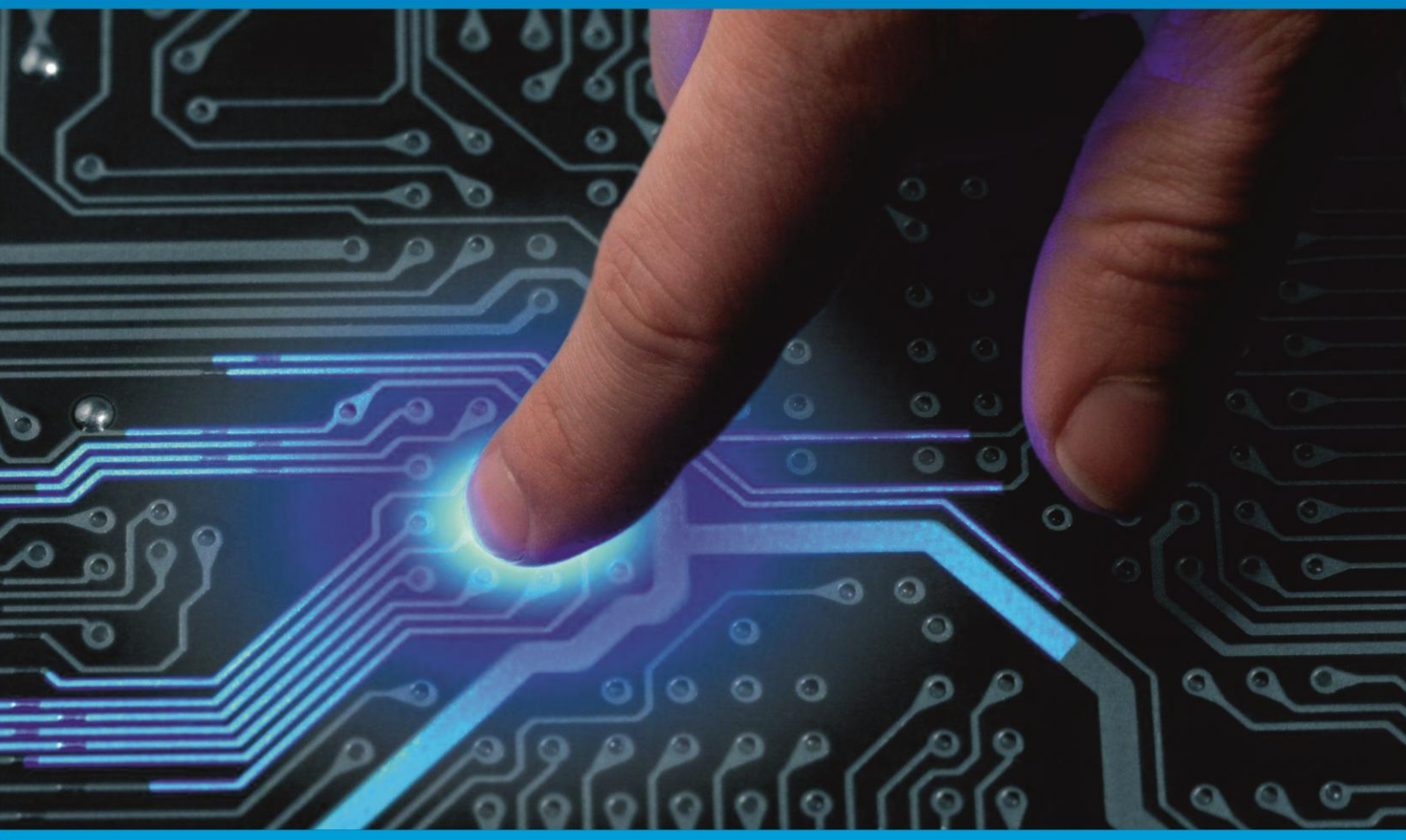
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |

Probable Distance Measure for Multi-Level Time Series Clustering

Mr.T.Rajesh, Dr.K. Venugopalrao

Asst. Professor, Dept. of CSE., G.Narayanamma Institute of Technology and Science, Telangana, India

Professor, Dept. of CSE., G.Narayanamma Institute of Technology and Science, Telangana, India

ABSTRACT: Time series clustering is a very effective approach in discovering valuable information in various systems such as finance, embedded bio-sensor and genome. However, focusing on the efficiency and scalability of these algorithms to deal with time series data has come at the expense of losing the usability and effectiveness of clustering. In this paper a new multi-level approach is proposed to improve the accuracy of clustering of time series data. In the first level, time series data are clustered approximately. Then, in the second level, the built clusters are split into sub-clusters. Finally, sub-clusters are merged in the third level. In contrast to existing approaches, this method can generate accurate clusters based on similarity in shape in very large time series datasets. The accuracy of the proposed method is evaluated using various published datasets in different domains.

KEYWORDS: Data mining, clustering, time series, large datasets

I. INTRODUCTION

A time series is a sequence of continuous values which is naturally high dimensional and large in data size. Clustering such complex data is particularly advantageous for exploratory data analysis, for summary generation, and as a pre-processing step for either another time series mining task or as part of a complex system. Researchers have shown that generally, clustering by using well-known conventional algorithms generate clusters with acceptable structural quality and consistency, and are partially efficient in terms of execution time and accuracy for static data. However, classic machine learning and data mining algorithms do not work well for time series due to their unique structure. The high dimensionality, very high feature correlation, and (typically) the large amount of noise that characterize time series data present difficult challenges for clustering. Accordingly, massive research efforts have been made to present an efficient approach for time series clustering. However, focusing on the efficiency and scalability of these methods to deal with time series data has come at the expense of losing the usability and effectiveness of clustering. That is, they suffer from either overlooking of data or inaccurate similarity computation. Overlooking of data is caused by dimensionality reduction (sometimes reduction to very low resolutions of data). Inaccurate similarity computation is due to use inappropriate metrics. For example, Euclidean distance (ED) is adopted as a distance metric on most of the existing works because it is very efficient whereas it is not accurate enough because it is only proper to calculate the similarity in time (i.e. similarity on each time step) . In contrast, it is shown that clusters generated based on similarity in shape (i.e. the time of occurrence of patterns is notImportant), are very accurate and meaningful clusters. For instance, Dynamic Time Warping (DTW) can compute the similarity in shape between time series. However, this metric is not usually used because it is very expensive in terms of time complexity.

In this paper, the problem of the low quality in existing works is taken into account, and a new Multi-level Time series Clustering model (MLTSC) is proposed which can make the clusters based on similarity in shape. This model facilitates the accurate clustering of time series datasets and is designed especially for very large time series datasets. It overcomes the limitations of conventional clustering algorithms in finding the clusters of similar time series in shape. In the first level of the model, data are pre-processed, transformed into a low dimensional space, andgrouped approximately. Then, the pre-clustered time series are split in the second level by using an accurateclustering method, and are represented by some prototypes. Finally, in the third level, the prototypesare merged to construct the ultimate clusters. To evaluate the accuracy of the proposed model, MLTSC is tested extensively by using published time series datasets from diverse domains. This model is more accurate than any existing works and is also scalable (on large datasets) due to the use of multi-resolution of time series in different levels of clustering. It is shown that using MLTSC, clustering of time series based on similarity in shape does not need to calculate the exact distances between all-time series in a dataset; instead, by using prototypes of similar time series, accurate clusters can be obtained.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |

II. RELATED WORK

There are two main categories in time series clustering, “Subsequence clustering” and “Whole time series clustering”. Subsequence clustering is based on sliding window extractions of an individual time series and aims to find similarity and differences among different time windows of an individual time series. Time series clustering has become an important topic, motivated by the challenge of developing methods to recognize dynamic change and similarity search of sequences. “Whole time series clustering” is the clustering performed on many individual time series to group similar series into clusters.
The focus of this research is on whole time series clustering with a short or modest length, not on long time series because comparing time series that are too long is usually not very meaningful. For long time series clustering, some global measures (e.g., seasonality, periodicity, skewness, chaos, etc.) which are obtained by statistic operations, are more important. In general, whole time series clustering can be broadly classified into five groups according to conventional forms of clustering algorithms: Partitioning, Hierarchical, Grid-based, Model-based and Density-based clustering algorithms. In hierarchical clustering of time series, nested hierarchy of similar groups is generated based on a pairwise distance matrix of time series. Hierarchical clustering has a great visualization power intime series clustering. This characteristic of hierarchical clustering leads to be used for time series clustering to a great extent. Additionally, in contrast to most algorithms, hierarchy clustering does not require the number of clusters as an initial parameter which is a well-known and outstanding feature of this algorithm. It is also a strength point in the time series clustering, because usually it is hard to define the number of clusters in real world problems. However, essentially hierarchical clustering cannot deal effectively with large time series datasets due to its quadratic computational complexity.

Accordingly, it leads to be restricted to the small datasets because of its poor scalability. A partitioning clustering method, makes k groups from n unlabelled objects such that each group contains at least one object. One of the most used algorithms of partitioning clustering is k-Means where each cluster has a prototype which is the mean value of its members. However, when it comes to the time series clustering, constructing an effective prototype is a challenging issue. Another member of partitioning family is k-Medoids algorithm, where the prototype of each cluster is one of the nearest objects to the centre of the cluster. In both k-Means and k-Medoids clustering algorithms,number of clusters, k, has to be pre-assigned, which is not available or feasible to determine for manyapplications, so it is impractical in obtaining natural clustering results and is known as one of their drawbacks in static objects and also time series data. Even it is worse in time series because the datasets are very large and diagnostic checks for determining the number of clusters is not easy. However-Means and k-Medoid are very fast compared to hierarchical clustering and it has made them very suitable for time series clustering. Therefore, they have been used in many works either in their “crispy” manner or “fuzzy” manner (Fuzzy c-Means and Fuzzy c-Medoids).

III. CONCEPTS AND DEFINITIONS

The key terms that are used in this work are presented in this section. The objects in the dataset related to the problem at hand are time series of similar length.

Definition *Time series clustering*, given a dataset of n objects $D = \{F_1, F_2, \ldots, F_n\}$, where $F_i = \{f_1, \ldots, f_t, \ldots, f_T\}$ is a time series, the process of unsupervised partitioning of D into, $C = \{C_1, C_2, \ldots, C_k\}$ in such a way that homogenous time series are grouped together based on their similarity in shape, is called time series clustering. Then, $C_i$is called a cluster, where $D = \cup_{i=1}^{k} C_i$and $C_i \cap C_j = \emptyset$for $i \neq j$.

Representation method
Many researches have been carried out focusing on the representation or dimensionality reduction of time series. Considering all these works, it is understood that dimension reduction is necessary to some extent, however it is undeniable that as more dimensionality reduction occurs, moredata is lost and becomes inaccurate. In fact, it is a trade-off between the accuracy and speed which is a controversial and non-trivial task in representation methods. However, among all these representation methods which have their strong points and weaknesses, in this paper, the focus is on Symbolic Aggregate Approximation (SAX) representation because of its strength in the representation of time series.

Brief review of SAX

Two different studies, separately approximated the time series using segmentation approach. They use the mean value of the equal-length segmentations of time series as the approximate value of that part. This technique is called

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

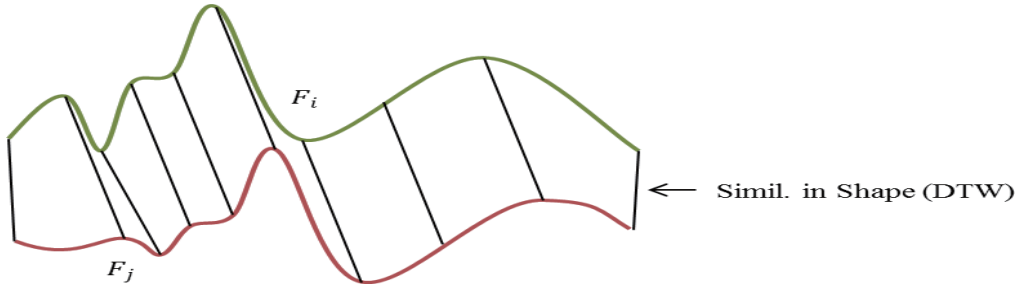
|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |

Piecewise Aggregate Approximation (PAA), and is quite fast can be used for arbitrary length queries. If $F_i = \{f_1, \ldots, f_t, \ldots, f_T\}$ is considered as a time series, then, PAA discretized time series to $\bar{F}$ where $\bar{F} = \overline{f1}, \ldots, \overline{fw}$. Each segment of $\bar{F}$, i.e., $\overline{fi}$, is a real value which is the mean of all data points in the *i*th segment of *F*. Symbolic Aggregate ApproXimation (SAX) maps the PAA coefficients to symbols. SAX was developed by Keogh et al. in 2003 and has been used by more than 50 groups in different data mining researches [58]. Considering $\bar{F}$ as discretized time series by PAA transformation, then, $\hat{F}$ where $\hat{F} = \hat{f}_1 \ldots \hat{f}_w$ is defined by mapping the PAA coefficients to 'a' SAX symbols. 'a' is the alphabet size (e.g., for the alphabet = *{a, b, c, d, e, f}*, 'a' = 6), which are defined by "breakpoints". Based on Keogh definition, a list of numbers $B = \beta_1 \ldots \beta_{a-1}$ is defined as "breakpoints" to determine the area of each symbol in SAX transformation.

Similarity measure

Time series clustering relies on distance measures to a high extent. There are many distance measures proposed by researchers in the literature such as ED, DTW, ERP, TQuEST ,LCSS , EDR and CDM . However, it can be drawn from literature that ED and DTWare the most common methods in the time series clustering because of the efficiency of ED and the effectiveness of DTW in similarity measurement. ED is a one-to-one matching measurement which is used in most of the works (about 80%) in the literature. ED is simple, fast and is used as a benchmark in many works, because it is parameter free. However, ED is not the best choice as a distance function because it is dependent on the domain of the problem in hand and the characteristics of the dataset's time series. In fact, it is veryweak and sensitive to small shifts across the time axis which make it proper for finding time series which are *similar in time*. For example it is not accurate enough for calculating similarity of sequences such as: <abaa>, <aaba>. In contrast to ED which proposes a one-to-one matching, DTW is suggested as a one-to-many metric. DTW is a generalization of ED which solves the local shift problem in the time series to be compared. Local shift problem is a time scale issue which is a characteristic of most time series. Handling local shifts allows similar shapes to be matched even if they are out of phase in the time axis, i.e., *similarin shape*. Using this definition, clusters of time series with similar patterns of change are onstructed



Regardless of time points, for example, to cluster share price related to different companies which have a common pattern in their stock independent of its occurrence in time series. It makes DTW superior to ED which is only able to find time series which are *similar in time*. DTW uses "warping" the time axis in order to achieve the best alignment between the data points within the series (Fig.). Given two time series, $F_i = \{f_1, f2. \ldots fx\}$, $F_j = \{f_1, f_2, \ldots, f_y\}$, a $x \times y$matrix is defined where the (s, k) element of the matrix is the Euclidean distance, $dis(f_s, f_k)$, between two time points $f_s$and f_k. The warping path $w_u$forms a set of warping paths = $\{w_1, w_2, \ldots, w_u\}$, that has the minimum distance between the two series of $F_i$and $F_j$is of interest.

$$dis_{DTW}(F_i, F_j) = \min\left(\sum_{u=1}^{U} Wu/U\right)$$

Generally, dynamic programming is used in order to find the warping path effectively. However, it causes scalability problem which is a big challenge for DTW because it requires quadratic computation. As a result, many researchers, try to speed it up usually by proposing an efficient lower bound approximations of the DTW distance to reduce its complexity. Nevertheless, most of these works are under the classification problem (the search area is pruned using a lower bound distance of DTW) and is not suitable for the clustering problem where the distance between all objects should be calculated. However, in this paper DTW can be adopted whiteout violating the efficiency of clustering using a novel approach explained further.

IV. PROPOSED MULTI LEVEL TSC

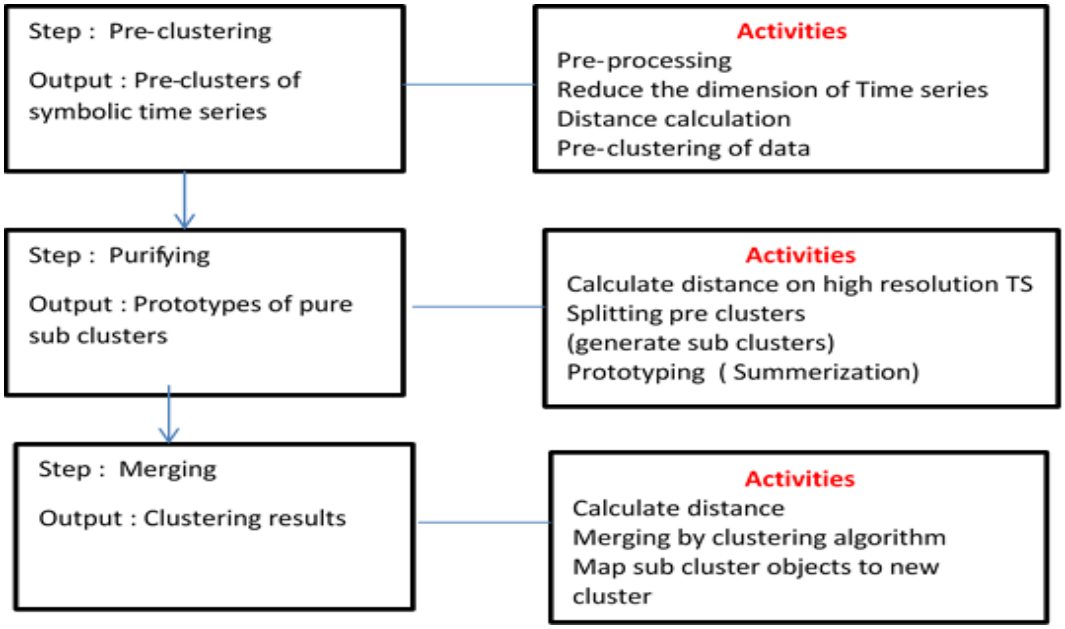
MLTSC includes three steps: pre-clustering, splitting and merging. Figure shows the overall view of the process in MLTSC briefly. In the first level (i.e. pre-clustering step), data are clustered to reduce the search space. In this step, time series data are used in a low-resolution mode. In the second level, time series data are used in their high-resolution

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |

mode. A new approach is designed to split the pre-clusters into sub-clusters. Finally, in the third level, sub-clusters are merged using an arbitrary scheme.



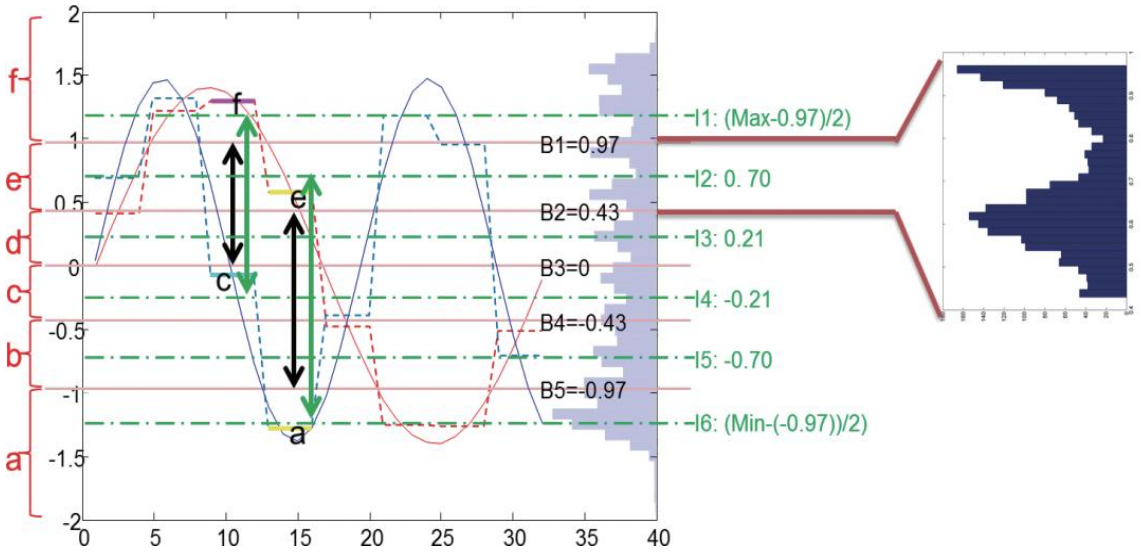
Pre-clustering

Pre-clustering is performed to group the low-resolution time series (which can fit in memory) rather than original (raw) time series dataset. It reduces the search area for the second level of MLTSC. At first time series are standardized using z-score (z-Normalization) which make time series invariant to scale and offset. Then, SAX is adopted as a representation method to reduce the dimensionality of normalized time series. The major reasons for using dimensionality reduced time series in the preclustering step is the problem of disk I/O constraint especially in the large datasets. The generic solution for this problem is to create an approximation of the data, which will fit in the main memory, yet retains the essential features of interest. As a result, the whole data can be loaded into main memory and the problem at hand is solved approximately. However, the size of a dataset can be very large such that even with dimensionality reduction, it cannot fit in the memory. The solution for this case is utilizing an incremental clustering where clusters are updated (or expanded) incrementally. Moreover, when we talk about large datasets, in addition to a large number of instances, it could include a long duration in the time axis. However, because MLTSC is a multi-step algorithm, the first level can be done by different distance measures depend on the nature of time series. For example, the methodproposed in can be adopted which takes into account dynamic changes in the data characteristics. In order to make the pre-clusters, an appropriate distance measure compatible with SAX is desirable. In the literature, Euclidean or MINDIST measure are used in order to calculate the similarity between two time series represented by SAX. Lin et al. Introduced MINDIST as a compatible distance metric for SAX. However, MINDIST has been introduced to address the indexing problem in the time series domain and is not enough accurate for calculation of distance among time series in the clustering problem, as a result, in this paper, a new approach, Probable Distance (PDIST) is introduced.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |



As mentioned, symbols in SAX are defined based on the location of PAA coefficients in some equipropbable regions made by some break points. MINDIST considers the distance between the neighbour symbols as zero, and ignores the maxima and minima points of time series. Instead, in PDIST, the distance between regions is calculated as the distance between indicators of the regions. The indicator of a region is defined in such a way that the closeness of PAA coefficients (in the region) to the indicator is the highest in that region. For defining the indicator, the arithmetic mean of each area (minimum and maximum) is defined as an indicator of the area as the best estimator of the regions as:

$$Ind_i = (\beta i - 1 + \beta i) / 2$$

Where $\beta 0$ is the global minimum and βa is the global maximum. Figure 3 illustrates a visual intuition of the PDIST metric. The black line shows the MINDIST between two symbols, while dashed line indicates the PDIST distance between the same symbols. Based on this definition, the PDIST between each pairs of symbolized time series is defined as following:

$$dis_{PDIST}(\bar{F}_x, \bar{F}_y) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} \left(dis(Indi, Indj)\right)^2}$$

where*Indi* is the indicator of *i*th region and the dis() function is calculated by

$$dis(\mathrm{Ind}i, \mathrm{Ind}j) = \begin{cases} 0 & i = j \\ \left| \dfrac{\beta_j + \beta_{j-1} - \beta_i - \beta_{i-1}}{2} \right| & otherwise \end{cases}$$

Which is pre-computed and read-off by a lookup table. In this part, PDIST is evaluated against the MINDIST and ED. To make a comparison, some datasets are adopted from UCR. At first, time series of each dataset is transferred into discrete space (SAX). Then, three distance matrices are made by each distance measure (i.e., PDIST, MINDIST and ED). At that point, the difference between the obtained distance matrices and the distance matrix calculated by ED using raw time series (ED_{RAW}) is computed as tightness of each metric to ED_{RAW}. The tightness of the metrics, for example PDIST, to $ED_{RAW}$is calculated by the following equation.

$$Tightness\ (\hat{F}_i \hat{F}_j) = 1 - \frac{\left| ED_{RAW}(F_i, F_j) - PDIST(\hat{F}_i, \hat{F}_j) \right|}{ED_{RAW}(F_i, F_j)}$$

Where $\hat{F}_i$*is* the SAX representation of time series F_i.The average tightness of all UCR datasets over the cross product of the alphabet and compression-ratio is shown in the Fig. The bigger tightness indicates more accurate approximation. As the results show, PDIST is closer to *EDRAW*rather than the other two approaches for most of the datasets. It implies that PDIST is more accurate than other approaches. To cluster the approximated data, k-Modes is used in order to divide *N* time series into *k* partitions. K-Modes is based on k-Means family algorithm, and is used for clustering categorical data. Considering that time series represented by SAX is categorical data in each segment, and k-Modes

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |

work fine with categorical data, it is a good choice for pre-clustering. Moreover, similar to other partitioning algorithms, it has a high speed (especially in large datasets) and provides a good quality by choosing the centroids on a low dimension approximation of data (e.g. SAX) which increases the quality in the partitioning clustering.

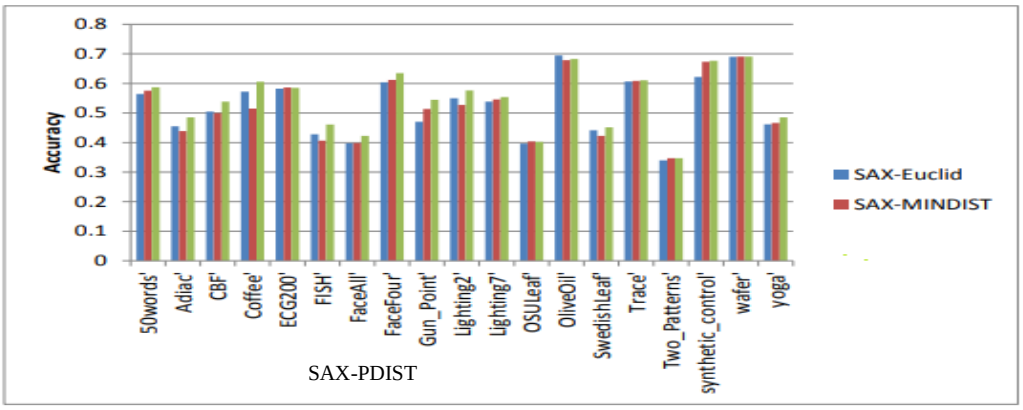
V. EXPERIMENTAL EVALUATION

Datasets

The proposed model is experimented with 19 different datasets from the UCR Time series Data Mining Archive where the numbers of clusters, dimensionality, and number of instances have been explained. This set is chosen because it is of various numbers of clusters, different cluster shapes and density contains noise points, and used in many articles in the literature as a benchmark.

Evaluation method

In general, evaluating of extracted clusters (patterns) is not easy in the absence of the data labels. However, all these datasets have class labels and can be used for evaluation of MLTSC using external indices such as Dunn, DB, etc. Complete reviews and comparisons of some popular techniques exist in the literature. However, these indices are dependent on the adopted clustering algorithm and structure of clustering, and there is not a compromise and universally accepted technique to evaluate clustering approaches. Therefore, the most common-used external indices in the time series clustering domain are used for evaluation of accuracy of MLTSC, i.e., Rand Index. To avoid biased evaluation, the conclusions are drawn based on the average value of the indices. Considering $G = \{G1, G2. . . GM\}$ as ground truth clusters, and $C = \{C1C2, . . . , CM\}$ as the clusters made by MLTSC under evaluation using accuracy.



SAX-PDIST

Moreover, to report the accuracy of SAX-PDIST is calculated to prevent the bias of random initialization .Although the focus of this study is on improving the accuracy of PDIST, the scalability of the proposed model is also calculated to prove its feasibility theoretically.

VI. CONCLUSION

In this paper, we provided a detailed overview of various techniques used for clustering of time series data. Focusing on whole time series clustering, the results obtained by applying first level of MLTSC on different datasets were evaluated visually and objectively. It was visually shown that clustering can be applied on large time series datasets to generate a hierarchy of meaningful and accurate clusters using First level of MLTSC. In order to evaluate the validity of the clusters formed, different evaluation methods were used to show the accuracy of first level of MLTSC. The accuracy of the proposed method was evaluated using various datasets. Moreover, First level ofMLTSC was applied on large time series datasets. Finally, the time complexity of the proposed model was computed.Experimental results show that first level of MLTSC outperforms other conventional distance measures experimenting on various datasets. In future work we will be extending to implement the second level, third levels of MLTSC to consider long time series datasets to evaluate the approach.

REFERENCES

[1] J. Aach and G.M. Church, Aligning gene expression time series with time warping algorithms, *Bioinformatics* 17 (2001), 495.

[2] S. Aghabozorgi, M.R. Saybani and T.Y.Wah, Incremental clustering of time-series by fuzzy clustering, *Journal of Information Science and Engineering* 28 (2012), 671–688.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 1, January 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0901008 |

[3] S. Aghabozorgi, T.Y. Wah, A. Amini and M.R. Saybani, A new approach to present prototypes in clustering of time series, in: *The 7th International Conference of Data Mining*, Las Vegas, USA, (2011), 214–220.
[4] R. Alcock and Y. Manolopoulos, Time-series similarity queries employing a feature-based approach, in: *7th Hellenic Conference on Informatics*, Ioannina, Greece (1999), 1–9.
[5] T. Rajesh, Y. S. Devi and K. V. Rao, "Hybrid clustering algorithm for time series data — A literature survey," *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, Chirala, 2017, pp. 343-347, doi: 10.1109/ICBDACI.2017.8070861
[6] A. Banerjee and J. Ghosh, Clickstream clustering using weighted longest common subsequences, in: *Proc of the Workshop on Web Mining, SIAM Conference on Data Mining*, Citeseer, (2001), 33–40.
[7] T.Rajesh,Dr.K.V.GRao, Time series clustering- Introduction to Healthcare system,International Journal of Innovative Technology and Exploring Engineering vol 9 , issue 1 (2019) 5786-5794.

REFERENCES

Mr.T.Rajesh is graduated with B.Tech in 2006 from JNT University, India and completed M.Tech from CBIT, India during 2009. He is presently working as Assistant Professor of the Dept. of Computer Science and Engineering, G. Narayanamma Institute of Technology & Science for women College, India. He has about 7 years of teaching experience. His areas of interest include Data Mining, Machine Learning, Software Engineering.

Dr. K. VenuGopalaRao is presently working as Professor of the Dept. of Computer Science; G. Narayanamma Institute of Technology & Science for women for women College, India .He has published more than twenty papers in national/International journals. His areas of interest include E-Learning, Software Engineering, Data Mining, Networking and etc. He has about 25 years of teaching experience.

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details