IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 2 (Feb. 2013), ||V2|| PP 06-13

Noise Removal in Distributed Time Series Database Using Predominant Pattern Distribution Model

B.Sujatha[1], Dr. S.Chenthur Pandian[2]
[1] Dept.of CSE, Sengunthar Engineering College, Tiruchengode, Tamilnadu, India.
[2] Principal, Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India.

***Abstract*:** A time series database is a collection of well-defined data sets obtained through repeated measurements of time. The data to be examined are regularly noisy and diverse periodicity types. The existing suffix tree based periodic pattern mining algorithm can detect symbol, sequence and segment periodicity in time series data with noise filters for diverse noise kinds. But the running time desired to identify the patterns without redundancy is high. To overcome this issue, in this paper, predominant pattern distribution model is introduced with which redundant and unwanted noisy patterns are identified and discarded from it. Predominant patterns are extracted with automatic or user defined threshold of pattern of interest, generated from the dynamic online time series data. Performance of proposed framework is measured and evaluated in terms of periodic pattern mining accuracy, noise distribution rate, and predominant pattern occurrence.

***Keywords*:** *Time series, periodic pattern mining, periodicity types, suffix tree, predominant pattern distribution model.*

I. INTRODUCTION

A time series is a collection of data sets accumulated at regular interval time to indicate certain behavior of an entity. Real time examples of time series data, e.g., meteorological data, network delays, power consumption, stock growth, gene expression data analysis, etc. Data mining is the process of extracting patterns and trends from large amounts of data stored in data repositories that uses statistical and mathematical techniques. Study in time series data mining has concentrated on extracting different types of patterns.

Periodicity detection is a process of discovering temporal regularities of data within the time series and the objective of analyzing a time series is to find whether and how a periodic pattern is repeated within the series. A time series is described by a set of repeating cycles.

A time series is discretized [1],[3],[8] before it is analyzed. Let T= $e_0,e_1,e_{2,...,}e_{n-1}$ be a time series with n events, where e_i denotes the event recorded at time i, time series T may be discretized by considering m different ranges such that all values in the same range are represented by ∑. For example, consider the time series including the hourly number of transactions in a superstore; the discretization procedure may define the following mapping by taking the distinguished possible range of transactions; {0} transactions: a, {1-100} transactions: b,{101-200} transactions: c,{201-300} transactions :d,{>300} transactions: e. Based on this mapping, the time series T=147,162,185,296,76,0,0,95 can be discretized into T' = cccdbaab. In general, three types of periodic patterns can be detected in a time series:

1) symbol periodicity,
2) sequence periodicity or partial periodic patterns, and
3) segment or full-cycle periodicity.

A time series is known as symbol periodicity if at least one symbol is repeated periodically. For e.g., in time series T= abd acb cab abc, symbol a is periodic with periodicity p=3, starting at position zero (stPos=0).

Next, a pattern consisting of more than one symbol may be periodic in a time series and this is known as partial periodic patterns.

If the whole time series can be represented as a repetition of pattern or segment, and then this kind of periodicity is called segment or full-cycle periodicity. For example, the time series T= abcab abcab abcab has segment periodicity of 5. (p=5) starting at the first positions (stops=0).That is ,T consists of only three occurrences of the segment abcab.

To conclude, time series exists often in our daily life and their analysis could lead to valuable discoveries. Those are identifying three types of periodic patterns, handling asynchronous periodicity by locating periodic patterns that may drift from their expected positions and finding periodic pattern in the whole time series as well as in a subsection of the time series using suffix tree. The algorithm proposed in this paper can identify and discard the unwanted and redundant data from the time series data and predominant pattern are extracted with automatic or user defined threshold. It is applicable to biological datasets and it has been analyzed for periodic pattern mining accuracy, noise distribution rate, and predominant pattern occurrence.

II. LITERATURE REVIEW

Existing work on time series examination approximately faces two kinds of algorithms. The primary group comprises algorithms that need the user to identify the period, and then appear only for patterns happening with that period. The second class, conversely, is algorithms which appear for all probable periods in the time series. Our algorithm group; it performs more than the other algorithms by appearing for all probable periods initiating from all potential locations inside a pre-specified range, whether the complete time series or a section of the time series. In [1], the author presented an algorithm which can identify sign, series (partial), and section (full cycle) periodicity in time series. The algorithm employs suffix tree as the fundamental data structure; this permits us to plan the algorithm such that it's most terrible case difficulty of the time series. The algorithm is noise pliant; it has been productively established to effort with substitute, addition, crossing out, or a combination of these kinds of noise.

Classification has been employed for representing numerous types of data sets, counting deposits of items, text credentials, graphs, and systems. Representing such data is helpful with the budding GPS and RFID knowledge and is significant for efficient haulage and traffic setting up. In [2], the author considered techniques for categorizing routes on road networks. Periodicity mining is employed for forecasting development in time series data. Determining the time at which the time series is cyclic has always been an obstruction for completely mechanized periodicity mining. In [3], the crisis of noticing the periodicity time of a database is addressed. Two types of periodicities are distinct, and a scalable, computationally proficient algorithm is planned for every type. Conventional outline growth-based strategies for chronological pattern mining obtain patterns supported on the estimated databases recursively. At every level of recursion, they unidirectional produce the span of noticed patterns by one all along the suffix of noticed patterns, which wants k stages of recursion to discover a pattern [4].

To decrease the number of models and develop the efficiency of the algorithm, Lo et al. have also commenced mining blocked iterative patterns, i.e., maximal patterns devoid of any super-pattern containing the similar support. In [5], to officially intensify study on iterative pattern mining, the authors initiate mining iterative generators, i.e., negligible patterns devoid of any sub-pattern containing the similar support. Periodic Pattern Mining or Periodicity Detection has numerous applications such as Prediction, Forecasting, Detection of Unusual events, etc. The periodic patterns are detected in a Time-Series database depending on the time intervals [6].

In [7], the author described a dynamic periodicity discovery algorithm to determine periodicity in DNA series. Our algorithm supported suffix tree as the fundamental data structure. The proposed strategy in [8] believes the periodicity of substitute substrings, besides allowing for active window to notice the periodicity of convinced illustrations of substrings. Nevertheless, even devoid of nested patterns, the lingo is influential sufficient to detain different forward-temporal stipulations from numerous release source requests [9], [10]. In the Move Mine system [11], a position of normally employed touching object mining purposes are constructed and a user-friendly boundary is presented to make possible interactive examination of touching object data mining and bendable alteration of the mining restraints and parameters. Analyzing such data consumes more time and noises on the set of data proceed in it.

Research in time series data mining has concerted on determining diverse kinds of patterns: chronological patterns sequential patterns, episodic connection rules, incomplete cyclic patterns and astonishing patterns to forename a few. This periodicity pulling out methods need the user to identify a time that decides the time at which the time series is cyclic. They imagine that users either recognize the value of the period earlier or are enthusiastic to attempt different period values in anticipation of acceptable cyclic patterns emerge. Since the mining process must be performed frequently to attain good results, this trial-and-error method is obviously not competent. Even in the study of time series information with a priori recognized periods, there might be incomprehensible periods and, as a result, motivating periodic patterns that will not be exposed. The clarification to these troubles is to invent methods for determining possible periods in time series data. In this work, a highly efficient periodic pattern mining model is presented for determining the unwanted data in the distributed time series data and processed the database in terms of user specified threshold value.

III. PROPOSED PREDOMINANT PATTERN DISTRIBUTION MODEL FOR NOISE DISTRIBUTED TIME SERIES DATABASE

3.1 Time series database

A Time-Series Database comprises of data values grouped at consistent period of time to produce convinced actions for an entity. There exist numerous examples for Time-Series Database, for instance weather circumstances of a distinct position, stock development and communication in a superstore, power utilization and earthquake occurrences. Predominant pattern distribution model is a procedure of identifying the temporal regularities involved in the Time-Series. The objective of examining a Time-Series Database is to discover how common a periodic prototype (full or partial) is repetitive in a particular time interval. Three types of periodic patterns are present in time series. They are given below:

- Symbol Periodicity
- Sequence Periodicity or Partial Periodic Patterns
- Segment or Full-Cycle Periodicity

3.1.1. Symbol Periodicity

A Time-Series is supposed to be a symbol periodicity, if no less than one symbol occurs repetitively. For instance, in a Time-Series, let T = abdacbabdabc, symbol ‘a’ is cyclic inside periodicity p = 4.

3.1.2. Sequence Periodicity

A Time-Series is supposed to be a Sequence Periodicity, if more than one symbol might be cyclic and it is also termed as limited periodic patterns. For example, in a TimeSeries Database let T = ‘bbaaabbdabcaabbcabcd’ then the series ‘ab’ is cyclic inside periodicity p = 4.

3.1.3. Segment Periodicity

A Time-Series is supposed to be a Segment Periodicity, if the entire Time-Series is typically symbolized as a replication of a model or segment and it is also recognized as full-cycle periodicity. For example, in a Time-Series Database let T=’abcababcababcab’, T comprises of only three incidences of the segment abcab.

It is not essential to always have ideal periodicity in a time series as in the preceding three examples. Generally, the degree of excellence is characterized by confidence factor, which is 100 percent in the three examples. The confidence of a prototype is termed as the part of its genuine occurrence (actual) in the series over its projected (predicted) occurrence in the series. The genuine (actual) and estimated (predicted) perfect frequency are both similar in the specified three instances. Nevertheless, in the time series T_x = ‘abefd abcde acbcd abefa’, the pattern ‘ab’ initiating at location 0 with p = 5 contain four and five as its genuine and estimated ideal frequencies, correspondingly; so the self-reliance of ab is 4 / 5.

The first part of the work concentrates on building highly efficient noisy removal to the unwanted data, distributed across the time series. The architecture diagram of the proposed predominant pattern distribution model for noise distributed time series database [PPDMND] is shown in fig 3.1.

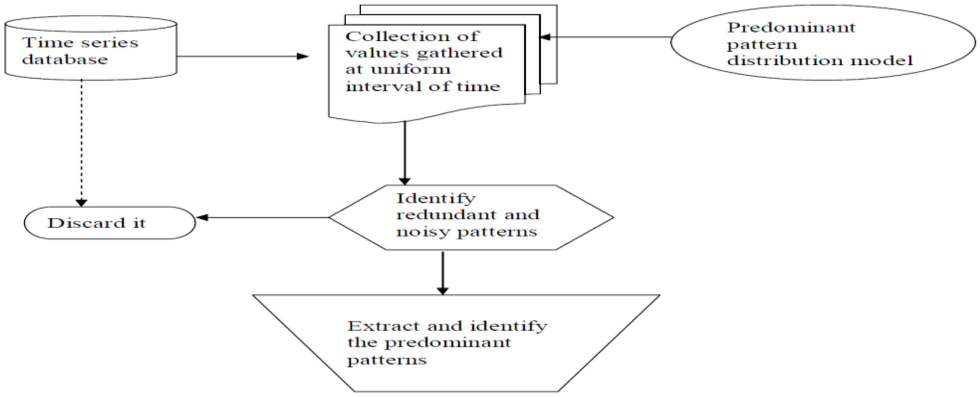

Time series database
Collection of values gathered at uniform interval of time
Predominant pattern distribution model
Identify redundant and noisy patterns
Discard it
Extract and identify the predominant patterns

Fig 3.1 Architecture diagram of the proposed PPDMND

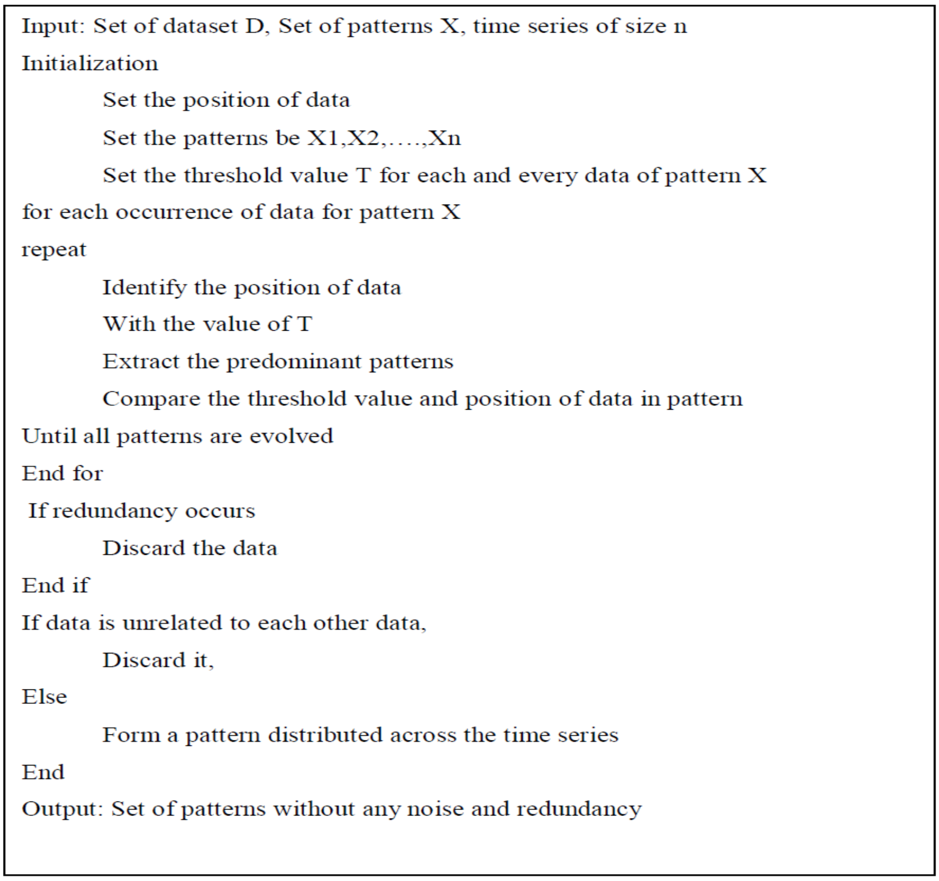
From the fig 3.1, it is being observed that predominant patterns are extracted with automatic or user defined threshold of pattern of interest generated from the dynamic online time series data. The predominant pattern distribution model is introduced with which redundant and unwanted noisy patterns are identified and discarded from the time series data.

3.2 Periodicity detection in presence of noise

Three types of noise usually measured in time series data are substitution, addition, and deletion noise. In substitution noise, various symbols in the dishonored time series are restored at arbitrary with further symbols. In case of addition and deletion noise, several symbols are interleaved or deleted, correspondingly, arbitrarily at diverse locations (or time values). Noise is also a combination of these three kinds, for example, substitution type noise resources the consistent combination of replacement (R) and addition (I) noise. When the time series is moreover completely cyclic or includes only substitution noise and achieves inadequately in the occurrence of addition or deletion noise. This is since insertion and deletion noise develops or deals the time axis important to move of the imaginative time series values.

For instance, time series T = abcabcabc after adding symbol b at locations 2 and 6 would be T' = abbcabcabbc. The event for sign a in T is (0; 3; 6), as it is (0; 4; 7) in T'. It is very obvious that when the time series is indistinct by addition and or removal noise, STNR do not execute well. To enhance the noise detection process, in this work, we present a predominant distribution model to remove the redundant and unwanted noisy distribution of data in the time series dataset.

With the set of patterns obtained from the time series database, the predominant patterns are identified. Before that, a user defined threshold value 'T' is set for all the patterns derived. The threshold values are used to identify the interesting patterns without any redundancy in data. After that predominant distribution model is presented to identify the unwanted data, distributed on the time series dataset by using the threshold values. The procedure below describes the predominant pattern distribution model.



```
Input: Set of dataset D, Set of patterns X, time series of size n
Initialization
        Set the position of data
        Set the patterns be X1,X2,….,Xn
        Set the threshold value T for each and every data of pattern X
for each occurrence of data for pattern X
repeat
        Identify the position of data
        With the value of T
        Extract the predominant patterns
        Compare the threshold value and position of data in pattern
Until all patterns are evolved
End for
 If redundancy occurs
        Discard the data
End if
If data is unrelated to each other data,
        Discard it,
Else
        Form a pattern distributed across the time series
End
Output: Set of patterns without any noise and redundancy
```

Fig 3.2 Process of PPDMND

With the above process, a highly efficient predominant pattern distribution model is presented for removal of noise and redundancy for the data being observed in the time series data.

IV. EXPERIMENTAL EVALUATION

In this section, we provide the outcome of numerous experiments processed using both synthetic and real data. The outcome of trying different individuality for the proposed PPDMND algorithm is compared against the existing algorithm STNR (Suffix-Tree-based Noise-Resilient algorithm). As mentioned in Section 2, there are two kinds of algorithms explained in the literature. Algorithms in the initial group discover periodic patterns for a precise periodic value and those in the other group ensure the time series for each and every time. The first set of tests reveals the comprehensive nature of PPDMND that it should be capable to discover a time once it exists in the time series. We perform an analysis to show how PPDMND suits this on both synthetic and real data.

The parameters proscribed through data creation are data allocation (uniform or normal), alphabet extent (number of exclusive signs in the data) and dimension of the data (quantity of symbols in the data), time period size, the style and quantity of noise in the data. A datum might hold substitution, addition, and removal noise or any combination of these kinds of noise. For real data experiments, the Packet data have been employed, where the authors established the intense periodic patterns, i.e., the area where the time series is typically cyclic; the three regions which were originated cyclic for symbol 'a' and its patterns are given as below:

Table 4.1 Periodic pattern in packet data

Pattern	Period	Confidence
aa	2	0.42
aa	2	0.42
aa	2	0.47
aaa	3	0.44
aaa	3	0.43
aaa	3	0.47
aaa*a	5	0.44
aaaa*	5	0.36
aaa**	5	0.44

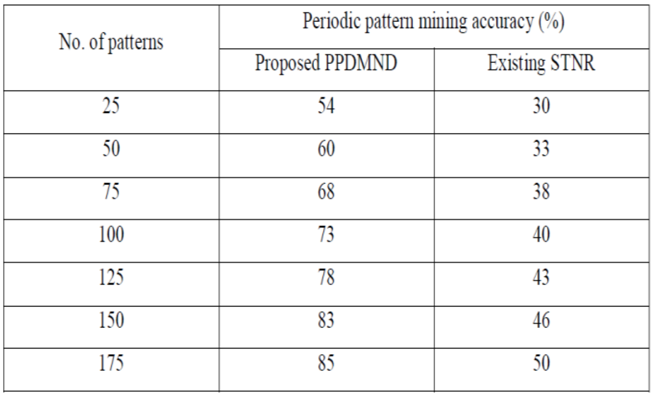
The PPDMND algorithm discover all the periods in the intense periodic section, generally at superior poise level. Since the confidence level of the periodicity patterns increases, the noise level in the periodic data is low compared to an existing STNR technique. The performance of the proposed predominant pattern distribution model for noise distributed time series database [PPDMND] is measured in terms of

- Periodic pattern mining accuracy
- Noise distribution rate
- Exécution time

V. RESULTS AND DISCUSSION

In this work, we have seen how the noise has been removed from the set of periodic patterns in the time series dataset. At first, periodic pattern distribution model is applied to the time series database to remove the redundant and unwanted noisy patterns present in it. Experiments are conducted with several time series dataset and the performance has also been evaluated. The table given below and graph describes the performance of the proposed predominant pattern distribution model for noise distributed time series database and compared the results with the existing STNR technique [1].

Table 5.1 No. of patterns vs. Periodic pattern mining accuracy



No. of patterns	Periodic pattern mining accuracy (%)	
	Proposed PPDMND	Existing STNR
25	54	30
50	60	33
75	68	38
100	73	40
125	78	43
150	83	46
175	85	50

The accuracy of periodic patterns in the time series dataset is analyzed based on the number of patterns formed is illustrated in the table 5.1 for the proposed PPDMND and existing STNR.

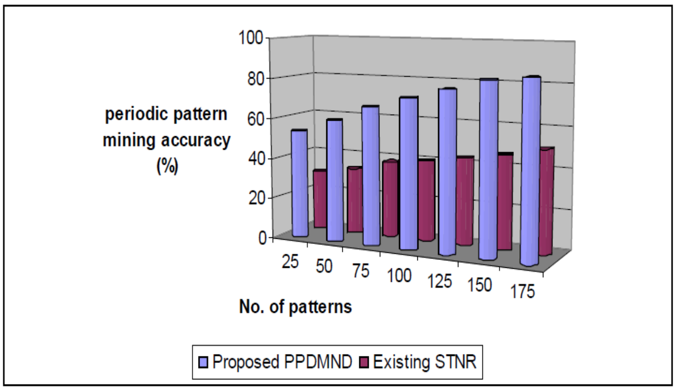


Fig 5.1 No. of patterns vs. Periodic pattern mining accuracy

Fig 5.1 describes the accuracy of the periodic patterns in the time series dataset is analyzed based on the number of patterns formed. The accuracy of the patterns is measured based on the confidence level of the patterns. The confidence of a pattern is termed as the ratio of its genuine occurrence in the series over its estimated ideal frequency in the series. Since the proposed PPDMND set a user defined threshold value for patterns generated from the dynamic time series data, the predominant patterns are automatically extracted. With that, redundant and unwanted data in the patterns are easily identified by building a predominant pattern distribution model. So, the accuracy of the periodic patterns in the dataset is high in the proposed PPDMND. Compared to the existing STNR (Suffix-Tree-based Noise-Resilient algorithm), the proposed PPDMND provides high level of accuracy on periodic patterns for the given dataset and the variance is 35-45% high in it.

Table 5.2 Periods vs. Noise distribution rate

Periods	Noise distribution rate	
	Proposed PPDMND	Existing STNR
20	10	20
40	15	25
60	13	30
80	18	28
100	20	35
120	25	33
140	23	40

The noise distribution rate for the periodic patterns present in the time series dataset is illustrated in the table 5.2.

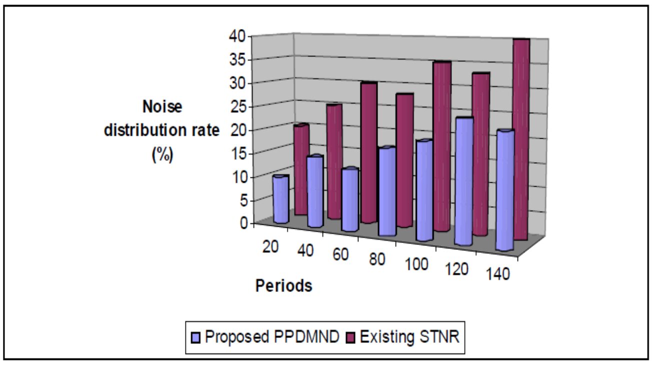


Fig 5.2 Periods vs. Noise distribution rate

Fig 5.2 describes the noise distribution rate for the periods obtained for assigning the patterns present in the time series dataset. The noise distribution rate is measured in terms of presence of noise in the specified data distributed in the given dataset. Since the proposed PPDMND presented a predominant pattern distribution model, the noisy and redundant data are removed from the distributed dataset by setting a user defined threshold. With the threshold value, the mechanism makes insertion, deletion and modification of patterns in the given dataset. While changing the patterns, the user defined threshold value might change. So, the process of noise distribution is less in the proposed PPDMND. Compared to the existing STNR (Suffix-Tree-based Noise-Resilient algorithm), the proposed PPDMND provides less noise distribution on the given dataset and the variance is 45-55% low in it.

Table 5.3 Noise ratio vs. Exécution time

Noise ratio	Execution time (secs)	
	Proposed PPDMND	Existing STNR
0.1	0.5	2
0.2	1	2.3
0.3	0.9	2.5
0.4	0.5	2
0.5	0.7	1.9
0.6	0.9	2.2
0.7	1	2.6

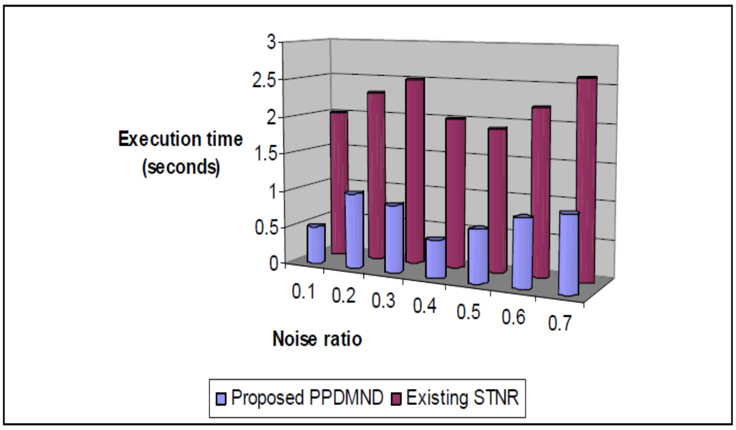


Fig 5.3 Noise ratio vs. Exécution time

The next set of researches determines the impact of noise ratio on the time presentation of the proposed PPDMND. For this trial, we set the time series span, size, alphabet size, and data allocation and considered the force of unstable noise ratio on time recital of the algorithm. The results are plotted in Fig. 5.3. These experiments have been processed employing the two algorithms: STNR, and PPDMND. It is true that PPDMND consumes more time when the noise ratio lies between 10-15 percent; but when the noise ratio is level is high, PPDMND tends to obtain the analogous time. As the results show, the noise ratio does not concern the time presentation of existing STNR.

Finally, it is being observed that the proposed PPDMND efficiently removed the noise and discard the unwanted data, distributed in the time series dataset by adapting the predominant pattern distribution model. At first, predominant patterns are extracted from the dataset by setting a user threshold pattern of interest, generated from the dynamic online time series data. Then the periodic patterns are extracted devoid of any noise in the generated patterns.

VI. CONCLUSION

In this paper, we have presented PPDMND, a predominant pattern distribution model for removal of noise in time series data. The PPDMND algorithm is noise-resilient and processes the time series dataset efficiently even in the worst case. The single algorithm discover three types of periodicity involved in time series which comprises of symbol, series (partial periodic), and segment (full cycle) periodicity. It also discovers the periodicity inside a part of the time series. An efficient noisy removal strategy is introduced to remove the unwanted data, distributed across the time series. Predominant patterns are extracted not only using automatic pattern of interest but also using a user defined threshold, generated from the dynamic online time series data. Several experiments were conducted to show the time behavior, accuracy, and noise resilience characteristics of the data. The algorithm is processed using both real and synthetic data. The results generated provide the significant of the efficiency and accuracy of predominant patterns involved in the time series dataset.

REFERENCES

[1] Faraz Rasheed et. Al., "Efficient Periodicity Mining in Time Series Databases Using Suffix Trees", IEEE

[2] TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011

[3] Jae-Gil Lee et. Al., "Mining Discriminative Patterns for Classifying Trajectories on Road Networks", IEEE

[4] TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 5, MAY 2011

[5] Mohamed G. Elfeky et. Al., "Periodicity Detection in Time Series Databases", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 7, JULY 2005

[6] Jinlin Chen et. Al., "An UpDown Directed Acyclic Graph Approach for Sequential Pattern Mining", IEEE

[7] TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 7, JULY 2010

[8] David Lo et. Al., "Mining Iterative Generators and Representative Rules for Software Specification Discovery",

[9] IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 2, FEBRUARY 2011

[10] O.Obulesu et. Al., " Finding Maximal Periodic Patterns and Pruning Strategy in Spatiotemporal Databases",

[11] International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012 ISSN: 2277 128X

[12] F. Rasheed, M. Alshalalfa, and R. Alhajj, "Adapting Machine Learning Technique for Periodicity Detection in

[13] Nucleosomal Locations in Sequences," Proc. Eighth Int'l Conf. Intelligent Data Eng. and automated Learning (IDEAL), pp. 870-879, Dec. 2007.

[14] F. Rasheed and R. Alhajj, "STNR: A Suffix Tree Based Noise Resilient Algorithm for Periodicity Detection in Time Series Databases," Applied Intelligence, vol. 32, no. 3, pp. 267-278, 2010.

[15] D. Lo, S. Maoz, and S.-C. Khoo, "Mining Modal Scenario-Based Specifications from Execution Traces of Reactive Systems," Proc. ACM/IEEE Int'l Conf. Automated Software Eng., 2007.

[16] D. Lo and S. Maoz, "Mining Scenario-Based Triggers and Effects," Proc. ACM/IEEE Int'l Conf. Automated Software Eng., 2008.

[17] Zhenhui Li, J.Han, Ming Ji, Lu-An Tang, Y. Yu, Bolin Ding,R.Kays and J.Lee, "MoveMine:Mining Moving Object Data for Discovery of Animal Movement Patterns:, ACM Journal Name,Vol.,No., 05 2010, Pages 111–077.