

Preventing Meaningless Stock Time Series Pattern Discovery by Changing Perceptually Important Point Detection

Tak-chung Fu^{1,2,†}, Fu-lai Chung¹, Robert Luk¹, and Chak-man Ng²

¹ Department of Computing, The Hong Kong Polytechnic University, Hong Kong.
{cstcfu, cskchung, csrluk}@comp.polyu.edu.hk

² Department of Computing and Information Management
Hong Kong Institute of Vocational Education (Chai Wan), Hong Kong.
cmng@vtc.edu.hk

Abstract. Discovery of interesting or frequently appearing time series patterns is one of the important tasks in various time series data mining applications. However, recent research criticized that discovering subsequence patterns in time series using clustering approaches is meaningless. It is due to the presence of trivial matched subsequences in the formation of the time series subsequences using sliding window method. The objective of this paper is to propose a threshold-free approach to improve the method for segmenting long stock time series into subsequences using sliding window. The proposed approach filters the trivial matched subsequences by changing Perceptually Important Point (PIP) detection and reduced the dimension by PIP identification.

1 Introduction

When time series data are divided into subsequences, interesting patterns can be discovered and it is easier to query, understand and mine them. Therefore, the discovery of frequently appearing time series patterns, or called surprising patterns in paper [1], has become one of the important tasks in various time series data mining applications.

For the problem of time series pattern discovery, a common technique being employed is clustering. However, applying clustering approaches to discover frequently appearing patterns is criticized as meaningless recently when focusing on time series subsequence [2]. It is because when sliding window is used to discretize the long time series into subsequences given with a fixed window size, trivial match subsequences always exist. The existing of such subsequences will lead to the discovery of patterns derivations from sine curve. A subsequence is said to be a trivial match when it is similar to its adjacent subsequence formed by sliding window, the best matches to a subsequence, apart from itself, tends to be the subsequence that begin just one or two points to the left or the right of the original subsequence [3]. Therefore, it is necessary to prevent the over-counting of these trivial matches. For example, in Fig.1, the shapes of S_1 , S_2 and S_3 are similar to a head and shoulders (H&S) patterns while the

[†] Corresponding Author.

shape of S_4 is completely different from them. Therefore, S_2 and S_3 should be considered as trivial matches to S_1 and we should only consider S_1 and S_4 in this case.

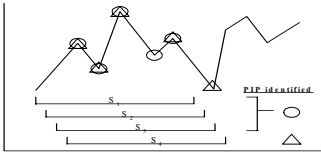


Fig. 1. Trivial match in time series subsequences (with PIPs identified in each subsequence)

References [3,4] defined the problem of enumerating the most frequently appearing pattern (which are called the most significant motifs, 1-motif, in reference [3]) in a time series P is the subsequence S_1 that has the highest count of non-trivial matches. Therefore, the K^{th} most frequently appearing pattern (significant motif, K -Motif) in P is the subsequence S_K that has the K highest count of non-trivial matches and satisfies $D(S_K, S_i) > 2R$, for all $1 \leq i < K$. However, it is difficult to define a threshold R , to distinguish trivial and non-trivial matches. It is case dependent and there is no general rule for defining this value. Furthermore, reference [2] suggested that applying a classic clustering algorithm in place of subsequence time series clustering to cluster only the motifs discovered from K -motif detection algorithm.

The objective of this paper is to develop a threshold-free approach to improve the segmentation method for segmenting long stock time series into subsequences using sliding window. The goal is redefined as to filter all the trivial matched subsequences formed by sliding window. The remaining subsequences should be considered as non-trivial matches for further frequently appearing pattern discovery process.

2 The Proposed Frequently Appearing Pattern Discovery Process

Given a time series $P = \{p_1, \dots, p_m\}$ and fixing the width of the sliding window at w , a set of time series segments $W(P) = \{S_i = [p_i, \dots, p_{i+w-1}] \mid i = 1, \dots, m - w + 1\}$ can be formed. To identify trivial matches from the matching process of subsequences formed by sliding window, a method based on detecting the changes of the identified Perceptually Important Points (PIPs) is introduced. PIP identification is first proposed in reference [5] for dimensionality reduction and pattern matching of stock time series. It is based on identifying the critical points of the time series as the general shape of a stock time series is typically characterized by a few points. By comparing the differences between the PIP identified between two consequent subsequences, a trivial match occurred if the same set of PIP is identified and the second subsequence can be ignored. Otherwise, both subsequences are non-trivial matched and should be considered as the subsequence candidates of the pattern discovery process. This process carries along from the starting subsequence of the time series obtained by using sliding window till the end of the series. In Fig.1, the same set of PIP is identified in the subsequence S_1 , S_2 and S_3 . Therefore, the matching of subsequence S_2 and S_3 with subsequence S_1 should be considered as trivial match and subsequence S_2 and S_3

should be filtered. On the other hand, the set of PIP obtained from subsequence S_4 is different from that of subsequence S_3 . This means that they are non-trivial match. Therefore, S_1 and S_4 are identified as the subsequence candidates.

After all these trivial matched subsequences are filtered, the remaining subsequences should be considered as non-trivial matches and served as the candidates for further discovery process on frequently appearing patterns. They will be the input patterns for the training of the clustering process. k -means clustering technique can be applied to these candidates. The trained algorithm is expected to group a set of patterns M_1, \dots, M_k which represent different structures or time series patterns of the data, where k is the number of the output patterns.

Although the clustering procedure can be applied directly to the subsequence candidates, it will quite time consuming when a large number of data points (high dimension) are considered. By compressing the input patterns with the PIP identification algorithm, the dimensionality reduction can be achieved (Fig.2).

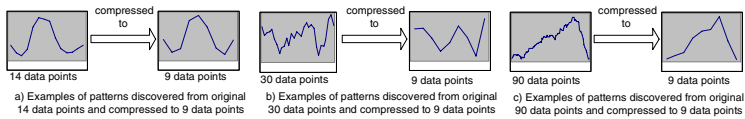



Fig. 2. Examples of dimensionality reduction based on PIP identification process

3 Experimental Result

The ability of the proposed pattern discovery algorithm was evaluated in this section. Synthetic time series is used which is generated by combining 90 patterns with length of 89 data points from three common technical patterns  in stock time series as shown in Fig.8 (i.e.30x3). The total length of the time series is 7912 data points. Three sets of subsequence candidate were prepared for the clustering process. They include the (i) original one, the subsequences formed by sliding window, where $w=89$. This is the set which is claimed to be meaningless in reference [2]; (ii) motifs, K -motifs [2,3] formed from the time series, where K is set to 500 based on the suggested method in paper [3] and (iii) proposed PIP method, the subsequence candidates filtered by detecting the change of PIPs and 9 PIPs are used.

The number of pattern candidates and the time needed for pattern discovery are reported in Fig.4a. Only 281 motifs were formed while half of the subsequences were filtered by detecting the change of PIPs. The proposed PIP method is much faster than the other two approaches because the subsequences are compressed from 89 data points to 9 data points. The dimension for the clustering process is greatly reduced. Fig.4b shows the final patterns discovered. Six groups were formed by each of the approaches and it shows that the set of the pattern candidates deducted from the proposed approach is the most similar set to the pattern templates used to form the time series. On the other hand, the patterns discovered from the original subsequences seem not too related to the patterns which are used to construct the time series. Although the motifs approach can also discover the patterns which used to construct

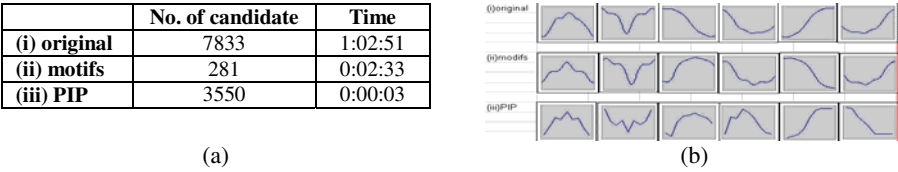


Fig. 3. (a) Number of patterns and time needed for pattern discovery by using different pattern candidates and (b) Pattern discovered (i) original, (ii) motifs and (iii) PIP

the time series, it smoothed out the critical points of those patterns. Also, uptrends, downtrends and a group of miscellaneous patterns are discovered in all the approaches.

To sum up, meaningless patterns are discovered by applying the clustering process on the time series subsequences (i) whereas both motifs and the proposed approach can partially solve this problem by filtering the trivial matched subsequences. However, it is still difficult to determine the starting point of the patterns and leads to the discovery of the shifting patterns. Additionally, the proposed approach can preserve the critical points of the patterns discovered and speed up the discovery process.

4 Conclusion

In this paper, a frequently appearing pattern discovery process for stock time series by changing Perceptually Important Point (PIP) detection is proposed. The proposed method tackles the main problem of discovering meaningless subsequence patterns with the clustering approach. A threshold-free approach is introduced to filter the trivial matched subsequences, which these subsequences will cause the discovery of meaningless patterns. As demonstrated in the experimental results, the proposed method can discover the patterns hidden in the stock time series which can speed up the discovery process by reducing the dimension and capturing the critical points of the frequently appearing patterns at the same time. We are now working on the problem of determining the optimal number of PIPs for representing the time series subsequences and the results will be reported in the coming paper.

References

1. Keogh, E., Lonardi, S., Chiu, Y.C.: Finding Surprising Patterns in a Time Series Database in Linear Time and Space. *Proc. of ACM SIGKDD* (2002) 550-556

2. Keogh, E., Lin, J., Truppel, W.: Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research. *Proc. of ICDM*, (2003) 115-122

3. Lin, J., Keogh, E., Lonardi, S., Patel, P.: Finding Motifs in Time Series. In: *Workshop on Temporal Data Mining*, at the *ACM SIGKDD* (2002) 53-68

4. Patel, P., Keogh, E., Lin, J., Lonardi, S, Mining Motifs in Massive Time Series Databases. *Proc. of the ICDM* (2002) 370-377

5. Chung, F.L., Fu, T.C., Luk, R., Ng, V., Flexible Time Series Pattern Matching Based on Perceptually Important Points. In: *Workshop on Learning from Temporal and Spatial Data* at *IJCAI* (2001) 1-7