ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Design & Implementation of Efficient Periodicity Mining Technique for Time Series Data

Prof. Saneep Khanna[1], Mr. Swapnil Kasurkar[2]

Assistant Professor, Department of Computer Engineering at Padm Dr VB Kolte COE Malkapur, India[1]

Dr VB Kolte COE Malkapur, India[2]

Abstract: In almost every scientific field, measurements are performed over time. These observations lead to a collection of organized data called time series. The purpose of time-series data mining is to try to extract all meaningful knowledge from the shape of data. Even if humans have a natural capacity to perform these tasks, it remains a complex problem for computers. The first paper was based on study of various periodicity detection techniques and extracts their advantages and disadvantages. In this paper we intend to provide the result of new efficient technique for periodicity detection. This paper includes finding of three type of periodic pattern symbol periodicity, sequence periodicity or partial periodic pattern and segment or full-cycle periodic. The degrees of perfection calculated by confidence, and are mostly characterized by the presence of noise in the data. In this paper, we address the problem of detecting the periodicity rate of a time series database. Three types of periodicities are defined, and a scalable, computationally efficient algorithm is proposed for each type. The algorithms perform in O(nlogn) time for a time series of length n. Moreover, the proposed algorithms are extended in order to discover the periodic patterns of unknown periods at the same time without affecting the time complexity. Experimental results show that the proposed algorithms are highly accurate with respect to the discovered periodicity rates and periodic patterns. Real-data experiments demonstrate the practicality of the discovered periodic patterns.

Keywords: Time series, Data Mining, confidence, efficient algorithm, Periodicity detection.

I. INTRODUCTION

The article provides a survey of the techniques applied for time-series data mining and new efficient technique that gives better result. The first paper is devoted to an overview of the tasks that have captured most of the interest of researchers. Considering that in most cases, time-series task relies on the same components for implementation, the literature depending on these common aspects is divided, namely representation techniques, distance measures, and indexing methods. The study of the relevant literature has been categorized for each individual aspect. Four types of robustness could then be formalized and any kind of distance could then be classified.

A time series is a collection of data values gathered generally at uniform interval of time to reflect certain behaviour of an entity [1]. Real life has several examples of time series such as weather conditions of a particular location, spending patterns, stock growth, transactions in a superstore, network delays, power consumption, computer network fault analysis and security breach detection, earthquake prediction, gene expression data analysis, etc. A time series is mostly characterized by being composed of repeating cycles [1][2][3].

A time series is mostly discretized before it is analyzed [4] [5] [6]. Several example of time series such as frequently sold products in a retail market, frequent regular interval pattern in DNA sequence, stock growth, power consumption, computer network fault analysis, transactions in a superstore, gene expression data analysis [1] [5] etc. In the above examples, we observe that the occurrence periodicity plays an important role in discovering some interesting frequent patterns in a wide variety of application areas. Identifying repeating (periodic) patterns could reveal important observations about the behaviour and future trends of the case represented by the time series [7], and hence would lead to more effective decision making. The goal of time series analysis is to find whether and how frequent a periodic pattern (full or partial) is repeated within the data. In time series is said to have three types of periodic patterns (symbol / Sequence / Segment) can be detected [1][2][3][4].

In time series periodicity can be full or partial[4]. Also several noises can be present in the time series. Methods used for finding full periodicity cannot be used for detecting partial periodicity. Partial periodicity is a looser kind of periodicity than full periodicity and it exist ubiquitously in the real world [4]. So partial periodic detection is a expensive mining process because it is the mixture of periodic and non-periodic events. Another problem occurs in periodic detection is the presence of noise. Most of the algorithms have poor resilience to noise. Another problem in periodicity detection is perfect periodicity [1]. All the periodicity in time series database is not perfect. The degree of perfection of time series can be represented in terms of confidence. The periodicity mining algorithm requires user to specify a periodic length that determines the rate at which the time series is periodic. This cannot be done in trial and error method. The solution of this problem is to devise a technique for discovering the potential periods in the time series data followed by the application of any existing pattern mining technique to determine the interesting pattern. [4]

DOI 10.17148/IJARCCE.2015.4595

To sum-up, time series exist frequently in our daily life and their analysis could lead to valuable discoveries. So there is need for noise resilience algorithm that can tackle the problem of i) Identifying three different type of periodic pattern ii) handling asynchronous periodicity. Discovering the rate at which the time series is periodic has always been an obstacle for fully automated periodicity mining. Existing periodicity mining algorithms assume that the periodicity rate (or simply the period) is user-specified. This assumption is a considerable limitation, especially in time series data where the period is not known a priori. In this paper, we address the problem of detecting the periodicity rate of a time series database. Two types of periodicities are defined, and a scalable, computationally efficient algorithm is proposed for each type. The algorithms perform in O(nlogn) time for a time series of length n. Moreover, the proposed algorithms are extended in order to discover the periodic patterns of unknown periods at the same time without affecting the time complexity. Experimental results show that the proposed algorithms are highly accurate with respect to the discovered periodicity rates and periodic patterns. Real-data experiments demonstrate the practicality of the discovered periodic patterns.

II. GLOSSARY OF TERMS RELATED TO TIME SERIES

A. Symbol periodicity

A time series T is said to have symbol periodicity for a given symbol s with period p and starting position stPos if the periodicity of s in T is either perfect or imperfect with high confidence, i.e., s occurs in T at most of the positions specified by stPos+i*p, where p is the period and integer i >= 0 takes consecutive values starting at 0. For example, in T= bacdabcaccabcabd, the symbol a is periodic with stPos =1 and p=3, i.e., a occurs in T at positions 1, 4, 7, 10 and 13[8][9][10].

B. Confidence

The confidence of a periodic pattern occurring in time series T is the ratio of its actual periodicity to its expected perfect periodicity.

Conf=actual frequency/expected frequency
For example, in T= abccaaccdbabcdbabbca, the pattern ab is periodic with stops=0, p=5, and conf=3/4. Note that the confidence is 1 when perfect periodicity is achieved [8][9][10].

C. Perfect Periodicity

Consider a time series T, a pattern X is said to satisfy perfect periodicity in T with period p if starting from the first occurrence of X until the end of T every next occurrence of X exists p positions away from the current occurrence of X. It is possible to have some of the expected occurrences of X missing and this leads to imperfect periodicity [8][9][10].

D. Sequence Periodicity

A time series T is said to have sequence periodicity or partial periodicity for a pattern X starting at position stPos, if |X|= 1 and the periodicity of X in T is either perfect or imperfect with high confidence, i.e., X occurs in T at most of the positions specified by stPos+i*p,

where p is the period and integer i >=0 takes consecutive values starting at 0.

For example, in T= acbcaaccdbaccdbacbca, the pattern ac is partially periodic starting from stP os= 0 and period value of 5[8][9][10].

E. Segment Periodicity

A time series T of length n is said to have segment periodicity for a pattern X with period p and starting position stPos if p=|X| and the periodicity of X in T is either perfect or imperfect with high confidence,

i.e., X occurs in T at most of the positions specified by stPos+i*p, where integer i >=0 takes consecutive values starting at 0 [8][9][10].

F. Periodicity in Subsection of a Time Series

A time series T possesses symbol, sequence, or segment periodicity with period p between positions stPos and endPos
(Where 0<=stPos < endPos <=|T|), if the investigated period satisfies above 3 concepts by considering only subsection [stPos,endPos] of T. For example, in T = babacbcaaccdbaccdbacbca, the pattern ac is periodic with p=5 in the subsection [3, 19] of T [8][9][10].

Above 4 concepts will not be applicable in case the time series contains some insertion or deletion noise.

Unfortunately, it is not possible to avoid noise in real-life applications.

Thus, to be able to tackle noisy time series, we introduce the concept of time tolerance, denoted tt, where a periodic occurrence is counted valid if it is found within +tt or -tt positions of the expected position.

G. Periodicity with Time Tolerance

Given a time series T which is not necessarily noise-free, a pattern X is periodic in subsection [stP os,endPos] of T with period p and time tolerance tt>=0 if X is found at positions stPos, stPos+p+tt , stPos+2p+tt…., endPos –p+tt or stPos+p-tt , stPos+2p-tt…., endPos -p-tt.[8]

H. Time Series Representation and Indexing

One of the major reasons for time series representation is to reduce the dimension (i.e.the number of data point) of

DOI 10.17148/IJARCCE.2015.4595

the original data. The simplest method perhaps is sampling (Astrom, 1969). [8]

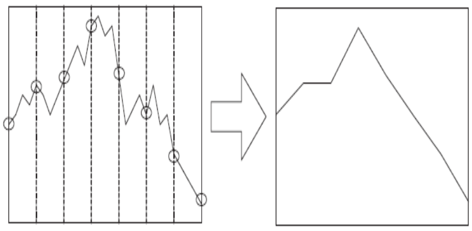


Fig 1 Time Series Dimensionality Reduction by Sampling the Time Series on the Left is Sampled Regularly (Denoted by Dotted Lines) and Displayed on the Right with a Large Distortion.

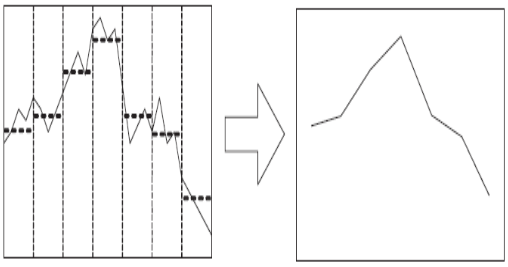


Fig. 2 Time Series Dimensionality Reduction by PAA the Horizontal Dotted Lines Show the Mean of Each Segment.

In this method, a rate of m/n is used, where m is the length of a time series P and n is the dimension after dimensionality reduction. However, the sampling method has the drawback of distorting the shape of sampled/compressed time series, if the sampling rate is too low. [9][10]

I.FFT

FFT re-expresses the discrete Fourier transform of an arbitrary composite size $N = N_1N_2$ in terms of smaller DFTs of sizes N_1 and N_2, recursively, in order to reduce the computation time to O($N \log N$) for highly-composite N (smooth numbers). Because of the algorithm's importance, specific variants and implementation styles have become known by their own names, as described below. FFT plays an important role in data mining for determining the periodic patterns.

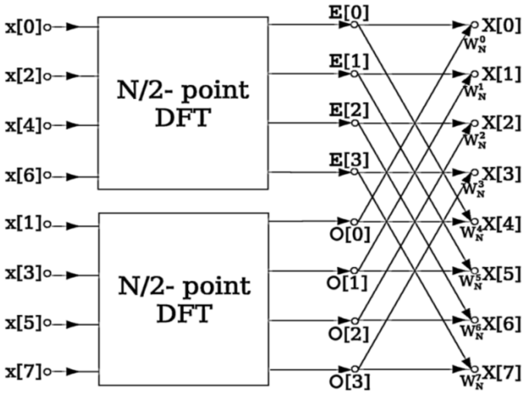


Fig 3 Data Flow Diagram for N=8: A Decimation-In-Time Radix-2 FFT

Breaks a Length-N DFT Into Two Length-N/2 Dfts Followed By a Combining Stage Consisting of Many Size-2 Dfts Called "Butterfly" [11]

J. Autocorrelation

Correlation is a mathematical function related with making data stream with desired application. In *circular* autocorrelation, the point at the end of the series is shifted out of the product in every step and is moved to the beginning of the shifting vector. Hence in every step we compute the following dot product for all N points:

$$r(k) = \frac{1}{N}\sum_{x\text{-}1}^{N} f(x)f(x+k)$$

Consider, the time series $T = abcabdacba$, peak represent frequency of symbol at that time. The first non-zero autocorrelation value represents fundamental frequency of symbol in entire time series. The peak difference in consecutive peaks guide to compute period. In below fig, the symbol ‘a’ has a periodicity 4 and period is 3.

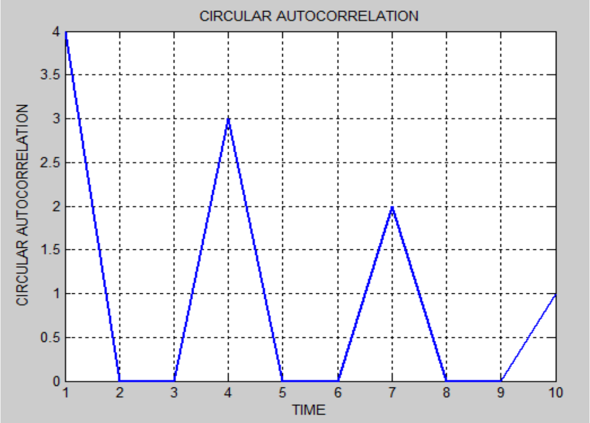


Fig 4Circular Autocorrelation Graph for Symbol ‘a’ [11]

III. SUFFIX TREE, ALGORITHM AND FLOWCHART

We will use Temperature and Weather data as input. And the input data will be processed using proposed algorithm in order to mine periodicity efficiently.The system will be operated by interactive GUI as a front end. Results will be displayed in tabular and graphical form.

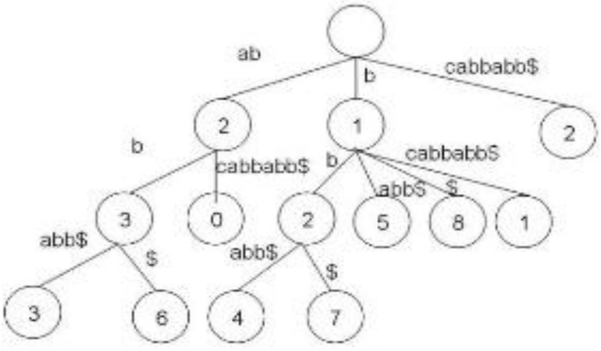


Fig 5 Suffix Tree for the String abcabbabb$[1]

The algorithm for finding the Symbol periodicity takes the symbolized time series as input and retrieves the symbols that are periodic. The suffix tree finds each and every token present in a given string [8].

A. Algorithm for symbol periodicity

Input*:* A symbolized time series sequence T = x1, x2 ....., xn, of length n.
Output*:* Symbol Periodic Patterns, Number of Occurrences, Perfect and Imperfect Periodic Rates.
Step-1) For the time series T, create a binary matrix M of size K*n values, in which every row represents binary vector to a particular symbol, where K is the Interval range specified by user and n is the size of the time series. The existence of a symbol is denoted by "1" and the non existence by "0" in binary vector for each symbol.
Step-2) Apply circular auto correlation to every row of the matrix M separately to find the circular auto correlation for every symbol using the formula,

$$r(k)=\frac{1}{N}\sum_{x\text{-}1}^{N} f(x)f(x+k)$$

Step-3) Every non-zero element of the resulted sequence represents the total number of occurrences of that symbol from that position. The first non-zero element represents the total number of occurrence of that symbol.
Step-4) The symbol that exceeds the minimum threshold percentage of occurrence is considered as a periodic symbol.
Step-5) The Index positions of the non-zero elements represent the starting position of the symbol pattern.
Step-6) The Period are derived from the index positions (PRi = Pi – Pi-1).
i) If the periodic rate of the symbol is the same in a minimum threshold percentage, it is considered as perfect periodic rate.
ii) If the periodic rate does not satisfy the above condition, it is considered as imperfect periodic rate. [11]

B. Algorithm representing workflow and input output
Input:

Let
TS[n] → Time-Series Sequence
SymN → Number of symbols for symbolization
Sym[SymN]→Symbol vector
stPos →Staring position of time series data
endPos → Ending position of time series data

Output: Periodic Patterns, Period, Confidence Measure

PerPattern[m] → Sequence of periodic patterns
Period→ period
ConfM(k)→ Confidence Measure of kth periodic pattern

Step 1) Symbolize the pre-processed input time series sequence by using SymN
Number of symbols.
Step 2) Select the time series frame from index stPos to endPos.

Step 3) Binarize the time series segment (frame) and get Symbol Representation
Matrix (SRM)
The SRM is SymN x (endPos-stPos) matrix
SRM(i,j) = {1 if SymN[i]=Sym[i]
{0 otherwise

Step 4) Deremine frequent patterns PerPattern[m]

Step 5) Calculate period and ConfM(k). [11]

C. Flowchart for Periodic Pattern Mining

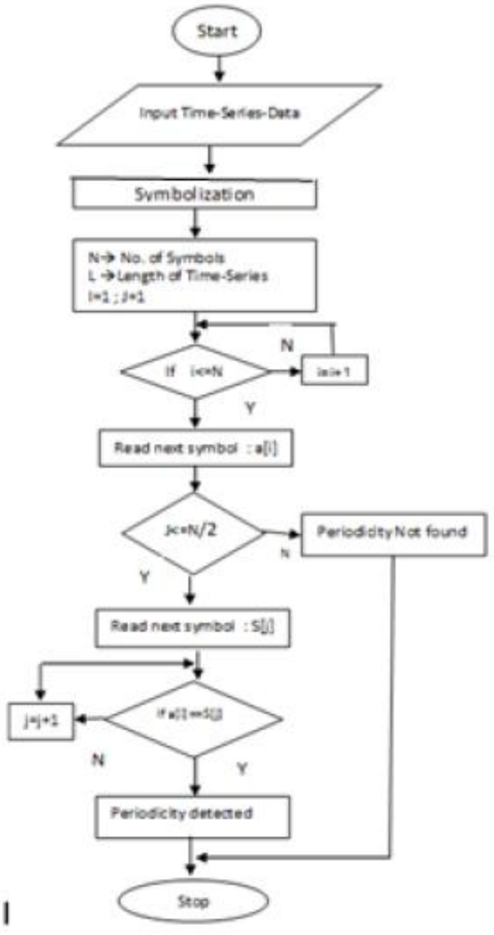


Fig. 6 Flowchart for periodicity detection [11]

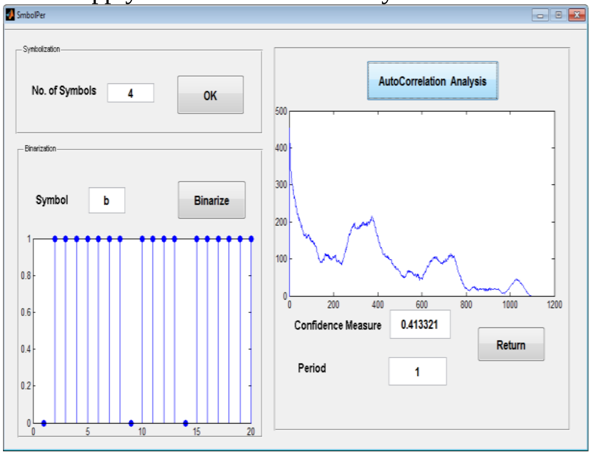
IV. IMPLEMENTATION

A. Symbol Periodicity

Symbol periodicity means to determine whether any particular symbol is repeating frequently or not i.e. to find if any symbol is periodic or not. This is accomplished as follows

DOI 10.17148/IJARCCE.2015.4595

- Enter number of symbols
- Binarize the symbol string
- Apply Auto-correlation analysis



No. of Symbols 4

Symbol b

Confidence Measure 0.413321

Period 1

Fig 7 Symbol Periodicity Detection

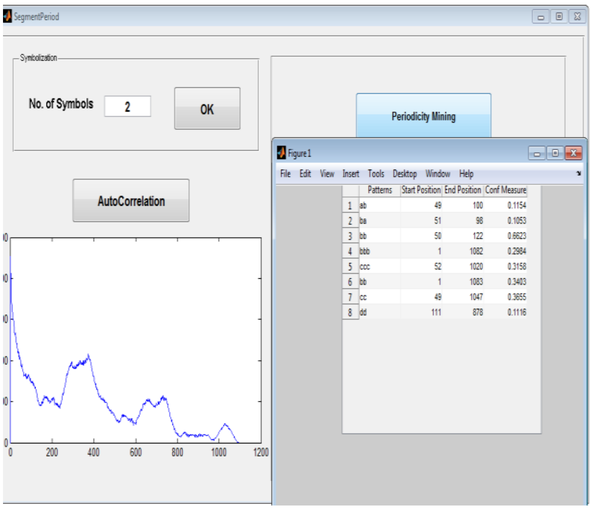
The output of this process is:

- Confidence measure
- Period

B. Segment Periodicity

Segment periodicity concerns the periodicity of the entire time series. Unlike symbol periodicity that focuses on the symbols (where different symbols may have different periods), segment periodicity focuses on the entire time series. A time series T is said to be periodic with a period p if it can be divided into equal-length segments, each of length p, these are "almost" similar. This is accomplished as follows

- Enter number of symbols
- Auto-Correlation
- Periodicity Mining



No. of Symbols 2

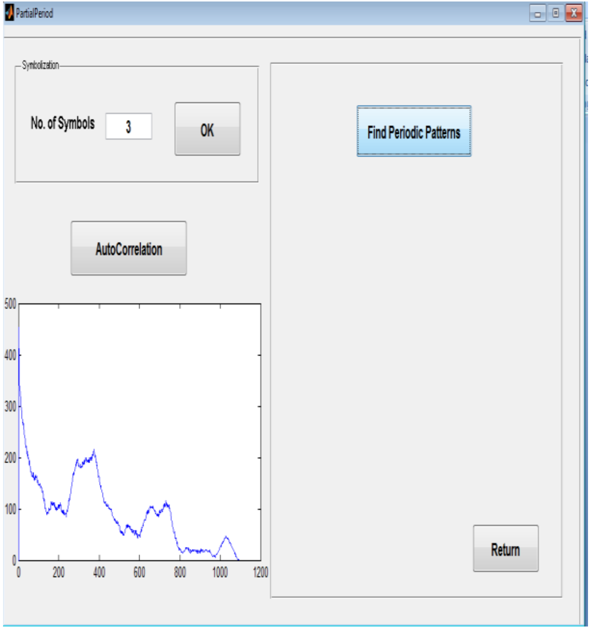
	Patterns	Start Position	End Position	Conf Measure
1	ab	49	100	0.1154
2	ba	51	98	0.1053
3	bb	50	122	0.6623
4	bbb	1	1082	0.2984
5	ccc	52	1020	0.3158
6	bb	1	1083	0.3403
7	cc	49	1047	0.3655
8	dd	111	878	0.1116

Fig 8 Segment Periodicity

C. Partial Periodicity

This module is used to determine partial periodic patterns from time-series data. This is accomplished as follows:

- Enter number of symbols
- Auto-Correlation
- Periodicity Mining



No. of Symbols 3

Fig 9 Partial Periodicity

V. RESULTS

A. Real Data

For real data analysis, we used the temperature data which contain daily temperature. The record contains temperature data of 36 months. The temperature data is discretized by interval width which is based on user input for interval. Here we find out max and min value of temperature, and then it is discretized into given user input intervals.

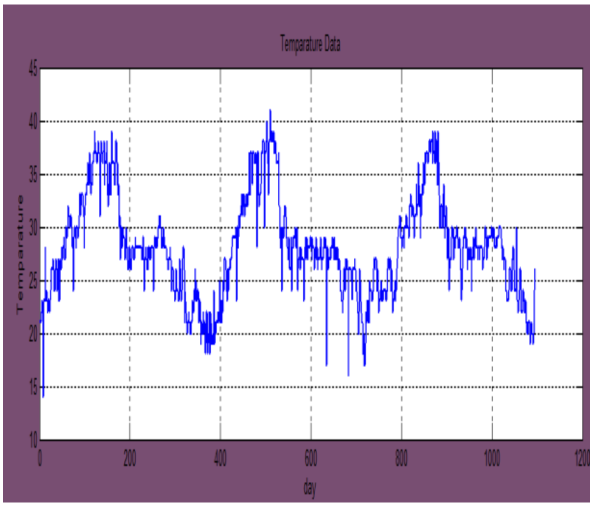


Fig 10 Temperature Data of Three Years

DOI 10.17148/IJARCCE.2015.4595

B. Symbolized String

The given set of elements converted to a symbolized string

aaaaaaaaaaaaaaaaaaabaabaaabbbabbbbbbbab baabbbbabbbbbcbbbbbbbbbbbbbbbbbbbbbbbcc bbbbbcbcbbbbbcbbcbbbaabcccccbbccccbbbcccc cccccc

C. Result of periodicity mining

```
---------------------------------------------------------
 Periodic   St_Pos   End_Pos  Period       Conf. Mes.
---------------------------------------------------------

  aaaa         1        12       4            0.62

  bbbb        13        29       4            0.56

  abbb         7        48       4            0.33

  babb         8       101       4            0.24

  a***         8       118       4            0.71
```

Fig 11 Result of Periodicity Mining

D. Comparison with Han's Algorithm

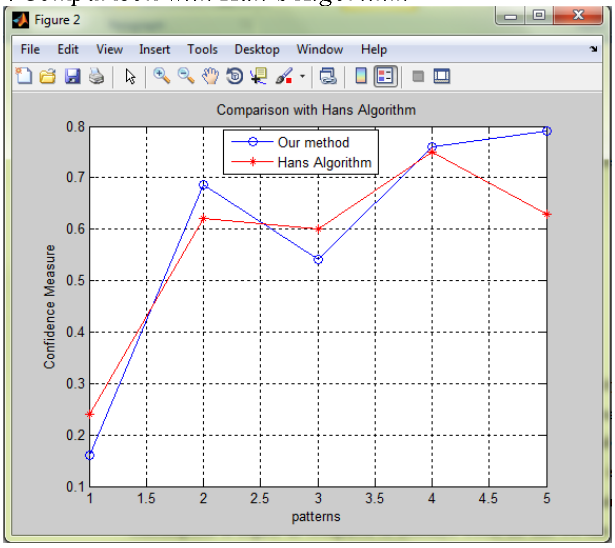


Fig 12 comparison with Hans Algorithm

From the above figure it is clear that the presented algorithm is more efficient than Han's algorithm.

VI. CONCLUSION

A great try to implement a simple but efficient periodicity mining technique System. The proposed method is mainly designed for study of temperature will result in required conclusion about weather. Proposed technique will temperature assumption for particular span of time. The reliability of the system has been shown with the high accuracy results reported in the previous sections. However, there are still some problems can not solved by this system.

ACKNOWLEDGMENT

We express our sincere gratitude to **Prof. B. K. Chaudhari**, Head of Computer Engineering Department, Padm Dr V. B. Kolte College of Engineering, Malkapur, for his generous guidance, help and useful suggestions. I also wish to extend my thanks to other staff members and colleagues for attending our seminar and for their insightful comments and constructive suggestions to improve the quality of this paper work. I am extremely thankful to **Dr. A. W. Kharche, Director**, Padm Dr V. B. Kolte College of Engineering, Malkapur, for providing me infrastructural facilities to work in, without which this work would not have been possible.

REFERENCES

[1] Faraz Rasheed, Mohammed Alshalfa, Reda Alhaji Associate Member,IEEE," Efficient periodicity Mining In Time Series Databases Using Suffix Trees" IEEE Trans. Knowledge and Data Eng., January2011,vol.23, no.1, pp.79-94.

[2] M.G. Elfeky, W.G. Aref, and A.K. Elmagarmid, "PeriodicityDetection in Time Series Databases," IEEE Trans. Knowledge andData Eng., July 2005,vol. 17, no. 7, pp. 875-887.

[3] M.G. Elfeky, W.G. Aref, and A.K. Elmagarmid, "WARP: TimeWarping for Periodicity Detection," Proc. Fifth IEEE Int'l Conf.Data Mining, Nov. 2005.

[4] J. Han, Y. Yin, and G. Dong, "Efficient Mining of Partial Periodic Patterns in Time Series Database," Proc. 15th IEEE Int'l Conf. DataEng., 1999, p. 106.

[5] E.F. Glynn, J. Chen, and A.R. Mushegian, "Detecting Periodic Patterns in Unevenly Spaced Gene Expression Time Series UsingLomb-Scargle Periodograms," Bioinformatics, Feb. 2006, vol. 22, no. 3 pp. 310316.

[6] Hiroshi Sugimura,Kazunori Matsumoto,"Classification System for Time Series Data Based on Feature Pattern Extraction", IEEE,2011.

[7] Elfeky MG, Aref WG, Elmagarmid AK (2004a) "Using convolution to mine obscure periodic patterns in one pass" In: EDBT 2004: Proceedings of the ninth international conference on extending database technology,2004,LNCS 2992, Berlin, Springer-Verlag, pp 605-620

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

[9] Yang J, Wang W, Yu PS "Mining asynchronous periodic patterns in time series data" IEEE Trans Knowl Data Eng, 2003, 15(3):613-628

[10] Huang K, Chang C SMCA: "A general model for mining asynchronous periodic patterns in temporal databases" IEEE Trans Knowl Data Eng,2005, 17(6):774-785

[11] Prof. Sandeep Khanna and Mr. Swapnil A. Kasurkar, "Astudy of various periodicity detection techniques for time series data and primary level plan for new technique", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 11, November 2013

DOI 10.17148/IJARCCE.2015.4595