Find the Best Rule of Time Series Data Mining with Cluster Analysis

Relita Buaton¹, Muhammad Zarlis², Herman Mawengkang³, Syahril Efendi²

¹Graduate Program Of Computer Science,
²Department Of Computer Science
Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Medan, Indonesia

Email: relitabuaton@usu.ac.id

Abstract. Data mining is an analytical process of knowledge discovery in large and complex data sets. Many studies wish to explore data, to find information so that knowledge can be obtained through the grouping process, classification, rules discovery, associations and data mining visualization which shows similarity. Periodic data often occurs in business applications and sciences that has big size, high dimension and continuously updated. The similarity in periodic data is based on several approaches. One of common approaches is to transform periodic series into other domains so that dimensions are reduced, followed by index mechanism. Many studies of time series do not give optimal result because limited to extracting data not able to represent time series and its pattern which is then change into rules. Rules can be found in time series data, but they are still constrained by over fitting and difficult to present. It causes time series data and non linear function of data mining decision can’t be optimal. The basic idea in the method proposed is to do periodic discretization for subsequential formation. These sub-sequences are grouped through a measure of similarity. The simple rule-finding technique is applied to obtain hidden rules in the temporal pattern. The optimal time series data expected to generate the uncertainty trend, previously unknown and can be used to make decisions or forecasting in the future.

1. Introduction
In periodic data mining (time series data mining), the fundamental problem is how to present periodic data. One common approach is to transform periodic series into other domains so that dimensions are reduced which are followed by an index mechanism, research on time series is not optimal because it is still limited to mining data that has not been able to represent time series [1], able to find patterns in time series data [2], this pattern needs to be developed to change the pattern into a rule. Rule can be found from time series data, but it is still constrained by over fitting [3]. Furthermore, the size of the similarity in the periodic sequence or its sub sequences and the segmentation process are the two main tasks for various tasks included in periodic mining. One of the mining tasks is the rule of discovery (rule discovery). A periodic series can be converted into a discrete representation by first forming a sub-sequence and then clustering these sub sequences using the appropriate size of the pattern of similarity. The development of computer systems is increasing very rapidly in generating and collecting data which can be seen from the side of the application of computerized systems that continue to increase transaction data in the business world and in government systems, as well as the
ability of hardware to store data with large capacity. Collection of databases produced from various sources in business management and government administration as well as the application of various other applications. With the development of a very fast database size that can affect the performance and capabilities of database systems. This creates a new technique in managing a collection or stack of data from large database transaction data to extract valuable information from previously unknown data bases. Therefore Data Mining is very important to be the object and study of current research [4]

2. Cluster Analysis
Cluster analysis is the process of partitioning data sets or observations into several subset. Each subset is a cluster, so objects that are in one cluster are similar to each other, but different from objects in other groups. Clusters of clusters generated from cluster analysis can be referred to as clustering, different grouping methods can produce different clustering in the same data set. Partitioning system is done by clustering algorithm. Therefore, grouping is useful because it can produce the discovery of previously unknown groups in the data. Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, web search, biology, and security. In business intelligence, grouping can be used to organize large numbers of customers into groups, where customers in groups have similar characteristics.

To find clusters in a graph that is by cutting the graph into several parts, each part becomes a cluster, formally for a graph $G = (V, E)$, for cut $C = (S, T)$, which is a partition of the vertex set $V$ in $G$, $V = S \cup T$ and $S \cap T = \emptyset$. The intended piece to produce a cluster is to use a minimum cut in accordance with graph theory. A piece is said to be minimum if the cut size is not greater than the size of the other pieces [1].

![Figure 1. Graph with 2 pieces](image)

3. Motives Discovery Algorithm
According to [6], provide an innovative perspective and provide optimal solutions using algorithms that are able to find time series motives or patterns through particle swarms. Taking into account data from various domains, the use of particle swarms is a competitive solution compared to previous studies because it can find comparable motives in a short time and minimal memory. Time Series is a sequence of real numbers measured sequentially, usually in the form of a regular time interval. Data in the form of time series includes data from economics to medicine, from biology to physics, and from social science to computers. Repetition of data is a fundamental characteristic of natural and artificial systems such as measurement of system activity, time series (time series) that become pairs of segments of data. This segment pair is usually called a motif [7] and their existence is impossible by chance. In fact, they usually carry important information about the system. So, the discovery of motives is very important to understand, characterize, model, and predict the system behind the time series. In addition, the discovery of motives is a core part of several high-level algorithms related to time series, in special classifications, grouping, summarization, compression, and rules of discovery algorithms. Identifying the same pair or segment motif means checking all paired comparisons between all possible segments in a time series. This, especially when dealing with time series lengths, results in too complex time and space. This study aims to provide an innovative perspective and provide optimal solutions using algorithms that are able to find time series motives / patterns through
Taking into account data from various domains, the use of particle swarms is a very competitive solution compared to previous research because it can find comparable motives in a short time and minimal memory, the right indication to find time series motifs is considered difficult to solve even for long time series. In the study, intelligent sequence sequences were combined with lower boundaries based on triangle inequality to produce the right, right, and most similar motives. The proposed algorithm is more efficient than the existing approach, including all precise and many estimates. Mueen has published an algorithm for finding long motives of variables that outperform iterative searches for optimal lengths and, from the numbers reported, also outperformed the approach further. This algorithm, called MOEN, is basically a free parameter, and is believed to be one of the most efficient motive discovery algorithms available today. However, the complexity is still quadratic within the series period, and hence its application to large-scale time series flows remains problematic. Furthermore, to reduce the lower limit used, a restricted algorithm related to measurement inequality is used to compare time series segments (Euclidean distance after z-normalization). In general, the algorithm of finding the right motive has important limitations related to the size of inequality, and is not effective because it has to change parameters at any time. In this article, it also displays the SWARM MOTIF algorithm which is another algorithm for finding the motives of a time series based on particle swarms optimization. SWARM MOTIF is very competitive when compared to previous research, because it gets comparable pairs of similarity motifs in much less time and with minimum storage requirements. In addition, SWARM MOTIF is very strong against different application choices. Particle Swarms Optimization (PSO) is a population-based stochastic approach to solve the continuous and discrete optimization problems that have been applied to multi-capital problems. This is a meta heuristic, which means that it cannot guarantee whether the solution found matches the global optimum. The original PSO algorithm cannot even guarantee local convergence to optimal, but a customized version has been proven to solve this problem. Another version guarantees optimal global convergence, but only with the number of iterations approaching infinity.

4. Research Framework

The main problem in this research is how to handle periodic data mining data series, because if only the pattern is displayed visually it will be very difficult because it includes more than thousands of observations [2]. Therefore in this study cluster analysis method was used in the time series sequence. With cluster analysis methods can find rules on periodic large-scale data mining series. The data used for the process and analysis of results in this study are numerical data. Numerical data which includes data of one unit in a periodic series. A Periodic T-sized series m is a sequence with a real value data sequence, where \( T = \{t_1, t_2, \ldots, t_m\}\) [3]. A sub-sequence with length n of the periodic series T is \( T_{i,n} = \{t_i, t_{i+1}, \ldots, t_{i+n-1}\}\), where \( 1 \leq n \leq n-m+1 \) [3]. Time series data is then processed using Matlab tools. According to [4], Detecting very similar patterns in time series is usually called a motif. Detection of repetitive or very similar patterns in time series, has proven to be very useful for researchers and practitioners. There are two motive definitions in time series, namely:

1. Based on the idea of frequency [5], interesting pattern if you have a number of significant repetitions.
2. Based on the idea of equality [6].

Both definitions are complementary, because very similar patterns are not always needed often. Apart from frequency-based definitions, ranking motives found in a time series are considered insignificant. The motives that are considered the most important are Motives with the highest count first, Motif with the second highest count, Motif with the third highest count and so on. Motives can also be assessed from statistical significance by comparing the observed numbers and expected counts under null which reflect some time series characteristics [7]. In this study found obstacles in the discovery of the motives of a time series series, namely if the motive pair of a time series has a different length because it cannot directly compare similarities or distances.
5. Clustering of Periodic Series

The grouping of periodic sequence sequences includes grouping a series of periodic sub-sequences extracted through sliding windows, which are grouping segments from one periodic to long-term series. In order to be extracted by sliding windows, the periodic series in question needs to be discretized first.[8] has shown that a periodic series with real number data can be drastically reduced without significantly affecting the available information. The method proposed for periodic discretization through windows grouping is stated as follows. Suppose that it is known as \( s \) and window with width \( w \). Given \( s = (x_1, ..., x_n) \) a window with \( b \) width \( w \) at \( s \) is sequential \( (x_s, ..., x_{s+w-1}) \).

From all windows formed (sequential) \( s_1, ..., s_{n-w+1} \) with width \( w \) where \( s_i = (x_i, ..., x_{i+w-1}) \). State \( W(s) = \{ s_i \mid i = 1, ..., n - w + 1 \} \).

Suppose there is a distance \( d(s_i, s_j) \) between 2 sequences \( s_i \) and \( s_j \) width \( w \). These distances can be used to group all subsets together \( W(s) \) into the set \( C_1, ..., C_k \). For each group \( C_k \) inserted symbol \( a_k \) and the critical version \( D(s) \) of sequences will include alphabet \( \sum = \{ a_1, ..., a_k \} \). Sequence \( D(s) \) obtained by searching for each sub sequence \( s_i \) group \( C_j(l) \) such that \( s_j \in C_j \) and by using related symbol \( a_j(l) \). So \( D(s) = a_j(l_1), a_j(l_2), ..., a_j(l_{n-w+1}) \).

Every symbol \( a_k \) presenting a basic form and what you want to get is a rule of discovery that includes a pattern formed from these basic forms. Note that the discretization process described above is very dependent on selection \( w \), selection of the periodic distance function and the type of grouping algorithm used.

6. Periodic Series Similarity

The most important issue in the discussion of periodic data mining series is the determination of similarity, that is, the degree to which a periodic series is considered resembles another periodic series. In fact the size of periodic similarity is very important for grouping [9][10][11][12]. Grouping sets \( W(s) \) it takes the meaning of distance to periodically line length \( w \). There are several possibilities and choices regarding the size of distance in the discovery of rules. The simplest choice is to treat the sequential with length \( w \) as an element of \( \mathbb{R}^w \) and then Euclid distance (ie, metric \( L_2 \) ) is used. [13] has empirically proven that Euclid's distance is hard to beat. Euclid distance is a parameter free method, fast computational time and suitable for various data mining optimizations such as indexing. The definition is, for \( \tilde{x} = (x_1, ..., x_w) \) and \( \tilde{y} = (y_1, ..., y_w) \) defined

\[
d(\tilde{x}, \tilde{y}) = (\sum (x_i - y_i)^2)^{1/2}
\]

As a metric in grouping. Other metrics include general \( L_p \) metrics defined by

\[
L_p(\tilde{x}, \tilde{y}) = (\sum (x_i - y_i)^p)^{1/p}
\]

For \( p \geq 1 \) and \( L_\infty = \max_i \mid x_i - y_i \mid \).

In a variety of uses, it is desirable to obtain a sequential form as the main factor in determining distance. That means, two subsectors can be essentially the same even though they have amplitude and different baselines. One way to achieve this is by normalizing subscribers and then applying the metric \( L_2 \) in a normalized subset. State the version of \( \tilde{x} \) the sequence normalized by \( \kappa(\tilde{x}) \), defined the distance between \( \tilde{x} \) and \( \tilde{y} \) by

\[
d(\tilde{x}, \tilde{y}) = L_2(\kappa(\tilde{x}) - \kappa(\tilde{y}))
\]
Normalization can be done by means of \( \kappa(x_i) = x_i - E\bar{x}_i \) (where \( E\bar{x}_i \) is the expectation value or average of the sequential value), which results in an average value of 0. It can also be used \( \kappa(\bar{x}) = (x - E\bar{x}) / D\bar{x} \) (where \( D\bar{x} \) is the diversity of sequences), which will force the average sequence of 0 and diversity 1.

7. Grouping Method
The first step in the discretization process is grouping. Note that \( w \) is one parameter to the system, it is used to define the set \( W(s) \). In essence, any grouping algorithm can be used to group subscriptions in \( W(s) \) as a point in \( \mathbb{R}^w \) and use \( L_2 \) metric as the distance between points. Take small constants \( d > 0 \) other parameters for grouping algorithms. For each point in \( W(s) \), the method can determine the center of grouping \( q \) so that \( d(p, q) \) minimum. If \( d(p, q) < d \) then \( p \) is added to the group whose center is \( q \), if not a new group with center \( p \) is formed. After the algorithm checks all points in \( W(s) \) suppose there is a group center \( q_1, \ldots, q_k \). It can easily be proven that the distance between the two center groups is the smallest \( d \) while the radius of each group is the largest \( d \).

8. Discussion
The optimization model used in maximizing time series data mining is cluster analysis in time series sequences. By applying the results model that has been found, the following is a discussion of the models found. As material for data simulation, data is obtained from the Time Series Data Library [14] by using variable years and temperatures, the data used from 1782 until 1988, according to the definition of data mining is a series of stages or processes to mine data so as to produce added value from a data collection in the form of knowledge that has not been known manually from a data collection or is a process to produce useful information from a large database warehouse. Then it is processed by plotting data into time series graphics

![Figure 2. Temperature Series Time Chart](image)

Figure 2 showing the results of plots of time series temperature data over time (years), it is clear that the graph is not linear with changes in shape over time each year. Visually it is very difficult to analyze if using pattern analysis, it is difficult to identify the rules contained in it and potentially interesting. The next step is dividing the graph into several windows called the time series sub sequences. Figure 3 shows that, the graph is divided into several windows, in this case divided into 10 windows and then the window will become the center of analysis.
Figure 3. Result of Window Analysis

Figure 3 shows the results of analysis for each window, each window produces a point that is obtained by calculating a similarity distance, this point is the result of the trend for the window as the time changes occur, the data for each analysis window is presented in Table 1.

### Table 1. Data Analysis Window

| Window | Year     | Similarity | Result                                                                 |
|--------|----------|------------|------------------------------------------------------------------------|
| I      | 1780-1800| 1784-15.6°C | Temperatures tend to be 15.3 °C up to 15.9 °C                           |
|        |          | 1796-15.9°C |                                                                         |
|        |          | 1789-15.3°C |                                                                         |
| II     | 1801-1820| 1816-14.6°C | Temperatures tend to be 14.6 °C up to 16.4 °C, there was a significant increase in temperature in 1801 is 16.4 °C |
|        |          | 1809-16.4°C |                                                                         |
|        |          | 1803-15.7°C |                                                                         |
| III    | 1821-1840| 1836-14.7°C | Temperatures tend to be 14.6 °C up to 15.42 °C                         |
|        |          | 1824-15.2°C |                                                                         |
|        |          | 1830-14.6°C |                                                                         |
| IV     | 1841-1860| 1856-15.1°C | Temperatures tend to be 14.4 °C up to 15.4 °C                         |
|        |          | 1842-14.4°C |                                                                         |
|        |          | 1849-15.4°C |                                                                         |
| V      | 1861-1880| 1863-15.1°C | Temperatures tend to be 14.7 °C up to 15.4 °C                         |
|        |          | 1876-14.4°C |                                                                         |
|        |          | 1870-15.4°C |                                                                         |
| VI     | 1881-1900| 1896-14.8°C | Temperatures tend to be 14.5 °C up to 14.8 °C                         |
|        |          | 1889-14.7°C |                                                                         |
|        |          | 1882-14.5°C |                                                                         |
| VII    | 1901-1920| 1900-14.4°C | Temperatures tend to be 14.4 °C up to 15.5 °C                         |
|        |          | 1902-15.5°C |                                                                         |
|        |          | 1916-14.6°C |                                                                         |
| VIII   | 1921-1940| 1923-14.6°C | Temperatures tend to be 14.6 °C up to 15.2 °C                         |
|        |          | 1936-14.9°C |                                                                         |
|        |          | 1929-15.2°C |                                                                         |
| IX     | 1941-1960| 1943-15.4°C | Temperatures tend to be 14.4 °C up to 15.5 °C                         |
|        |          | 1956-14.4°C |                                                                         |
|        |          | 1950-15.5°C |                                                                         |
| X      | 1961-1980| 1965-15.0°C | Temperatures tend to be 14.8 °C up to 15.3 °C                         |
|        |          | 1984-15.3°C |                                                                         |
|        |          | 1975-14.8°C |                                                                         |

Based on the analysis of knowledge from each window, a rule that can be potentially interesting can be determined and pruning the rules that are not interesting, rules that tend to have redundancy, rules
that do not have redundancy, then the best rule is IF Increasing Temperature THEN will decrease in 6 to 7 years by 14.5 °C to 15.5°C

9. Conclusion
The optimal of time series data through optimization of rule discovery by sequence time series data mining clusters analysis can be generated information or knowledge or trends and patterns in the database from uncertain time series data, which were previously not known and able to find interesting rules. These information can then help interested parties in decision making and forecasting in the future.

References

[1] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Third Edit., vol. 1. Elsevier, 2012.
[2] J. Lin, S. Lonardi, J. Lin, and S. Lonardi, “Visualizing and Discovering Non-Trivial Patterns In Large Time Series Databases Short running title : Time Series Visualization Visualizing and Discovering Non-Trivial Patterns In Large Time Series Databases,” 2005.
[3] S. Rodpongpun, V. Niennattrakul, and C. A. Ratanamahatana, “Selective Subsequence Time Series clustering,” Knowledge-Based Syst., vol. 35, pp. 361–368, 2012.
[4] J. Serr, “Ranking and significance of variable-length similarity-based time series motifs ,” 2015.
[5] J. Lin, E. Keogh, S. Lonardi, and P. Patel, “Finding motifs in time series,” Proc. 2nd Work. Temporal Data Min., pp. 53–68, 2002.
[6] A. Mueen, S. Cash, and B. Westover, “Exact Discovery of Time Series Motifs,” 2002.
[7] N. Castro and P. J. Azevedo, “Time Series Motifs Statistical Significance,” 2008.
[8] M. Shokoohi-Yekta, Y. Chen, B. Campana, B. Hu, J. Zakaria, and E. Keogh, “Discovery of Meaningful Rules in Time Series,” Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD ’15, pp. 1085–1094, 2015.
[9] E. Keogh and S. Kasetty, “On the Need for Time Series Data Mining Benchmarks : A Survey and Empirical Demonstration.”
[10] T. Fu, “Engineering Applications of Artificial Intelligence A review on time series data mining,” Eng. Appl. Artif. Intell., vol. 24, no. 1, pp. 164–181, 2011.
[11] J. Serra and J. L. Arcos, “An Empirical Evaluation of Similarity Measures for Time Series Classification.”
[12] O. Lauwers and B. De Moor, “A Time Series Distance Measure for Efficient Clustering of Input / Output Signals by Their Underlying Dynamics,” vol. 1, no. 2, pp. 286–291, 2017.
[13] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, “Querying and Mining of Time Series Data : Experimental Comparison of Representations and Distance Measures,” 2008.
[14] K.W. Hipel A.I McLeod, Time Series Modelling of Water Resources and Environmental Systems. Elsevier Science, 1994.