequence clustering **(Rebuge & Ferreira, 2012)**.

In the previous stage, each sequence had been calculated for the frequency of its appearance. Based on this frequency value the sequence will be grouped. The algorithm used in this study was the K-Means algorithm. To determine the right number of groups (K), internal validity is carried out to calculate how well the grouping has been done. Several methods can be used to determine the K value for data including DBI, Silhouette Coefficient, and Elbow method **(Prasetyo, 2014)**. Several experiments have been carried out to see which method is the most appropriate for determining the value of K.

For DBI values, the smaller the resulting value, the better the grouping results obtained **(Kumar Singh, Mittal, Malhotra, & Srivastava, 2020)**. Unlike the case with DBI in Silhouette Coefficient, the higher a value indicates that the data is right in a cluster **(Dinh, Fujinami, & Huynh, 2019)**. The third method used is the elbow method. Unlike the other two methods, if the DBI and Silhouette Coefficient, the writer can immediately see the best K value proposed by both methods, the elbow method will identify the best number of clusters by looking at the visualization provided by this method.

After each sequence has been allocated to the group, sequences will be separated into different CSV documents according to the available groups. This CSV document will then be converted into XES format (eXtensible Event Stream) using ProMimport **(van der Aalst, 2016)**. ProMimport is a tool that supports conversion from various data sources such as CSV documents to XES **(Duling & Gao, 2016)**.

3. RESULTS AND DISCUSSION

In this section, the results of the research are explained and at the same time is given a comprehensive discussion. The results can be presented in figures, graphs, tables and others that make the reader understands easily. The discussion can be made in several sub-chapters.

To assess how well the methodology was designed, two different data sets were used as shown in Table 1. Percetakan Bandung is one of the companies engaged in printing services. Transactions that occur in the printing, are recorded in a database. The event log was gained from a query in the application involved in the case study Percetakan Bandung. The issue tracking dataset was taken from Google Chromium Issue Tracking.

Table 1. Data Test

No	Data	Jumlah Case	Jumlah Event
1	Percetakan Bandung	30	131
2	Issue Tracking	1091	7924

In the sequence clustering stage, two attributes are used, that is caseId and activity. The activity attribute describes the activity that has been executed while the caseId attribute functions as an identifier where one caseId consists of several activities or events. The timeStamp attribute was used as a time indicator when the activity was executed. As preparation for the sequence clustering event log phase, it will be grouped based on caseId and activity sorted in ascending order or sorted by timeStamp. Below in Table 3 are the results of queries and manipulations that have been carried out on the event log which are presented in Table 2.

Table 2. Printing Bandung Event-Log

caseId	*eventId*	*activity*	*timeStamp*
1	1	Order (O)	2017-11-30;10:00
1	2	Setting (S)	2017-12-02;10:00
1	4	Production_kbr (K)	2017-12-05;16:00
1	8	Pengiriman (P)	2017-12-06;16:00
2	1	Order (O)	2017-11-30;09:00
2	2	Setting (S)	2017-11-30;10:00
2	4	Production_kbr (K)	2017-12-04;10:00
2	6	Finishing_kbr (X)	2017-12-06;10:00
2	8	Pengiriman (P)	2017-12-08;10:00
3	1	Order (O)	2017-11-30;10:00
3	2	Setting (S)	2017-12-02;10:00
3	5	Production_pgr (R)	2017-12-06;16:00
3	8	Pengiriman (P)	2017-12-07;16:00
....	...	...	...
30	1	Order (O)	2017-11-30;10:00
30	2	Setting (S)	2017-12-04;10:00
30	3	Production_jkt (J)	2017-12-05;10:00
30	6	Finishing_kbr (X)	2017-12-13;15:00
30	8	Pengiriman (P)	2017-12-13;16:00

Table 3. Presentation of Event Log in Horizontal Form

Case	Event
1	OSKP
2	OSKXP
3	OSRP
4	OSKXP
5	OSKXP
6	OSKXP
...	...
30	OSJXP

The next step was to calculate the probability of each sequence. To calculate the probability value of each sequence three main processes needed to be done, namely calculating the frequency of 1-sequence and 2-sequence, and calculating the probability value of 2-sequence that represents the transition matrix from the first-order Markov chain. The 1-sequence frequency had been obtained from the results of log inspection. Whereas the 2-sequence frequency is presented in Table 4.

Table 4. Two Consecutive Activities Appearance Frequency

2-sequence	KP	XP	SJ	KX	YP	RP	SR	SK	SX	RX	OS	JY	JX
Appearance	5	10	12	6	3	4	5	11	2	1	30	3	1

Based on the results obtained from the input log and Table 4, the next was to calculate the transition matrix using Equation (1). The following in Table 5 is a transition matrix for Percetakan Bandung case studies.

Table 5. Transition Matrix

	J	**R**	**S**	**P**	**K**	**Y**	**X**
O			1.0				
J				0.7		0.25	0.05
R				0.8			0.2
S	0.4	0.17			0.36		0.07
P							
K				0.45			0.55
Y				1.0			
X				1.0			

Using this transition matrix, the probability for each sequence can be calculated. To calculate the probability of each sequence, Equation (2) was used. Data that had a probability value were then grouped using the K-Means method. When using the K-Means method, the parameter that must be entered was the K value or the number of clusters **(Cornell & Sastry, 2015)**. In determining the right number of clusters, several methods were used including the DBI method, the Silhouette Coefficient, and the Elbow.

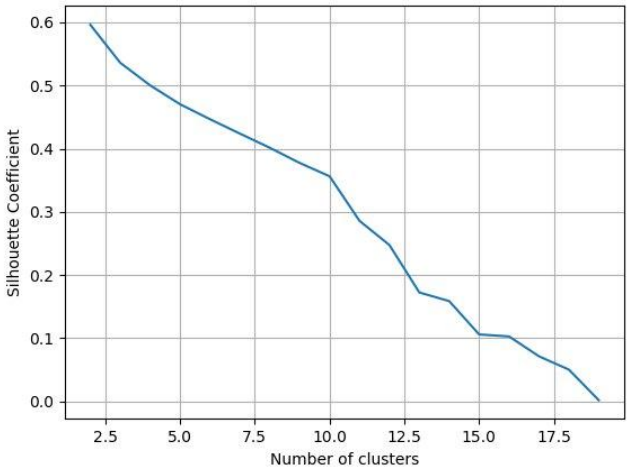

Figure 2. Silhouette Coefficient Method For K Value Optimizing

The K value was obtained from the Silhouette Coefficient calculation shown in Figure 2. It shows that the best K value is 2. In addition to using the Silhouette Coefficient, the DBI matrix is also used to evaluate the number of the best clusters where several experiments were carried out using different K values. The results of this experiment are presented in Table 6. Table 6 shows the lowest DBI value is a cluster with a value of K = 2.

Table 6. Davies-Bouldin Index Method For K Value Optimizing

No	K	DBI
1	2	2.0
2	3	6.96064555096535
3	4	6.7633971188310325
4	5	16.009365362500493

Previously, it has been discussed that three methods were used to determine the best K value. The Elbow method provides information in the form of a graph to determine the value of K. The following Figure 3 is a graph to identify the most optimal K value from this study case.

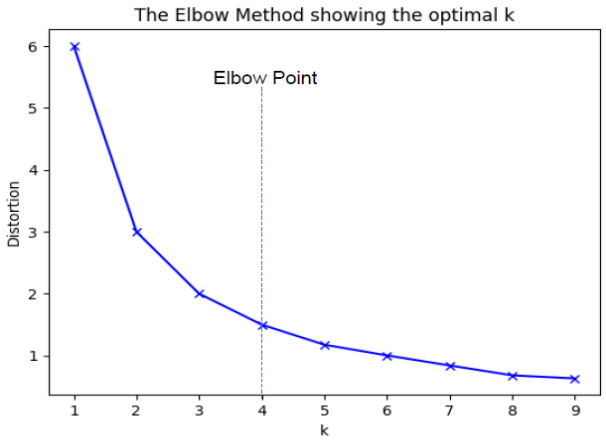

Figure 3. Elbow Method For K Value Optimizing

In Figure 3, the distortion goes down rapidly with K from 1 to 2, from 2 to 3, and from 3 to 4, and then the distortion goes down very slowly after that. Based on the explanation from Bholowalia et.al, the optimal number of clusters should be around 4 **(Bholowalia & Kumar, 2014)**. There was a difference when determining the value of K using the elbow method and the two previous methods DBI and Silhouette Coefficient. Therefore, this research compares the data set which is divided into 2 clusters and 4 clusters. After clustering on the event log was complete, the event log will split into four and two CSV documents.

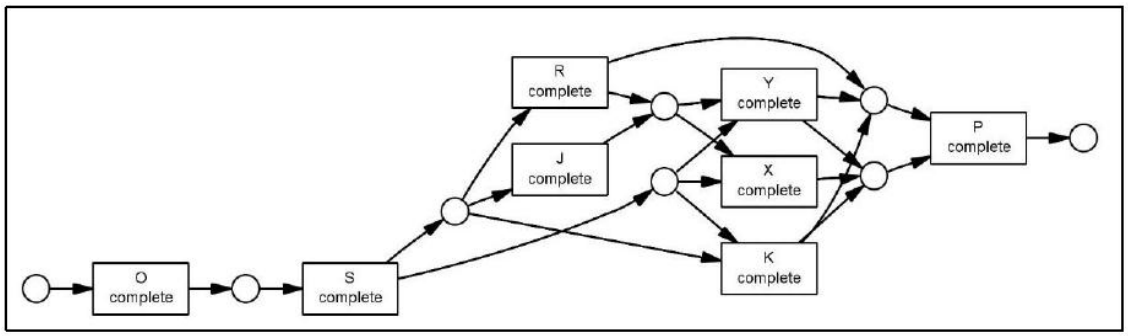

Figure 4. The Example Process Model Before Clustered

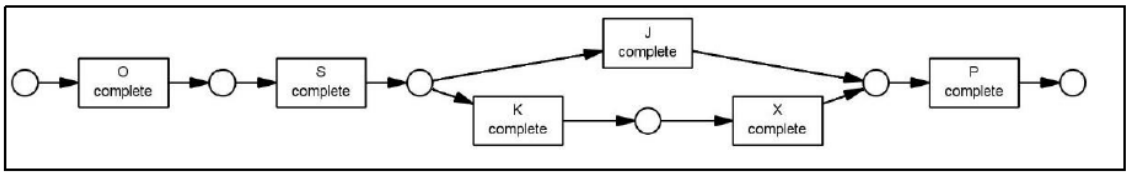

Figure 5. Process Model Base on Some Case After Implement Sequence Clustering

To test clustering results, first, the event log is modeled using the alpha miner method. This aimed to see whether the model generated from the results of this clustering was better when compared to the model generated from the event log that was not carried out by the previous clustering process. Based on the original event log, the alpha algorithm construct the model Petri net depicted in Figure 4 and Figure 5 shows the model process from one of the event log that has been clustered. To analyze the performance of the model generated by the alpha miner, further conformance checks can be executed.

The model and the event log are said to be "fit" or "fit" if the process model can replay each sequence in the event log **(Mannhardt, de Leoni, Reijers, & van der Aalst, 2016)**. In other words, the model must be able to describe each event in the sequence. However, a good fitness value did not imply conformance **(Rozinat & van der Aalst, 2008)**. The model that has a fitness value of 1 did not necessarily produce meaningful information. Therefore, in addition to calculating the fitness value, the precision value is also used. Precision or Behavioral Appropriateness aimed to evaluate how precisely a model describes the process observed. Precision evaluated how many behaviors were allowed by the actual model that never existed in the event log **(Premchaiswadi & Porouhan, 2015; Tax et al., 2018)**. The resulting model should not display actions that were not related to what was contained in the event log.

To evaluate the results of this study, a comparison was made with previous studies. The test was carried out on the data in Table 1 by using the sequence clustering method which had been used by other researchers namely the Markov Chain and Expectation-Maximization method **(Sübakan, Kurt, Cemgil, & Sankur, 2014; Veiga & Ferreira, 2010)**. Testing of previous studies was carried out using the plug-in sequence clustering in ProM 5.2. This plug-in implemented Markov Chain and Expectation Maximization to a cluster.

The following aspects are used to test the proposed methodology:

1. The consistency of clustering results carried out using the methodology proposed in this thesis compared to the clustering produced by the comparison method.

2. The quality of the process model generated by the event log, both those that have been through the clustering stage and those that have not been clustered. Testing was done using the fitness and precision dimensions.

3. Increased fitness value and precision generated by proposed methodology and EM methods.

4. Determination of the optimal number of clusters. The test is done by comparing data that has been grouped into 2 clusters based on the results of the Silhouette Coefficient calculation and data that has been grouped into 4 clusters based on the Elbow method to see which method can produce a better model.

The following are the results of testing based on the aspects mentioned:

1. Consistency of Clustering Results

At this stage, the consistency of the clustering results generated by the methodology proposed in this study will be tested, then compared with the clustering method using Markov Chain and EM. To test for consistency, 5 clustering experiments were performed on the data set using the same K values. The results of clustering using the EM method can be seen in Table 1.The results of grouping using the proposed methodology, getting consistent or the same results in each trial. The sequence distribution mapped into each cluster was always the same. The result of Sequence distribution in the five experiments with the same K value was always consistent which was indicated by 16 cases distributed on cluster 0 and 14 cases distributed on cluster 1.

The results of the proposed methodology are different from the results of sequence clustering using the EM method, the distribution results of the five experiments are not consistent as shown in Figure 6. The first experiment showed 19 cases in cluster 0 and 11 cases in cluster 1. However, the second experiment got 18 cases in cluster 0 and 12 cases in cluster 1. As indicated in other experiments, the results of every distribution were inconsistent. Inconsistent distribution of cases will make it difficult for researchers to choose which cluster results will be used to model the process.

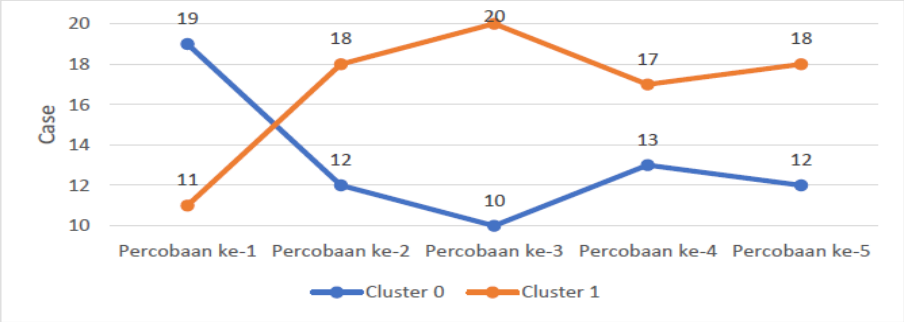

Figure 6. The Graph Shows the Inconsistency of the Results of Clustering with the EM Method for Cluster 0 and Cluster 1 in 5 Trials

2. Quality Model after Clustering

At this stage, the testing was carried out to see the effect of the application of sequence clustering with K-Means at the preprocessing stage on the resulting model presented in

Table 7. The event log was divided into two clusters by following the results of the calculation of the DBI and Silhouette Coefficient methods.

Table 7. The Results of testing the process Model of Sequence Clustering K=2

Data	Alpha		ILP		Inductive		Heuristic	
	f	***p***	***f***	***p***	***f***	***p***	***F***	***p***
Non Cluster	0.9	0.88	1	0.87	0.99	1	0.99	0.88
Cluster 0	0.92	0.77	1	0.67	1	0.92	1	0.97
Cluster 1	1	1	1	1	1	1	1	1

As observed from Table 7, the model generated from the original event log (non-cluster) has a fitness of 0.9 and the precision of the model is 0.88. However, the model produced by the event log that has been clustered in this case "cluster 0" has a smaller precision value compared to the previous model (non-cluster). This was caused by the selection of an incorrect K value.

3. Comparison of model quality based on number of clusters

Previously in Table 7, the test results were obtained where the event log was divided into two clusters. The results show a decrease in the precision value of cluster 0. In this section, the process model will come from the event log that was divided into four clusters. As discussed earlier, there was a difference between the results obtained from the three methods proposed to look for K.

Table 8. The Results of Testing the Process Model of Sequence Clustering K=4

Data	Alpha		ILP		Inductive		Heuristic	
	f	***p***	***f***	***p***	***f***	***p***	***f***	***p***
Non Cluster	0.9	0.8	1	0.87	0.99	1	0.99	0.87
Cluster 0	0.95	1	1	0.94	1	1	1	1
Cluster 1	1	1	1	1	1	1	1	1
Cluster 2	1	1	1	1	1	1	1	1
Cluster 3	0.94	1	1	0.95	1	1	1	0.95

The results of sequence clustering using K-Means will be presented. The value of K = 4 was obtained from the identification using the Elbow method as illustrated in Figure 3. The following data in Table 8 presents the results of testing the fitness and precision values for the 4 clusters. As can be seen from Table 8 fitness and precision value of each cluster have increased compared with the model from the original event log (non-cluster).

4. Comparison of EM and K-Means Results

In this section, the comparison results of the clustering process model using K-means and EM are presented. In this experiment, using the second data was Tracking Issues. The

experiment conducted had divided the event log into 4 clusters, this was based on the results obtained from the identification using the Elbow method. The result from the proposed methodology showed better performance compared with EM as shown in Figure 7. The precision value from the model was clustered with EM has decreased. However, that is not the case in the model generated by the proposed methodology.

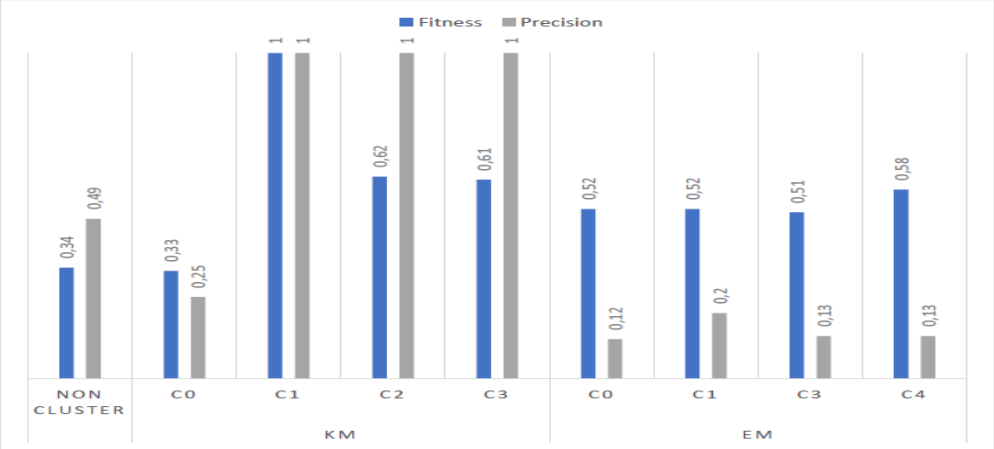

Figure 7. Result of Comparison Diagram K-Means dan EM Clustering

3.1 Discussion

Experiments that have been carried out using the proposed methodology have been shown to increase the fitness value and precision of the model generated by the discovery process. This can be seen in the results presented in Table 7 and Table 8. As shown in Table 7, there is a decrease in precision values. The decrease in precision values occurred in the model produced by almost all methods, that was Alpha, ILP, and Inductive Miner. However, in the Heuristic Miner method, the value of precision has increased. This was because Heuristic-miner was an improvement of the alpha-miner. Based on Table 8, showed that the application of sequence clustering with the value K = 4 in the pre-processing stage was proven to help improve the fitness value and precision of the resulting model. If Table 7 decreased the value of precision, this did not occur in the results presented in Table 8. The result had shown an increase. The conclusion inferred from this experiment is the fault in determining the value of K affects the results of clustering. Determining inappropriate K values can result in a lack of precision in the process model that results from the discovery process.

The graph shown in Figure 7 showed a significant increase in the results of sequence clustering using K-Means. Clusters 1, 2, and 3 had precision values increased from 0.49 to 1. However, cluster 0 experienced a decrease in fitness value from 0.34 to 0.32 and a decrease in precision from 0.49 to 0.24. Sequence clustering results using Markov chains and EM had decreased the precision value in all clusters. For cluster 0, it decreased by 0.37, cluster 1 by 0.30 last cluster 3 and 4 decreased by 0.366 points. However, the poor results on the event log clustered using EM are also caused by inconsistent clustering results, as shown by Figure 6. In this test, only do one experiment. If it was tested in another experiment, the results obtained would not be the same. Actual results could improve if the results of the EM distribution were taken in other trials, but this would be time-consuming as it is difficult to know which distribution would produce a model with good performance. The difference in the results of clustering in the EM method was because the determination of data points was

randomly selected as centroid from each cluster. This caused the transition matrix obtained to be different in each experiment, so the mapping of each data changed in each experiment.

4. CONCLUSION

In this study, a methodology was developed to perform sequence grouping. This methodology makes it easier for analysts to analyze the process model because it provides a simpler process model and is presented in models divided into several groups. Based on the implementation results of Percetakan Bandung and Tracking Issue data, it shows that adding sequence clustering before the discovery process can make a better process model. The process model generated in Table 7 shows an increase in the fitness value from 0.34 to 1 and the precision value from 0.49 to 1 in several clusters when compared to the model generated from the event log that did not go through the clustering process.

Models with high fitness and precision values indicate that the resulting model is able to generate event logs, in other words, the resulting model has a high level of accuracy. However, the success of implementing sequence clustering also depends on how to determine the correct number of clusters. Therefore, choosing the right method to determine the value of K is very important. This study focuses on the implementation of sequence clustering as a preprocessing step. It seems that there are still obstacles in determining the initial cluster for the K-means and EM methods. In the future, further methods should be found to solve this problem.

ACKNOWLEDGEMENT

I especially thank Prof. Dr. Ir. Jaka Sembiring, M.Eng. and Dr. Ir. Arry Akhmad Arman, M.T. for constructive criticism and advice of my thesis.

REFERENCES

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 975–8887. https://doi.org/10.5120/18405-9674

Bolt, A., de Leoni, M., & van der Aalst, W. M. P. (2018). Process variant comparison: Using event logs to detect differences in behavior and business rules. *Information Systems*, *74*, 53–66. https://doi.org/10.1016/j.is.2017.12.006

Cornell, D., & Sastry, S. (2015). *Performance Comparison of K-Means and Expectation Maximization with Gaussian Msixture Models for Clustering*. (May). Retrieved from https://dcornellresearch.org/2015/10/30/performance-comparison-of-k-means-and-expectation-maximization-with-gaussian-mixture-models-for-clustering/

Dinh, D.-T., Fujinami, T., & Huynh, V.-N. (2019). Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient. *Communications in Computer and Information Science*.

Duling, D. R., & Gao, E. Y. (2016). *Improve Your Business through Process Mining*. 1–17.

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). Fundamentals of Business Process Management. In *Quantitative Process Analysis*. https://doi.org/10.1007/978-3-642-33143-5

Ferreira, D., Zacarias, M., Malheiros, M., & Ferreira, P. (n.d.). *Approaching Process Mining with Sequence Clustering : Experiments and Findings*. (1), 1–15.

Fukui, K. ichi, Okada, Y., Satoh, K., & Numao, M. (2019). Cluster sequence mining from event sequence data and its application to damage correlation analysis. *Knowledge-Based Systems*, *179*, 136–144. https://doi.org/10.1016/j.knosys.2019.05.012

Gyu, Y., Soo, M., & Heo, J. (2014). ARTICLE ; BIOINFORMATICS Clustering performance comparison using K -means and expectation maximization algorithms. *Biotechnology & Biotechnological Equipment*, *28*(1), 44–48. https://doi.org/10.1080/13102818.2014.949045

Jung, Y. G., Kang, M. S., & Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnology and Biotechnological Equipment*, *28*(1), S44–S48. https://doi.org/10.1080/13102818.2014.949045

Kumar Singh, A., Mittal, S., Malhotra, P., & Srivastava, Y. V. (2020). Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*.

Le, R., Ku, H., & Jun, D. (2021). Sequence-based clustering applied to long-term credit risk assessment. *Expert Systems with Applications*, *165*(August 2020), 113940. https://doi.org/10.1016/j.eswa.2020.113940

Liu, X., Alshangiti, M., Ding, C., & Yu, Q. (2018). Log sequence clustering for workflow mining in multi-workflow systems. *Data and Knowledge Engineering*, *117*, 1–17. https://doi.org/10.1016/j.datak.2018.04.002

Lu, X., Fahland, D., & van der Aalst, W. M. P. (2015). Conformance checking based on partially ordered event data. *Lecture Notes in Business Information Processing*, *202*, 75–88. https://doi.org/10.1007/978-3-319-15895-2_7

Mannhardt, F., de Leoni, M., Reijers, H. A., & van der Aalst, W. M. P. (2016). Balanced multi-perspective checking of process conformance. *Computing*, *98*(4), 407–437. https://doi.org/10.1007/s00607-015-0441-1

Omar AlShathry. (2016). Journal of Computer Engineering & Information Technology Process Mining as a Business Process Discovery Technique. *Journal of Computer Engineering & Information Technology*.

Prasetyo, E. (2014). *DATA MINING Mengolah Data Menjadi Informasi Menggunakan MATLAB*.

Premchaiswadi, W., & Porouhan, P. (2015). Process modeling and bottleneck mining in online peer-review systems. *SpringerPlus*, *4*(1). https://doi.org/10.1186/s40064-015-1183-4

Putu, N., Merliana, E., Studi, P., Teknik, M., Industri, F. T., & Jaya, U. A. (2014). Analisa Penentuan Jumlah Kluster Terbaik Pada Metode K-means Klustering. *Prosiding Seminar Nasional Multidisiplin Ilmu Dan Call For Paper Unisbank*, 978–979.

Rebuge, Á., & Ferreira, D. R. (2012). Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems*, *37*(2), 99–116. https://doi.org/10.1016/j.is.2011.01.003

Rozinat, A., & van der Aalst, W. M. P. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, *33*(1), 64–95. https://doi.org/10.1016/j.is.2007.07.001

Song, M., Günther, C. W., & Van Der Aalst, W. M. P. (2009). Trace clustering in process mining. *Lecture Notes in Business Information Processing*, *17 LNBIP*, 109–120. https://doi.org/10.1007/978-3-642-00328-8_11

Sübakan, Y. C., Kurt, B., Cemgil, A. T., & Sankur, B. (2014). Probabilistic sequence clustering with spectral learning. *Digital Signal Processing: A Review Journal*, *29*(1), 1–19. https://doi.org/10.1016/j.dsp.2014.02.014

Tax, N., Lu, X., Sidorova, N., Fahland, D., & van der Aalst, W. M. P. (2018). The imprecisions of precision measures in process mining. *Information Processing Letters*, *135*, 1–8. https://doi.org/10.1016/j.ipl.2018.01.013

van der Aalst, W. M. P. (2016). Process Mining. In *Process Mining Data Science in Action Second Edition* (2nd ed., pp. 30–34). https://doi.org/10.1007/978-3-662-49851-4

Veiga, G. M., & Ferreira, D. R. (2010). Understanding spaghetti models with sequence clustering for ProM. *Lecture Notes in Business Information Processing*, *43 LNBIP*, 92–103. https://doi.org/10.1007/978-3-642-12186-9_10