Journal of Theoretical and Applied Information Technology
15th August 2020. Vol.98. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMIZATION OF FORECASTING TIME SERIES WITH RBT (RULE BEST TIME SERIES)

[1]RELITA BUATON, [2]MUHAMMAD ZARLIS, [3]HERMAN MAWENGKANG, [4]SYAHRIL EFENDI

[1]Graduate Program Of Computer Science, Department Of Computer Science, Universitas Sumatera Utara, Medan, Indonesia

[2]Department Of Computer Science, Universitas Sumatera Utara, Medan, Indonesia

[3]Department Of Mathematics, Universitas Sumatera Utara, Medan, Indonesia

[4]Department Of Computer Science, Universitas Sumatera Utara, Medan, Indonesia

E-mail: [1]bbcbuaton@gmail.com, [2]m.zarlis@yahoo.com , [3]hmawengkang@yahoo.com , [4]syahrilkom1@yahoo.com

ABSTRACT

This research is motivated by the abundance of time series data stack found, often regarded as garbage and neglected due to the inability to find knowledge or interesting patterns from the data pile. Time series is one of the topics that is often associated with forecasting through a series of data that depends on time periods. The basic problem in time series data mining is how to present the knowledge contained therein, then how to find the rules of periodic data series and how to optimize the decision of the resulting time series data so that it can be used to predict in the future. Based on previous papers, there is no model to present knowledge in the form of rules in time series. In this paper the proposed model is RBT (Rule Best Time Series). The main process in RBT is to discretize periodic series to form sub-sequences, then these sub-sequences are grouped through measures of similarity with distance using euclidean, then the discovery of rules is applied to obtain hidden rules on temporal patterns and to rank with J-measures. From the results of this study time series data can be optimized, new knowledge or trends and patterns in time series databases that are uncertain and previously unknown can be generated. The decision or information can be used to display decisions, or forecasting in the future with an accuracy rate of the model mean absolute deviation (MAD) of 73%, forecasting accuracy of the mean squared deviation (MSE) of 87% and the percentage of the mean absolute percentage error of the MAPE of 4,7%.

Keywords: *Forecasting Time Series Data Mining, Rule Best Time Series*

1. INTRODUCTION

Many abundant time series data piles are found and are considered junk and neglected due to the inability to find knowledge or interesting patterns from the data pile. One of the important topics of discussion in the data mining community is forecasting. Many studies appear in the literature relating to discrete objects [1], for example, the emphasis on predictions, database queries, weather forecasting, prediction of academic achievement, and others. In statistics, time series is one of the topics that is always associated with forecasting through a series of data that depends on time periods. Time series data often occur in business applications and in science. Time series is a collection of observations made chronologically. Time series data has characteristics such as large size, high dimension, and continuous updating, and has numerical and continuous properties that are always seen as an entity rather than an individual numeric, unlike traditional databases where similarity searches are based on matching, similarity searches in periodic data are performed based on approach. Notable examples include daily stock price predictions on the stock exchange, the number of hours of mobile phone usage per hour, and daily sea surface temperature readings in the Pacific Ocean. Many studies have been conducted on a periodic series on the basis of similarity. The decomposition of time series into a time-frequency spectrum can determine the dominant mode of variability and how it varies in time. Wavelet transform has been used in various studies

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

including in geophysical research, tropical convection, El Nino-South Oscillation and sea wave dispersion [2].

The fundamental problem in time series data mining is how to present periodic data series. One common approach is to transform periodic series into other domains so that the reduced dimensions are followed by an indexing mechanism, research on time series is not optimal because it is still limited to mining data not being able to represent time series [3], able to find patterns in time series data [4], but the pattern needs to be developed to change the pattern into a rule. The rule can be found from time series data, but it is still constrained by overfitting [1]. In time series data sometimes there are data values that are far different from other data and do not reflect the characteristics of a set of data called outliers. In the analysis of time series data using methods such as ARIMA, ARMA, and others [5]. Some research on the discovery of the series rules periodically. However, the proposed algorithm is only evaluated for processing speed and then only for random data. It is not shown whether this algorithm finds general rules in periodic series by using linear part-by-part representation to support the rules of discovery in periodic series. Their algorithm is tested on financial data, with a 68% accurate prediction. The method most frequently referred to for rule discovery in the literature is that quantifying data by grouping K-means from all training datasets and entering symbolic data into the classical association algorithm discovery rules, the quality of rules induced from time series data is influenced by the number and parameters of the cluster. The success of a rule is measured by using a score called the J-measure. However, by [4] It has been shown that the quantification step which includes grouping all the subsequences will not be able to produce a group center.

The main problem that needs to be dealt with in series data mining is that if only the visualization of periodical series that can cover more than thousands of observations will be very difficult [6]. Multi-objective reactive power dispatch (MORPD) is a non linear optimization problem multi-objective that has non-convex, multi-constraint, and characteristics multi-variable consisting of a mixture of solutions that have discrete and continuous variables[7]. The Laplacian similarity graph method is a method for semi-supervised classification based cluster analysis group solutions. This method needs to be optimized with solutions in a probabilistic context, to get an estimate of the quality of classifications and speed up calculations, solve systems of linear equations[8]

Working with very high-dimensional raw data will be very expensive in terms of processing and storage costs [3]. Because of that high level data representation or abstraction is needed. The discovery of rules is one way of the intended representation. Difficult to present time series data in multi dimensions to be mined, how to find rules from periodic data series and how to optimize time series data decisions generated by data mining so that it can be used to predict the future [9]. This study aimed to create a rule discovery model on large-scale periodic data mining series by discretizing periodic series for sub-sequence formation. Then this sub-sequence is grouped through measures of similarity, then the rule discovery technique is applied to obtain hidden rules on temporal patterns that can be used as a tool for decision support and forecasting for the future. With the development of the model carried out on sub sequence time series, the contribution of this study is the discovery of a new model in the discovery of new knowledge in time series data without visual patterns with the discretization approach of sliding window series, the determination of periodic similarity based on the size of the similarity distance between the series other periodicals, and make rules by presenting them based on J-Measure size. The results of the model can be used to explore time series data that are not linear or uncertain with a large capacity so as to produce new knowledge or rules that can be used to make decisions or predictions in the future.

2. RELATED WORK

2.1 Motive Discovery

Several studies have suggested motive discovery algorithms such as [10][11][12]. In previous studies [13], the algorithm can extract motifs from multi-dimensional time series data. The first step is the motive discovery algorithm to change the multidimensional time series into one dimensional time series data by using Principal Component Analysis (PCA) to reduce the dimensions of the data. The second step then extracts the patterns that appear in the one-dimensional time series data into a sequence of symbol shapes. The final step is to find a motive by calculating the 'description length' length of the pattern description based on the Minimum Description Length (MDL), assuming that the time series sequence (TSS) is the same, called the SL pattern (same length pattern), even though in reality

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

there are actually differences the length of the pattern, hereinafter referred to as the DL pattern (different length pattern), as appears in the electrocardiogram as shown below

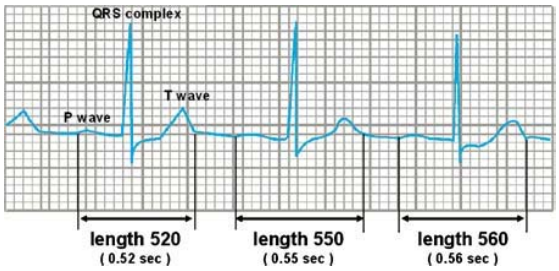

Figure 1: Pattern In The Electrocardiogram [13]

Each TSS represents a series of waves, but the time series is slightly different from one another, for example, the first TSS has a length of 520, the second 550 and the third 560 (the frequency of the time series data sampling is 1kHz). In this case, the extracted pattern is a DL and SL pattern as an extension of the previous algorithm [12]. Some research on data mining, the purpose of which is to reduce calculation time, the proposed approach is to obtain a discrete representation of time series is indexing, clustering, and discovery of association rules. The proposed approach is to change the time series data into sequence symbols with the SAX algorithm, approaches that are similar to SAX are:

- Time series active search, i.e. with time series data in the form of sequence symbols based on histograms obtained by short time spectrum analysis. This algorithm is used for discrete representations in the frequency domain,
- Recognition of patterns based on episodes. In this case, time series data is divided into episodes. Each episode represents the basic behavior of the data and produces a sequence of symbols or sequence symbols and assigns unique symbols to each episode, but this representation is easily influenced by noise in the data, also creates a problem that is ambiguous in terms of defining the distance between episodes,
- Some other approaches that are similar are that there are several approaches to extract patterns or motifs that are not known beforehand is the Enumeration of Motifs through Matrix Approximation (EMMA) algorithm. The EMMA algorithm finds a motive by obtaining all possible TSS of a user-specified length,
- Motive discovery algorithm using random projection is the development of the EMMA algorithm by adopting random projections to find motives more efficiently.

According to [14], pattern detection that is very similar in time series is usually called a motif. A new approach to finding similar motifs of different lengths has been proposed. In this study, it was shown that the motives were obtained based on the similarity of variable lengths that could not be compared directly but could be normalized. Specifically, it was found that the mismatch of normalized motif lengths still has an intrinsic dependency on motif lengths, and the lowest difference is strongly influenced by this dependency. In addition, it was found that the dependencies are generally not linear and change with data sets that are considered not to have the same size. Based on these findings, a solution is proposed to rank the motives and measure their significance. This solution relies on a compact but accurate space dissimilarity model, using a beta distribution with three parameters that depend on the length of the motif in a non-linear way, it is believed that the mismatch of variables can exceed the time series space and that similar modeling strategies such as those users can help in a broader context.

Detection of repetitive patterns or very similar in time series has proven to be very beneficial for researchers and practitioners. There are two definitions of motifs in time series, namely:

- Based on the idea of frequency, an interesting pattern if it has a number of significant repetitions,
- Based on the idea of similarity: an interesting pattern, if the events are identical or too similar, occurs randomly.

Both definitions are complementary because very similar patterns are not always needed often. Beyond frequency-based definitions, the ranking of motifs found in a time series is considered to be unimportant. The motifs that are considered the most important are the motif with the highest count first, the motive with the second-highest count, the motif with the third-highest count and so on. Motives can also be assessed from statistical significance by comparing the number observed and the expected count below null which reflects some of the characteristics of time series [15]. This study found obstacles in finding the motives of a time series, that is if the pair of motifs of a time series have different lengths because they cannot directly compare the similarity or distance. The solution that can be done is:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Calculate the ranking of each motif length that allows covering motives that are not appropriate [16],
- Normalize a measure of inequality with the length of the motif, or use a measure that has combined several ideas of normalization. For example, by dividing the Euclidean distance by the square root of its length, or considering the size of the Pearson correlation. In terms of the significance of motives, the similarity-based approach is less developed than the frequency-based one,
- Develop a space comparison model of motif dissimilarity, using a beta distribution whose parameters are not linear depending on the length of the motif.

The main contribution of this algorithm is to show that the inequality of time series motifs cannot be compared directly and thus cannot be ranked. Through both examples of motivation and formal quantitative analysis, it can be shown that the inequality of normalized motives lengths has a non-linear dependence on motive lengths, that these dependencies change according to the data set and the size of the inequality, that the data set greatly influences the lowest difference, which is precisely the focus of attention on similarity based on the discovery of motifs. Another contribution of this algorithm is the spatial inequality model which allows comparing motives of different lengths and assessing their significance with respect to the overall inequality distribution, proposing procedures to fit the three-parameter beta distribution when taking into account local continuity and non-linearity of spatial inequality.

2.2. Measure of Interest

Genetic Algorithm (GA) is an algorithm used to optimize fuzzy prediction rules to convince users. GA is also used to find a population of rules that are more dynamically interesting. The rules identified by genetic algorithms are compared to the rules selected by expert domains, there are several ways to statistically test the strength of patterns. An approach by Hilderman using heuristic-based methods to determine the usefulness of data mining patterns.

1) Measure of Shanon Entropy: measuring the entropy of information theory by calculating average information content using formulas:

$$-\sum_{i=1}^{m} P_i Log_2 P_i \qquad (1)$$

Where m: number of target attribute values (number of class classifications), pi: number of samples for class-i.

2) Measure of Lorenzo: statistical curve, calculates the probability of association of data with formulas:

$$\bar{q}\sum_{i=1}^{m}(m-i+1)P_i \qquad (2)$$

Where m: total tuples, Pi: the probability of tuples $\bar{q}$: probability of tuple distribution.

3) Measure of Gini: measuring inequality based on the Lorenz curve by using the ratio of the Lorenz curve and the total area using the formula:

$$\frac{\bar{q}\sum_{i=1}^{m} q\sum_{j=1}^{m}\left|P_i - P_j\right|}{2} \qquad (3)$$

Where m: total tuples, Pi: the probability of tuples i, Pj: tuple probability j $\bar{q}$: the probability of tuple distribution.

4) Measure of Kullback Leiber: use Shannon's size and distance to calculate the difference between the actual distribution and the uniform distribution with the formula:

$$\log_2 m - \sum_{i=1}^{m} p_i \log_2 \frac{P_i}{\bar{q}} \qquad (4)$$

Where m: total tuples, Pi: the probability of tuples i $\bar{q}$: the probability of tuple distribution.

5) Measure of Atkinson: Measuring inequality in terms of economy and population distribution with the formula:

$$\max\left(p_i\right) \qquad (5)$$

Where Pi: probability of tuples,

Measurement results objectively do not highlight the most important patterns produced by data mining systems, because these techniques are generally used to compare the level of user confidence in patterns generated by data mining algorithms [17].

2.3. Selection of Rules

The method of finding rules that have been proposed in the previous section can produce a number of rules with various beliefs. For interactive knowledge discovery, one strategy is to give users the flexibility to choose their own set of rules and provide tools for the selection of interesting rules. Likewise, there are no criteria that may be enough to be able to choose the most appropriate rules. For that, the user needs instructions in determining

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

which rules have high confidence. There are various metrics that can be used to rank the rules, in this study the J-measure is used, which is defined as follows:

$$J(B_T;A) = p(A)*(p(B_T \mid A)\log(\frac{(B_T / A)}{p(B_T)}) + (1-p(B_T \mid A)\log(\frac{1-p(B_T / A}{1-p(B_T)})) \quad (6)$$

Where, in the context of the sequence rule, the probability of occurrence of symbol A at random locations in the sequence, is the probability of occurrence of at least one B in a randomly selected window with duration t and p (BT | A) is the probability of occurrence of at least one B in the window randomly selected with duration T with know that the window already contains A. So the first term of this size form, p (A) is a bias against the rules that often occur. The second term is known as cross-entropy, which is information obtained (or the degree of shock) to move from the probability of prior p (BT) to the posterior probability of p (BT | A). From a practical point of view, this measure produces a useful method for ranking rules in which there is a balance between the frequency of the rules and the beliefs of the rules[18].

2.4. Model Testing Method

The model found will be tested using cluster analysis, which measures variance. Variance measurement is useful for assessing the spread of similarity of data formed using variance within-clusters and variance between clusters. A good cluster is when a cluster member has a high level of similarity for each sequence (internal homogeneity) and not like other cluster members (external homogeneity). Internal homogeneity is called variance within cluster (V_w), while external homogeneity is called variance between clusters (V_b). Ideal clusters are found to have a minimum V_w value to express homogeneity and V_b maximum to express external homogeneity [19].

Cluster variance is calculated by the formula:

$$v_c^2 = \frac{1}{n_c - 1}\sum_{i=1}^{n}\left(d_i - \overline{d}_i\right)^2 \quad (7)$$

Where v_c=variance on the cluster c, k = number of similarity, n_c = Amount of data on similarity c, d_i = data to -i on a similarity, $\overline{d_i}$ = average data on similarity

Variance within similarity is calculated by formula:

$$v_w = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1).v_i^2 \quad (8)$$

Where V_w = Variance within similarity, N = Amount of all data.

Variance between similarity is calculated by the formula:

$$v_b = \frac{1}{K-1}\sum_{i=1}^{k} n_i\left(\overline{d}_i - \overline{d}\right)^2 \quad (9)$$

Where V_b = variance between similarity, k = number of similarity, n_c = smount of data on similarity c, $\overline{d}_i$ = average data on similarity, $\overline{d}$ = average $\overline{d_i}$.

Variance for all similarity is calculated by formula:

$$v = \frac{v_w}{v_b} \quad (10)$$

3. RESULT AND DISCUSSION

The discovery of rules is basically the essence of knowledge discovery through data mining. In other words, we want to indicate the relationship between variables, sequences or periodic series patterns that typically appear in very large databases. The basic idea in the proposed method is to discretize the periodic sequence for sub-sequence formation. Then this sub-sequence is grouped through measures of similarity, then simple rules finding techniques are applied to obtain hidden rules on temporal patterns. The models found to optimize the discovery of rules in time series data mining that is used to predict are:

$$RBT = \sum_{i\in T}\sum_{j\in w} W_{ij}\sum_{k\in s} C_{jk} + \sum_{i\in\alpha}\sum_{j\in\beta} R_{ij} + \sum_{i\in\mu}\sum_{j\in\rho} B_{ij} \quad (11)$$

The notation used is:

RBT = rule best time series, T = Set of time periods T, Ti, n = {ti, ti + 1, ..., ti + n-1}, where 1 ⩽ n ⩽ n-m + 1, W = width of the analysis window, where $s_i = (x_i,...,x_{i+w-1})$. State $W(s) = \{s_i \mid i = 1,...,n-w+1\}$, S = The set of time series subsequences, μ = confidence set, ρ = rule set, α = Similarity set, β = window pattern set, W_{ij} = Parameter sliding window, C_{jk} = Cluster window parameters, R_{ij} = rule time series parameters, B_{ij} = rule ranking parameter.

3.1. Definitions and Framework

Before proposing the method for rule discovery, several related definitions are put

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

forward. Definition 1 (time series). A time series T of size m is a sequence of real value data sequences, where T = {t1, t2, ..., tm}[20]. Definition 2 (Sub-sequence). A sub-sequence with length n of the periodic series T is Ti, n = {ti, ti + 1, ..., ti + n-1}, where 1 ⩽ n ⩽ n-m + 1 [20]. So a sub-sequence is a data sequence arrangement that ignores several elements without exchanging the order of the remaining elements. The time series grouping technique is divided into three, namely, the grouping of all time series, the grouping of time series subsequences and grouping of time series [21]. This study uses periodic sub-sequence grouping to obtain the rules of discovery.

3.2. Time Series Discretization by Clustering

Grouping of subsequence time series includes a grouping of subsequence time series extracted through sliding windows, which is the grouping of segments from one long-term time series. In order to be extracted by sliding windows, the intended time series needs to be discretized. [22] has shown that a time series with real number data can be drastically reduced without significantly affecting the information available. The method proposed for time series discretization through Windows grouping is stated as follows. Suppose that known sequence s and window with width w. Given $s = (x_1, ..., x_n)$ a window with b width w on s is a subsidiary $(x_i, ..., x_{i+w-1})$. From s formed all windows (subscriptions) $s_1, ..., s_{n-w+1}$ 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 subscriptions s_i and s_j with width w. These distances can be used to group all subscriptions of $W(s)$ into groups of $C_1, ..., C_k$. For each group C_k the a_k symbol is entered and the $D(s)$ version is criticized from the sequence s will include the alphabet $\sum = \{a_1, ..., a_k\}$. Sequence $D(s)$ is obtained by searching for each subsection s_i group $C_{j(i)}$ such that $s_i \in C_{j(i)}$ and by using the symbol associated $a_{j(i)}$.

$$D(s) = a_{j(1)}, a_{j(2)}, ..., a_{j(n-w+1)} \qquad (12)$$

Each a_k symbol presents a basic form and what we want to obtain is a rule of discovery which includes a pattern formed from these basic forms. Note that the discretization process described earlier is very dependent on the selection of w, the choice of the periodic series distance function and on the type of clustering algorithm used.

3.3. Time Series Similarity

The most important issue in the discussion of periodic data mining is the determination of similarity, which is the degree to which a periodic series that is considered to resemble another period series. In fact, the size of the periodic similarity is very important to the grouping [3] [23]. Grouping $W(s)$ sets the required distance understanding for periodic series with length w. There are several possibilities and choices regarding distance measurement in rule discovery. The simplest choice is to treat subscriptions of length w as elements of R^w and then Euclid's distance (ie, metric L_2) is used. It has been proven empirically that Euclid's distance is hard to beat. Euclid's distance is a parameter-free, fast computational time method and is compatible with various data mining optimizations such as indexing. The meaning is, for $\hat{x} = (x_1, ..., x_w)$ and $\hat{y} = (y_1, ..., y_w)$ defined:

$$d(\hat{x}, \hat{y}) = (\sum_i (x_i - y_i)^2)^{1/2} \qquad (13)$$

As a metric in the grouping. Other metrics include general metrics L_p which are defined by:

$$L_p(\hat{x}, \hat{y}) = (\sum_i (x_i - y_i)^2)^{1/p} \qquad (14)$$

While $p \geq 1$ and $L_\infty = \max_i \mid x_i - y_i \mid$.

In a variety of uses, we want to obtain the form of subscriptions as the main factor determining distance. That is, two subscriptions are essentially the same shape even though they have different amplitudes and baselines. One way to achieve this is to normalize subscriptions and then apply a metric L_2 to normalized subscriptions. Express the normalized version of $\hat{x}$ with $\kappa(\hat{x})$, defined by the distance between $\hat{x}$ and $\hat{y}$ by:

$$d(\hat{x}, \hat{y}) = L_2(\kappa(\hat{x}) - \kappa(\hat{y})) \qquad (15)$$

Normalization can be done by means of $\kappa(\hat{x})_i = x_i - E\hat{x}_i$ (where $E\hat{x}_i$ is the expected value or average of the sequence value), which results in the average value of the sequence is 0. It can also be used $\kappa(\hat{x})_i = (x_i - E\hat{x}) / D\hat{x}$ (where $D\hat{x}$ is the diversity of the sequence), which will force the average sequence 0 and diversity 1. Although according to [24], Manhattan distance is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

11% more efficient compared to Euclidean, distance and 7% are more efficient than Manhattan. Whereas for efficiency it needs further observation about the behavior of the algorithm in the standard grouping of dataset types.

3.4. Discovery of Rules

The simplest rule format is if A occurs, then B occurs in time T. Here A and B are the basic forms. Low support rules may not be interesting in every transaction, this states how many itemsets appear in the dataset. Meanwhile, minimum trust is a guideline on how often rules are proven correct. For example, suppose N becomes the set of items, N => M is the association rule and Z is the transaction set. N support about Z is defined as part of the z transaction in a dataset that contains items set N, support (N). The trust value of a rule (N => M), regarding a set of Z transactions, is part of a transaction containing N which also contains M [25]. The rules in mathematical logic notation can be written as follows:

$$c(A \overset{T}{\Rightarrow} B) = \frac{F(A,B,T)}{F(A)} \qquad (16)$$

Generally, A is called antecedent and B is called consequent,Where

$$F(A,B,T) = \left| \left\{ i \mid a_i = A \wedge B \in \left\{ a_{i+1}, \ldots, a_{i+w+T-1} \right\} \right\} \right|$$

, the number of occurrences of A followed by B in T, that is, counted only the occurrence of B that appears after w units of time

3.5. Time Series Analysis

Testing data using Hourly Interstate 94 traffic volume data for MN DoT ATR 301 station, around midway between Minneapolis and St. Paul, MN. Daily weather features and holidays including for the impact on traffic volume, data sets from 2012 to 2018 with a total data set of 44439 records, can be downloaded from https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume.

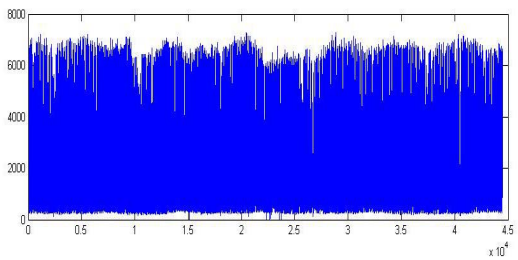

Figure 2: Traffic Time Series Traffic Graph

Figure 2 shows the results of a time series data plot of traffic against time (hours), it is clear that the graph does not linearly change shape as the traffic changes every hour. Visually it is very difficult to analyze if using pattern analysis, it is difficult to identify the rules contained therein and find potentially interesting rules. The next step divides the traffic graph into several windows called subsequence time series.

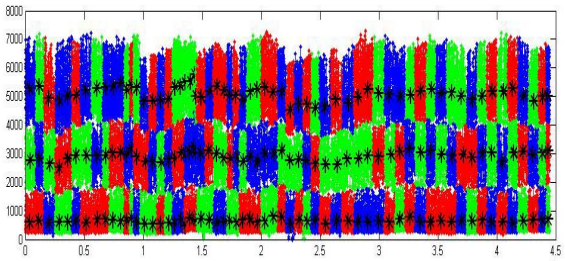

Figure 3: Traffic Window Analysis Results

Figure 3 shows the results of the analysis for each window, from s formed all windows (subscriptions) with width w where, s consists of 63 windows with width w = 1 month, using equation 13 ie Euclid distance, each window generates points obtained through the calculation of similarity distances, these points are the result of similarity to changes in traffic for each window along with the change in time. Each window produces 3 central points of similarity, information on each window analysis is presented in table 1.

Table 1: The Pattern Found In Each Window Analysis

Window/ Month	Pattern Found		
	Traffic Trend	Traffic Changes	Time (Day)
9 (1-790)	19-10-2012,04:00:00 19-10-2012,06:00:00 19-10-2012,21:00:00	- Increase Decrease	1
10 (791-1575)	14-11-2012,01:00:00 15-11-2012,06:00:00 15-11-2012,21:00:00	Decrease Increase Decrease	2
11 (1576-2544)	14-12-2012,07:00:00 16-12-2012,06:00:00 16-12-2012,21:00:00	Increase Decrease Increase	3
12 (2545-3250)	13-1-2013,07:00:00 13-1-2013,10:00:00 14-1-2013,14:00:00	Decrease Increase Increase	2
1 (3251-3967)	14-2-2013, 13:00:00 15-2-2013, 01:00:00 15-2-2013, 20:00:00	Increase Decrease Increase	2
2 (3968-4682)	15-3-2013, 22:00:00 16-3-2013, 04:00:00 16-3-2013, 09:00:00	Increase Decrease Increase	2
3 (4683-5621)	15-4-2013, 16:00:00 15-4-2013, 22:00:00 16-4-2013, 21:00:00	Increase Decrease Increase	2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4 (5622-6561)	17-5-2013, 07:00:00 18-5-2013, 07:00:00 19-5-2013, 05:00:00	Increase Decrease Decrease	3
5 (6562-7105)	14-6-2013, 01:00:00 14-6-2013, 22:00:00 15-6-2013, 16:00:00	Decrease Increase Increase	2
6 (7106-7779)	13-7-201323:00:00 16-7-2013, 01:00:00 17-7-2013, 12:00:00	Decrease Decrease Increase	5
7 (7780-8395)	15-8-2013, 01:00:00 15-8-2013, 02:00:00 15-8-2013, 12:00:00	Decrease Decrease Increase	1
8 (8396-8849)	15-9-2013, 20:00:00 17-9-2013, 02:00:00 17-9-2013, 13:00:00	Decrease Decrease Increase	3
9 (8850-9181)	10-10-2013,03:00:00 10-10-2013,19:00:00 11-10-2013,14:00:00	Increase Decrease Increase	2
10 (9182-9755)	18-11-201310:00:00 18-11-2013,23:00:00 19-11-2013,20:00:00	Decrease Decrease Increase	2
11 (9756-10483)	13-12-2013,08:00:00 16-12-2013,14:00:00 19-12-2013,22:00:00	Increase Decrease Increase	7
12 (10484-11182)	14-01-2014,12:00:00 15-01-2014,01:00:00 16-01-2014,17:00:00	Decrease Decrease Increase	3
1 (11183-11843)	14/02/2014 6:00:00 15/02/2014 5:00:00 16/02/2014 18:00:00	Increase Decrease Increase	3
2 (11844-12483)	13/03/2014 12:00:00 15/03/2014 8:00:00 16/03/2014 4:00:00	Increase Decrease Decrease	4
3 (12484-12937)	12/04/2014 15:00:00 12/04/2014 18:00:00 13/04/2014 9:00:00	Increase Decrease Increase	2
4 (12938-13470)	16/05/2014 8:00:00 18/05/2014 6:00:00 20/05/2014 20:00:00	Increase Decrease Increase	5
5 (13471-13680)	08/06/2014 1:00:00 08/06/2014 15:00:00 10/06/2014 21:00:00	Decrease Increase Decrease	3
6 (13681-14282)	12/07/2014 22:00:00 14/07/2014 4:00:00 19/07/2014 20:00:00	Increase Decrease Increase	7
7 (14283-14416)	04/08/2014 19:00:00 05/08/2014 3:00:00 05/08/2014 8:00:00	Decrease Decrease Increase	2
8 (14417-14552)	27/06/2015 8:00:00 27/06/2015 14:00:00 28/06/2015 3:00:00	Decrease Increase Decrease	2
9 (14553-15275)	15/07/2015 2:00:00 16/07/2015 1:00:00 17/07/2015 8:00:00	Increase Decrease Increase	3
10 (15276-15953)	13/08/2015 22:00:00 16/08/2015 0:00:00 19/08/2015 6:00:00	Decrease Decrease Increase	7
11 (15954-16570)	13/09/2015 15:00:00 15/09/2015 4:00:00 15/09/2015 14:00:00	Decrease Increase Increase	3
12 (16571-17124)	11/10/2015 17:00:00 13/10/2015 22:00:00 15/10/2015 8:00:00	Decrease Decrease Increase	5
1 (17125-17612)	16/11/2015 11:00:00 16/11/2015 23:00:00 17/11/2015 22:00:00	Decrease Decrease Increase	2
2 (17613-18182)	14/12/2015 17:00:00 18/12/2015 0:00:00 23/12/2015 18:00:00	Increase Decrease Increase	7
3 (18183-18784)	14/01/2016 1:00:00 16/01/2016 20:00:00 17/01/2016 13:00:00	Decrease Increase Decrease	4
4 (18785-19400)	12/02/2016 3:00:00 13/02/2016 9:00:00 13/02/2016 16:00:00	Decrease Increase Increase	2
5 (19401-20004)	16/03/2016 16:00:00 17/03/2016 19:00:00 18/03/2016 3:00:00	Increase Decrease Decrease	3
6 (20005-20676)	16/04/2016 16:00:00 16/04/2016 21:00:00 17/04/2016 7:00:00	Increase Decrease Decrease	2
7 (20677-21427)	15/05/2016 15:00:00 16/05/2016 1:00:00 19/05/2016 22:00:00	Increase Decrease Increase	5
8 (21428-22117)	16/06/2016 1:00:00 16/06/2016 11:00:00 16/06/2016 22:00:00	Decrease Increase Decrease	1
9 (22118-22857)	15/07/2016 1:00:00 16/07/2016 13:00:00 17/07/2016 6:00:00	Decrease Increase Decrease	3
10 (22858-23566)	15/08/2016 19:00:00 17/08/2016 0:00:00 17/08/2016 23:00:00	Increase Increase Decrease	3
11 (23567-24252)	11/09/2016 21:00:00 15/09/2016 23:00:00 16/09/2016 13:00:00	Increase Decrease Increase	6
12 (24253-24967)	16/10/2016 15:00:00 16/10/2016 20:00:00 16/10/2016 23:00:00	Decrease Decrease Decrease	1
1 (24968-25946)	18/11/2016 9:00:00 20/11/2016 19:00:00 20/11/2016 23:00:00	Increase Decrease Decrease	3
2 (25947-26881)	14/12/2016 11:00:00 17/12/2016 1:00:00 17/12/2016 21:00:00	Increase Decrease Increase	4
3 (26882-27907	15/01/2017 10:00:00 18/01/2017 1:00:00 19/01/2017 6:00:00	Increase Decrease Increase	5
4 (27908-28621)	14/02/2017 14:00:00 15/02/2017 8:00:00 16/02/2017 19:00:00	Decrease Increase Decrease	3
5 (28262-29486)	17/03/2017 8:00:00 17/03/2017 21:00:00 18/03/2017 2:00:00	Increase Decrease Decrease	2
6 (29487-	15/04/2017 7:00:00 16/04/2017 1:00:00	Increase Decrease	2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

30449)	16/04/2017 19:00:00	Increase	
7 (30450-31401)	15/05/2017 15:00:00	Decrease	6
	16/05/2017 10:00:00	Decrease	
	20/05/2017 11:00:00	Increase	
8 (31402-32218)	15/06/2017 2:00:00	Decrease	3
	15/06/2017 21:00:00	Increase	
	17/06/2017 9:00:00	Increase	
9 (32219-33068)	14/07/2017 23:00:00	Decrease	5
	17/07/2017 13:00:00	Decrease	
	19/07/2017 8:00:00	Increase	
10 (33069-34028)	16/08/2017 0:00:00	Decrease	2
	16/08/2017 8:00:00	Increase	
	17/08/2017 5:00:00	Decrease	
11 (34029-34833)	14/09/2017 20:00:00	Increase	5
	15/09/2017 2:00:00	Decrease	
	18/09/2017 9:00:00	Increase	
12 (34834-35761)	15/10/2017 2:00:00	Decrease	2
	16/10/2017 7:00:00	Increase	
	16/10/2017 20:00:00	Decrease	
1 (35762-36605)	14/11/2017 17:00:00	Increase	2
	14/11/2017 23:00:00	Decrease	
	15/11/2017 20:00:00	Increase	
2 (36606-37486)	13/12/2017 13:00:00	Increase	7
	16/12/2017 2:00:00	Decrease	
	21/12/2017 5:00:00	Increase	
3 (37487-38370)	15/01/2018 8:00:00	Decrease	2
	15/01/2018 22:00:00	Decrease	
	16/01/2018 13:00:00	Increase	
4 (38371-39174)	15/02/2018 6:00:00	Increase	3
	16/02/2018 21:00:00	Decrease	
	17/02/2018 7:00:00	Decrease	
5 (39175-40041)	14/03/2018 23:00:00	Decrease	4
	15/03/2018 19:00:00	Increase	
	17/03/2018 15:00:00	Increase	
6 (40042-40943)	14/04/2018 9:00:00	Decrease	3
	14/04/2018 23:00:00	Decrease	
	16/04/2018 8:00:00	Increase	
7 (40944-41814)	15/05/2018 14:00:00	Increase	5
	16/05/2018 2:00:00	Decrease	
	19/05/2018 8:00:00	Increase	
8 (41815-42684)	15/06/2018 21:00:00	Increase	3
	17/06/2018 6:00:00	Decrease	
	17/06/2018 12:00:00	Increase	
9 (42278-43554)	14/07/2018 7:00:00	Decrease	4
	15/07/2018 5:00:00	Decrease	
	18/07/2018 6:00:00	Increase	
10 (43555-44196)	12/08/2018 1:00:00	Decrease	1
	12/08/2018 8:00:00	Increase	
	12/08/2018 15:00:00	Increase	
11 (44197-44493)	11/09/2018 18:00:00	Increase	2
	12/09/2018 14:00:00	Decrease	
	12/09/2018 15:00:00	Increase	

Table 1 contains information about the patterns generated based on the graphical analysis in figure 3. The number of rows in the window column corresponds to the number of sequences, as well as the year column based on the predetermined window interval. Each window has a month interval resulting from the calculation of the distance of similarity and each month interval has an increase or decrease in traffic data, this data is obtained from the points of similarity in figure 3. Changes in the pattern occur along with changes in time in units of days that are shown in the time column. With an increase and decrease in traffic on each window, this information includes interesting things to be used as a rule in making decisions or predictions in the future.

Based on table 1 generated 63 rules, each window has a rule, the format of the rules used is the format using the proposition logic and AND if A and B occur, then C occurs in time T. Here AB and C are basic forms, i.e. points similarity produced by discretization in table 1. Using mathematical notation in equation 13, 63 rules are generated The rule is then searched for strong correlations and tends to calculate high levels of confidence by using equation 16, to obtain the following rules:

Table 2: Rule With Confidence Based On The Pattern Found

No	Rule	Confidence
1	IF Vehicle volume increases and decreases THEN will increase in 2 to 7 days with a volume of 2000 to 5000 vehicles	77%
2	IF Vehicle volume increases and decreases THEN will decrease in 2 to 3 days in the morning and evening with a volume of 536 to 700 vehicles	23%
3	IF Vehicle volume decreases and increases THEN will increase in 1 to 4 days in the morning, afternoon and evening with a volume of 4000 to 5000 vehicles	41%
4	IF vehicle volume decreases and increases THEN will decrease in 1 to 4 days in the morning, afternoon and evening with a volume of 2000 to 5000 vehicles	59%
5	IF Vehicle volume decreases and decreases THEN will increase in 1 to 5 days at 6 to 22 with a volume of 2000 to 5000 vehicles	93%

To get rules that are close to the original data, then try to trim the rules by determining the level of confidence for each rule obtained. 5 The above rules are the result of merging the month

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

intervals along with changes in the pattern found from the original rules of 63 rules. 7 rules are sorted based on the level of confidence or confidence in each rule. With the discovery of the rules that have been obtained by 5 rules and different levels of confidence. For the discovery of knowledge that is more maximally done ranking rules by using J-measure.

Table 3: Results of ranking rules with J-Measure

No	Rule	J-Measure
1	IF Vehicle volume increases and decreases THEN will increase in 2 to 7 days with a volume of 2000 to 5000 vehicles	-2,37
2	IF Vehicle volume decreases and decreases THEN will increase in 1 to 5 days at 6 to 22 with a volume of 2000 to 5000 vehicles	-1,95
3	IF vehicle volume decreases and increases THEN will decrease in 1 to 4 days in the morning, afternoon and evening with a volume of 2000 to 5000 vehicles	-1,66
4	IF Vehicle volume increases and decreases THEN will decrease in 2 to 3 days in the morning and evening with a volume of 536 to 700 vehicles	-1,35
5	IF Vehicle volume decreases and increases THEN will increase in 1 to 4 days in the morning, afternoon and evening with a volume of 4000 to 5000 vehicles	-1,35

Ranking results with measurement J are sorted by the smallest results. As a summary of the results and compare them with the results of measurement J for each rule.

3.6. Best Rule Time Series

Based on the analysis of knowledge from each window it can be determined rules that have the potential to be attractive and cut rules that are not interesting, rules that tend to have redundancy, rules that do not have redundancy, then the best rule is IF Vehicle volume increases and decreases THEN will increase in 2 to 7 days with a volume of 2000 to 5000 vehicles. The rule is used to predict the volume level of vehicles that pass every day.

3.7. Testing and Validation of Data Sets

For daily data set testing using 111 data samples as follows:

Table 4: Testing Data Sets

	Date/Time	Traffic		Date/Time	Trafik
1	03/10/2012 0:00	506	56	05/10/2012 13:00	5967
2	03/10/2012 1:00	321	57	05/10/2012 14:00	6108
3	03/10/2012 2:00	273	58	05/10/2012 15:00	6128
4	03/10/2012 3:00	367	59	05/10/2012 16:00	6069
5	03/10/2012 4:00	814	60	05/10/2012 17:00	6169
6	03/10/2012 5:00	2718	61	05/10/2012 18:00	5384
7	03/10/2012 6:00	5673	62	05/10/2012 19:00	4063
8	03/10/2012 8:00	6511	63	05/10/2012 20:00	3364
9	03/10/2012 9:00	5471	64	05/10/2012 21:00	2971
10	03/10/2012 12:00	5097	65	05/10/2012 22:00	2450
11	03/10/2012 13:00	4887	66	05/10/2012 23:00	1832
12	03/10/2012 14:00	5337	67	06/10/2012 0:00	1231
13	03/10/2012 15:00	5692	68	06/10/2012 1:00	718
14	03/10/2012 16:00	6137	69	06/10/2012 2:00	545
15	03/10/2012 18:00	4623	70	06/10/2012 4:00	411
16	03/10/2012 19:00	3591	71	06/10/2012 5:00	688
17	03/10/2012 20:00	2898	72	06/10/2012 6:00	1352
18	03/10/2012 21:00	2637	73	06/10/2012 7:00	2072
19	03/10/2012 22:00	1777	74	06/10/2012 8:00	3342
20	03/10/2012 23:00	1015	75	06/10/2012 9:00	4258
21	04/10/2012 0:00	598	76	06/10/2012 10:00	4861
22	04/10/2012 1:00	369	77	06/10/2012 11:00	5191
23	04/10/2012 2:00	312	78	06/10/2012 12:00	5669
24	04/10/2012 3:00	367	79	06/10/2012 13:00	5074
25	04/10/2012 4:00	835	80	06/10/2012 14:00	5025
26	04/10/2012 5:00	2726	81	06/10/2012 15:00	5153
27	04/10/2012 6:00	5689	82	06/10/2012 16:00	5024
28	04/10/2012 7:00	6990	83	06/10/2012 17:00	4779
29	04/10/2012 8:00	5985	84	06/10/2012 18:00	4362
30	04/10/2012 9:00	5309	85	06/10/2012 19:00	3860
31	04/10/2012 10:00	4603	86	06/10/2012 20:00	3160
32	04/10/2012 11:00	4884	87	06/10/2012 21:00	2992
33	04/10/2012 12:00	5104	88	06/10/2012 22:00	3202
34	04/10/2012 13:00	5178	89	06/10/2012 23:00	1941
35	04/10/2012 14:00	5501	90	07/10/2012 0:00	1233
36	04/10/2012 15:00	5713	91	07/10/2012 3:00	323
37	04/10/2012 16:00	6292	92	07/10/2012 4:00	312
38	04/10/2012 17:00	6057	93	07/10/2012 5:00	627
39	04/10/2012 18:00	4907	94	07/10/2012 6:00	1508
40	04/10/2012 19:00	3503	95	07/10/2012 7:00	1539
41	04/10/2012 20:00	3037	96	07/10/2012 8:00	2232
42	04/10/2012 21:00	2822	97	07/10/2012 9:00	4568
43	04/10/2012 22:00	1992	98	07/10/2012 10:00	4624
44	04/10/2012 23:00	1166	99	07/10/2012 11:00	4529
45	05/10/2012 0:00	627	100	07/10/2012 12:00	5139
46	05/10/2012 1:00	388	101	07/10/2012 13:00	6640
47	05/10/2012 3:00	368	102	07/10/2012 14:00	6589
48	05/10/2012 5:00	2489	103	07/10/2012 15:00	4429
49	05/10/2012 6:00	5537	104	07/10/2012 16:00	4329
50	05/10/2012 7:00	6628	105	07/10/2012 17:00	3833
51	05/10/2012 8:00	5534	106	07/10/2012 18:00	4015
52	05/10/2012 9:00	5403	107	07/10/2012 19:00	3382
53	05/10/2012 10:00	4770	108	07/10/2012 20:00	2815
54	05/10/2012 11:00	5217	109	07/10/2012 21:00	2110
55	05/10/2012 12:00	5870	110	07/10/2012 22:00	1555
			111	07/10/2012 23:00	959

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Graphically, similarity results are displayed as follows:

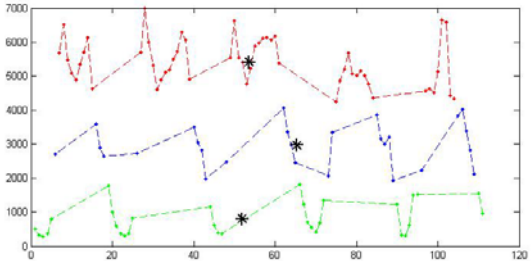

Figure 4: Results of the similarity data set

Clusters of variance data set daily using equations 7,8,9 and 10

V_i^2=23,900,057.48

V_2^2=7,109,285.91

V_3^2=9,971,080.03

Next, calculate the variance within the similarity

V_w=379, 448.365

Variance between similarity

V_b=225,041,580.1

Then the variance of all similarity is

$V_{minimum}$=0,001=0,1%

Based on the calculation above, it can be concluded that for testing the daily data set the minimum value is 0.1%. Then an error evaluation is performed to find out the model error level found, while the method used is Mean absolute deviation (MAD) to measure accuracy, Mean squared deviation (MSD) is used to measure the accuracy of the time series value, Mean absolute percentage error (MAPE) is used to look for errors / absolute errors in each period. The measurement results are presented in the following table

Table 5: The results of validation and error evaluation

No	Y_t	$\widehat{Y}_t$	$\lvert Y_t - \widehat{Y}_t \rvert$	$\lvert Y_t - \widehat{Y}_t \rvert^2$	$\frac{\lvert Y_t - \widehat{Y}_t \rvert^2}{Y_t}$
1	15.25	15	0.25	0.0625	0.016393
2	15.35	15	0.35	0.1225	0.022801
3	15.75	15.6	0.15	0.0225	0.009524
4	14.96	15.6	0.64	0.4096	0.042781
5	15.72	15	0.72	0.5184	0.045802
6	16.27	15	1.27	1.6129	0.078058
7	15.56	15.6	0.04	0.0016	0.002571
8	15.9	15	0.9	0.81	0.056604
9	16.58	15	1.58	2.4964	0.095296
10	16.09	15.6	0.49	0.2401	0.030454
11	14.88	15	0.12	0.0144	0.008065
12	15.97	15	0.97	0.9409	0.060739
13	16.02	15.6	0.42	0.1764	0.026217
14	15.55	15	0.55	0.3025	0.03537
15	14.84	15	0.16	0.0256	0.010782
16	15.99	15.6	0.39	0.1521	0.02439
17	14.29	15.6	1.31	1.7161	0.091672
18	15.72	15	0.72	0.5184	0.045802
19	17.43	15	2.43	5.9049	0.139415
20	16.94	15.6	1.34	1.7956	0.079103
21	15.73	15.6	0.13	0.0169	0.008264
22	15.87	15	0.87	0.7569	0.05482
23	17.21	15.6	1.61	2.5921	0.09355
24	15.43	15	0.43	0.1849	0.027868
25	14.1	15	0.9	0.81	0.06383
26	13.72	15.6	1.88	3.5344	0.137026
27	14.65	15.6	0.95	0.9025	0.064846
28	13.34	15	1.66	2.7556	0.124438
29	15.04	15	0.04	0.0016	0.00266
30	14.93	15.6	0.67	0.4489	0.044876
31	15.68	15.6	0.08	0.0064	0.005102
32	14.5	15.6	1.1	1.21	0.075862
33	14.13	15	0.87	0.7569	0.061571
34	16.17	15.6	0.57	0.3249	0.03525
35	15.22	15	0.22	0.0484	0.014455
36	14.71	15	0.29	0.0841	0.019714
37	15.16	15	0.16	0.0256	0.010554
Σ	570.65	564.6	27.23	32.3035	1.766524

$$MAD = \frac{1}{n}\sum_{t=1}^{n}|Y_t - \widehat{Y}_t| = \frac{1}{37}x27.23 = 0.73 = 73\%$$

$$MSD = \frac{\sum_{t=1}^{n}|Y_t - \widehat{Y}_t|^2}{n} = \frac{32.3035}{37} = 0.87 = 87\%$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \widehat{Y}_t|}{Y_t} = \frac{1}{37}x1.766 = 0.047 = 4,7\%$$

Based on the calculation above, the accuracy rate of the model is 73%, forecasting accuracy is 87% and the percentage of errors is 4.7%.

4.CONCLUSIONS

The new model in optimizing rule discovery in large-scale data mining time series is by using the RBT (Rule Best Time series) model, this model discretizes periodic series for sub-sequence formation. Grouping of periodic sub-sequences includes grouping of periodic sub-sequence sequences extracted through sliding windows, ie grouping segments from a long-term periodic sequence, then determining similarity, which is the level to which a periodic sequence is considered to resemble other periodic series using distance Euclid. Then the rule discovery is applied to obtain hidden rules on the temporal pattern and rank with the J-measure. From the results of this study time series data can be optimized, new knowledge or trends and patterns in time series databases that are uncertain and previously unknown can be generated. The decision or information can be used to display decisions, or forecasting in the future with an accuracy rate of the model mean absolute deviation (MAD) of 73%, forecasting accuracy of the mean squared deviation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(MSE) of 87% and the percentage of the mean absolute percentage error of the MAPE of 4,7%.

ACKNOWLEDGMENT

I would like to thank Mr. Herman Mawengkang, Mr. Muhammad Zarlis and Mr. Syahril Efendi as a dissertation promotor of Doctoral Program of Universitas Sumatera Utara and thank to an honorable of The Ministry of Research and Technology Indonesian for funding this research.

REFERENCES:

[1] 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.

[2] S. A. P. Rosyidi, M. R. Taha, and Z. Chik, "Acta Geophysica Signal Reconstruction of Surface Waves on SASW Measurement Using Gaussian Derivative Wavelet Transform," vol. 57, no. 3, pp. 616–635, 2009.

[3] 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.

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

[5] A. S. Ahmar, S. Guritno, and A. Rahman, "Modeling Data Containing Outliers using ARIMA Additive Outlier (ARIMA-AO) Modeling Data Containing Outliers using ARIMA Additive Outlier (ARIMA-AO)," 2018.

[6] 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.

[7] W. Java, "MULTI-OBJECTIVE OPTIMAL REACTIVE POWER DISPATCH USING HYBRID TIME VARYING PARTICLE," vol. 97, no. 19, pp. 5103–5114, 2019.

[8] V. Berikov, Y. Amirgaliyev, L. Cherikbayeva, D. Yedilkhan, and B. Tulegenova, "CLASSIFICATION AT INCOMPLETE TRAINING INFORMATION : USAGE OF GROUP CLUSTERING TO," vol. 97, no. 19, pp. 5048–5060, 2019.

[9] R. Buaton, H. Mawengkang, M. Zarlis, and S. Effendi, "Time Series Optimization on Data Mining," 2019.

[10] C. Berberidis, "Principles of Data Mining and Knowledge Discovery," *Conf. Pap.*, vol. 2431, no. September 2014, 2014.

[11] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," *Proc. ninth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '03*, no. May, p. 493, 2003.

[12] Y. Tanaka and K. Uehara, "Discover motifs in multi-dimensional time-series using the principal component analysis and the MDL principle," *Mach. Learn. Data Min. Pattern Recognition, Proc.*, vol. 2734, pp. 252–265, 2003.

[13] 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.

[14] J. Serr, "Ranking and significance of variable-length similarity-based time series motifs ′," 2015.

[15] N. Castro and P. J. Azevedo, "Time Series Motifs Statistical Significance," 2008.

[16] A. Mueen and N. Chavoshi, "Enumeration of time series motifs of all lengths," *Knowl. Inf. Syst.*, pp. 105–132, 2015.

[17] A. C. Onate, O. Adebimpe, and B. O. Adebesin, "Information-theoretic measure of the hyperbolical exponential-type potential," pp. 402–414, 2018.

[18] R.M. Konijn, "Detecting interesting differences: Data mining in health insurasnce data using outlier detection and subgroup discovery.," Vrije Universiteit, Netherland, 2016.

[19] C. J. Veenman, M. J. T. Reinders, and E. Backer, "A Maximum Variance Cluster Algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1–19, 2002.

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

[21] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, "A Review of Subsequence Time Series Clustering A Review of Subsequence Time Series Clustering," no.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

June, 2014.

[22] 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.

[23] 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.

[24] J. A. Tolentino and B. D. Gerardo, "Enhanced Manhattan-based Clustering using Fuzzy C-Means Algorithm for High Dimensional Datasets," vol. 9, no. 3, pp. 766–771, 2019.

[25] Z. A. Othman, N. Ismail, A. A. Bakar, M. T. Latif, S. Mastura, and S. Abdullah, "Rules Discovery of High Ozone in Klang Areas using Data Mining Approach," vol. 8, no. 6, pp. 2683–2689, 2018.