№ 3 (3), 2021 ISSN: 2682 – 9517 (print) ISSN: 2683 – 0930 (online)
INSTITUTE OF INFORMATICS AND INNOVATIVE TECHNOLOGIES

FORECAST TIMESERIES BASED ON MATRIX PROFILE

Dung Tran Thi[1] **Nha Tran Phong**[2] **Ngoc Dung Bui**[3]

[1,2]Division of Information Technology, Campus in Ho Chi Minh City, University of Transport and Communications, Ho Chi Minh City, Viet Nam *dungtt_ph@utc.edu.vn; nhatp_ph@utc.edu.vn*

[3]Faculty of Information Technology, University of Transport and Communications, Ha Noi, Viet Nam *dnbui@utc.edu.vn*

Abstract. ***Time series prediction has many practical applications, and is therefore an area of interest to scientists. Many methods have been proposed, such as the linear prediction model, self-regression model, moving average model, artificial neural network model, and hidden Markov model, to name a few. However, these methods have the disadvantage of long calculation time and not many experimental cases to compare optimal results. In this paper, we propose a new method of time series prediction using matrix profile, which is a distance vector of pairs of motifs or adjacent pairs. With the application of Consecutive Neighborhood Preserving (CNP) property, the experimental results show that the proposed method has higher accuracy and faster calculation time than previous methods.***

Keywords: time series, motif, time series forecast, forecast, Consecutive Neighborhood Preserving.

I. INTRODUCTION

Time series forecasting is indispensable need for human activities in the context of an information explosion. Prediction will provide the necessary rationale for planning and it can be said that, without predictive science, the future human intentions outlined would not have significant convincing. Applications of time series forecasting are used in the fields of finance for stock price prediction [1], petroleum business forecast [2], university enrollment forecast [3], forecasting population report [3].

Time series forecasting is quickly becoming an indispensable need for in the face of the explosion on human activities information. Prediction will provide the necessary inputs for planning and it can be said that, without prediction science, outlines of our future intentions outlined would not be convincing. Thus far, time series forecasting is heavily utilized for stock price prediction [1], petroleum business forecast [2], university enrollment forecast [3], and forecasting population growth and change [3].

Many time series prediction methods have been proposed by researchers in recent years. In 2009, Jiang proposed a method of forecasting stock time series based on motif information [4]. In 2007, Lora used the closest number of neighbor techniques to forecast data [5]. In 2015, a new approach based on semantic hegemony algebra in the fuzzy time series prediction problem was introduced by Hieu and his colleagues [6]. In 2016, Tung et al. Used fuzzy time series following the incremental algebra approach to predict time series [7]. The prediction method by matrix profile is a new method applied to time series. This is the method of finding the closest neighbor of each subsequence in time series. Based on the closest proximity property, one can make predictions about the continuing values in the time series. The method tested on the neuroscience data set gave better results in terms of accuracy and timing

II. BACKGROUND

A. Definitions

Time series: If *T* is a time series then $T = (t_1, t_2, \ldots, t_n)$ include set n real-valued numbers [8].

Subsequence: A time series $T = (t_1, t_2, \ldots, t_n)$, a subsequence has length *n* of *T* is a subsequence $T_{i,m} = (t_i, t_{i+1}, \ldots, t_{i+m-1})$, $1 \le i \le m - n + 1$ [8].

B. Matrix profile definitions

Definition 1: A Matrix distances D_i corresponds to subsequence $T_{i,m}$ and time series *T* is a vector of the

Euclidean distance between a subsequence $T_{i,m}$ and each subsequence in time series T. Or $D_i = (d_{i,1}, d_{i,2}, ..., d_{i,n-m+1})$, where $d_{i,j}$ $(1 \leq j \leq n - m + 1)$ is the distances of $T_{i,m}$ and $T_{j,m}$ [9].

Definition 2: A Matrix profile *P* of the time series *T* is a vector of the range of Euclid's between each subsequence of *T* and nearest neighbor in *T*, the concept of closest proximity means that two pairs of subsequences have the smallest distance from other subsequences. Or,

$$P = [\min(D_1), \min(D_2), ..., \min(n - m + 1)],$$

where D_i $(1 \leq i \leq n - m + 1)$ is Matrix distances D_i corresponding to the query $T_{i,m}$ and time series *T* [9].

In Fig 1 shows the relationship between matrix distances, Matrix distances, and Matrix profile. Each component of the distance matrix $d_{i,j}$ is the distance between $T_{i,m}$ and $T_{j,m}$ (1 ⩽ i, j ⩽ n-m+1) in time series *T*.

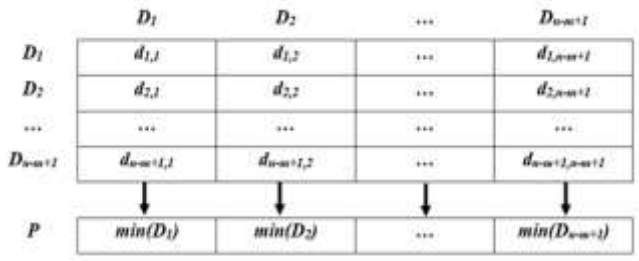

Fig. 1. Relationship between matrix distances, Matrix distances and Matrix profile ([9]).

The index *i* in the Matrix profile *P* tells us about the Euclidean distance between the subsequence $T_{i,m}$ and the closest neighbor in the *T* time series. However, it does not say the position of the nearest neighbors, so the Matrix profile indicators are given:

Definition 3: Matrix profile index *I* of time series *T* is a vector of integers: $I = [I_1, I_2, ... I_{n-m+1}]$, where $I_i = j$ if $d_{i,j} = \min(D_i)$ [9].

Index	1	2	3	4	...	7	8	9	...	24	25	...
I	56	57	112	113	...	116	133	134	...	149	150	...

Fig. 2. Matrix profile index example of a time series [9].

The minimum value position in each column is stored with the Matrix profile index.

C. *Algorithm*

The SCRIMP ++ algorithm is an algorithm that combines two algorithms: PreSCRIMP and SCRIMP. Algorithm PreSCRIMP is algorithm of finding approximate motif method, it has the complexity of $O(n^2 \log n/s)$. SCRIMP algorithm is an algorithm of exact search method and it has $O(n^2)$ complexity. SCRIMP algorithm uses PreSCRIMP algorithm as time series preprocessing, it is able to detect motif in time series and it only finds approximate Matrix Profile. From that approximate Matrix Profile will be input to the SCRIMP algorithm to find the correct Matrix Profile. I.e. the idea of SCRIMP++ algorithm [9]. For problems with large processing data, the SCRIMP++ algorithm can still be performed, we can stop at any time to find the motif without necessarily traversing the time series.

According to Consecutive Neighborhood Preserving (CNP) in preSCRIMP algorithm, if *i* and *j* are neighbors, then *i* + 1 is also a neighbor of *j* + 1. In Fig 3, the 11, 12, 13 and 14 subqueries correspond to their nearest neighbors, with 136, 137, 138 and 139 chains.

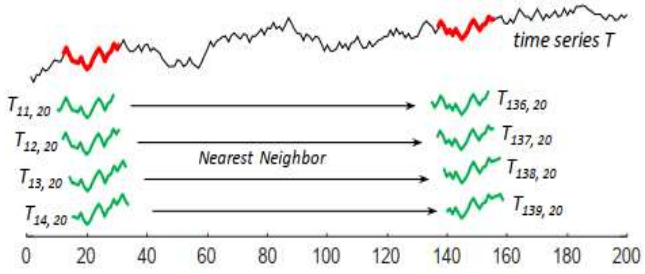

Fig. 3. Conseceutive Neighborhood Preserving (CNP) attribute [9]

D. *Proposed method*

Based on the results of SCRIMP ++ algorithm, we can use matrix profile to perform the prediction problem. Once we know the Matrix profile, we can see the most similar string pairs. Then based on the CNP property introduced in the preSCRIMP algorithm, we can predict the sequence appearing behind the last series in the time series, which is also the time series to predict.

However, it is because the sub-chains have different amplitudes in positions. Therefore, to make prediction series, we need to put prediction series with amplitude close to the actual series amplitude. To perform the normalization of the amplitude we do the following: Subtract the last point of the neighboring series from the last point in the time series, then we

will have a value of the deviation. of these 2 points. Next, we proceed to find the predicted series and normalize the amplitude by: Take each point of the series after the neighboring series plus the deviation to find the predicted series.

III. EXPERIMENT RESULTS

The Neuroscience dataset is a data set about human neuroscience. Scientists researched and introduced on many sources such as books, newspapers, websites.

Running experiments on Neuroscience data in these cases gives the results as shown below:

+ Case 1: Time series length: 512 points, subsequence length: 80 points, prediction string length: 20 points.

TABLE 1. RESULTS OF CASE EXPERIMENT 1

Points	Real value	Predictive value
513	-55,66406374	-55,32837038
514	-55,69458132	-55,02319459
515	-55,66406374	-54,80957154
516	-55,84716922	-54,68750123
517	-55,93872195	-54,56543091
518	-55,69458132	-54,59594849
519	-55,66406374	-54,80957154
520	-56,21338016	-54,80957154
521	-56,21338016	-54,62646607
522	-56,18286258	-54,65698365
523	-56,24389774	-55,26733522
524	-56,48803837	-55,38940554
525	-56,42700321	-55,48095828
526	-56,42700321	-55,41992312
527	-56,51855595	-55,48095828
528	-56,70166142	-55,57251101
529	-56,64062627	-55,78613406
530	-56,70166142	-56,12182743
531	-56,97631963	-56,42700322
532	-57,22046026	-56,51855596

According to the results in Table 1, the value difference between the predicted series and the real series is not too large. The largest difference in this case is approximately 1.5. Follow the chart shown in Fig 4 below to observe the difference (the blue series is the actual series and the red series is the predicted series).

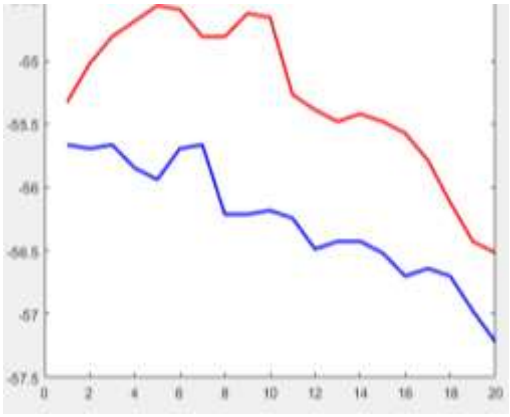

Fig. 4. Results of experimental series of predictions in case 1.

+ Case 2: Time series length: 1024 points, subsequence length: 100 points, predicted string length: 25 points.

TABLE 2. RESULTS OF CASE EXPERIMENT 2

Points	Real value	Predictive value	Difference
1025	-55,7556	-56,0913	0,335693
1026	-55,4504	-55,8472	0,396729
1027	-55,2979	-55,4504	0,152588
1028	-55,1453	-55,2368	0,091553
1029	-55,0232	-54,9011	0,12207
1030	-54,8706	-54,5959	0,274658
1031	-54,8706	-54,657	0,213623
1032	-54,7791	-54,5959	0,183105
1033	-54,5654	-54,4739	0,091553
1034	-54,5654	-54,4434	0,12207
1035	-54,7485	-54,6875	0,061035
1036	-54,5959	-54,718	0,12207
1037	-54,1992	-54,3518	0,152588
1038	-54,1687	-54,0161	0,152588
1039	-53,894	-54,1077	0,213623

1040	-53,5584	-53,9246	0,366211
1041	-53,1616	-53,5278	0,366211
1042	-52,948	-53,6499	0,701904
1043	-52,9175	-53,7415	0,823975
1044	-52,7649	-53,894	1,12915
1045	-52,8564	-53,9856	1,12915
1046	-53,009	-54,1382	1,12915
1047	-53,0701	-54,1992	1,12915
1048	-52,948	-54,2603	1,312256
1049	-53,1616	-54,3213	1,159668

According to the results in Table 2, the difference between predicted series and actual series is not too large. The biggest difference is 1.15. The graph in Fig 5 visually shows the experimental results.

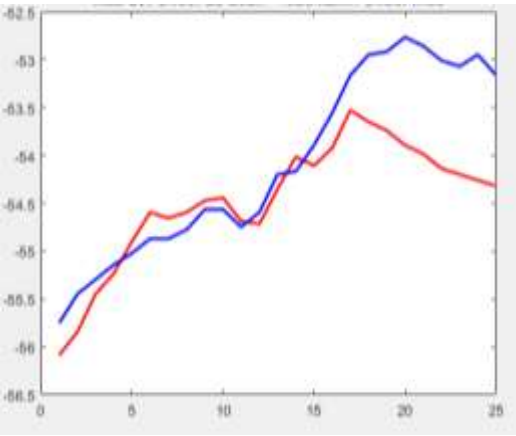

Fig. 5. Results of experimental series of predictions in case 2

+ Case 3: Time series length: 2048 points, subsequence length: 40 points, prediction string length: 10 points.

IV. CONCLUSIONS

Through the experimental cases, it shows that the predicted results are approximately equal to the real value. The up and down trends over the time series in the forecast results are similar to the real value. Since then, the method of finding motifs using matrix profiles to predict has a relatively good result and can be applied in practice. The next development direction may be to rely on different subsequence to predict the time series, not consider a subsequence.

ACKNOWLEDGMENT

This research is funded by University of Transport and Communications (UTC) under grant number T2020-PHII-003.

REFERENCES

[1] N. M. Dung, "Dự báo giá chứng khoán bằng phương pháp chuỗi thời gian", University of Science, Ha Noi, Viet Nam, 2014.

[2] T. V. T. Em, "Nghiên cứu ứng dụng chuỗi thời gian trong việc dự báo kinh doanh xăng dầu", Lac Hong Univesity, Viet Nam.

[3] N. V. Tinh and N. C. Dieu, "Dự báo chuỗi thời gian mờ dựa trên nhóm quan hệ mờ phụ thuộc thời gian và tối ưu bầy đàn", The 9th National Conference on Fundamental and Applied IT Research, Can Tho University, Viet Nam, 125-133.

[4] Z. Q. Jiang, W. J. Xie, Trading networks abnormal motifs and stock manipulation, Quantitative Finance Letters, 4 (2013) 1-8.

[5] A. T. Lora, J. M. R. Santos, A. G. Expósito, J. L. M. Ramos, Electricity Market Price Forecasting Based onWeighted Nearest Neighbors Techniques, IEEE TRANSACTIONS ON POWER SYSTEMS, 22 (2007), 1294-1301.

[6] N. D. Hieu, V. N. Lan, N. C. Ho, "Dự báo chuỗi thời gian mờ dựa trên ngữ nghĩa", The 8th National Conference on Fundamental and Applied IT Research, 2015, Ha Noi National University, Viet Nam, 232-243.

[7] H. Tung, N. Đ. Thuan, V. M. Lọc, "Phương pháp dự báo chuỗi thời gian trên chuỗi thời gian mờ theo tiếp cận đại số gia tử", The 9th National Conference on Fundamental and Applied IT Research , 2016, Can Tho University, Viet Nam, 161-177.

[8] A. Mueen, E. Keogh , Q. Zhu, S. Cash and B. Westover, Exact Discovery of Time Series Motifs, Proceedings of the SIAM International Conference on Data Mining, 2009, IEEE, Nevada, 473-484.

[9] Y. Zhu, C. M. Yeh, Z. Zimmerman, K. Kamgar, E. Keogh, Matrix Profile XI: SCRIMP++: Time Series Motif Discovery at Interactive, IEEE International Conference on Data Mining, 2018, IEEE, Singapore, 837-846.

[10] Y. Zhu, Z. Zimmerman, Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins, 2016 IEEE 16th International Conference on Data Mining, 2016, IEEE, Barcelona, 739-748