Time Series Optimization on Data Mining

Relita Buaton$^{1,2}$, Herman Mawengkang$^{1}$, Muhammad Zarlis$^{1}$, Syahril Effendi$^{1}$, Akim Manaor Hara Pardede$^{1,2}$, Yani Maulita$^{2}$, Ahmad Fauzi$^{2}$, N Novrijenni$^{2}$

$^{1}$Program Studi Doktor Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi, Jl. Universitas No. 9A, Medan, 20155, Universitas Sumatera Utara, Indonesia
$^{2}$STMIK Kaputama, Jl. Veteran No. 4A-9A,Binjai- Sumatera Utara, Indonesia

bbcbuaton@gmail.com

Abstract. Forecasting is one of the important topics in the data mining field, such as, predictions, weather forecasting, predictions of academic achievement. Another topic associated with forecasting through a series of data that depends on the time period is called time series. The problem in data mining time series is how to present the data. A common approach is to transform periodic series into other domains so that the reduced dimensions are followed by an index mechanism. However, research on time series has not been successful optimal yet, because it is still limited to mining data yet to represent time series, this pattern needs to be developed to change the pattern into a rule. The main problem that needs to be addressed in a periodic series is to present the results of visualization which includes more than thousands of observations are very difficult in order to present time series data in multi-dimensional to be mined. Working with high-dimensional data will be very expensive in terms of process and storage costs, because it requires high-level data representation or abstraction. The method used is the cluster window. The results obtained are the discovery of patterns based on the frequency of symbol behavior

1. Introduction

Data mining is an analytical process of knowledge discovery in large and complex data sets, data mining is a discipline residing in statistical and computer science intersection. More precisely, data mining is the result of hybridization of statistics, computer science, artificial intelligence and machine learning [1]. Many data scientists want to explore the data, seeking information for knowledge acquired through grouping process, classification, rule discovery, association and visualization in data mining. Forecasting becomes one of the important talking topics in the data mining community. Most of the research that appears in the literature relates to discrete objects [2], for example on prediction suppression, data base queries [3], weather forecasts, predictions of academic achievement, and others. In statistics, Time Series is one topic that is always associated with forecasting through a series of data that depends on the time period. A periodic series is a collection of observations made chronologically. Data from periodic series have characteristics such as large, high dimension and continuous updating. The next characteristic is that the numerical and continuous nature of the data is always viewed as a whole rather than an individual numeric. Therefore, unlike traditional databases where similarity search is based on matching, the search for similarity in the periodic series data is based on the approach. The famous examples include daily stock prices on the Jakarta Stock Exchange, the number of hourly mobile phone usage in Medan, and daily sea-level readings in the Pacific Ocean. Many studies have been conducted on a periodic basis on the basis of similarity.[4].
In periodic data mining, the fundamental issue is how to present periodic data series. One common approach is to transform periodic series into other domains so that reduced dimensions are followed by index mechanisms. Research on time series is not optimal because it is still limited to mine data that has not been able to represent time series [5], able to find patterns in time series data [6], the pattern needs to be developed to change the pattern into a rule. Rules can be found from the time series data, but are still constrained by overfitting [2]. Furthermore, the size of similarity in the periodic series or sub sequences and the segmentation process are the two main tasks for the various tasks covered in periodic sequential mining. One of these mining tasks is the discovery rule. [7] examines the mechanisms of finding rules for the periodic series. However, their algorithm is only evaluated for process speed and then only on random data. It is not shown whether this algorithm finds the rule in general in the periodic series. [8] using a linear part-by-section representation to support the discovery rules in the periodic series. Their algorithm is tested on financial data, with a precise 68% of predictions. The most widely used method for the discovery of rules in the literature is from [9] They quantify the data by grouping K-means from all dataset training and entering symbolic data into the classic association algorithm of the invention of the rule, the quality of the rule induced from the time series data is influenced by the number and cluster parameters. The success of a rule is measured by using a score called J-size. But by [6] it has been shown that the quantification step which includes grouping all sub sequences will not be able to produce a group center. The main issue that needs to be addressed in periodic data mining is if only with a visualization of the periodic series that can cover more than thousands of observations would be very difficult [10]. Working with very high-dimensional raw data will be very expensive in terms of process and storage costs [5]. Therefore, we need high level data representation or abstraction. Rule discovery is one way against the representation. Difficult to present time series data in multi dimension to be mined

2. Data mining

The ability of computer systems to generate and collect data increases rapidly. Millions of databases are generated in business management, government administration, and in many other applications. The rapid development of database size can be due to the ability of its database system. This condition raises important new needs, namely: a new technique that performs the transformation process from the large transactional database to obtain important information needed. So Data Mining becomes an important research material nowadays[11]. Data mining is the process of finding interesting patterns and knowledge of large amounts of data. Data sources can be databases, data warehouses, web, other information repositories, or data that are streamed into the system dynamically. The process of information discovery is done by several stages: Data Cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation[12]

2.1. Time Series

Time series $T$ with size $m$ is a set of real data values, where $T=(t_1,t_2...t_m)$ . Subsequence with length $n$ time series $T$ is $T_{i,real}=(t_i,t_{i+1},...,t_{i+n-1})$, where $1 \leq i \leq m-n+1$ [13]. Clustering time series is used in various fields, such as e-commerce, outlier detection, speech recognition, biological systems, DNA recognition, and text mining. In addition, time series clustering is also applied to one field in the domain clustering time series that is pattern recognition using sequence time series[14]. T time dimensional time series data with length $n$ is represented by $T = x_1, ..., x_n$. Each $x_i$ is a real valued variable. Data time series $T_{m, m}$ dimension with length $n$ is a sequence / sequence $m$ set of real valued variables, represented by $T_m = (x_{11}, ..., x_{m1}), ..., (x_{in}, ..., x_{mn})$. Time series subsequence(TSS) $C$ with length $q$ part of $T$, represented by $C_{p,q} = x_p, ..., x_{p+q-q}(q \leq n, 1 \leq p \leq n - q + 1)$ [15]

2.2. Subsequence Time Series Clustering

An important issue in the subsequence time series clustering is how to apply a method so as to generate meaningful outcomes from a large amount of time series data. The following describes some important methods in subsequence time series clustering Hierarchical Clustering. Hierarchical Clustering has a powerful visualization compared to other clustering approaches. Hierarchical Clustering creates a nested hierarchy associated with a paired distance matrix object. One of the
3. Subsequent Time Series Clustering Algorithm

- Cluster quality (Accuracy).
  
  To evaluate the quality of clustering using cross entropy is stated as follows
  \[
  Cross\ entropy = \sum_{i=1}^{K} \left( \frac{n_j}{M} \right) \left( - \sum_{i=1}^{m} p_{ij} \log(p_{ij}) \right)
  \]
  
  Where \( K \) is the number of clusters, \( n_j \) is the number of sequences in a cluster \( j \), \( m \) is the number of classes in the database sequence, \( p_{ij} \) is the random probability \( j \) cluster, \( SDB \) is the sequence database.

- J-Measure
  
  Used for rule rankings, defined as follows
  \[
  J(B_T; A) = p(A) \cdot \left( \frac{p(B_T | A) \log \left( \frac{p(B_T | A)}{P(B_T)} \right) + (1 - p(B_T | A)) \log \left( \frac{1 - p(B_T | A)}{1 - p(B_T)} \right)}{1} \right)
  \]
  
  \( p(A) \) is the probability of symbol \( A \) in a random location in sequence. \( p(B_T) \) is the probability of at least one \( B \) occurring in a random selection over the duration \( t \). \( p(B_T | A) \) is the probability of at least one \( B \) occurring in a random selection over the duration of \( T \) given by process \( A \).

- Normalized Mutual Information (NMI).
  
  NMI is one of the significant comparison measures to evaluate cluster or algorithm results. This can help researchers to assess the performance and analysis of improvements of an algorithm. Specifically determined as follows. Let \( C_T \) and \( C_E \) be the set of class labels and cluster label sets calculated using clustering algorithms, then NMI between \( C_T \) and \( C_E \) is
  \[
  NMI(C_T, C_E) = \frac{H(C_T) + H(C_E) - I(C_T; C_E)}{H(C_T) + H(C_E)}
  \]
  
  Where \( H(P), H(P, Q) \) and \( I(P; Q) \) is to represent entropy, joint entropy and variable information variable \( P \) dan \( Q \). When \( C_T \) and \( C_E \) is free from one variable with another then, \( NMI(C_T, C_E) = 1 \), because \( H(C_T, C_E) = H(C_T) + H(C_E) \). Must be fulfilled. The greater it is \( NMI(C_T, C_E) \) the more accurate the cluster results [21].

- Performance Algorithm
  
  The performance of the algorithm is usually evaluated by test environment and dataset and scalability test.

  Runtime and Memory usage. By measuring the runtime and memory usage, it can compare algorithms with time and memory usage. Next can calculate the minimum support threshold limit in the algorithm. The runtime and memory usage of an algorithm increases exponentially as the minimum threshold limit decreases.[14]

4. Data Mining and Knowledge Discovery

Data mining is a data mining process that is perfectly capable of identifying valid, new, potential knowledge to be used, finding understandable knowledge and can be used to make critical business decisions, including impractical and not simple calculations, reveals hidden patterns within the data. Applicable, the patterns found must be true as new data or knowledge. Novelty, the pattern found should be new to the organization. Useful, organizations must be able to act on patterns that are found to be more profitable and efficient. Understood, new patterns must be understandable and increase user knowledge[1]. Data mining and knowledge discovery from the database have gained much attention in recent years. However, most of his research concentrates on how to produce accurate models and little research on the assessment or ranking of the resulting patterns. Much of the data mining literature concentrates on the accuracy of discovery and understanding of patterns found, while
it is necessary how to convince users with the discovery of patterns or knowledge. The important task of data mining algorithms is to look for hidden patterns in addition to being accurate and understandable. [22]. There are four steps to determine the best and effective rule for measuring the correlation between antecedents and consequences. RI = 0 if A&B = |A||B| N. RI monotonically increases with |A&B| when other parameters are fixed. RI monotonically decreases with |A| or |B| when other parameters are fixed.[23]. RI=rule interest, Where N is number of pattern A and B is tupels that satisfy A dan B when RI=0, then A and B are independent and an uninteresting rule. Then major and Mangano propose step 4 and RI monotonically increases with |A| with fixed confidence factor Cf > C0. To assess / rank the rule that is by using the factor of domination in accordance with the position of the rule in a hierarchy, based on the dominance of potential interesting, technically interesting dan genuinely interesting. To achieve this by designing criteria such as performance, simplicity and significance in the database.[24]

5. Motif Discovery Algorithm
Motives invented algorithm dynamically to detect motifs of one-dimensional time series data. Some of the algorithms used are Minimum description length(MDL)[25], Akaike’s information criterian(AIC)[26], Bayesian Information criterion(BIC)[27]. AIC produces the best model based on predictive capability, in this case AIC finds patterns that often appear rather than time series predictions. BIC produced the best model based on bayes theorem and MDL produced the best model by minimizing the overall length description of the data set. MDL is the best model to summarize or summarize the times series data.

- Change the time series into a sequence of symbols
- Extraction pattern with MDL Principle has a weakness that is the same pattern almost does not appear in time series and there is noise in time series. For that reason, time series data is converted into a sequence symbol by using the dimensionality algorithm based on the PAA representation, shown in the following figure 1

![Figure 1](image)

**Figure 1.** The visualization algorithm converts the time series into a sequence of symbols
(a) Obtain TSS by sliding window analysis, (b) Converts TSS to SAX symbol sequence, (c) Generate a symbol behavior for each SAX symbol sequence

Provides window analysis (Figure 1.a) to generate \( T_{\text{min}} \). \( T_{\text{min}} \) is the minimum length of the motive generated from the data. By shifting the analysis window then we get TSS with length of \( T_{\text{min}} \) from data, then each TSS is represented by a PAA representation and converted into a sequence PAA symbol (figure 1.b). The PAA representation is the vector expression generated by dividing the time series data into several segments and calculating the average value of each segment, in Figure 1.b The TSS is divided into 4 segments. Using PAA time series representation \( T = x_1, \ldots, x_n \) with \( n \) length can be represented as \( w \) dimension space by vector \( \tilde{C} = c_1 \ldots c_w \). In Figure 1.b. The PAA
representation on each TSS is represented by a vector $\tilde{C}$, then break points are determined to convert the vector $w$ dimension into SAX symbol sequence. The break point provides several areas of PAA representation that can be enlarged under the Gaussian distribution [28]. In Figure 1.b, provided 2 break points and 3 regions, each region is given a unique SAX symbol, and each PAA coefficient is converted to SAX symbol where PAA coefficient is found. For example on $\tilde{C}$ from the first TSS is transformed into SAX sequence symbol ‘cbba’. Each TSS gets the SAX sequence symbol. Each SAX symbol, and each SAX symbol is converted into a unique symbol called the behavior symbol (BS), as in figure 1.c BS ‘A’ is assigned to for the sequence of SAX cbba symbols B for SAX beba symbols and so on. Then in detail in figure 1.b symbol SAX obtained cbba, where c = high, b = middle and a = low. So the cbba string can be assigned to the letter A which is conceptually conceptual meaning the time series has two peaks, the second is lower than the first. The beba string is dedicated to the letter B which conceptually means starting from the middle range, then the high peak and then down. So in this way can be detected TSS with the same behavior and reduce the memory space to find the motive.

Estimate the extraction pattern with the MDL principle. BS subsequence is part of the BS sequence called the BSS (behavior symbol subsequence). BS subsequence pattern is called BSS pattern. Let $n_p$ be the length of the BSS SC pattern, $s_p$ is the unique number of BBS SC, then it takes a number of $\log_2 n_p$ bits to determine the BSS on the SC, thus can be defined as $DL(SC) = \log_2 n_p + n_p \log_2 s_p$. In addition to $DL(SC)$, it is also necessary to define $DL(\tilde{C}|SC)$, where $\tilde{C}$ is a BS sequence representation. SC represents one symbol, it is assumed that the length of the sequence is $n_a$, the number of BSS uniquely in $\tilde{C}$ is $s_a$ and the frequency of the occurrence of SC in $\tilde{C}$ is $q$. The length description of $DL(\tilde{C}|SC)$ is as $DL(\tilde{C}|SC) = \log_2 n_a + n_a \log_2 (s_a + q)$ . In this case, $\log_2 n_a$ is the number of bits needed to express the number of BSS in $\tilde{C}$, $n_a \log_2 (s_a + q)$ is the number of bits needed to encode the unique number of BSS from $\tilde{C}$. So the MDL estimation with the $MDL(\tilde{C}|SC)[15]$

6. Time Series Optimization

- Input data time series. T time dimensional time series data with length $n$ is represented by $T = x_1, \ldots, x_n$, Each $x_i$ is a real valued variable [15]
6

Figure 3. Time Series Optimization

- Figure 3.a. Change the time series data into graphic form. Figure 3.b. Divide the graph into several segments, the minimum length of the motif generated from the graph is called with $T_{min}$. By shifting the analysis window then we get TSS with length $T_{min}$ of data. Figure 3.c. Clustering window. Figure 3.d, e extract the cluster into a symbol. Figure 3.e. Extract symbol becomes behaviour and finding symbols by frequency (A B C)

7. Conclusion

Based on the discussion of time series optimization, it can be concluded that the discovery of rules in time series data can be optimized by dividing time series into several window analyzes, then each windows analysis is clustered with proximity distance to euclidean distance or manhattan city. The cluster results are then rules or information found from previously unknown time series

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[2] M. Shokoohi-Yekta, J. Wang, and E. Keogh, “On the Non-Trivial Generalization of Dynamic Time Warping to the Multi-Dimensional Case,” *Data Min. Proceeding 2015 …*, pp. 289–297, 2015.

[3] Y. Li, J. Lin, and T. Oates, “Visualizing Variable-Length Time Series Motifs.,” *Sdm*, pp. 895–906, 2012.

[4] R. Agrawal, C. Faloutsos, and A. Swami, “Efficient Similarity Search In Sequence Databases,” *Springer*, 1993.

[5] 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.

[6] E. Keogh, “Clustering of Time Series Subsequences is Meaningless : Implications for Previous and Future Research,” 2005.

[7] S. Park and wesley Chu, “Discovery and matching elastic rules from sequence database,” *Fundam. Inform.*, 2001.

[8] H. Wu, B. Salzberg, and D. Zhang, “Online Event-driven Subsequence Matching over Financial Data Streams,” *Sigmod Conf.*, 2004.

[9] G. Das, K. Lin, and H. Mannila, “Rule discovery from time series,” *Am. Assoc. Artif. Intell.*,
1998.

[10] 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.

[11] L. W. Santoso, “PEMBUATAN PERANGKAT LUNAK DATA MINING UNTUK METODE APRIORI,” J. Pendidik. Teknol. Inf. dan Komun., pp. 49–56, 2013.

[12] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Third Edit., vol. 1. Elsevier, 2012.

[13] S. Rodpongpun, V. Niennattrakul, and C. A. Ratanamahatana, “Selective Subsequence Time Series clustering,” Knowledge-Based Syst., vol. 35, pp. 361–368, 2012.

[14] S. Zolhavari, S. Aghabozorgi, and Y. W. Teh, “A review of subsequence time series clustering,” Sci. World J., vol. 2014, 2014.

[15] K. U. Yoshiki Tanaka, Kazuhisa Iwamoto, “Discovery of Time-Series Motif from Multi-Dimensional Data Based on MDL Principle,” Springer Sci. + Bus. Media, Inc. Manuf. Netherlands, no. 2000, pp. 269–300, 2005.

[16] J. Lin, E. Keogh, and W. Truppel, “Clustering of streaming time series is meaningless,” Proc. 8th ACM SIGMOD Work. Res. issues data Min. Knowl. Discov. - DMKD ’03, p. 56, 2003.

[17] T. Warren Liao, “Clustering of time series data - A survey,” Pattern Recognit., vol. 38, no. 11, pp. 1857–1874, 2005.

[18] T. Oates, “Identifying Distinctive Subsequences in Multivariate Time Series by Clustering,” Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’99, pp. 322–326, 1999.

[19] T. Rakthanmanon et al., “Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping,” Trans. Knowl. Discov. from Data (TKDD), vol. 7, no. 3, pp. 3047–3051, 2013.

[20] C. S. Möller-levet and F. Klawonn, “Fuzzy clustering of short time-series and unevenly distributed sampling points,” vol. 2810, no. June 2014, 2003.

[21] R. Fujimaki, S. Hirose, and T. Nakata, “Theoretical analysis of subsequence time-series clustering from a frequency-analysis viewpoint,” Sdm Siam, no. Figure 1, pp. 506–517, 2008.

[22] K. MCGARRY, “A Survey of Interestingness Measures for Knowledge Discovery,” Knowl. Eng. Rev., vol. 0, no. January, pp. 1–24, 2004.

[23] W. J. Frawley, G. Piatetsky-shapiro, and C. J. Matheus, “Knowledge Discovery in Databases : An Overview,” AI Mag., vol. 13, no. 3, pp. 57–70, 1992.

[24] J. A. Major and J. J. Mangano, “Selecting among rules induced from a hurricane database,” J. Intell. Inf. Syst., vol. 4, no. 1, pp. 39–52, 1995.

[25] J. Rissanen, “Stochastic Complexity in Statistical Inquiry,” World Sci. Ser. Comput. Sci., 1989.

[26] H. Akaike, “Fitting Autoregressive Models For Prediction,” Ann. Inst. Stat. Math., vol. 21, pp. 243–247, 1969.

[27] G. Schwarz, “Estimating the Dimension of a Model.pdf,” vol. 6, no. 2, pp. 461–464, 1978.

[28] J. Lin, E. Keogh, S. Lonardi, and P. Patel, “Finding motifs in time series,” Proc. 2nd Work. Temporal Data Min., pp. 53–68, 2002.