Database Security via ML Algorithms

by

Ali Arda Akkız

A report submitted for EE492 senior design project class
in partial fulfillment of the requirements for the degree of
Bachelor of Science
(Department of Electrical and Electronics Engineering)
in Boğaziçi University

January 13th, 2023

Principal Investigator:
Prof. Emin Anarım

ACKNOWLEDGMENTS

Principal investigator Prof. Emin Anarım and research assistant Çağatay Ateş have been contributed to this project.

ABSTRACT

Matrix Profile is a powerful Machine Learning method that can be used to reveal patterns in massive datasets. In this paper, the dataset consisting of 15-days of monitoring data regarding application servers has been analyzed with the purpose of extracting useful information from them, more specifically, detecting the anomalous events present in them. Both constrained and unconstrained search models have been tested and analyzed. In the conclusion part, several real-life examples together with other potential application areas have been discussed as well as the constrains and limitations of the method Matrix Profile.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1
INTRODUCTION

Data has become one of the most valuable assets for companies, hence, unauthorized operations on data should be prevented to preserve the integrity, confidentiality, and availability of the data. Failure to do so may result in financial loss and reputational harm [1]. One may recall the incident took place in March 2021 when a Chinese hacker group exploited a security flaw in a Windows server causing extreme damage to many individuals and organizations. Even though Windows Security Department employs highly qualified staff and implements various kinds of security schemes, the security of the data can never be achieved completely. However, raising the CTB (Cost to Break) to a reasonable level will discourage ill-intended users to attempt to break the system whereas monitoring the data is crucial for the detection of any malicious events. Traditional security mechanisms such as masking the sensitive information are not sufficient to protect the data, hence, more sophisticated security mechanisms should be employed. Machine Learning algorithms can be implemented to enhance the traditional security mechanisms. Classification of sensitive data to ensure that proper access controls are in place or risk assessment regarding security incidents are examples of many use cases. In this project, we will be focusing on anomaly detection on database servers to identify potential suspicious activity.

CHAPTER 2
REVIEW OF LITERATURE

According to the IBM's guide on database security [1], database security depends on the security of its components which are listed below:

Data

Database Management System

Applications associated with the database

The physical and virtual database server and the underlying hardware

The computing and/or network infrastructure used to access the database

In this paper we focused on the analysis of the time series data gathered from the servers of a large internet company for the purpose of detecting anomalies. The method that will be used is Matrix Profile, a powerful time-series data mining tool developed by Keogh research group at UC-Riverside [3]. The Matrix Profile has been chosen for its simplicity and scalability since the data is growing exponentially imposing a need for robust and efficient algorithms.

The library "Stumpy" will be used in the calculation of the Matrix Profile which is developed by an independent contributor named Sean Law [4]. Stumpy has been chosen for its ease of implementation and well-formed documentation.

CHAPTER 3
METHODOLOGY

Matrix Profile is a time-series data mining tool mainly used in the detection of motifs and discords based on similarity join. The **Matrix Profile** consists of a **Distance Profile** and a **Profile Index**. Distance Profile is a vector of distances between all pairs of subsequences of the data whereas the Profile Index stores the distance to the nearest neighbour. The standard Euclidian distance has been used for calculations, however, MP supports other kinds of distance calculations as well. The distance profile is calculated for each subsequence in a sliding window manner see appendix A for visualizing the calculation. The lowest value in the distance profile is the trivial match of the subsequence with itself (zero distance) which is excluded when constructing the matrix profile. One may examine the figure below to visualize the Matrix Profile and Distance Profile.

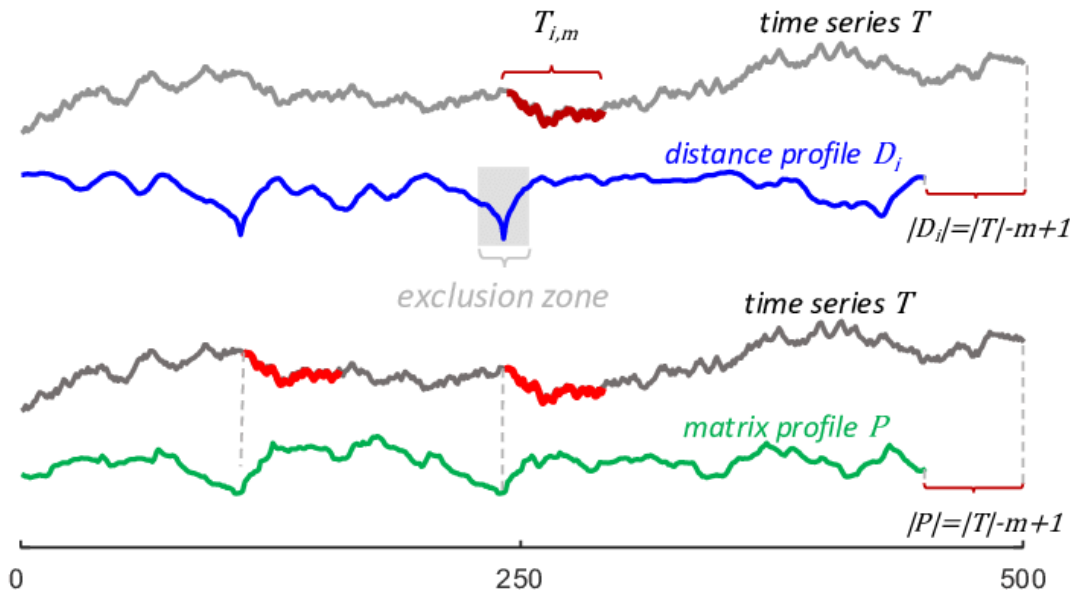


Fig. 1. Time series data with its Distance Profile and Matrix Profile

Note that the low values in the Matrix Profile corresponds to the motifs whereas high values corresponds to the discords present in the time series.

Besides its simplicity and scalability there are many other advantages of using MP to analyze time series as introduced in the paper [5]:

1. PROPERTIES OF THE MATRIX PROFILE

Exactness: no false positives and no false dismissals

Parameter-free: The only parameter to be tuned is the length of the subsequence

Space-efficiency: Space overhead of MP is linear in the time series length paves the way for analyzing massive datasets

Anytime-property: when the calculation has been interrupted before it ends, the method delivers an approximate solution which rapidly converges to the exact solution as the iteration continues.

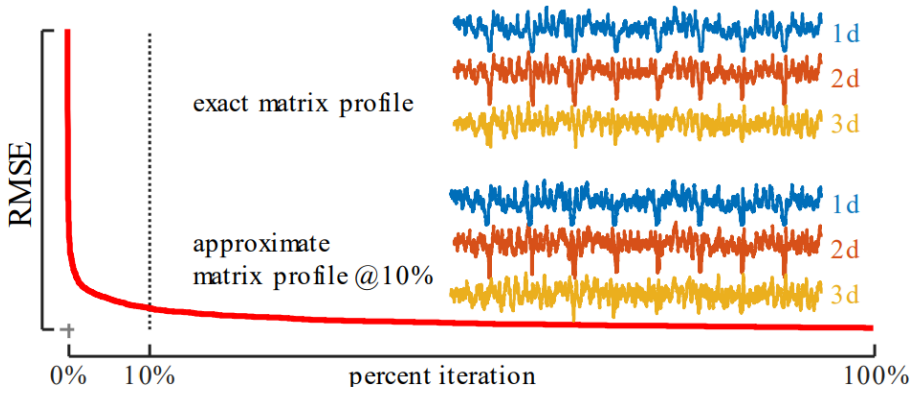


Fig. 2. Convergence of multi-dimensional MP after %10 percent of the total iterations

Incrementally maintainable: It is possible for MP to be used in online processing since when the MP has been constructed once, it can be updated with the incoming data.

No threshold value: There is no need to specify a threshold value since MP utilizes full joins supporting our statement regarding the simplicity of the method.

Deterministic time: The time it takes to construct the MP does not change with different datasets of equal length but rather relies on some consant multiplicative of the time series data length.

Hardware leverage: Calculation of the MP can be leveraged by parallel processing both on multicore processors and in distributed systems

Handling of missing data: Even in the presence of missing data in the time series, the MP can be constructed without false negatives.

Time-series semantic segmentation: MP can partition the time series into different modes of operation.

Time complexity constant in subsequence length: Most of the algorithms used in detection of motifs and discords scale poorly when the subsequence length increases. This is not true for MP since it is possible to work with subsequence length upto 10^6 data points.

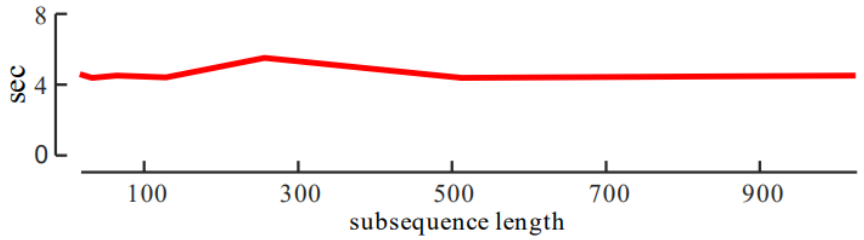


Fig. 3. Runtime vs subsequence length

2. MULTI-DIMENSIONAL MATRIX PROFILE

One unmentioned advantage of MP is that it is possible to analyze as many dimensions as needed since the time it takes to compute MP is linear with the size of the dimensions of the dataset as the figure below suggest:

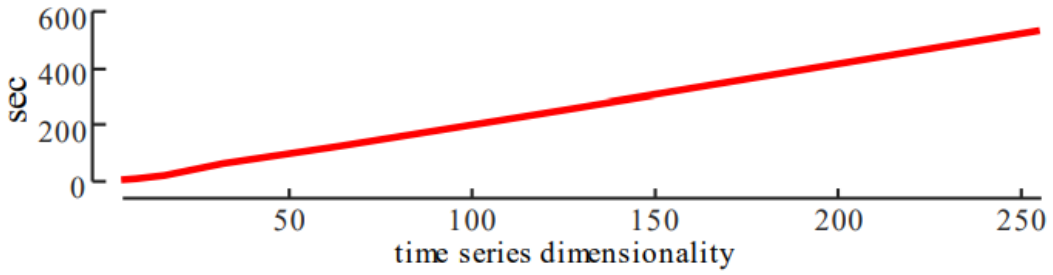


Fig. 4. Runtime vs dimensionality

Note that, the overall MP is not constructed with the addition of the individual 1-D MPs on top of each other but rather in a more sophisticated manner [6]. The algorithm for the multi-dimensional MP construction involves the filtering of the relevant dimensions that produces semantically meaningful motifs and/or discords. The filtering process is crucial since the motifs present in all dimensions rarely produce meaningful motifs. Therefore, the subset of dimensions which actually contribute to the overall motifs/discords present in the time series should be determined. This could

be done in two ways: Converting this problem into an elbow/knee finding problem or computing the MDL (Minimum Description Length) which is presented in the original paper written by the Keogh Team [7].

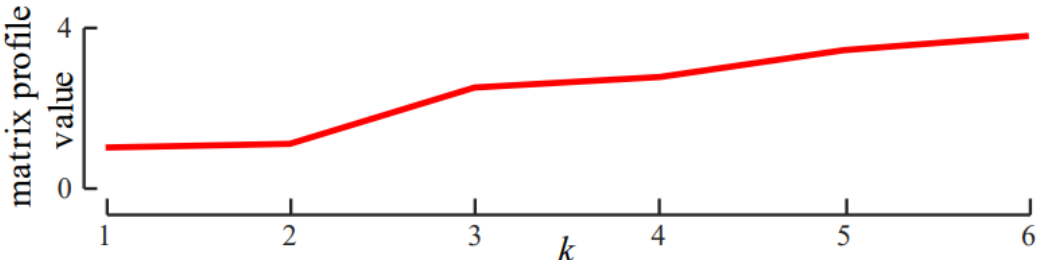


Fig. 5. An example of an MDL plot. k value (zero-based) refers to the # of dimensions included in the calculation of the MP.

Notice the dramatic increase in the MP value when the 3rd dimension is included. Selecting k to be equal to 2 corresponds to the minimum bitsize (maximum compression) of the data which indicates that this is the best model to be used according to the MDL approach.

MDL approach returns how many dimensions should be used as well as which dimensions should be chosen in order to produce meaningful motifs. This subset of dimensions can be used in the construction of the multidimensional MP which then can be used to detect meaningful motifs and discords present in the time series.

CHAPTER 4
EXPERIMENTS AND ANALYSIS

The time series data to be analyzed consists of 12 different entities (servers) with 19 different metrics including CPU-related metrics, memory-related metrics, network metrics, etc. The data points are equally-spaced 5 minutes apart with a length of 4320 for each metric accounting for 15 days of monitoring data in total. The regions of anomalous events have been tagged by domain experts. One may examine the figure below to visualize the multidimensional time series data. Note that, the anomalous events have been plotted on the dimensions that are contributing to the event.

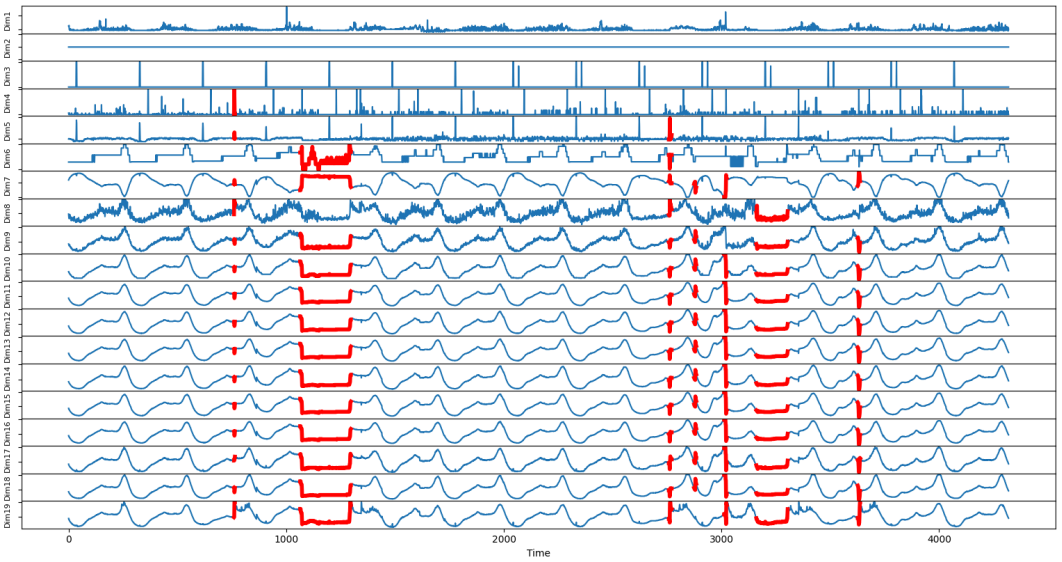


Fig. 6. Multi-dimensional Time Series Data. Anomalous events are indicated as red on the relevant dimensions that contribute to the event.

MP supports both constrained and unconstrained search. In the case that the user does not know how many and which dimensions should be involved, unconstrained search should be implemented. If the user knows which dimensions are relevant, MP can be constructed by explicitly including or excluding specific dimensions.

PART A: UNCONSTRAINED SEARCH

To obtain the best model for MP, one should compute the MDL and look for the minimas to determine the subset of dimensions which will produce meaningful discords. The result of MDL approach can be found in the figure below for a subsequence length of 72 which corresponds to a window of 6-hours.

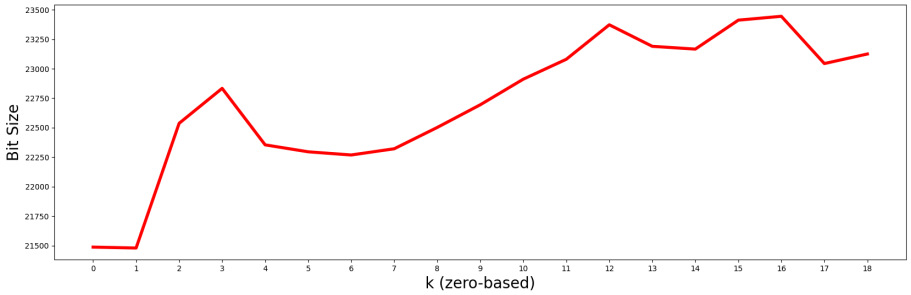


Fig. 7. The result for the MDL approach.

The global minima occurs at k = 1 (zero based) suggesting that 2 dimensions should be included for the calculation of multi-dimensional MP. However, since the number of dimensions are relatively high, a local minima at a larger k will also be considered as a candidate to analyze the performance. The multi-dimensional MPs for each k can be found from the figure below plotted together with the original time series data.

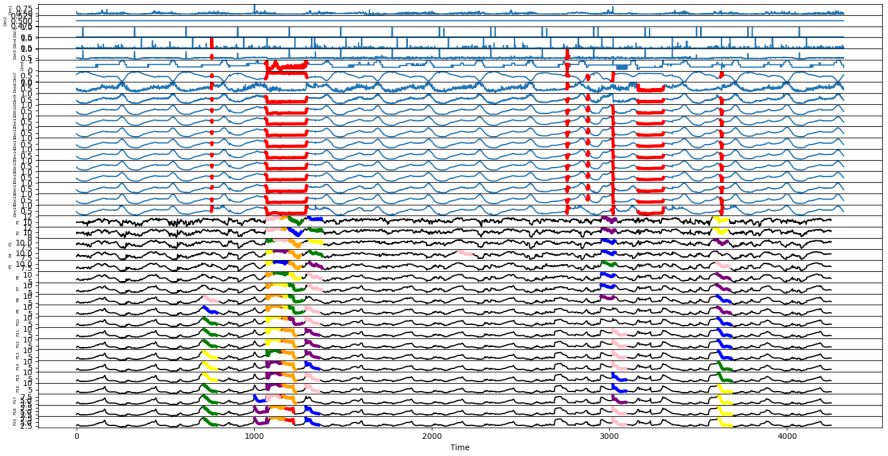


Fig. 8. The multi-dimensional MPs for each k. The colors from red to pink (ordered by their wavelength) indicates the top 7 discords. Subsequence length is equal to 6-hours.

Top 7 discords have been plotted on the MPs for each k as the reader may review Appendix B to find out how this is done.

The accuracy of the MP has been calculated as follows: The high values of MP should correspond to the subsequences that contain anomalies. Therefore, if the subsequence indicated in the MP includes an anomaly then this is accurate for the index having the high value, if not, this is identified as false positive. We iterate this for the 434 highest values for MP since there are exactly 434 data points belonging to an anomaly. The accuracy for different subsequence lengths (m) and different subsets (k) can be found from the table below:

TABLE I. ACCURACY RESULTS FOR THE UNCONSTRAINED CASE

k \ m (hours)	3	6	12	24
1	0.4078341	0.48847926	0.34331797	0.73502304
2	0.38709677	0.49078341	0.41705069	0.63364055
3	0.37788018	0.49078341	0.47465438	0.99769585
4	0.38248848	0.51382488	0.53917051	1
5	0.40552995	0.53917051	0.64976959	1
6	0.43317972	0.56451613	0.73963134	1
7	0.44700461	0.63133641	0.85483871	1
8	0.47465438	0.70276498	0.89631336	1
9	0.48847926	0.76958525	0.91705069	1
10	0.48387097	0.80875576	0.91935484	1
11	0.47235023	0.81336406	0.92165899	1
12	0.46774194	0.82488479	0.92396313	1
13	0.46082949	0.83410138	0.92626728	1
14	0.45391705	0.84101382	0.93087558	1
15	0.44009217	0.84331797	0.93087558	1
16	0.41474654	0.85483871	0.93548387	1
17	0.40552995	0.85483871	0.93778802	1
18	0.40552995	0.85253456	0.94009217	1
19	0.40552995	0.85253456	0.94009217	1

The accuracy results colored red indicates the suggested subset of dimensions regarding the global minima in the MDL plot. The blue ones indicate the candidate subset of dimensions regarding the local minima in the MDL plot.

One can observe the high accuracy as the subsequence length increases which is not surprising since the probability of containing an anomaly within a larger window is higher.

Another observation to be made is that increasing the dimensions to be included in the construction of MP increases the accuracy of detecting anomalies in general. However, this is not true for the case m = 3 (hours) since the accuracy result starts to drop after including more dimensions. This is an important result since it emphasizes the importance of selecting the most natural subset of dimensions. Upon close examination of the original time series data, one may observe that dimensions through 10-18 are contributing to all of the discords. k = 9 corresponding to the best accuracy result for m = 3 confirms our statement regarding the importance of selecting relevant dimensions.

PART B: CONSTRAINED SEARCH

Dimensions 1-3 do not contribute to any of the anomalous events, hence, they are excluded from the search whereas dimensions 10-18 contribute to all anomalous events, hence, they are explicitly included in the search and the construction of the MP. The MDL plot for m = 6 (hours) can be found from the figure below:

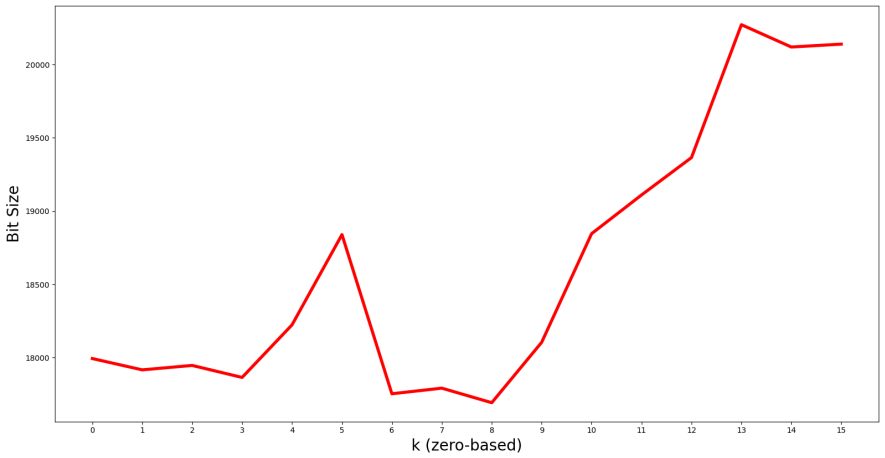


Fig. 7. The result for the MDL approach.

Notice that the global minimum has changed from k = 1 to k = 8 when the constraints have been added. The resulting MPs can be found from the figure below.

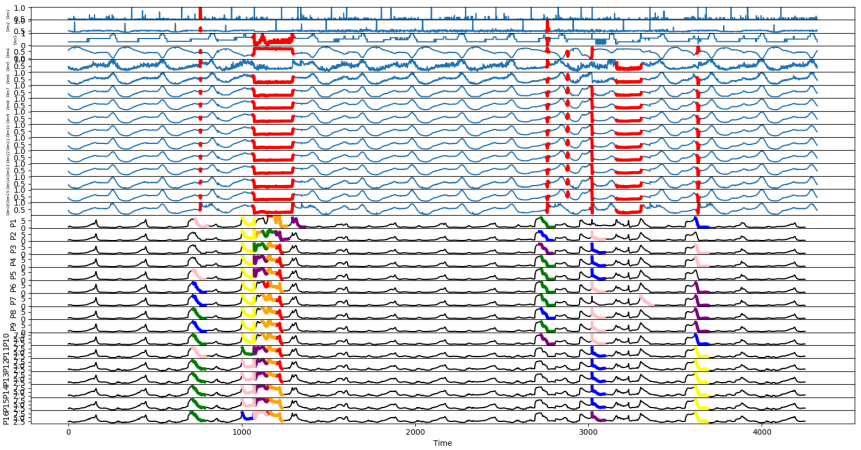


Fig. 8. The multi-dimensional MPs for each k. The colors from red to pink (ordered by their wavelength) indicates the top 7 discords. Subsequence length is equal to 6-hours.

Upon close examination of the MPs, the false positives present in the unconstrained case have been disappeared from the plot regarding the constrained case. The test results for the constrained case with the same subsequence length as before can be found from the table below:

TABLE II. ACCURACY RESULTS FOR THE CONSTRAINED CASE

k \ m (hours)	3	6	12	24
1	0.20737327	0.87096774	0.91474654	1
2	0.22580645	0.87788018	0.94700461	1
3	0.28801843	0.88018433	0.96082949	1
4	0.3156682	0.87327189	0.96082949	1
5	0.33410138	0.88479263	0.96774194	1
6	0.33410138	0.87557604	0.96774194	1
7	0.35023041	0.87327189	0.96774194	1
8	0.36175115	0.88479263	0.95852535	1
9	0.39170507	0.88479263	0.96082949	1
10	0.39861751	0.88248848	0.95852535	1
11	0.4124424	0.87788018	0.9516129	1
12	0.41705069	0.87096774	0.9516129	1
13	0.41474654	0.8640553	0.94700461	1
14	0.42857143	0.86175115	0.94009217	1
15	0.43087558	0.85483871	0.93778802	1
16	0.44239631	0.859447	0.93778802	1

By comparing the results for the constrained and unconstrained case, it can be observed that by specifying the subset of dimensions, one can achieve significantly higher accuracy since the irrelevant dimensions are no longer affecting the MP to detect false multi-dimensional anomalies. However, this is not true for the case m = 3 (hours). This is because the subsequence length of 3 hours is relatively small and susceptible to the high frequency changes in the data causing false alarms. Choosing the subsequence length is crucial in the construction of the MP since it is the only parameter to be tuned.

CHAPTER 5
CONCLUSION

It is important the remention the core superiorities of the Matrix Profile compared to other machine learning algorithms before concluding this paper. MP is simple to use, robust, domain agnostic and extremely efficient. The property of being domain agnostic, paves the way for applications that extends to all industries. Robustness, extreme efficiency and leveraging hardware for parallel processing makes it possible to analyze massive datasets. Since MP is an open source software and it is easy to implement the costs of integration is minimum. These core qualities also determines its limitations as well. MP is great for processing massive datasets, however, it is limited to provide basic insights regarding the dataset. The best use case for MP is to obtain a general sense about the data which then can be used to gain further information with the aid of other feature extracting machine learning algorithms.

A. APPLICATION AREAS

In the paper presented by Keogh Team [7] several real-life cases has been studied with domain experts which are explained briefly as follows:

Case study #1: Motion Capture

The team managed to find a latent motif present in the subset of 3 dimensions among the 38 dimensions in total, allowing to process the dataset to obtain higher quality meaningful motifs compared to including all the dimensions which is also done by using the MP.

Case study #2: Music Processing

The researchers managed to find the chorus of a song with the inclusion of all 32 dimensions as well as the percussion motifs when only searched in subspaces including one or two dimensions.

Case study #3: Electrical Load Measurement

When applied to the electrical load measurement dataset from UK households, the motif present in the 2-dimensional subspace of total 5 dimensions revealed the pattern related to the usage of dryers after washers.

Case study #4: Physical Activity Monitoring

The dataset containing physical activities which are lying, sitting, standing, walking, running, cycling, Nordic walking, climbing stairs, vacuum cleaning, ironing, rope jumping, had been analyzed and the retrieved motifs corresponded to the dynamic activities rather than passive activites which are the first three activites just listed. This is not surprising since MP does not identify the stationary data as a motif but rather favor the dynamic motifs.

Apart from these case studies, MP can be applied to many more industries after processing the obtained data into a time-series one. To name a few, one may consider the analysis of the genetic information contained in the DNA to find common genes across species or the analysis of seismic activity to detect earthquakes.

B. Realistic Constraints

The constraints for MP is minimum thanks to its ease of implementation. All the codes are open sourced ensuring accessibility. It is easy to construct since there is no threshold value to specify or parameters to be tuned except the subsequence length and the dimensions to be searched in the multi-dimensional case. MP has extensive application area but with limited functionality.

C. Social, Environmental And Economic Impact

Patterns contain information which is valuable. Discovering them is crucial whether it is related to economics or diseases or anything else. People understood the importance of data long time ago even before the invention of writing, however, with the digital revolution the size of the data expanded exponentially to the level where our brains can not keep up anymore, imposing a need for fast processing of massive datasets to obtain insights about them to solve problems of any kind. The Matrix profile provides us these insights regarding the latent structure in any kind of data. One may leverage these insights to design better systems for the sake of other folks.

D. Cost Analysis

The only costs are associated with the salary of the data scientists and the costs of measuring data regarding the costs of sensors, encoders, etc. No significant costs associated with manufacturing or maintainability since the space overhead is small.

E. Standards

“IEEE P2830™ - Standard for Technical Framework and Requirements of Shared Machine Learning” should be reviewed when processing confidential encrypted data from multiple sources.

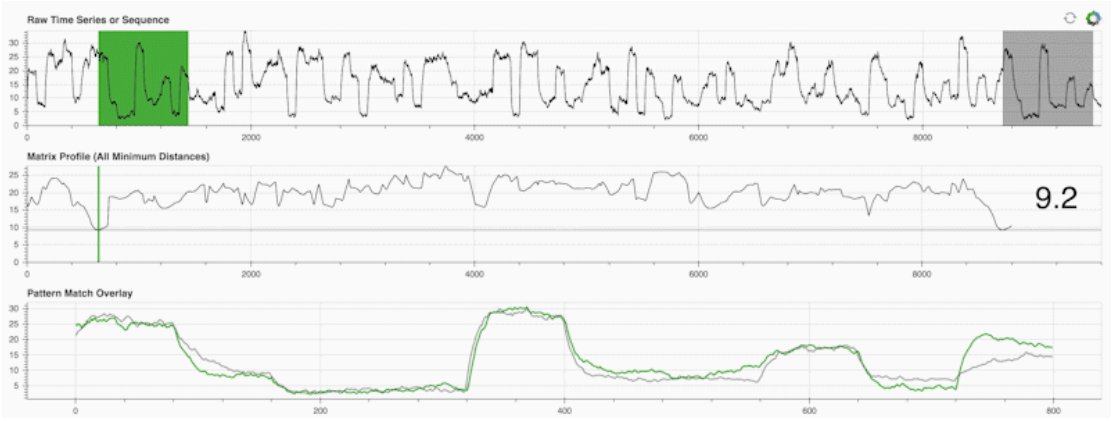
“P2986™ - Recommended Practice for Privacy and Security for Federated Machine Learning” & “IEEE P3652.1™ - Guide for Architectural Framework and Application of Federated Machine Learning” should be reviewed when designing a federated ML system.

“P3123™ - Standard for Artificial Intelligence and Machine Learning (AI/ML) Terminology and Data Formats” should be reviewed to have a clear understanding for relevant terms and data formats.

APPENDICES

APPENDIX A: Distance Profile

The calculation of Distance Profile can be found from the gif below.



stumpy_demo.gif

APPENDIX B: Top-K Discords

Top discord starting at index $t_{0,}$ which has the maximum value in the MP, has been plotted for each k, then the values in the zone with the range (t_0 – m/2, t_0 +m/2) are set to 0 since indexes that are close to the top discord will have high value as well which should be excluded from the next iteration in order to find the next discord. The iteration con

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] IBM Cloud Education, 27 Aug 2019, "*What is database security?",* accessed 13.11.2022, <https://www.ibm.com/cloud/learn/database-security>

[2] F. Alotaibi and A. Lisitsa, "Matrix profile for DDoS attacks detection," 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), 2021, pp. 357-361, doi: 10.15439/2021F114.

[3] UCR, *UCR Website*, accessed 13.11.2022, https://www.cs.ucr.edu/~eamonn/MatrixProfile.html

[4] Law, Sean M., 2019, "STUMPY: A Powerful and Scalable Python Library for Time Series Data Mining" accessed 13.11.2022 < https://github.com/TDAmeritrade/stumpy>

[5] C. -C. M. Yeh et al., "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 2016, pp. 1317-1322, doi: 10.1109/ICDM.2016.0179.

[6] Sean Law, 2019, "Multidimensional Motif Discovery", accessed 12.01.2023, https://stumpy.readthedocs.io/en/latest/Tutorial_Multidimensional_Motif_Discovery.html

[7] C. -C. M. Yeh, N. Kavantzas and E. Keogh, "Matrix Profile VI: Meaningful Multidimensional Motif Discovery," 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 2017, pp. 565-574, doi: 10.1109/ICDM.2017.66