The 3rd International Conference on Computer Engineering and Mathematical Sciences (ICCEMS 2014)

Online Pattern recognition in subsequence time series clustering

Seyedjamal Zolhavarieh[1], **Saeed Aghabozorgi**[2] and **Teh Ying Wah**[2]

1School of Computer and Mathematical Sciences, Faculty of Design and Creative Technologies, Auckland University of Technology (AUT), 1142 Auckland, New Zealand
[2]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia
szolhava@aut.ac.nz, saeed@um.edu.my, tehyw@um.edu.com

Abstract. One of the open issues in the context of subsequence time series clustering is online pattern recognition. There are different fields in this clustering such as e-commerce, outlier detection, speech recognition, biological systems, DNA recognition, and text mining. Among these fields pattern recognition is one the essential concept. To implement the idea of online pattern recognition, we choose sequences of ECG data as a subsequence time series data. Additionally, using ECG data can help to interpret heart activity for finding heart diseases. This paper will offer a way to generate online pattern recognition in subsequence time series clustering in order to have a runtime results.

Keywords: Subsequence time series clustering, pattern recognition, time series, subsequence time series

1. Introduction

Recently, high speed growth of computer and internet technology leads to appear huge amount of time series data in different fields such as shopping transaction, climate change, web click stream, outlier detection [1, 2], speech recognition [3], biological sequences (e.g. ECG), DNA optimization [4], text mining and so on. Among these different fields pattern recognition is essential. This paper will examine ECG data on the dynamic clustering algorithm to reach online pattern recognition.

Despite of the existence of a lot of general time series clustering algorithms and methods, subsequence time series clustering leads to many interesting new kinds of knowledge including sequential patterns, motifs, periodic patterns, partially ordered patterns, approximate biological sequence patterns, and so on; Around 10 years of active research on subsequence time series mining by data mining, machine learning, statistical data analysis, and bioinformatics researchers, it is time to present a systematic introduction and comprehensive overview of the state-of-the-art technology in this area. This paper integrates the methodologies of subsequence time series clustering developed in multiple disciplines, including data mining, machine learning, statistics, bioinformatics, genomics, and financial data analysis, into one comprehensive and easily accessible introduction.

Regarding to interpret heart activity over a period of time, we use electrodes on the human skins and record the electrical activities by an external machine. In this case, these activities save as Electrocardiography (ECG) data. This data shows rate and regularity of heartbeats and can be used for measuring damages and illnesses of people's hearts [5]. Most of the ECG data is utilized for research purposes on human hearts, usually for distinguishing heart abnormal heart rhythms. Figure 1 shows a sample of ECG data as a time series.

In this research, an ECG signal is selected as Time Series data and some clustering algorithms will be utilized for categorization of extracted subsequences. Many researches on time series clustering regards to the clustering of individual time series such as individual heartbeats [6]. There are few attempts on the concept of clustering time series streams. Inside of this ECG data, there are some subsequences which repeated time by time. The structure of ECG data will be discovered by extracting from time series and categorizing based on their similarity. Finding these categories can help physicians to categorize repeated and unrepeated patterns in an ECG data (e.g. recognizing normal and abnormal patterns) that can lead to distinguishing different illnesses.

Our idea is about recording fixed amount of subsequence time series data in a buffer and analyze it time by time. It this way the new data will be assessed with the previous one and then they will merge together in order to reach the best results. In the other hand, they can upgrade each other if they cannot merge. While data produce forever, we can keep clusters in a file. Here, the daily periodicity is obvious, and a more careful inspection reveals very similar patterns.

The rest of this paper is organized as follows. In part 2, we clarified some main definitions and background in pattern recognition and subsequence time series clustering. Section 3 includes the related works according to this paper. Decision and framework will be provided in section 4 and 5 respectively and , we conclude all the proposed concepts in this paper as a glance at the end.

2. Backgrounds and definitions

In this section, we provide definitions and background knowledge used in this work.

2.1 Definitions

Definition 1. A time series T of size m is an ordered sequence of real value data, where T= (t_1, t2 … t_m) [7]. A time series is a sequence of data points, measured typically at

successive points in time spaced at uniform time intervals. In the context of signal processing, control engineering and communication engineering it is used for signal detection and estimation, while in the context of data mining, pattern recognition and machine learning time series analysis can be used for clustering, classification, query by content, anomaly detection as well as forecasting."

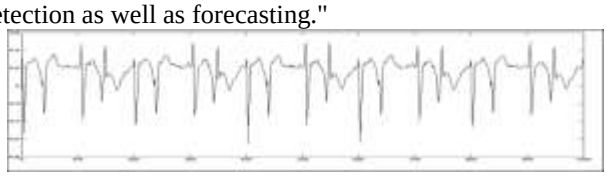

Figure 1. A sample of ECG data

Definition 2. A subsequence of length n of time series T is $T_{i,n}=(t_i, t_{i+1},...,t_{i+n-1})$, where $1 \leq i \leq m-n+1$ [7]. A subsequence is an arranged sequence data which omits some elements without changing the order of the remaining elements. [8D,E,F] [9].

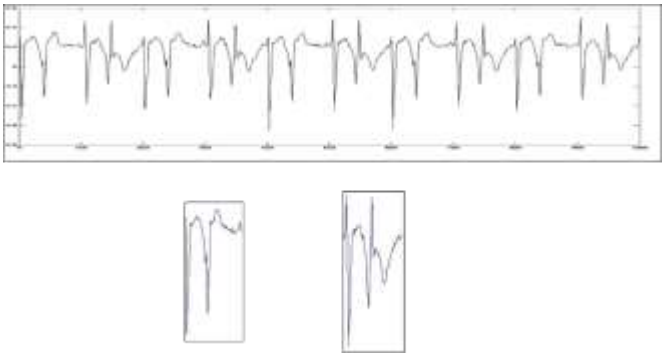

Figure 2. Clusters of ECG data by using subsequence time series clustering

Definition 3. One of the main tasks of data mining technique is Clustering. This function groups more similar objects in the same group which is called cluster. It is the most prevalent task for analyzing statistical data in different aspects. In the cluster analysis, most of the similar data objects will be discovered based on some criteria to compare with the other ones. The goal of clustering is having high efficiency of similarity among members of clusters [10].

Definition 4. Mining Time Series data has a tremendous growth of interest in today's world. To provide an indication various implementations are studied and summarized to identify the different problems in existing applications. Clustering time series is a trouble that has applications in an extensive assortment of fields and has recently attracted a large amount of research. Time series data are frequently large and may contain outliers. In addition, time series are a special type of data set where elements have a temporal ordering. Therefore clustering of such data stream is an important issue in the data mining process.

Definition 5. In the context of time series mining, Subsequence Time Series Clustering is proposed to group interesting subsequences time series data in a same cluster. Subsequence time series clustering is used for discovering structures or patterns in a time series data. In this type of clustering, firstly, subsequences are extracted from whole time series data, then, one of the clustering techniques such as partitional clustering, model-based clustering, grid-based clustering, and hierarchical clustering will be used [7].

2.2 Pattern recognition

One of the interesting job in the time series clustering is pattern discovery which contains two major fields: frequently appearing [11] and surprising patterns [12]. These methods are also called motif discovery [13, 14] and anomaly detection [15, 16] or finding discords [17], respectively.

The pattern discovery is interesting in a variety of domains as one of the significant task in data mining [18, 19]. In 2003, Ma and Perkins [20] have offered a support vector regression (SVR)-based online novelty detection algorithm which applied pattern discovery method for clustering temporal sequences data. Respectively, Chan and Mahoney [15] in 2005, have suggested an online anomaly detection approach based on the Gecko algorithm. It produces a sequence of minimal bounding boxes with the training trajectories. For the problem of time series pattern discovery, a common group of techniques being employed is distance-based clustering [21-23].

2.3 Euclidean distance

Let x_i and v_j each is a P-dimensional vector. The Euclidean distance is [24]:

$$d_E = \sqrt{\sum_{k=1}^{p}(x_{ik} + v_{jk})^2} \qquad (1)$$

3. Related works

One of the most challengeable issues in time series data mining community [25-30] is time series clustering [31-35]. Large number of algorithms have been working on time series clustering area [34, 35]. The presented paper in 1998 [21] proposed the main referenced technique for a single time series clustering which is used for the subroutine rule discovery of time series. In this paper, K-means clustering algorithm has applied for extracting subsequences clusters. The method of this paper only shows the sine waves and also had independent output to compare with input.

Therefore, some of the papers were presented to improve this problem [31-33, 36]. They assume that subsequences clusters have equal intervals with a correct lue of K for clustering. This method helps them to solve some fatal error in the main paper. Between 1998 and 2004, most of the papers try to explain algorithms for Time Series Clustering, but they have some problems that one of the critical things was meaningless results [21, 37-41].

Moreover, researchers prove their claim about meaningless problem in 2005 and explain that all of those works which have done till now is meaningless because at the end they have same results and it is not acceptable [33]. Hence, researchers try to find the solution of this issue and answer to the question of why it is meaningless (From 2005 to 2011). Since the problem was posed in 2005, so many authors try to solve the problem [32, 36, 42-48], however, they could not.

After that, in recent years (from 2011), three main papers has proposed to solve the problem of meaningless result for Time Series Clustering and finally they explained how we can have meaningful Time Series Clusters [6, 7, 49]. In the following, these papers will be explained briefly.

In the first article [6], authors make two fundamental contributions. First, they illustrate that the problem definition for time series clustering from streams currently used is inherently imperfect. Second, they employ the Minimum Description Length (MDL) framework that is efficient, productive and parameter-free method for time series clustering. They prove that their method produces correct outputs on a wide variety of datasets from medicine, zoology and industrial process analyses. In this research authors produce a clustering representation that has great ability on ignoring some of the data for grouping different length subsequences.

Another research has suggested a new Subsequence Time series clustering named Selective Subsequence Time Series (SSTS) clustering. Their authors believe that if some noise or unimportant subsequences be ignored and different lengths of member subsequences are allowed, then, clustering on time series subsequences can be meaningful. They promised the efficiency of their algorithm by testing in different data domains such as ECG data. [7]. However, their approaches do require some predefined constraint values, such as the width of the subsequences that are in fact subjective and sensitive.

Third article has explained how to cluster multiple time series. Additionally, Subsequence Time Series (STS) clustering that is a clustering of subsequences within a single time series have described. New research has proposed a novel parameter free clustering technique to remove this deficiency by appending a discovery algorithm and some statistical principles to determine these parameters. Their output from datasets proves the efficiency of the method in selecting the proper subsequence width [49]. However, it is offline and very expensive in term of complexity.

Although these three related articles try to solve the problem of subsequence time series clustering, there are still some problem which include lack of online time series clustering and high complexity. It is very essential in ECG data domain to diagnose online heart illnesses. Hence, we proposed online time series clustering to improve efficiency of clustering tasks for ECG data.

In this part, convention approach will be explained. First of all, we have an ECG time series data without any segmentation. Then, time series data will be segmented to the subsequence data. Continually, the required details of subsequences are put on the similarity matrix in order to apply for clustering. In the next step, subsequences will be clustered according to the similarity matrix and finally, results of clustering will be appeared.

Around 10 years ago, some of the researchers have done time series clustering, but, it was shown that the results are meaningless [33]. For example, in this case, the reason is the same results of ECG cluster types for each patient at the end. By passing the time, researchers found that, there are some ways to solve this problem and then proposed some methods and published some papers in accordance with their results.

In this paper we focus on applying online Subsequences Time Series Clustering for creating runtime results. For implementing this methodology we concentrate on some parts of time series instead of whole one. Regarding to this issue partial result will be upgraded time by time. We hope to obtain an appropriate result which can have high accuracy to compare with previous researches. In this methodology we choose different size of window side. According to window side, we will divide whole time series for generating continues subsequences .Then Subsequences Time Series Clustering algorithm will be applied for each subsequences. Additionally, the result of this clustering will be merged and updated dynamically. In this way, there is one matrix for saving all of the results of subsequence clustering. Then the result will be compared with main result which was obtained from main algorithm.

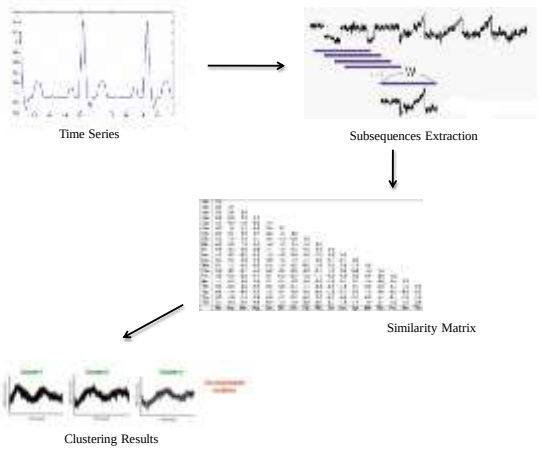

Figure 3. Subsequences Time Series clustering framework

4. Decision

The first decision in this paper is about the base subsequence clustering algorithm to use. There is ML clustering algorithm which used for creating the best clusters from subsequence data by using fixed slid window [45, 50, 51]. This slid window use for dividing time series data and also the length of clusters which we want to use. This only leaves the question of which distance measure to use. There is increasing empirical evidence that the best distance measure for time series is either Euclidean Distance (ED), or its generalization to allow time misalignments, Dynamic Time Warping (DTW) [45]. DTW has been shown to be more accurate than ED on some problems; however, it requires a parameter, the warping window width, to be carefully set using training data, which we do not have. Because ED is parameter-free, computationally more tractable, allows several useful optimizations in our framework (triangular inequality etc.), and works very well empirically [45, 51], we use it in this work. However, nothing in our overarching architecture specifically precludes other measures.

5. Framework

There are some assumptions in this framework:

- We assume we have a dynamic data stream TS.

TS could be each type of time series data such as an audio stream, a video stream, a text document stream, multi-dimensional time series telemetry, etc. Moreover, TS could be a combination of any of the above. In this work, we use ECG data in order to find some solution for heart anomaly discovery.

- Given TS, we assume we can generate a subsequence stream STS that is "subsequence" of TS.

STS is simply a subsequence time series that has fixed slid window (which is smaller than the whole time series length, therefore, it is easy to analyze in real time). In some situations, STS may be a parts of TS which is created time by time. In our framework, we consider, TS is an ECG data and STS is subsequence of TS which change time by time. In this way, the clusters information about this STS will be recorded in a matrix and then the results comprise and merge together to reach a final result. Note that our framework includes the possibility of the special case where TS = STS, as in Figure 2.

Given the sparseness of our assumptions and especially the generality of our work, we wish to produce a very general framework in order to address a wealth of domains. However, many of these domains come with unique domain specific requirements. Thus, we have created the framework outlined in Figure 3, which attempts to divorce the domain dependent and domain independent elements.

Recall that STS itself may be the subsequence of interest, or it may just be a proxy for a higher dimensional stream TS, such as an ECG data in Figure 3.top.right

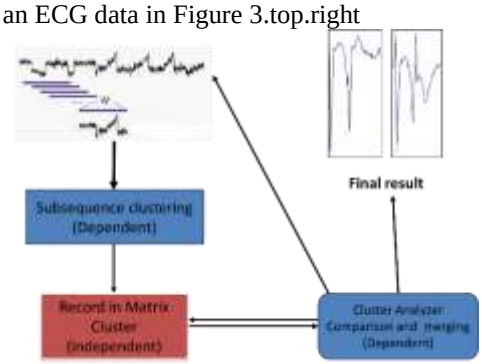

Figure 4. System framework

Our framework is further explained at a high level in Table 1. We begin in Line 1 by initializing the cluster matrix, at the beginning it is empty and by passing time, it will complete. We then initialize a slid window size of TS, w. This slid window is initialized with random data, but as we shall see, these random data are quickly replaced with subsequences from STS as the algorithm runs.

Table1. Online pattern recognition framework

Algorithm: Online Pattern Recognition (TS,STS,w)	
1	Matrix ← initializing the cluster matrix
2	w = slid window(e.g. between 200_t to 600_t)
3	**For ever**
4	STS ← subsequence from TS
5	STS clustering (TS,STS,w)
6	Comparison analyzing (new clusters with previous ones)
7	Updating cluster matrix
8	**end**

After these initialization steps, we enter an infinite loop in which we repeatedly extract the next available subsequence from the time series TS (Line 4), then pass it to a module for analyzing subsequence clusters. In this unit, domain dependent normalization may take place (Line 5), and we will attempt to classify the subsequence using the matrix. If the subsequence is not classified and is regarded as valid, then it is passed to the frequent pattern maintenance algorithm in Line 6, which attempts to maintain an approximate history of all data seen thus far. If the new subsequence is similar to previously seen data, this module may signal this by returning a new 'top' motif. In Line 7, the active learning module decides if the current top motif warrants seeking a label. If the motif is labeled by a cluster, the current matrix is updated to include this new known cluster.

Subsequence clustering refers to any domain specific preprocessing that must be done to prepare the data for the next stage (frequent pattern mining). In this clustering z-normalization is necessary [45]. More generally, this step could include down sampling, smoothing, wandering baseline removal, taking the derivative of the signal, filling in missing values, etc. In some domains, very specialized processing may take place. For example, for ECG datasets, robust beat extraction algorithms exist that can detect and extract full individual heartbeats, and converting from the time to the frequency domain may be required [52].

As shown in Table 2-Line 3, after processing, we attempt to classify the subsequence by comparing it to each subsequence time series in our matrix and assigning its information to matrix, if and only if it is within the appropriate threshold. If that is the case, we increment the class counter and the subsequence is simply discarded without passing it to the next stage.

Table2. Subsequence clustering framework

Algorithm : subsequence clustering (TS,STS, result)	
1	STS ← subsequence of TS
2	sim ← check similarity with the recorded clusters by Euclidean distance
3	**If** sim < 0.7 (70 percent accuracy)
4	Counter ++
5	Save the result
6	Empty STS
7	**End**

6. Conclusion

We clarified the online pattern recognition framework by using subsequence time series clustering for real valued time series data. We have shown our framework overview in order to apply in the next generation of online subsequence time series clustering which is robust to significant noise. Moreover, by applying it to diverse domains, it is a very general and flexible framework. In future work, we hope to remove the few assumption/parameters we have and apply our ideas to year-plus length streams. And also we hope to explain more deeply about this regards. We will made all

codes and data freely available and hope to see our work built upon and applied to an even richer set of domains such as healthcare.

7. References

[1] J. Bao, "The Outlier Interval Detection Algorithms on Astronautical Time Series Data," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[2] G. Ren, "Detection of Outliers in a Time Series of Available Parking Spaces," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[3] S. Fong, "Using hierarchical time series clustering algorithm and wavelet classifier for biometric voice classification," *BioMed Research International*, vol. 2012, 2012.

[4] X.-H. Yang and Y.-Q. Li, "DNA optimization threshold autoregressive prediction model and its application in ice condition time series," *Mathematical Problems in Engineering*, vol. 2012, 2011.

[5] S. Pandey, W. Voorsluys, S. Niu, A. Khandoker, and R. Buyya, "An autonomic cloud environment for hosting ECG data analysis services," *Future generation computer systems*, vol. 28, pp. 147-154, 2012.

[6] T. Rakthanmanon, E. J. Keogh, S. Lonardi, and S. Evans, "Time series epenthesis: Clustering time series streams requires ignoring some data," *IEEE 11th International Conference on Data Mining (ICDM)*, pp. 547-556, 2011.

[7] S. Rodpongpun, V. Niennattrakul, and C. A. Ratanamahatana, "Selective subsequence time series clustering," *Knowledge-Based Systems*, vol. 35, pp. 361-368, 2012.

[8] R. Akkiraju, J. Farrell, J. Miller, M. Nagarajan, M.-T. Schmidt, A. Sheth, *et al.* (2005). *Web service semantics-wsdl-s*.

[9] A. Gorbenko and V. Popov, "On the Longest Common Subsequence Problem," *Applied Mathematical Sciences*, vol. 6, pp. 5781-5787, 2012.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, pp. 264-323, 1999.

[11] T.-c. Fu, F.-l. Chung, V. Ng, and R. Luk, "Pattern discovery from stock time series using self-organizing maps," in *Workshop Notes of KDD2001 Workshop on Temporal Data Mining*, 2001, pp. 26-29.

[12] E. Keogh, S. Lonardi, and B. Y.-c. Chiu, "Finding surprising patterns in a time series database in linear time and space," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 550-556.

[13] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 493-498.

[14] Y. Tanaka, K. Iwamoto, and K. Uehara, "Discovery of time-series motif from multi-dimensional data based on MDL principle," *Machine Learning*, vol. 58, pp. 269-300, 2005.

[15] P. K. Chan and M. V. Mahoney, "Modeling multiple time series for anomaly detection," in *Data Mining, Fifth IEEE International Conference on*, 2005, p. 8 pp.

[16] L. Wei, N. Kumar, V. N. Lolla, E. J. Keogh, S. Lonardi, and C. Ratanamahatana, "Assumption-Free Anomaly Detection in Time Series," in *SSDBM*, 2005, pp. 237-242.

[17] E. Keogh, J. Lin, S.-H. Lee, and H. Van Herle, "Finding the most unusual time series subsequence: algorithms and applications," *Knowledge and Information Systems*, vol. 11, pp. 1-27, 2007.

[18] J. P. Caraça-Valente and I. López-Chavarrías, "Discovering similar patterns in time series," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 497-505.

[19] A. Lerner, D. Shasha, Z. Wang, X. Zhao, and Y. Zhu, "Fast algorithms for time series with applications to finance, physics, music, biology, and other suspects," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 2004, pp. 965-968.

[20] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 613-618.

[21] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth, "Rule Discovery from Time Series," in *KDD*, 1998, pp. 16-22.

[22] T. Oates, "Identifying distinctive subsequences in multivariate time series by clustering," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 1999, pp. 322-326.

[23] H. Wang, W. Wang, J. Yang, and P. S. Yu, "Clustering by pattern similarity in large data sets," in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, 2002, pp. 394-405.

[24] T. Warren Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, pp. 1857-1874, 2005.

[25] R. de A Araújo, "A class of hybrid morphological perceptrons with application in time series forecasting," *Knowledge-Based Systems*, vol. 24, pp. 513-529, 2011.

[26] Y.-S. Lee and L.-I. Tong, "Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming," *Knowledge-Based Systems*, vol. 24, pp. 66-72, 2011.

[27] H. Li and C. Guo, "Piecewise cloud approximation for time series mining," *Knowledge-Based Systems*, vol. 24, pp. 492-500, 2011.

[28] V. Niennattrakul, D. Srisai, and C. A. Ratanamahatana, "Shape-based template matching for time series data," *Knowledge-Based Systems*, vol. 26, pp. 1-8, 2012.

[29] X. Weng and J. Shen, "Classification of multivariate time series using locality preserving projections," *Knowledge-Based Systems*, vol. 21, pp. 581-587, 2008.

[30] X. Weng and J. Shen, "Classification of multivariate time series using two-dimensional singular value decomposition," *Knowledge-Based Systems*, vol. 21, pp. 535-539, 2008.

[31] J. R. Chen, "Making subsequence time series clustering meaningful," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Texas, USA, 2005, pp. pp. 114-121.

[32] A. M. Denton, C. A. Besemann, and D. H. Dorr, "Pattern-based time-series subsequence clustering using radial distribution functions," *Knowledge and Information Systems*, vol. 18, pp. 1-27, 2009.

[33] E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: implications for previous and future research," *Knowledge and Information Systems*, vol. 8, pp. 154-177, 2005.

[34] C.-P. Lai, P.-C. Chung, and V. S. Tseng, "A novel two-level clustering method for time series data analysis," *Expert Systems with Applications*, vol. 37, pp. 6319-6326, 2010.

[35] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Mining and Knowledge Discovery*, vol. 13, pp. 335-364, 2006.

[36] J. R. Chen, "Useful clustering outcomes from meaningful time series clustering," in *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, 2007, pp. 101-109.

[37] T. Bastogne, H. Noura, A. Richard, and J. Hittinger, "Application of subspace methods to the identification of a winding process," in *Proc. of the 4th European Control Conference, ECC*, 1997, pp. 1-4.

[38] D. J. Cook and L. B. Holder, "Substructure discovery using minimum description length and background knowledge," *arXiv preprint cs/9402102*, 1994.

[39] H. Li and N. Abe, "Clustering words with the MDL principle," in *Proceedings of the 16th conference on Computational linguistics-Volume 1*, 1996, pp. 4-9.

[40] M. Li, *An introduction to Kolmogorov complexity and its applications*, 2nd ed.: Springer Verlag, 1997.

[41] E. P. Pednault, "Some Experiments in Applying Inductive Inference Principles to Surface Reconstruction," in *IJCAI*, 1989, pp. 1603-1609.

[42] V. Athitsos, H. Wang, and A. Stefan, "A database-based framework for gesture recognition," *Personal and Ubiquitous Computing*, vol. 14, pp. 511-526, 2010.

[43] D. Bouchard and N. Badler, "Semantic segmentation of motion capture using laban movement analysis," *Intelligent Virtual Agents*, pp. 37-44, 2007.

[44] Z.-J. Chuang, C.-H. Wu, and W.-S. Chen, "Movement epenthesis generation using NURBS-based spatial interpolation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, pp. 1313-1323, 2006.

[45] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: experimental comparison of representations and distance measures," *Proceedings of the VLDB Endowment*, vol. 1, pp. 1542-1552, 2008.

[46] S. C. Evans, A. Kourtidis, T. S. Markham, J. Miller, D. S. Conklin, and A. S. Torres, "MicroRNA target detection and analysis for genes related to breast cancer using MDLcompress," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, p. 4, 2007.

[47] A. Mueen, E. Keogh, and N. Bigdely-Shamlo, "Finding time series motifs in disk-resident data," *ICDM'09*, pp. 367-376, 2009.

[48] S. Papadimitriou, J. Sun, C. Faloutsos, and S. Y. Philip, "Hierarchical, parameter-free community discovery," in *Machine Learning and Knowledge Discovery in Databases*, ed: Springer, 2008, pp. 170-187.

[49] N. Madicar, H. Sivaraks, S. Rodpongpun, and C. A. Ratanamahatana, "Parameter-free subsequences time series clustering with various-width clusters," *2013 5th International Conference on Knowledge and Smart Technology (KST)*, pp. 150-155, 2013.

[50] B. Hu, Y. Chen, and E. Keogh, "Time series classification under more realistic assumptions," in *Proceedings of the thirteenth SIAM conference on data mining (SDM)*, 2013.

[51] A. Mueen, E. Keogh, and N. Young, "Logical-shapelets: an expressive primitive for time series classification," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1154-1162.

[52] G. E. Batista, E. J. Keogh, A. Mafra-Neto, and E. Rowton, "SIGKDD demo: sensors and software to allow computational entomology, an emerging application of data mining," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 761-764.