ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 6 | Issue: 2 | February 2020 || Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 5.614||ISI Value: 1.188

FAST AND ACCURATE SIMILARITY DETECTION IN TIME SERIES REPRESENTATION MODEL

A.Angeline Valentina Sweety[1]
[1,] PGStudent,
Francis Xavier Engineering College,
Department of Computer Science and Engineering

R.Muthulakshmi [2]
[2] PGStudent,
Francis Xavier Engineering College,
Department of Computer Science and Engineering

R.Uma Maheshwari[3]
[3]UGStudent,
Francis Xavier Engineering College,
Department of Computer Science and Engineering

N.Raja Priya[4]
[4]Assistant Professor,
Francis Xavier Engineering College,
Department of Computer Science and Engineering
Tamilnadu,
India

ABSTRACT

Likeness search and location is a focal issue in time arrangement information preparing and the executives.In this paper, two novel approaches to perform similarity detection efficiently and effectively. The new information portrayal model depends on the pattern data of time arrangement, which can give succinct yet include rich portrayal of time arrangement. FAD_DTW can adjust the portions of time arrangement in direct time, which significantly quickens the closeness discovery process.It extensively compare FAD_DTW with state-of-the-art time series representation models and similarity .One is composed of a new time series representation model and a corresponding similarity measure, which is called Fragment Alignment Distance(FAD);the other applies dynamic time warping to the representation.It efficiency and effectiveness validations on various data sets demonstrate that FAD_DTW can achieve fast and accurate similarity detection.

KEYWORDS: ***Time series data mining ,Time series , Similarity measures , Data representation models, Clustering.***

I. INTRODUCTION

A time series is a sequence of ordered numeric values between which an interval of points is defined.Time series are generally used to indicate the object with time;hence,large amounts of such data are available from many domains,including speech recognition,financial and market data analysis ,biomedical measurement,sensor networking and moving-object trajectory tracing.Time-series data mining unveils numerous facets of complexity. The most unmistakable issues emerge from the high dimensionality of time-arrangement information and the trouble of characterizing a type of similitude measure dependent on human observation.To normalize the similarity detection problem and guide the research work,many scholars have noted various benchmarks for similarity measurement algorithms.Most of them can be classified as one of two types:
Data representation models.

Representing data in a form that can be effectively processed is the first step of data mining. The ideal representation of time series not only can maintain the original features of the data but also has a simple format. Hence, the representation model should be realized in a low-dimensional space and consider the basic distribution of the data.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 6 | Issue: 2 | February 2020 || Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 5.614||ISI Value: 1.188

Similarity measures.

Similarity measurement is the central technique of similarity search and detection. Distinguishing between two time series or formalizing the difference between two series in accordance with human common sense is the crucial problem. Thus, a reasonable similarity measure should have the following characteristics: consistency with human cognition, consideration of the most prominent features on both the local and global spaces, and the capability to unconditionally identify arbitrary objects.

II. RELATED WORKS

Phongsakorn Sathianwiriyakhun[2] has proposed a fast,accurate,parameter-free shape averaging method that can automatically discover the proper number of subclasses within the training data, and then globally average sequences within these subclasses to generate multiple templates for classification task. The investigation results show that our proposed work can accelerate the general order assignments by huge edge, while having the option to keep up high exactnesses, contrasting and the cutting edge NCC approach. It is likewise seen that our proposed highlight can accomplish similar grouping correctnesses (little lower in a few and minimal higher in a few, with no measurably noteworthiness).In any case, in some datasets where the mistake rates drop down beneath noteworthy degree of 5%, great speedups on all different datasets are accomplished, showing the tradeoff between the characterization correctnesses and the running time it could spare.

Donald J. Bemdt, James Clifford [3] has proposed the Knowledge discovery in databases presents many interesting challenges within the content of providing computer tools for exploring large data archives. Electronic information storehouses are developing rapidly and contain information from business, logical, and different spaces. A lot of this information is naturally worldly, for example, stock costs or NASA telemetry information. Distinguish £bug designs in such information streams or time arrangement is a significant information disclosure task. This work portrays some esseWntial investigations with a dynamic prograrning way to deal with the issue. The example discovery calculation depends on the dynamic time traveling system utilized in the discourse acknowledgment field.

Johannes ABfalg, Hans-Peter Kriegel..en has proposed the most conspicuous work has concentrated on closeness search thinking about either complete time arrangement or comparability as indicated by subsequences of time arrangement. For some, spaces like monetary examination, medication, ecological meteorology, or natural perception, the recognition of worldly conditions between various time arrangement is significant. Rather than customary methodologies which consider the course of the time arrangement to coordinate, coarse pattern data about the time arrangement could be adequate to take care of the previously mentioned issue. Specifically, worldly conditions in time arrangement can be recognized by deciding the purposes of time at which the time arrangement surpasses a particular edge. In this job, we begin the book thought of threshold queries in time series databases which statement those time series greater than a user-defined query threshold at similar time frames compared to the query time series. We present a new resourceful access method which uses the fact that only partial information of the time series is necessary at query time. The performance of our solution is demonstrated by an extensive investigational evaluation on real world and an artificial time series data.

Gustavo E.A.P.A. Batista, Xiaoyue Wang has proposed The pervasiveness of time arrangement information across practically all human undertakings has delivered an incredible enthusiasm for time arrangement information mining in the most recent decade.While there is a plenty of characterization calculations that can be applied to time arrangement, the entirety of the current exact proof proposes that basic closest neighbor order is particularly hard to beat. The choice of distance measure used by the nearest neighbor algorithm depends on the invariance's required by the domain. In this work we make a surprising claim. There is an invariance that the network has missed, multifaceted nature invariance. Instinctively, the issue is that in numerous spaces the various classes may have various complexities, and sets of complex articles, even those which abstractly may appear to be fundamentally the same as the human eye, will in general be extra separated under present separation measures than sets of basic items. This reality presents blunders in closest neighbor order, where complex items are inaccurately appointed to an easier class.

Xiaoyue Wang, Abdullah Mueen has proposed the exploration endeavors right now centered around presenting new portrayal techniques for dimensionality decrease or novel comparability occasions for the central information. In the limitless best piece of cases, every individual work presenting a demanding strategy has made explicit cases and beside the incidental hypothetical supports, gave quantitative exploratory perceptions.In any case, generally, the similar parts of these trials were excessively barely centered around showing the advantages of the proposed strategies over a portion of the recently presented ones. In sort to give a total approval, we led a general test study re-actualizing eight diverse time arrangement portrayals and nine likeness measures

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 6 | Issue: 2 | February 2020 || Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 5.614||ISI Value: 1.188

and their variations, and giving their effectiveness a shot 38 minutes in time arrangement informational indexes from a wide scope of use areas.In this editorial, we give a summary of these different techniques and present our relative experimental findings regarding their effectiveness.

III. SYSTEM ARCHITECTURE

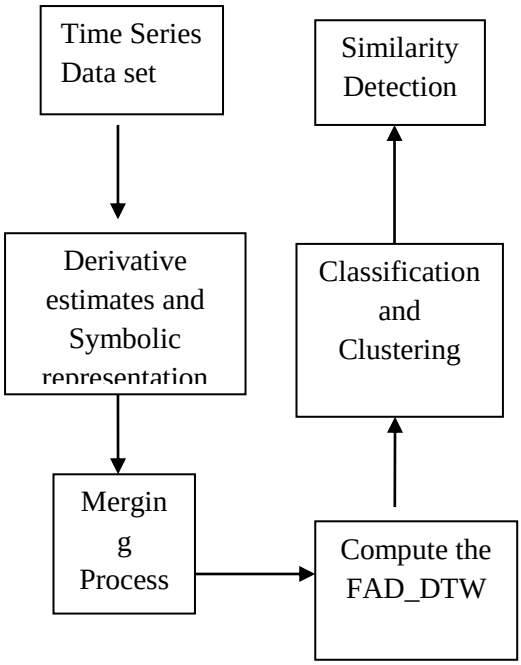

Fig.1 System Architecture for Time Series Representation Model

IV PROPOSED WORK

The contributions of our work as follows:

Like other data representation models, Dynamic Time Warping can be applied to the representation model of FAD directly; FAD_DTW is devised based on the concept that similar time series have similar change trends. It is consistent with human cognition and available for various data types; FAD_DTW is time-warping-aware and can deal with data of unequal length in linear time; FAD_DTW transforms the comparison between points into the comparison between change trends, which can address the high dimensionality as well as capture the essential features of time series.

To evaluate the performance of FAD_DTW, we conducted an extensive experiment by using clustering and classification frameworks. This evaluation inevitably involved prominent state-of-the-art methods for both the time series representation models and similarity measures. Experimental evidence that FAD_DTW is an effective and efficient method in similarity detection is presented.

V. METHODOLOGY

5.1 Derivative Estimation Model

Gullo et al. presented a more accurate way to approximate the first derivative of a time series;The DSA estimation model only considers the slope of the line from the left adjacent point to the right adjacent point to the right adjacent point.The derivatives of the first and last points in the series are calculated by their adjacent points as well.Formally,

$$xh = \begin{cases} xh+1-xh & h = 1 \\ \frac{1}{2}(xh+1-xh-1) & h \in [\ldots,n-1] \\ xh-xh-1 & h = n \end{cases} \quad (1)$$

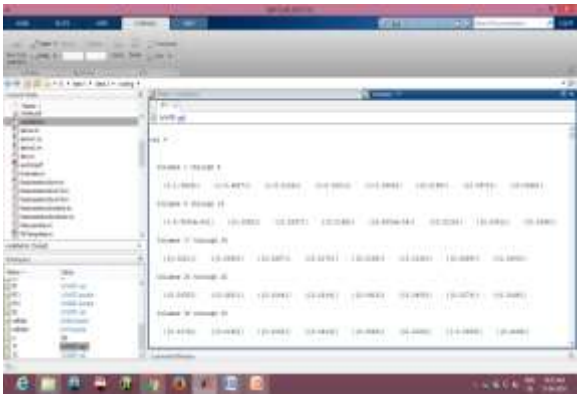

Fig.2 Derivative Estimates of Time Series

5.2 Symbolic Representation Sequences

The key idea of the segmentation step is to divide a time series according to the derivative estimation value of the points.Specifically,it set a threshold ε to judge the change magnitude of the data.If the derivative estimation value of a point is less than ε,that poimt has little change compared to previous one.In this way,FAD can transform the derivative sequence into a symbolic representation sequence.Formally.

$$Rh = \begin{cases} \lambda & xh > \lambda.\ \varepsilon \\ \ldots & \ldots \\ 1 & \varepsilon < xh \le 2.\ \varepsilon \\ 0 & |xh| \le \varepsilon \\ -1 & -2.\ \varepsilon < xh \le -\varepsilon \\ \ldots & \ldots \\ -\lambda & xh < -\lambda.\ \varepsilon \end{cases} \quad (2)$$

.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 6 | Issue: 2 | February 2020 || Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 5.614||ISI Value: 1.188

Rh is the Symbolic representation of xh.
ε is a threshold value for the change trend.
λ indicates the number of symbols used to represent the time series.
λ is an integer and is not less than one.

The original series is a sinusoidal signal with random noise.The series with different scales can be obtained by adjusting the value of ε.The series become flatter as the threshold value increases.The values are 0.08 , 0.10 , 0.12 , 0.20.If the derivative estimation value of a point has little change compared to the previous one.The change trends of series with threshold values.

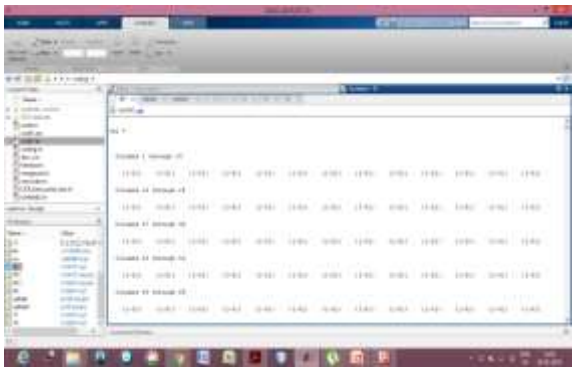

Fig.3 Symbolic Representation Sequences

5.3 Merging Process

A fragment of the time series is obtained by merging adjacent points with the same symbos. To avoid the loss of time axis information, every fragment also records the number of adjacent points that have the same symbols during the process of merging them. For example, the first subsequence of R in Fig. 4 is composed of four zero symbols and can be represented by (0, 4) in T1.

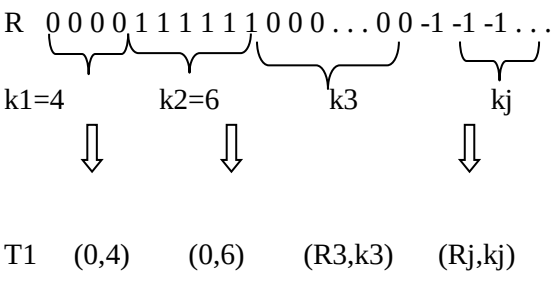

Figure 4. Symbolic representation of time series.

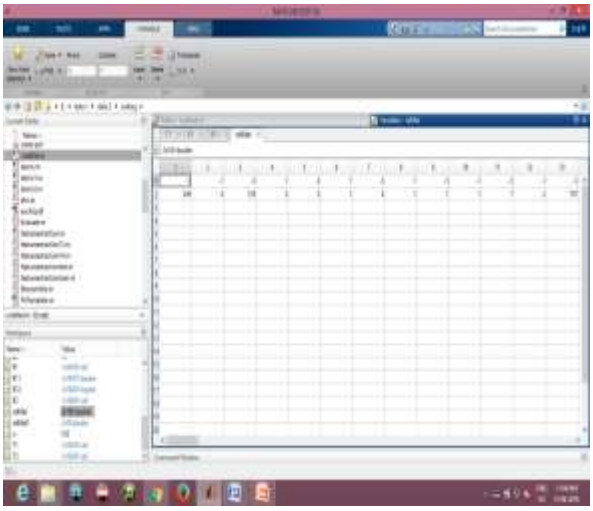

Fig.5 Merging process of Time Series

5.4 Fragment Alignment Distance

The purpose of this subsection is to give clear definitions of the terms used throughout this article. Definitions 1) A time series T consisting of n ordered numerical points can be defined as:

$T=(x_1,x_2,....,x_n)$, $x_i \epsilon R$

A time series is usually an observation result of an underlying process.It can thus be defined as a set of successive time instants.

Definitions 2) Given a time series T of length n,a subsequence S of T is a part of the series T that consists of contiguous time instants of length m(m<=n):

$S = (xk,xk+1, \ldots ,xk+m-1)$, $1 \leq k \leq n-m+1$.

Definition 3) The similarity measure D(T,U) of time series T and U is a function that measures the distance between them.D(T,U) takes two time series as inputs and returns the distance between them.This distance cannot be a negative value,that is,$D(T,U) \geq 0$.

5.5 Similarity Measure

In the following part, It propose a new computational method that can find such a warping path in linear time. To align the two series, three cases need to be considered satisfied in the course of similarity measurement.

1)The mapped fragments in T11 and T22 have same symbols,which indicates they have similar change trends.Hence,the distance between them mainly depends on the difference in their lengths.It compute the distance between them by Eq.(3).

$D(S1i,s2j)=\gamma x(\max\{k1i,k2j\}/\min\{k1i,k2j\}-1)$, ifR1i=R2j (3)

Where k1i and k2j are the quantities of focuses for S1i and S2j,respectively,and γ is a movable parameter to change the separation proportion of

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 6 | Issue: 2 | February 2020 || Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 5.614||ISI Value: 1.188

same images to various images..Intuitively, the distance between the same symbol fragments must be less than the distance ones.Thus,It have $0 \leq D(S1i,s2j)<1$ and $\gamma \epsilon(0,1)$.

2)Due to the time series' unequal lengths and the time warping awareness of FAD, some Fragments usually remain in one of the series with no fragments in the other series for mapping.Such fragments can be viewed as not being similar to any fragments, which correspond to case 1).It define the distance as Eq.(4)

$$D(-,Si)=1 \quad (4)$$

Algorithms

Since the objective of our work is to access the ability of FAD in time series data mining tasks,It employed standard classification and clustering framework for assessment,which include K-means clustering and nearest neighbor classification.

K-Means Clustering

The K-means algorithm,one need to specify the number of output clusters;for simiplicity,it adopted data sets for which relavant classifications are available.Thus,It set the number of output equal to the actual number of classes in each clustering evaluation.Moreover,since the initial cluster cendroids greatly affect the performance of the K-means algorithm.

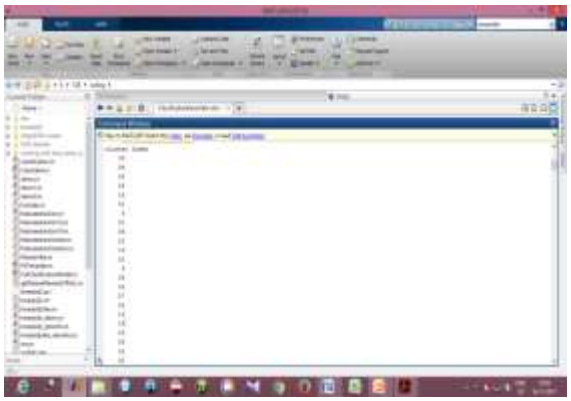

Fig.6 K-means clustering

One Nearest Neighbor Classification

Nearest-neighbor classification is widely known to be a straightforward and effective method to access the performance of various algorithms. The one nearest-neighbor classifications classifies each data instance according to the most similar instance to it.

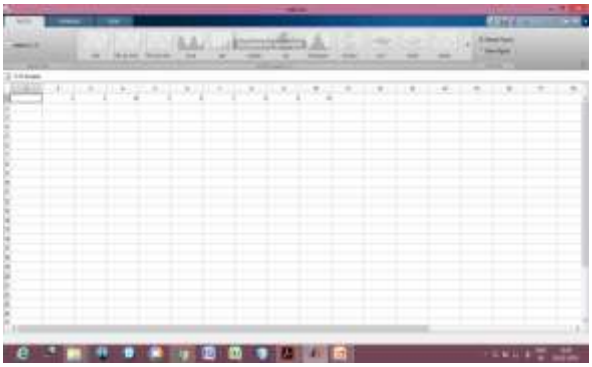

Fig.7 One Nearest Neighbor Classification

No	Data sets	Classes	Size of training/test set	Time series length	Type
1	OSU Leaf	6	200/242	427	Shape
2	Lighting7	7	70/73	319	Real
3	ECG200	2	100/100	96	Real
4	Plane	7	105/105	144	Shape

TABLE 1.Data sets.

Data Description

The 4 time series data sets employed in this experiment are collected from the UCR Time series Classification/Clustering in Homepage. The data resource is provided by Keogh et al.,which is derived from a variety of applications.Table 2.The type labels require a brief explanation.Some data sets are real,which simply means they were recorded as natural time series from some physical process,for instance,the earthquake signal from sensor reading.Some datas are shapes:these are one dimensional time series that were extracted by processing some two-dimensional shapes,such as leaf profiles or the silhouettes of planes.

Performance in Time Series Clustering

FAD is the fastest method among the competing methods. Moreover ,FAD_DTW,SAX_DTW and PAA_DTW all have relatively high performance in run time.Due to the high complexity of DTW,the methods based on it show inferior performances compared with FAD.FAD,SAX and PAA representation methods can greatly improved the time performance of basic DTW.Thus, FAD will demonstrate great superiority in terms of run time as the size of the data increases.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 6 | Issue: 2 | February 2020 || Journal DOI: 10.36713/epra2013 || SJIF Impact Factor: 5.614||ISI Value: 1.188

Table 2. Ranking of different methods for k-means clustering results

DATA SIZE	PAA_DTW	SAX_DTW	FAD_DTW
10	0.705	0.666	0.322
30	0.805	0.827	0.48
50	0.555	0.845	0.376
40	0.994	0.777	0.698
90	0.666	0.699	0.393
60	0.589	0.643	0.484

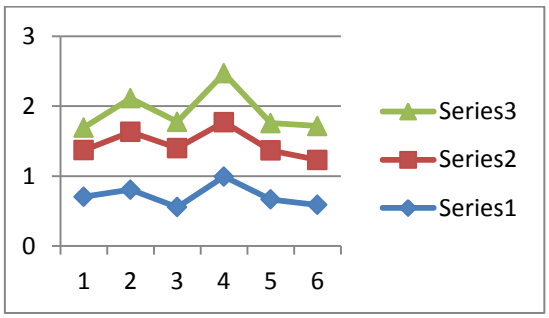

Fig.8 Time performance in the k-means clustering tasks

VI CONCLUSION

FAD_DTW and FAD are more accurate than other methods in both classification and clustering tasks. Uncommonly, FAD_DTW is more precise than FAD in the 1-NN characterization and K-implies bunching errands.transform time series into compact yet feature-rich symbolic sequences by extracting trend information of data and diagonally mapping the similar change trends between series. FAD is devised based on the notion that similar time series have similar trends. Thus, it is a technique dependent on the pattern of the information. FAD_DTW applies Dynamic Time Warping on the representation model of FAD; thus, it can extract the same features of time series as FAD. FAD_DTW has high accuracy in classification and clustering tasks.

VII REFERENCES

1. *Fujii, K. Yamamoto, and S. Nakagawa(2011), ``Automatic speech recognition using hidden conditional neural _elds," in Proc. IEEE Int. Conf. Acoust.,Speech, Signal Process., , pp. 5036_5039.*
2. *D. J. Berndt and J. Clifford, ``Using dynamic time warping to _nd patterns in time series," in Proc. Knowl. Discovery Databases Workshop, 1994, pp. 359_370.*
3. *J. Aßfalg, H.-p. Kriegel, P. Kroger, P. Kunath, A. Pryakhin, and M.Renz, (2006)"Similarity search on time series based on threshold queries,"in Proc. 10th Int. Conf. Adv. Database Technol., , pp. 276–294.*
4. *G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in Proc. SDM, 2011, pp. 699–710.*
5. *X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. Keogh,(2012) "Experimental comparison of representation methods and distance measures for time series data," Data Mining Knowl. Discovery, vol. 26, no. 2, pp. 275–309, Feb.*
6. *E. J. Keogh and M. J. Pazzani, ``Derivative dynamic time warping,(2001)" in Proc. SIAM Int. Conf. Data Mining, , pp. 1_11.*
7. *C. Ratanamahatana and E. J. Keogh(. 2004), ``Making time-series classi_cation more accurate using learned constraints," in Proc. SIAM Int. Conf. Data Mining, Lake Buena Vista, FL, USA, Apr, pp. 11_22.*
8. *E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra(2001), ``Dimensionality reduction for fast similarity search in large time series databases," Knowl. Inf. Syst., vol. 3, no. 3, pp. 263_286,.*
9. *Q. Li, B. Moon, and I. F. V. Lopez(2004), ``Skyline index for time series data," IEEE Trans. Knowl. Data Eng., vol. 16, no. 6, pp. 669_684, Jun..*