An Automated Matrix Profile for Mining Consecutive Repeats in Time Series

Mahtab Mirmomeni[1], Yousef Kowsar[1,2], Lars Kulik[1], and James Bailey[1]

[1] The University of Melbourne
[2] Microsoft Research Centre for Social NUI,
{m.mirmomeni, y.kowsar}@student.unimelb.edu.au, {lkulik, baileyj}@unimelb.edu.au

Abstract. A key application of wearable sensors is remote patient monitoring, which facilitates clinicians to observe patients non-invasively, by examining the time series of sensor readings. For analysis of such time series, a recently proposed technique is Matrix Profile (MP). While being effective for certain time series mining tasks, MP depends on a key input parameter, the length of subsequences for which to search. We demonstrate that MP's dependency on this input parameter impacts its effectiveness for finding patterns of interest. We focus on finding consecutive repeating patterns (CRPs), which represent human activities and exercises whilst tracked using wearable sensors. We demonstrate that MP cannot detect CRPs effectively and extend it by adding a locality preserving index. Our method automates the use of MP, and reduces the need for data labeling by experts. We demonstrate our algorithm's effectiveness in detecting regions of CRPs through a number of real and synthetic datasets.

1 Introduction

Activity and exercise detection using wearable sensors helps clinicians to remotely monitor and better diagnose patients' movements non-invasively [1]. A key task is to find patterns of interest in the time series data generated by wearable sensors. These patterns can be indicative of the patient's status, for example, showing the manner in which a patient is performing a rehabilitation exercise.

Recently, a technique known as Matrix Profile (MP) has been proposed to mine time series data. Despite MP's advantages, exactness, space efficiency and tolerance to missing data, it faces two fundamental challenges: its sensitivity to a key input parameter and its inability to detect CRPs. The authors of the MP assume that the method is effectively parameter-free: "In contrast, our proposed algorithm has zero parameters to set" [2] and the input parameter is based on "user choice" [3], and "our algorithm is insensitive to the value of the only input parameter" [4]. In Section 2, we demonstrate that, MP is highly sensitive to its input parameter, the length of subsequence used for searching and MP is limited in detecting CRPs.

Figure 1 shows an example of a time series interval (known as epoch) recorded by an accelerometer and collected from a patient performing rehabilitation exercises, moving their leg relative to the ground whilst wearing an ankle cuff with an embedded

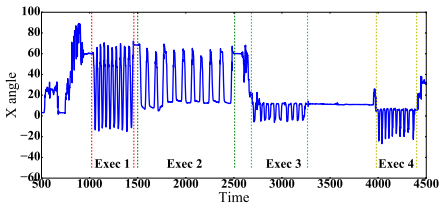


Fig. 1: A time series epoch captured from post-rehabilitation exercises, where 4 different exercises (CRPs) have been performed by a particular patient.

accelerometer. The repeating patterns correspond to exercises being repetitively performed. Regions between the exercises indicate other activities or breaks between the exercises.

For MP to be able to detect the rehabilitation exercises in Figure 1, it is not reasonable to expect the user (e.g. a physiotherapist) to set the input parameter for MP. Rather, it must be done automatically. To this end, we develop a technique to automatically select the subsequence length for MP that can accurately identify the CRPs. To overcome the MP's limitation in detecting CRPs, we extend MP by adding a new index that preserves the locality of the repeats.

We provide a theoretical justification for how our new index can be used to detect the regions of CRPs from a given time series. We show that using our method, we can automatically set MP's input parameter, so that it can accurately identify CRPs in a number of synthetic and real datasets.

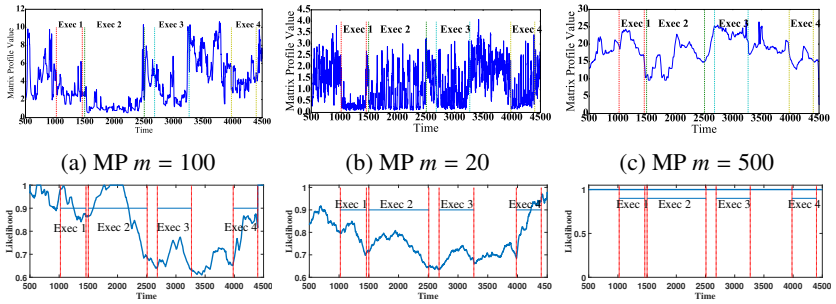
2 Related Work and limitations

Activity and exercise detection background Activity recognition using wearable sensors has gained a lot of popularity given their rise in the consumer space [1, 5, 6]. A common approach to detect activity and exercises is to use supervised or semi-supervised methods, such as statistical, hidden Markov or mixture models [7, 6, 8]. The supervised and semi-supervised methods, however, need domain expertise to label the data, which makes their use in real world applications limited.

Another common approach to detect activity and exercises is through motif discovery techniques on time series data, which are of an unsupervised nature [9, 10, 4]. Motifs are previously unknown subsequences that have been repeated over time [11]. Our problem of finding CRPs in a time series is different to motif discovery algorithms: we aim to find a burst of repeating patterns that happen at a specific point in the time series, e.g., when a patient performs an exercise.

The Matrix Profile (MP) is the newest technique in discovering similar patterns in a time series [2]. We explore the use of the MP for detecting CRPs, corresponding to an exercise in the time series generated by a wearable sensor.

Matrix Profile background Matrix profile (MP) [2] is the most recent technique for all-pair-similarity-search within a time series. A *Time Series* is a sequence of real value



(a) MP $m = 100$ (b) MP $m = 20$ (c) MP $m = 500$

(d) Arc Curve $m = 100$ (e) Arc Curve $m = 20$ (f) Arc Curve $m = 500$

Fig. 2: MP (Top row) and Arc Curve (Bottom row) with various input parameter m for time series in Figure 1. Both MP and Arc Curve values are heavily dependant on the value of the input parameter m, length of subsequence searching.

numbers observed in time $T = \{t_1, t_2, ..., t_n\}$, where n is the length of T. A *subsequence* of a given time series T is a time series starting from t_i with length $m < n+1-i$. In an all-pair-similarity-search, the distances between every subsequence in a time series with all subsequences are calculated. By definition MP is a vector that for each subsequence of a specified length (m) in the time series, stores the smallest Euclidean distance between that subsequence and its nearest neighbour in the time series [2]. The index of the nearest neighbour is stored in another vector called the Index Profile (IP) [2].

Sensitivity to key input parameter: Figure 2a shows the MP that corresponds to the time series in Figure 1, with input parameter $m = 100$. In the exercise regions, marked on the Figure 2a, the MP's values drop to a local minimum (often close to zero). Figures 2b and 2c show MP for the same time series in Figure 1 with $m = 20$ and $m = 500$. Comparing the behaviour of MP in these 3 figures (Figure 2a, 2b and 2c), we see that MP significantly depends on the value of subsequence length m. m can be seen as the granularity level for subsequence searching in a time series. Setting m too large (in our example 500) results in comparing long subsequences from the time series that reduces the chance of finding similar subsequences (Figure 2c). Setting m too small (20 in our example) results in most of the subsequences being assessed similar with each other, which can show itself as sudden fluctuations in the resulting MP (Figure 2b).

The closest application of MP to our problem is the semantic segmentation of a time series [4]. The authors introduced Arc Curves, a transformation of the time series into a new plot that at each point annotates the time series with likelihood of regime change, using IP. We investigated the Arc Curve algorithm [4] for detecting exercise regions of CRPs and applied the code provided by the authors to our rehabilitation exercise dataset. Figure 2 (2d, 2e, 2f) shows the Arc Curves for the time series in Figure 1 using different subsequence lengths, and demonstrates that Arc curves are also highly sensitive to changes in the input parameter. Despite our best efforts, we were not able to produce segments corresponding to regions of CRPs, using the provided code.

MP's Limitations for detecting CRPs: We define *Region of CRPs* to be the region where the same pattern is consecutively repeated. Let *Region of CRP*, RS, to be the region that $\exists f$ (function of repeat), $\Delta t > 0 | \forall x \in RS, f(x) = f(x+\Delta t)$. We call the pattern

that is repeating consecutively inside a RS, the Signal of Repeat. As shown in Figure 2a, MP value drops close to zero when detecting a repeating pattern. However, the repeating patterns do not necessarily need to be consecutive for this to happen. Figure 3 shows a time series with non-CRPs and the corresponding MP. The value of MP is close to zero for a repeating pattern, although the patterns are not CRPs. To solve this problem, we need to determine whether the most similar patterns are also temporally close together. We next outline how to define an index that preserves the locality of the repeats.

3 Problem Statement

To overcome MP's limitation of preserving the locality of repeating patterns, we define Distance Index (DI) as a vector that at each point stores the distance between the index of any subsequence of length m of a time series to the index of its nearest neighbour. We formally define DI as follows:

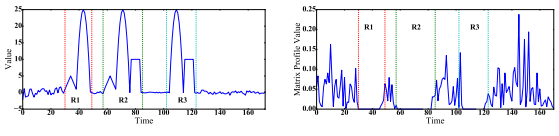
Definition 1. {*Distance Index*} *is a vector of distances, where* $DI[i] = i - Index\ Profile[i]$

Figure 4a shows the corresponding DI with $m = 100$ of the time series in Figure 1 and MP in Figure 2a. In Section 3, we show that the value of DI in the repeat section is equal to the period of the repeating pattern.

The most repeat-sensitive DI for detecting CRPs In Section 2, we showed that MP is sensitive to its key input parameter m. Thus finding the right value for the subsequence length m is crucial for using MP in different applications. To find this value, we need to determine which input parameter m results in a DI that best detects the area of CRPs. We define the most repeat-sensitive DI as a DI that aligns to the period of the repeating pattern and stays flat for the duration of the repeat.

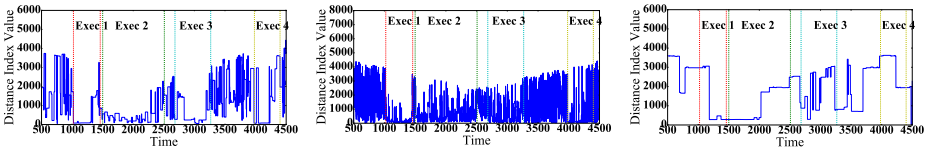
Definition 2. {*Most Repeat-sensitive Distance Index(DI))*} *The most repeat-sensitive Distance Index for finding region of CRPs,* RS*, is a DI, such that* $\forall x \in RS : DI = \Delta t$.

We can show that setting m to the smallest subsequence that is not repeating within the signal of repeat, results in a DI with a flat region equal to the repeat period, in the regions of CRPs.



(a) Time series (b) MP

Fig. 3: Motivation example for extending MP with Distance Index: Time series with repeating patterns that are not consecutive with the corresponding MP, which still drops to zero at repeats.



(a) DI $m = 100$ (b) DI $m = 20$ (c) DI $m = 500$

Fig. 4: Distance Index (DI) with various input parameter m for time series in Figure 1. In the repeating regions, DI stays flat at the value of the repeat period for the duration of the repeat period. Although susceptible, DI is more robust to changes in m.

Lemma 1. *Given a time-series T with region of CRPs, RS, function of repeat $f(t)$ and period of repeat Δt, $RS = \{t_i | f(t_i + \Delta t) = f(t_i)\}$. A MP with $m = \Delta t \rightarrow DI = C$ in the corresponding region of CRPs, where $\forall t_i \in RS | C = \Delta t$.*

Proof. The proof results from the definition of CRPs region, RS, that matches the periodic function definition. For any subsequence inside RS (except the first and last subsequences) the closest repeating pattern is one period away from the subsequence. The DI values for the first and last subsequences depend on the preceding and succeeding subsequences of RS, which are not part of the CRPs.

Lemma 2. *Given a Region of Repeat, RS, with period Δt and signal of repeat S, the DI value for any subsequence within the signal of repeat is equal to Δt iff the subsequence is not repeating within the signal of repeat.*

A direct conclusion from Lemmas 1, 2 is that the input parameter m (subsequence length) for finding the most repeat-sensitive DI is bounded by the period of the region of repeat. We include these findings in the following theorem:

Theorem 1. *Given a time-series T and a region of CRPs, RS, with period Δt, the best value of m for finding the region of repeat is equal to the length of the shortest subsequence that is not repeating within the signal of repeat.*

Automatically determining the best m In an unsupervised scenario, where the ground truth is not known, we need to define a mechanism to find the most-repeat sensitive DI from a pool of DIs calculated using a range of subsequence length ms. In the regions of CRPs the value of MP must be close to zero, and the area under MP can be used to find the most repeat-sensitive DI. Thus, for each DI, we find the flat segments and nominate them as regions of CRPs. We calculate the area under the corresponding MPs for those regions and select the DI corresponding to the MP with the minimum area as the most repeat-sensitive DI. Theorem 1 shows that the upper bound value for m to produce a repeat sensitive DI is equal to the period for that region. Thus a brute force search to find best m is of $O(n^2 \times$ estimated period) in time. We use the inverse of the most dominant frequency from time series' Fourier transformation to estimate the period.

4 Experimental Evaluation

The purpose of our experiments[3] is to evaluate how accurately we can identify the CRP regions of a time series (the regions of exercises in our physiotherapy dataset) using MP, when automatically setting the value for input parameter m, length of subsequence searching. To evaluate our algorithm, we use ground truth provided by experts on the location of exercises in an epoch.

We calculate DI using a range of subsequence length ms and find flat segments of each DI. We set the value of all points on the flat segment of DI to 1 and the rest to 0. For the ground truth, we set the value of all points in the repeating region RS to 1 and the rest to 0. We define *true positive* as the number of points with value 1 on DI that align with the region corresponding RS and *false positive* as the number of points on DI that have value 1, but are outside of RS. For each DI, we report the F1-Score and the Adjusted Mutual Information (AMI) between DI with the ground truth using the True/False positives.

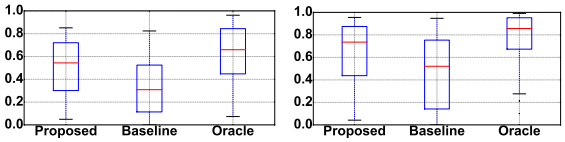
We created two synthetic datasets. The first synthetic dataset is a simple sine waveform with a period of 180. The sine waveforms are preceded and succeeded by Gaussian noise. The second synthetic dataset contains repeats within repeats. The repeating sequence is a waveform with a period of 30, which contains two repetitive waveforms with a period of 10.

We used a real Physiotherapy dataset, collected by the Physiotherapy department at the University of Melbourne, of patients with chronic knee pain performing 4 rehabilitation exercises, while wearing an ankle-cuff with an embedded accelerometer [12]. We used data from 10 patients. The accelerometer's angle with respect to the x axis is of interest, in this context. To detect regions of CRPs in an epoch, we use a tumbling window of size 1000 to segment the epoch and search within that tumbling window. We perform our search for m in the range of $m \in [2, period]$, according to the upper limit set for m described in Section 3. The brute-force algorithm searches for a best m in the range of $m \in [2, \text{window length}/3 \approx 300]$. The cut off value for the brute-force search is set so that we have at least three CRPs in our window.

Results and Discussion We evaluate how accurately we can find the region of CRPs using the proposed input parameter m. Our synthetic data has a fixed period of repeats. According to Theorem 1 the best value of m for finding region of repeats is equal to the length of the shortest subsequence that is not repeating within the signal of repeat. For the first Synthetic dataset, subsequence length m was found to be equal to 4 using both our proposed algorithm and the algorithm that searches for the best m, which results in AMI and F1-Score of 0.99. For the second synthetic dataset, subsequence length m was found to be 11, where 10 is the length of the subsequence repeating in the repeat signal, using both algorithms. The subsequence of length 11 results in AMI and F1-Score of 0.98. Both results agree with Theorem 1.

For our real dataset, since each repeat of an exercise varies slightly from its surrounding repeats, the period for the region of CRPs is unknown and can only be estimated. As a result, instead of observing one flat line in DI, corresponding to a region of

[3] Our code is available on `http://goo.gl/TLfCLp`



(a) AMI (b) F1-Score

Fig. 5: Boxplot of comparison for finding region of CRPs using our proposed input parameter vs domain-knowledge (Baseline) and Oracle m for the Physiotherapy dataset.

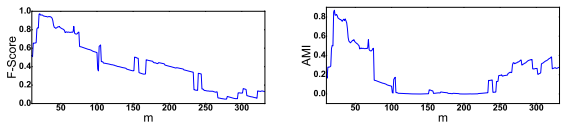


Fig. 6: F1-Score (left) and AMI (right) changes over changes in input parameter m.

repeat, we observe discontinuity in the flat line corresponding to the region of repeat. We estimate the period by taking the mean from the discontinued flat line, which in turn causes m found by our proposed algorithm to be different to the best m, found using a supervised approach. It is not possible to stipulate a single, universal value for m.

To set a baseline, we used a randomly generated m based on domain knowledge. According to experts, each exercise set in our Physiotherapy dataset on average takes between 30 sec-1 min. There are 10 repeats on each exercises set, therefore each repeat takes 3-6 sec (on average). Data sampling rate is 12 data points per second, therefore each region's period is expected to be between 36 ($3 * 12$) to 72 ($6 * 12$) points. We set m to a random number between 36 to 72 for each epoch as the baseline.

Figure 5a and 5b depict AMI (median=55%) and F1-Score (median=74%) of our proposed algorithm for finding CRPs vs the AMI (median=61%) and F1-Score (median=38%) for the baseline algorithm. In both box-plots we observe that our proposed method for selecting m results in a more accurate detection of region of CRPs. In both cases our proposed method significantly outperforms the baseline (AMI paired t-test p-value$<$ 0.0002 and F1-score paired t-test p-value$<$ 0.0001). In both diagrams, the baseline method has a wider range of values. This results from the sensitivity of MP to the input parameter m and that setting m based on domain knowledge cannot overcome this sensitivity. We have included the best m results as an oracle (AMI median=66%, F1-Score median=86%) for comparison. Finding the best m results requires labeled data. Since we are using an unsupervised approach and we want our method not to be dependant on labelled data, the best m is unknown.

We investigate the effect of changes in m on F1-Score and AMI for finding CRPs in Figure 6. In these plots, we are evaluating Theorem 1 for a randomly selected window from our Physiotherapy dataset. From both plots, it is evident that setting m too small results in a poor detection of regions of CRPs. However, the accuracy of the proposed method surges in both evaluation metrics as input parameter m increases and drops as input parameter m gets too large, which fully agrees with Theorem 1.

5 Conclusion

We explore the use of the MP to detect CRPs, that translate to exercises in our Physiotherapy dataset. We show that MP has two fundamental limitations: it is sensitive to a key input parameter, the subsequences length for which to search, and does not detect if repeating patterns are consecutive. We introduce a new index, DI, to preserve the locality of repeats. The most repeat-sensitive DI can accurately identify the region of CRPs. We prove that to achieve the most repeat-sensitive DI, the input parameter m has to be set to the length of the shortest subsequence that is not repeating within the signal of repeat. We compare the accuracy of our unsupervised algorithm in finding the CRP regions to the results from applying the best m, calculated using a supervised method. The comparison shows that we can, with high accuracy (Average difference AMI 12%), automatically and without a priori knowledge about the regions, find the regions of CRPs (exercises in our Physiotherapy dataset). Our proposed method finds the input parameter m that outperforms selecting m using domain knowledge by 15%.

References

1. J. Andreu-Perez, D. R. Leff, H. M. D. Ip, and G. Z. Yang, "From wearable sensors to smart implants toward pervasive and personalized healthcare," *IEEE Trans. on Biomedical Engineering*, vol. 62, no. 12, pp. 2750–2762, 2015.
2. C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh, "Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in *Proc. of ICDM 2016*.
3. Y. Zhu, Z. Zimmerman, N. S. Senobari, C. C. M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh, "Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins," in *Proc. of ICDM 2016*.
4. S. Gharghabi, Y. Ding, C. C. M. Yeh, K. Kamgar, L. Ulanova, and E. Keogh, "Matrix profile viii: Domain agnostic online semantic segmentation at superhuman performance levels," in *Proc. of ICDM 2017*.
5. S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 9, no. 1, p. 21, Apr 2012.
6. J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.
7. Y. Kowsar, M. Moshtaghi, E. Velloso, L. Kulik, and C. Leckie, "Detecting unseen anomalies in weight training exercises," in *Proc. of OzCHI 2016*.
8. D. Minnen, C. L. Isbell, I. Essa, and T. Starner, "Discovering multivariate motifs using subsequence density estimation and greedy mixture learning," in *Proc. of the 22Nd National Conference on Artificial Intelligence - Volume 1*. AAAI Press, 2007, pp. 615–620.
9. D. Minnen, T. Starner, I. Essa, and C. Isbell, "Improving activity discovery with automatic neighborhood estimation," in *Proc. IJCAI 2017*, pp. 2814–2819.
10. A. Vahdatpour, N. Amini, and M. Sarrafzadeh, "Toward unsupervised activity discovery using multi-dimensional motif detection in time series," in *Proc. IJCAI 2009*, pp. 1261–1266.
11. B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," in *Proc. of SIGKDD 2003*, pp. 493–498.
12. K. Bennell. Adherence to home exercises in the treatment of knee osteoarthritis. [Online]. Available: https://healthsciences.unimelb.edu.au/research-groups/physiotherapy-research/chesm/more