Pattern Recognition Letters 34 (2013) 361–371

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A representation of time series based on implicit polynomial curve

Gang Wu [*], Jiwen Yang

Key Lab of Electronic Business, Nanjing University of Finance and Economics, Nanjing 210003, China

ARTICLE INFO

Article history:
Available online 4 October 2012

Keywords:
Implicit polynomial curve
Time series
Similarity measure
Dimension reduction

ABSTRACT

Implicit polynomial (IP) curve is applied to represent data set boundary in image processing and computer vision. In this work, we employed it to reduce dimensionality of time series and produce similarity measure for time series mining. To use IP curve, time series was transformed to star coordination series. Then the star coordination series was fitted by implicit polynomial curve. That is, IP curve approximated (IPA) time series. Lastly, similarity measure of the time series was produced from the fitting implicit polynomial curve. To guarantee no false negatives, the lower bounding lemma for the similarity measure based on IP curve (IPD) was proved. We extensively compared IPA with other similarity measure and dimension reduction techniques in classification frameworks. Experimental results from the tests on various datasets indicate that IPA is more efficient than other methods.

1. Introduction

A time series is a collection of observation made chronologically, which can be easily obtained from scientific and economic application, e.g., daily fluctuations of stock market, electrocardiogram, medical and biological experimental observations etc. There are various kinds of time series data related research, one of which is time series data mining. This research includes indexing, classification, clustering and representation of time series. In the context of time series data mining, the key problem is how to represent and measure the time series data (Fu and Chung, 2011) because the time series is highly dimensional and needs to find representation techniques to reduce its dimensionality and still preserve its fundamental characteristic.

In the past decade, various representation methods and similarity measures for time series can be found in the literature (Ding and Trajcevsk, 2008). Based on the underlying approximation schemes, these studies can be divided into two categories: piecewise approximation and continuous approximation. Of these two approximation representations, the former uses the segmented means to represent the time series, such as Piecewise Aggregate Approximation (PAA) (Keogh and Chakrabarti, 2001), Adaptive Piecewise Constant Approximation (APCA)(Keogh and Chakrabarti, 2002), Symbolic Aggregate approXimation (SAX)(Lin and Keogh, 2007), Indexable Piecewise Linear Approximation (PLA) (Chen and Chen, 2007), Derivative time series Segment Approximation (DSA) (Gullo and Ponti, 2009) and Perceptually Important Points (PIP) (Fu and Chung, 2008), Discrete Wavelet Transformation (DWT) (Pong Chan and Fu, 1999) etc., and the latter uses a low-order continuous function to represent the time series, such as Chebyshev Polynomials (CHEBY) (Cai and Ng, 2004), Discrete Fourier Transformation (DFT), Single Value Decomposition (SVD) (Faloutsos and Ranganahan, 1994.) and Discrete Cosine Transformation (DCT) (Korn et al., 1997) etc.

In conjunction with the above representations, there are many distance measures for similarity of time series data, such as Euclidean Distance (ED) (Faloutsos and Ranganahan, 1994), Dynamic Time Warping (DTW) (Keogh and Ratanamahatana, 2005), distance based on Longest Common Subsequence (LCSS) (Vlachos et al., 2002), Edit Distance with Real Penalty (ERP) (Chen and Ng, 2004), Edit Distance on Real sequence (EDR) (Chen, 2005), Sequence Weighted Alignment model (Swale) (Morse and Patel, 2007), Spatial Assembling Distance (SpADe) (Chen and Nascimento, 2007) and similarity search based on Threshold Queries (TQuEST) (Aßfalg and Kriegel, 2006) etc. After representing the time series, we can use all these similarity measures to evaluate the representation method.

Piecewise approximation represents time series with discontinuous function, but it suffers from the three major problems: (1) Piecewise approximation methods lead to the unnecessary error or deviation after dimensionality reduction. The methods represent a time series by dividing it into segments and recording the mean value of the data points that fall within each segment. However, the sensitivity to length of each segment can seriously reduce the accuracy of the representation model or dimensionality method. For instance, PAA approximates a time series by dividing it into equal length segments and using the average values of each segment as its representation. APCA divides the time series into disjoint segments of different lengths according to the shape of time series. However, these representation methods may distort the shape of time series because the piecewise approximation ignores most of

[*] Corresponding author.
E-mail address: wugang69@gmail.com (G. Wu).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.09.016

time series information. (2) There is no correlation among the elements of segments representation dataset. Each segment of time series is fitted by constant, line or low degree curve, which are used to represent the time series. It is clear that mean value (e.g. PAA, APCA) or coefficients of fitting curve (e.g. PLA) of each segment dot not have any correlation, which leads to loss of the general information of time series; (3) Piecewise approximation methods are not efficient for dimension reduction. Many segments of piecewise approximation are still in high dimensionality after representation of time series. Therefore, piecewise approximation representation methods cannot efficiently reduce the dimensionality of time series.

Continuous approximation minimizes the maximum deviation (minimax approximation) between time series data and continuous function that often is polynomial curve, e.g. Chebyshev Polynomials. Polynomial approximation usually suffers the range oscillation phenomenon, especially to high degree polynomial. Other continuous approximations e.g. DWT, DFT are not able to minimize the maximum deviation from the original data points.

In order to overcome the above weaknesses, in this paper, we present a time series representation model that is conceived to support accurate and fast similarity detection. This model is called IP curve approximation (IPA), the advantage of which is due to representing the time series using continuous function and Least Squares (LS) approximation. Therefore, the weaknesses from piecewise approximation can be avoided. In the other hand, the range oscillation of the polynomial curve (e.g. CHEBY) is overcome due to the time series representation with IP curve that is function curve of two variables. In particularly, the LS approximation is better than minimax approximation. Actually, the reason of IPA outperforming other representation methods is that IPA can describe both the whole information and the local information of time series best. Specifically, the coefficients of IP curves when represent to time series appear to be relatively insensitive to noise or to modest changes (Subrahmonia et al., 1996.), because of this stability, IPA can ignore minor or noisy information of time series. In the other hand, IP curve can accurately represent time series with small number of coefficients and capture the important tends of time series.

As a preview, we make the following contribution: (1) Transform the time series into star coordinate series, which is a bounded data set. Besides, the lower bounding lemma for the data set is proved; (2) Fit the star coordinate series data with IP curve, the coefficients of which are used to represent the time series; (3) Present the similarity measure based the IP curve, and the lower bounding lemma for this measure is proved.

This paper is organized as follows. Below we discuss the related work. In the next section, we review the IP curve, and focus on the Min–Max algorithm. In the Section 3, we show how to transform the time series to star coordinate series. The methods of transformation and inverse transformation are given. Finally, the lower bounding lemma is proved. In Section 4, we show how to fit the star coordinate series with IP curve and give an example. In Section 5, we present the definition of IPD, and its lower bounding lemma is proved. In Section 6, we present our experimental setup and result.

2. Related work

Time series can be regarded as a discrete function, as the domain is a set, rather than an interval, which can be written as follows:

$$T = \{(t_i, v_i) | i = 1, 2, \cdots n\} \tag{1}$$

Where t_i is time, v_i is time series data. If t_i only expresses the order of time series data, then Eq. (1) can be simplified as $T = v_i | i = 1, 2, \cdots n$. The fundamental problem of time series data representation is to determine a function to approximate Eq. (1) for dimensionality reduction.

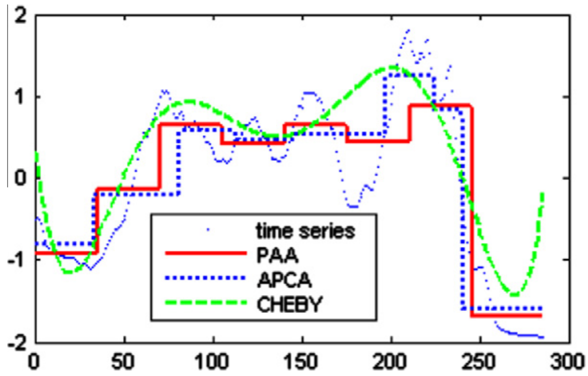


Fig. 1. Representation with PAA, APCA and CHEBY.

There are mainly two basic approaches as we mentioned above: piecewise approximation and continuous approximation. The first approach includes PAA, APCA, PLA, DSA and PIP. Among these methods, PAA is the simplest method. In this method, the average value of each segment is used to represent the corresponding set of data points. Again, with time series $\hat{T} = \{\hat{v}_i | i = 1, 2, \cdots n\}$, the times series can be represented by

$$\hat{v}_k = \frac{1}{w} \sum_{i=e_k}^{e_k+w-1} v_i, k = 1, 2, \cdots m$$

Where e_k denotes the starting data points of the kth segment in time series $\hat{T}$. w is the length of each segment ($w = n/m$) and m is the dimension after dimensionality reduction. Fig 1 shows a time series data representation with PAA, the length of which is 286. The time series is divided into 8 segments, the length of which is about 35. It is clear from Fig. 1 that the PAA algorithm uses the fixed length of segment, which distorts the shape of the times series and produces much more error. In order to overcome the disadvantage of APP, the APCA algorithm is proposed. A major difference from PAA is that APCA can identify segments of variable length. That is, the length of each segment is not fixed, but adaptively to the shape of the time series. In contrast to the PAA, the APCA is able to produce higher quality approximations of a time series. Both PAA and APCA approximate the time series with constant. Furthermore, the straight lines are used to approximate the time series instead of constant e.g. PLA. This representation tends to closely align the endpoint of consecutive segments, giving the piecewise approximation with connected line.

Furthermore, to reduce the dimensionality of time series, preserving the salient points and featuring derivative estimation of time series are two promising methods. The former is called as perceptually importance points (PIP) representation, and the latter is called as derivative time series segment approximation (DSA) representation. In the PIP representation, all the data points are found and reordered by its importance by going through the PIP identification process, which can be found in (Fu and Chung, 2008). DSA representation features derivative estimation, segmentation and segment approximation to provide high sensitivity in capturing the main trends of time series. Both of the two methods involve a segmentation scheme that employs the paradigm based on a piecewise discontinuous function. Hence, they still fall in the category of piecewise approximation representation method. In particular, the two methods cannot prove the lower bounding lemma for corresponding similarity measure.

Time series representation methods with a continuous polynomial include SVD, DFT, and CHEBY etc. SVD is computationally more expensive than the other methods for dimensionality reduction due to its space rotation and truncation applied on a data

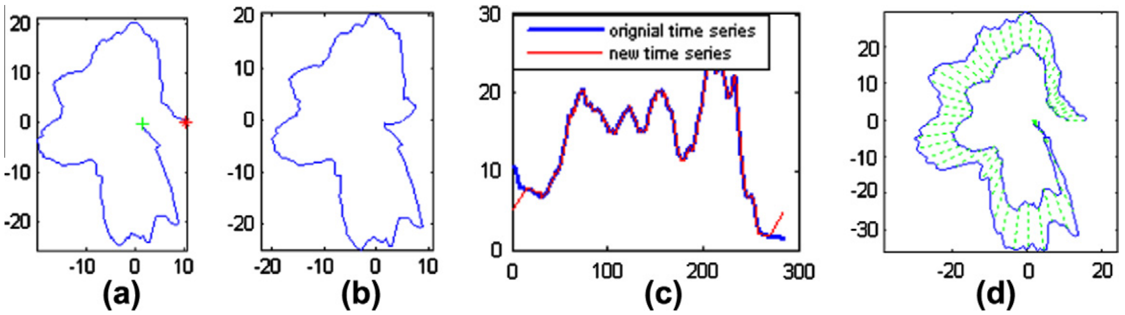


Fig. 2. Star coordinate series transformation.

matrix. Both DFT and CHEBY are based on the use of a set of orthonormal functions and the time series is represented by their relating coefficients. Major differences between DFT and CHEBY is that CHEBY approximates the time series with minimizing the maximum deviation from the times series data.

The similarity measures often are presented based on the times series representation. Faloutsos and Ranganahan (1994) points out that all similarity measures need to meet the lower bounding lemma in order to guarantee no false dismissals. Supposed that Q_1 and Q_2 are times series, $F(Q_1)$ and $F(Q_2)$ are representations of Q_1 and Q_2, respectively. Then the lower bounding lemma is

$$D_{eu}(F(Q_1),F(Q_2)) \leqslant D_{eu}(Q_1,Q_2) \quad (2)$$

Where $D_{eu}(F(Q_1), F(Q_2))$ is the distance measure between the two representations $F(Q_1)$ and $F(Q_2)$. $D_{eu}(Q_1,Q_2)$ is the distance measure between the time series Q_1 andQ_2. That is, the distance between two transformed data in the reduced space should be a lower bound of their actual distance in the original space. The lemma is critical in guaranteeing no false negatives in similarity search, index, classification and other mining tasks of the time series. The tighter the lower bound, the smaller is the number of false positives.

Euclidean Distance is the most straightforward similarity measure for time series, the disadvantages of which are very sensitive to noise and misalignments in time, and are unable to handle local time shifting. Therefore, some other similarity measures are presented: DTW performs a non-linear mapping of sequence to the other one by minimizing the total distance between them. LCSS uses the length of the longest common subsequence of two sequences to define the distance between them, which can process noise time series by performing approximate matching rather than exact matching of time series. ERP supports local time shifting. DTW is usually chosen as similarity measure to evaluate representation methods.

For time series representation methods, nearest neighbor classifier is often chosen to evaluate the efficacy of them. Ding and Trajcevsk, (2008) discuss the advantages with this approach. For example, the accuracy of the 1NN classifier directly reflects the effectiveness of the similarity measure. In addition, the 1NN classifier is easy to implement and is parameter free, which is helpful to fairly compare various representation methods.

3. IP curve representation of a time series

In this section, we discuss how to transform the time series to star coordinate series and how to preprocess the star coordinate series so that it can be fitted by the IP curve efficiently. In addition, the key properties of the star coordinate series and IP curve are discussed. In addition, the lower bounding lemma for star coordinate series is proved.

3.1. IP curve approximation

Implicit polynomial (IP) curve is a planar curve, and it can be specified by the zero set of a 2D polynomial of degree n given by

$$f(x,y) = \sum_{i+j\leqslant n, i,j\geqslant 0} a_{ij}x^i y^j = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \cdots + a_{n0}x^n + a_{n-1,1}x^{n-1}y + \cdots + a_{0n}y^n = 0 \quad (3)$$

The polynomial $f(x,y)$ can also be represented in the coefficient vector form as follows

$$f(x,y) = X^T A \quad (4)$$

where $A^T = (a_1\, a_2 \cdots a_{m-1}\, a_m)$ and $X^T = [1\, x\, y \cdots x^n\, x^{n-1}y \cdots y^n]$. m is the number of coefficients of $f(x,y)$ and $m = (n+1)(n+2)/2$.

The IP curve has been widely applied to many fields, such as object recognition (Taubin et al., 1994; Kautsky and Flusser, 2007; Subrahmonia et al., 1996; Tarel and Cooper, 2000; Oden et al., 2001), pose estimation (Yazicioglu et al, 2009; Zheng et al., 2009), coding (Helzer et al., 2000), boundary estimation from intensity or color images (Tasdizen and Cooper, 2000), symmetry detection (Wu and Li, 2002; Lebmeir and Richter Gebert, 2008;

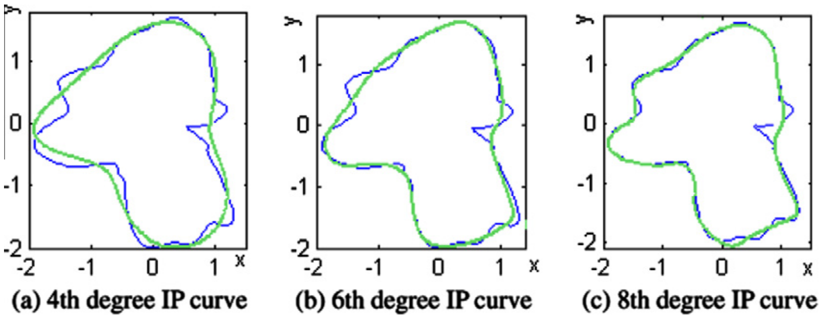


Fig. 3. IP curve fitting star coordinate series.

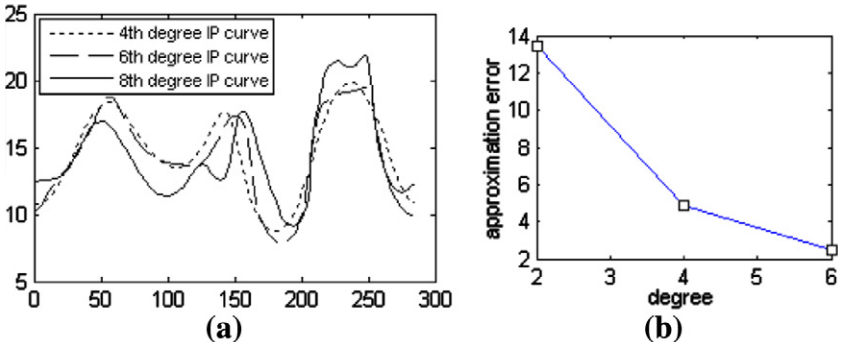


Fig. 4. IP curve representation for the time series.

Marola, 2005), image database indexing (Jiang and Xu, 2007) etc. Compared with the other models (e.g., B-spline, Fourier descriptors) to represent data set the elements of which are usually points of outline of object, IP curve has many merits. (1) IP curve can accurately describe data set and all parameters to describe it depend on coefficients of IP; (2) IP curve can be easily operated and used due to its good analytic expression. In addition, fitting IP curve is very robust to noise and can fill in missing part of data set.

The representation of IP curve is to describe data set points S (e.g. object boundary for 2D objects) by the zero set of IP, which is the fitting problem of IP curve. The fitting problem is to find a coefficient vector that leads to an IP curve $f(x,y) = 0$ that best fits the data set points under a criterion to be specified. In the past decades, many methods have been developed, such as 3L method (Blane and Lei, 2000), RR method (Tasdizen and Tarel, 2000), Min–Max method (Helzer et al., 2004) etc. Among them, the Min–Max method is much more stable, which further improves the RR and other algorithm by constraining the gradient vector along the zero set of IP to have a norm for each point of the data set. In order to obtain more stable IP curve, we usually give the points of data set a weight to balance the effect of the constraining data set.

In this paper, we use Min–Max algorithm to fit IP curve, which is reviewed as follows. We define X_x and X_y as the derivatives of the monomials vector X, with respect to x and y, respectively and denote the local normal unit vector at each of the data set points by (v_k,w_w), where $k = 1, 2, \cdots, N$ and N is the number of points in the data-set.

We construct the matrix of monomials and the matrices of monomials partial derivative in the following way:

$$M_0 = [\,X(x_1,y_1) \quad X(x_2,y_2) \quad \cdots \quad X(x_N,y_N)\,]$$

$$M_x = \left[\, \frac{X_x(x_1,y_1)}{||X(x_1,y_1)||_1} \quad \frac{X_x(x_2,y_2)}{||X(x_2,y_2)||_1} \cdots \frac{X_x(x_N,y_N)}{||X(x_N,y_N)||_1} \,\right]$$

$$M_y = \left[\, \frac{X_y(x_1,y_1)}{||X(x_1,y_1)||_1} \quad \frac{X_y(x_2,y_2)}{||X(x_2,y_2)||_1} \cdots \frac{X_y(x_N,y_N)}{||X(x_N,y_N)||_1} \,\right]$$

and the following gradient coordinates vectors:

$$v = [\,v_1 \quad v_2 \quad \cdots \quad v_N\,]$$

$$w = [\,w_1 \quad w_2 \quad \cdots \quad w_N\,]$$

Defining $M^T = [\,M_0 \quad M_x \quad M_y\,]$ and $b^T = [\,0 \quad v \quad w\,]$, we can formulate the fitting criteria as

$$MA = b$$

We use a linear least squares algorithm to solve the above equation, and can obtain the coefficients vector of fitting IP curve:

$$A = (M^TM)^{-1}M^Tb$$

Wu and Li, (2004), Wu, (2007) propose the determination algorithms of degree, closeness and connectivity of IP curve, by which IP curve with the adaptive degree is selected to fit data set points with closeness and connectivity of constraint condition.

3.2. Star coordinates transformation for time series

The initial motivation of the star coordinates work is to gain insight and numerical details for further analysis by visualizing multidimensional datasets (Tan et al., 2006). The basic idea of star coordinates is to arrange the coordinate axes on a circle on a two-dimensional plane with equal angles between the axes with origin at the center of the circle. The data points from datasets are scaled to the length of the axis. In this paper, we can efficiently use an IP curve representation of the star coordinate for the indirect analysis of time series, such as dimension reduction and feature extraction, because IP coefficients are the features that are extracted from the star coordinate transformed from the corresponding to the time series. The transformation formula for star coordinates from the times series (1) is the following (Li, 2011)

$$\begin{cases} x_i = v_i \cos(\frac{2\pi}{l} t_i) \\ y_i = v_i \sin(\frac{2\pi}{l} t_i) \end{cases} i = 1, 2, \cdots n \tag{5}$$

Where $l = 2\pi/(t_n - t_1)$, (x_i,y_i) is star coordinate point mapping from time series point (t_i,v_i). It is clear from (5) that the time series data points is converted to star coordinate ones which shape closed curve around origin (Fig. 2a). The angle between the two adjacent points (x_{i+1},y_{i+1}) and (x_i,y_i) corresponding to origin is $2\pi(t_{i+1} - t_i)/l$. We can obtain inverse transform formula from (5) as follows.

$$\begin{cases} \qquad v_i = \sqrt{x_i^2 + y_i^2} & \\ & i = 1, 2, \cdots n \\ t_i = \begin{cases} l\arccos(x_i/v_i)/2\pi & y_i > 0 \\ l[\arccos(-y_i/v_i) + \pi]/2\pi & y_i \leqslant 0 \end{cases} & \end{cases} \tag{6}$$

Table 1
Datasets used in the experiments.

Dataset	Size	# of classes	Length
Adiac	781	37	176
Beef	60	5	470
Coffee	56	2	286
ECG200	200	2	96
Mixed-BagShapes	160	9	1614
Trace	200	4	275
Diatom	322	4	345
FaceAll	2250	14	131
ECGFive	884	2	136
GunPoint	200	2	150
Haptics	463	5	1092
SwedishLeaf	1125	15	128

Table 2
Summary of quality results (error rate) for 1-NN classification.

	Adiac	Beef	Coffee	ECG200	Mixed Bag Shapes	Trace
L2	0.389	0.467	0.250	0.120	0.156	0.240
DTW	0.332	0.500	0.191	0.230	0.156	0
L2 on DFT	0.271	0.533	0.321	0.120	0.066	0.150
L2 on DWT	0.445	0.500	0.250	0.130	0.156	0.270
L2 on CHEBY	0.427	0.467	0	0.120	0.156	0.280
L2 on PAA	0.404	0.500	0.250	0.140	0.156	0.290
L2 on APCA	0.563	0.433	0.250	0.200	0.378	0.060
L2 on PL A	0.371	0.533	0.071	0.190	0.267	0.220
L2 on IPA (IPD)	0.327	0.300	0.036	0.120	0.133	0.030
DTW on PAA	0.402	0.533	0.286	0.250	0.156	0.070
DTW on APCA	0.660	0.600	0.214	0.230	0.311	0.010
DTW on PL A	0.371	0.533	0.071	0.200	0.289	0.160
DTW on IPA	0.333	0.300	0.036	0.150	0.178	0.100
	Diatom	FaceAll	ECGFive	GunPoint	Haptics	SwedishLeaf
L2	0.065	0.286	0.203	0.087	0.630	0.211
DTW	0.033	0.192	0.232	0.093	0.623	0.208
L2 on DFT	0.065	0.314	0.006	0.033	0.607	0.150
L2 on DWT	0.065	0.312	0.142	0.093	0.627	0.208
L2 on CHEBY	0.062	0.282	0.131	0.100	0.643	0.213
L2 on PAA	0.069	0.302	0.142	0.093	0.643	0.205
L2 on APCA	0.562	0.459	0.281	0.133	0.721	0.357
L2 on PL A	0.134	0.385	0.236	0.093	0.620	0.224
L2 on IPA (IPD)	0.052	0.268	0.026	0.047	0.610	0.200
DTW on PAA	0.069	0.301	0.276	0.080	0.701	0.194
DTW on APCA	0.461	0.364	0.312	0.080	0.718	0.243
DTW on PL A	0.134	0.453	0.318	0.147	0.617	0.213
DTW on IPA	0.062	0.348	0.022	0.060	0.614	0.221

Definition 1. The series $(x_i, y_i), i = 1, 2 \cdots n$ is called star coordinate series if it is transformed from the time series (1) according to (5).

Fig. 2. (a) shows one star coordinate series transforming from the time series in Fig. 1, in which the first and the last star coordinate series points are marked with "$*$" and "+", respectively. Generally, there exists a gap between the two points, which will result in the instability of fitting IP curve. In order to overcome this problem, the gap is removed by smoothing the two points and their neighbor points. Specially, Denoting the first point by (x_1, y_1) and the last point by (x_n, y_n), we can smooth these points (x_{l-1}, y_{l-1}), $(x_{l-2}, y_{l-2}), \ldots, (x_1, y_1)$, (x_n, y_n), $(x_{n-1}, y_{n-1}), \ldots, (x_{n-l}, y_{n-l})$ with 5-point moving average, where $l(l > 0)$ can be chosen according to size of the gap. Fig. 2 (b) shows the result of smoothing these points with $l = 10$. Obviously, the gap disappears. Fig. 2(c) shows the original time series with dotted line and the inverse transformation time series from star coordinate series with solid line, in which the difference between beginning and ending part of them can be found, but it cannot affect the IP representation.

Fig. 2(d) shows two star coordinate series of length 386 and lines between their corresponding points. From Definition 1, we can find that all of these lines pass through the origin. Specially, supposed $S = \{(x_i, y_i)|i = 1, 2, \ldots n\}$ and $\hat{S} = \{(\hat{x}_i, \hat{y}_i)|i = 1, 2, \cdots n\}$ are two star coordinate series, then (x_i, y_i) and $(\hat{x}_i, \hat{y}_i)$ which belong to the two series respectively lie on the straight line $y = k_i x$ (k_i is slope of the line).

Lemma 1. Let $T = \{(t_i, v_i)|i = 1, 2, \ldots n\}$ and $\hat{T} = \{(\hat{t}_i, \hat{v}_i)|i = 1, 2, \cdots n\}$ be two time series, and $S = \{(x_i, y_i)|i = 1, 2, \ldots n\}$ and $\hat{S} = \{(\hat{x}_i, \hat{y}_i)|i = 1, 2, \cdots n\}$ be their corresponding star coordinate series as defined in Eq. (5). Then:

$$D_{euc}(S, \hat{S}) \leqslant D_{euc}(T, \hat{T}) \tag{7}$$

Proof. According to the Euclidean distance formula, we have

$$D^2_{euc}(S, \hat{S}) = \sum_{i=1}^{n} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]$$

Substituting (x_i, y_i) in Eq. (5) for the above calculus, then we have

$$D^2_{euc}(S, \hat{S}) = \sum_{i=1}^{n} \left\{ \left[v_i \cos\left(\frac{2\pi}{l} t_i\right) - \hat{v}_i \cos\left(\frac{2\pi}{l} \hat{t}_i\right) \right]^2 + \left[v_i \sin\left(\frac{2\pi}{l} t_i\right) - \hat{v}_i \sin\left(\frac{2\pi}{l} \hat{t}_i\right) \right]^2 \right\}$$

$$= \sum_{i=1}^{n} \left[v_i^2 \cos^2\left(\frac{2\pi}{l} t_i\right) + \hat{v}_i^2 \cos^2\left(\frac{2\pi}{l} \hat{t}_i\right) - 2 v_i \hat{v}_i \cos\left(\frac{2\pi}{l} t_i\right) \cos\left(\frac{2\pi}{l} \hat{t}_i\right) \right]$$

$$+ \sum_{i=1}^{n} \left[v_i^2 \sin^2\left(\frac{2\pi}{l} t_i\right) + \hat{v}_i^2 \sin^2\left(\frac{2\pi}{l} \hat{t}_i\right) - 2 v_i \hat{v}_i \sin\left(\frac{2\pi}{l} t_i\right) \sin\left(\frac{2\pi}{l} \hat{t}_i\right) \right]$$

Noting that $v_i^2 \cos^2(\frac{2\pi}{l} t_i) + v_i^2 \sin^2(\frac{2\pi}{l} t_i) = v_i^2, \hat{v}_i^2 \cos^2(\frac{2\pi}{l} t_i) + \hat{v}_i^2 \sin^2(\frac{2\pi}{l} t_i) = \hat{v}_i^2$

$$2 v_i \hat{v}_i \cos\left(\frac{2\pi}{l} t_i\right) \cos\left(\frac{2\pi}{l} \hat{t}_i\right) + 2 v_i \hat{v}_i \sin\left(\frac{2\pi}{l} t_i\right) \sin\left(\frac{2\pi}{l} \hat{t}_i\right) = 2 v_i \hat{v}_i \cos\left[\frac{2\pi}{l}(t_i - \hat{t}_i)\right]$$

Then we have

$$D^2_{euc}(S, \hat{S}) = \sum_{i=1}^{n} \left\{ v_i^2 + \hat{v}_i^2 - 2 v_i \hat{v}_i \cos\left[\frac{2\pi}{l}(t_i - \hat{t}_i)\right] \right\}$$

$$\leqslant \sum_{i=1}^{n} [v_i^2 + \hat{v}_i^2 - 2 v_i \hat{v}_i] \leqslant \sum_{i=1}^{n} (v_i - \hat{v}_i)^2 \leqslant D^2_{eu}(T, \hat{T})$$

That is $D_{euc}(S, \hat{S}) \leqslant D_{euc}(T, \hat{T})$ □

Table 3
Summary of quality results (F measure and accuracy in the parentheses) for 1-NN classification.

	Adiac	Beef	Coffee	ECG200	Mixed Bag Shapes	Trace
L2	0.593 (0.611)	0.505 (0.533)	1 (1)	1 (1)	0.755 (0.844)	0.747 (0.760)
DTW	0.640 (0.669)	0.470 (0.500)	1 (1)	1 (1)	0.755 (0.844)	1 (1)
L2 on DFT	0.717 (0.729)	0.447 (0.467)	1 (1)	1 (1)	0.921 (0.933)	0.838 (0.850)
L2 on DWT	0.544 (0.555)	0.482 (0.500)	1 (1)	1 (1)	0.755 (0.844)	0.710 (0.730)
L2 on CHEBY	0.559 (0.573)	0.546 (0.533)	1 (1)	1 (1)	0.755 (0.844)	0.713 (0.720)
L2 on PAA	0.581 (0.596)	0.42 (0.500)	1 (1)	1 (1)	0.755 (0.844)	0.700 (0.710)
L2 on APCA	0.418 (0.437)	0.385 (0.400)	1 (1)	1 (1)	0.517 (0.622)	0.938 (0.940)
L2 on PL A	0.595 (0.629)	0.533 (0.467)	1 (1)	1 (1)	0.595 (0.733)	0.785 (0.780)
L2 on IPA (IPD)	0.644 (0.672)	0.694 (0.700)	1 (1)	1 (1)	0.849 (0.867)	0.969 (0.970)
DTW on PAA	0.582 (0.599)	0.447 (0.467)	1 (1)	1 (1)	0.755 (0.844)	0.925 (0.930)
DTW on APCA	0.314 (0.340)	0.362 (0.367)	1 (1)	1 (1)	0.587 (0.689)	0.990 (0.990)
DTW on PL A	0.601 (0.629)	0.446 (0.467)	1 (1)	1 (1)	0.569 (0.711)	0.844 (0.840)
DTW on IPA	0.640 (0.668)	0.694 (0.700)	1 (1)	1 (1)	0.790 (0.822)	0.900 (0.900)
	Diatom	FaceAll	ECGFive	GunPoint	Haptics	SwedishLeaf
L2	0.883 (0.937)	0.734 (0.714)	0.749 (0.769)	0.913 (0.913)	0.344 (0.370)	0.782 (0.789)
DTW	0.942 (0.967)	0.815 (0.808)	0.763 (0.768)	0.907 (0.907)	0.379 (0.377)	0.787 (0.792)
L2 on DFT	0.866 (0.937)	0.679 (0.686)	0.722 (0.754)	0.967 (0.967)	0.385 (0.393)	0.848 (0.85)
L2 on DWT	0.883 (0.937)	0.675 (0.688)	0.717 (0.751)	0.907 (0.907)	0.347 (0.373)	0.789 (0.792)
L2 on CHEBY	0.900 (0.938)	0.726 (0.718)	0.773 (0.797)	0.900 (0.9)	0.336 (0.357)	0.783 (0.787)
L2 on PAA	0.886 (0.931)	0.690 (0.698)	0.743 (0.768)	0.907 (0.907)	0.332 (0.357)	0.792 (0.795)
L2 on APCA	0.412 (0.438)	0.525 (0.541)	0.509 (0.54)	0.866 (0.867)	0.271 (0.279)	0.631 (0.643)
L2 on PL A	0.821 (0.866)	0.609 (0.615)	0.657 (0.673)	0.907 (0.907)	0.378 (0.38)	0.773 (0.776)
L2 on IPA (IPD)	0.900 (0.948)	0.720 (0.733)	0.769 (0.804)	0.953 (0.953)	0.381 (0.39)	0.797 (0.8)
DTW on PAA	0.886 (0.931)	0.688 (0.699)	0.723 (0.724)	0.920 (0.92)	0.267 (0.299)	0.804 (0.806)
DTW on APCA	0.511 (0.539)	0.629 (0.636)	0.684 (0.688)	0.920 (0.92)	0.267 (0.282)	0.754 (0.757)
DTW on PL A	0.822 (0.867)	0.538 (0.547)	0.678 (0.682)	0.853 (0.853)	0.366 (0.383)	0.783 (0.787)
DTW on IPA	0.875 (0.938)	0.618 (0.652)	0.978 (0.978)	0.940 (0.94)	0.375 (0.386)	0.771 (0.779)

Theorem 1 shows that the star coordinate transformation satisfies the lower bounding theorem, which provides the basis for the IP representation of the time series.

3.3. Representation of star coordinate series with IP curve

After obtaining the star coordinate series by transforming the time series, we can fit the star coordinate series using the IP curve. Currently, the state-of-the-art methods for fitting dataset with IP curve include 3L, Gradient and Min–Max, etc. Especially, Min–Max applies a linear least squares solution to the fitting problem, appear to have much better performance. Hence, in our work, we use Min–Max method to represent star coordinate series.

Fig. 3 (a), (b) and (c) show the fitting result of the star coordinate series in Fig. 2 (b) with IP curve of degree 4, 6 and 8 respectively. It is clear from Fig. 3 that the IP curve of degree 8 represents the star coordinate series most accurately, however, the IP curve of degree 4 and 6 have also a relatively good fit. Generally, we can choose the degree of IP curve according to accuracy requirement. In addition, the number of coefficients of the IP is important in dimensionality reduction. According to (4), the number of coefficients of 4th degree, 6th degree and 8th degree IP are 14, 27 and 44 respectively. Actually, the 6th degree IP curve with 27 coefficients can efficiently represent the time series whose length is 286 shown in Fig. 1. Clearly, ratio of dimensionality reduction is $27/286 \approx 0.31$, namely 31%, which exhibits the powerful representation of IP curve for the time series.

In order to evaluate the representation of IP curve, we define the distance between the IP curve and star coordinate series as representation error ε of IP curve, which can be written as follows.

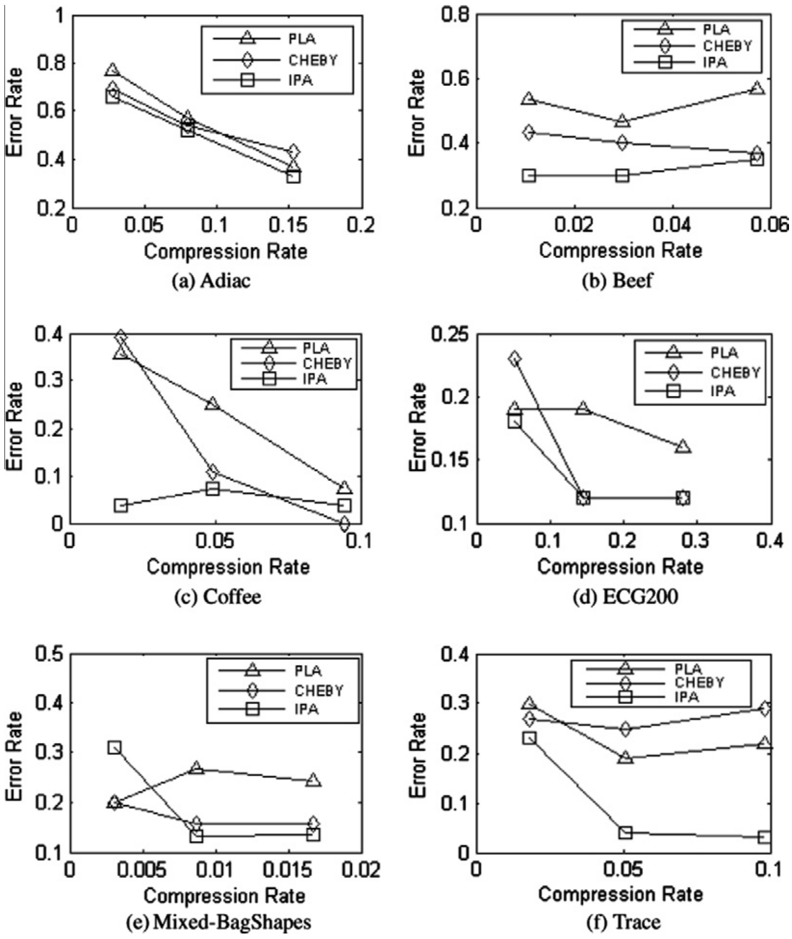


Fig. 5. Stability of representation comparison.

$$\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\frac{f^2(x_i,y_i)}{\|\nabla f(x_i,y_i)\|^2} \tag{8}$$

Fig. 4 (a) shows the representation results of IP curve of degree 4, 6 and 8 for the same time series shown in Fig. 1 respectively, which we can obtain by the following step: Firstly, fit the star coordinate series shown in Fig 2 (b) with IP curve of degree 4, 6 and 8 respectively. Secondly, obtain the zero sets of the three IP. Lastly, use the inverse transformation (6) to convert the three zero sets into three time series respectively. Fig. 4(b) shows the representation errors with Eq. (8) for IP curve of degree 4, 6 and 8 respectively. It is clear from the Fig. 4(b) that the representation error decreases rapidly as the degree of IP curve increases. Moreover, we can reach the same conclusion from the three time series of the IP curve representation in Fig. 4(a). In addition, although the degree of IP is very large, we cannot find the representation of IP curve suffers from the range oscillation phenomenon.

4. Similarity measure for IP coefficients

Given two time series T_1 and T_2, whose corresponding star coordinate series are denoted by S_1 and S_2, respectively. Then, these two star coordinate series are fitted by two IP curves whose coefficients vector are denoted by A and $\bar{A}$, respectively. The next task is to define a distance function between the two vectors. In our work, we adopt the Euclidean distance, because this distance function is simple and it is natural for many applications. It is also the distance function adopted by most studies on analyzing time series (Keogh and Chakrabarti, 2001; Keogh and Chakrabarti, 2002; Chen and Chen, 2007). Furthermore, for other more advanced distance function such as DTW, LCSS, ERP, we only apply them over the IP coefficients to compare with the ability of state-of-the-art representation methods.

Definition 2. Let $T_1$and T_2 be two time series of length N, and let A and $\bar{A}$ be the corresponding vectors of IP coefficients. Specifically, let $A^T = [a_1, a_2, ...a_m]$ and $\bar{A}^T = [b_1, b_2, \cdots b_m]$. Define:

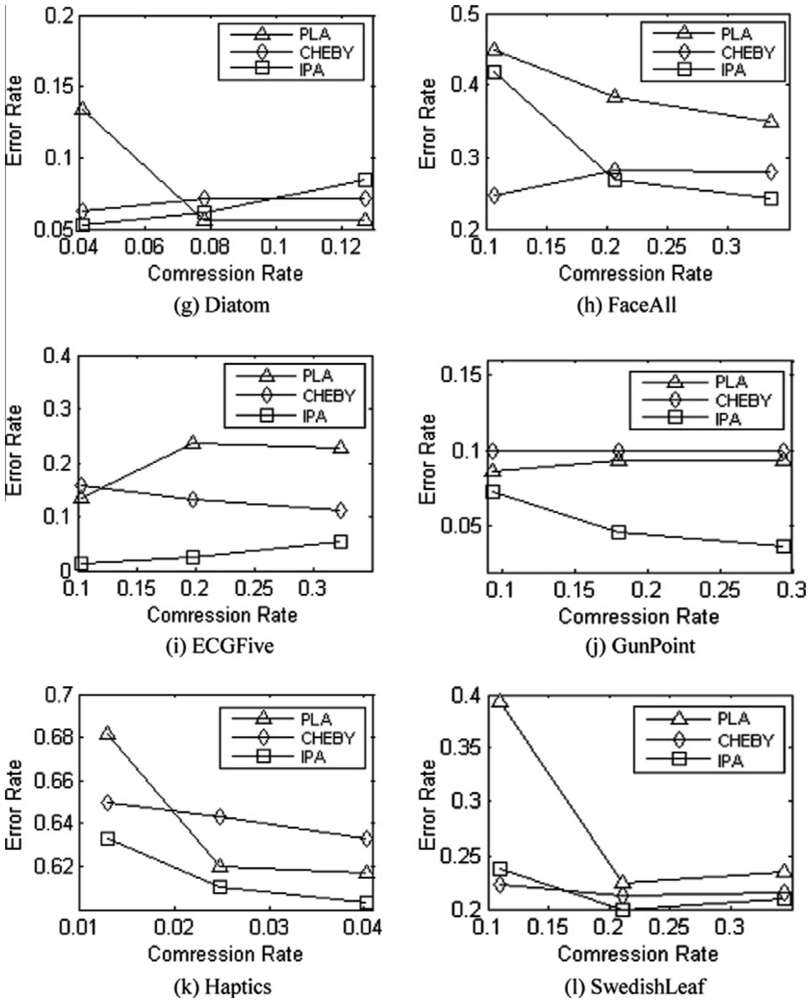
$$D_{IPD}(A^T, \bar{A}^T) = \sqrt{\sum_{i=1}^{m}(a_i - b_i)^2} \tag{9}$$

The distance function D_{IPD} called similarity measure based on IP curve (IPD) is a basic Euclidean distance function on the coefficients. It is clear that the distance function satisfies the three conditions: non-negativity, identity of indiscernible and triangle inequality.

Lemma 2. let $S = \{(x_i, y_i)|i = 1, 2, ...N\}$ and $\bar{S} = \{(\bar{x}_i, \bar{y}_i)|i = 1, 2, \cdots N\}$ be two star coordinate series of length N. Supposed that the point (x_i, y_i) in S and $(\bar{x}_i, \bar{y}_i)$ in $\bar{S}$ lie on the straight line $y = k_i x$, then we have

$$D_{euc}(S, \bar{S}) = \sqrt{\sum_{i=1}^{N}(1 + k_i^2)(x_i - \bar{x}_i)^2} \tag{10}$$

Proof. Computing the Euclidean distance between the two vectors S and $\bar{S}$, we have



(g) Diatom

(h) FaceAll

(i) ECGFive

(j) GunPoint

(k) Haptics

(l) SwedishLeaf

Fig. 5. (*continued*)

$$D_{euc}(S,\bar{S})=\sqrt{\sum_{i=1}^{N}[(x_i-\bar{x}_i)^2+(y_i-\bar{y}_i)^2]}$$

Substituting $y = k_ix$ for the y of the above equation, it follows Eq. (10). □

Lemma 3. Let $S = \{(x_i, y_i)|i = 1, 2, ...N\}$ and $\bar{S}=\{(\bar{x}_i,\bar{y}_i)|i=1,2,\cdots N\}$ be two star coordinate series of length N, which are represented by the two IP curves of degree n $f(x, y) = AX = 0$ and $\bar{f}(x,y)=\bar{A}X=0$, respectively. Denoting m by the number of coefficients of the IP curve, we have

$$D_{IPD}(A^T,\bar{A}^T)\leqslant D_{euc}(S,\bar{S}) \tag{11}$$

Proof. According to Eq. (4), $a_1, a_2, ..., a_{m-1}, a_m$ are coefficients of IP curve monomial $1,x,y,...x^n,x^{n-1}y,...,y^n$ respectively, where $m = (n + 1)(n + 2)/2$. Regarding x,y and a as variables of $f(x, y)$ and expanding $f(x, y, a)$ in the point (x_i, y_i, a_j) by a first order Taylor series approximation, it follows that

$$\begin{aligned}f(x_i+\Delta x_i,y_i+\Delta y_i,a_j+\Delta a_j)&=f(x_i,y_i,a_j)+f_x(x_i,y_i,a_j)\Delta x_i\\&\quad+f_y(x_i,y_i,a_j)\Delta y_i+f_a(x_i,y_i,a_j)\Delta a_j\end{aligned}$$

But $f(x_i,y_i,a_j)=0$, $f_a(x_i,y_i,a_j)=x_i^py_i^q$, where a_j is coefficient of the monomial $x_i^py_i^q$, so we can obtain

$$f_x(x_i,y_i,a_j)\Delta x_i+f_y(x_i,y_i,a_j)\Delta y_i+x_i^py_i^q\Delta a_j=0$$

Without loss of generality, suppose (x_i, y_i) lies on the straight line $y = k_ix$. Then we have

$$f_x(x_i,y_i,a_j)\Delta x_i+k_if_y(x_i,y_i,a_j)\Delta x_i+k_i^qx_i^{p+q}\Delta a_j=0$$

That is

$$[f_x(x_i,y_i,a_j)+k_if_y(x_i,y_i,a_j)]\Delta x_i=-k_i^qx_i^{p+q}\Delta a_j,$$

and hence

$$\frac{|\Delta a_j|}{|\Delta x_i|}=\frac{|f_x(x_i,y_i,a_j)+k_if_y(x_i,y_i,a_j)|}{|k_i^qx_i^{p+q}|}\leqslant\frac{|f_x(x_i,y_i,a_j)|+|k_i||f_y(x_i,y_i,a_j)|}{|k_i^qx_i^{p+q}|}.$$

By the change of x_i and y_i according to (5), it follows that

$$f_x(x_i,y_i,a_j)=-\frac{2\pi}{N}v_i\sin\left(\frac{2\pi}{N}t_i\right),\quad f_y(x_i,y_i,a_j)=\frac{2\pi}{N}v_i\cos\left(\frac{2\pi}{N}t_i\right).$$

From (6), we have $v_i=\sqrt{x_i^2+y_i^2}=\sqrt{1+k_i^2}x_i$. Hence, it follows that

$$\begin{aligned}\frac{|\Delta a_j|}{|\Delta x_i|}&=\frac{|\frac{2\pi}{N}v_i\sin(\frac{2\pi}{N}t_i)|+|k_i||\frac{2\pi}{N}v_i\cos(\frac{2\pi}{N}t_i)|}{|k_i^qx_i^{p+q}|}\\&\leqslant\frac{2\pi}{N}\frac{\sqrt{1+k_i^2}(1+|k_i|)|x_i|}{|k_i^qx_i^{p+q}|}\end{aligned}\tag{12}$$

Since the length of star coordinate series is very large, so N is much greater than m and 2π. Hence we may choose the point (x_i, -y_i),j = 1, 2, ..., m in S which corresponds to a_j, j = 1, 2, ..., m, such that $|x_i| \geqslant 1$, $|y_i| \geqslant 1$, $|y_i| \geqslant |x_i|$ for p = 0 or 1 and $|y_i| \leqslant |x_i|$ for q = 0 or 1. Consequently, we have

$$\frac{(1+|k_i|)|x_i|}{|k_i^q x_i^{p+q}|} = \frac{|x_i|+|y_i|}{|x_i^p y_i^q|} = \frac{1}{|x_i^{p-1} y_i^q|} + \frac{1}{|x_i^p y_i^{q-1}|} \leqslant 1.$$

Thus, from (12), it follows that

$$\frac{|\Delta a_j|}{|\Delta x_i|} \leqslant \frac{2\pi}{N} \frac{(1+|k_i|)|x_i|}{|k_i^q x_i^{p+q}|} \sqrt{1+k_i^2} \leqslant \sqrt{1+k_i^2}.$$

That is

$$|\Delta a_j|^2 \leqslant (1+k_i^2)|\Delta x_i|^2.$$

Furthermore, for j = 1, 2, ..., m, there exist m inequalities. Summing all these inequalities together, we have

$$\sum_{j=1}^{m} |\Delta a_j|^2 \leqslant \sum_{i=1}^{m} [(1+k_i^2)|\Delta x_i|^2] \leqslant \sum_{i=1}^{N} [(1+k_i^2)|\Delta x_i|^2].$$

Observe that if the two star coordinate series is similar, then Euclidean distance between them is very small. Hence, Euclidean distance between $f(x, y) = 0$ and $\bar{f}(x,y) = 0$ also is very small. Consequently, we may let $\Delta a_j = a_j - \bar{a}_j$ and $\Delta x_i = x_i - \bar{x}_i$. Then, it follows that

$$\sum_{j=1}^{m} |a_j - \bar{a}_j|^2 \leqslant \sum_{i=1}^{N} [(1+k_i^2)|x_i - \bar{x}_i|^2].$$

From (9) and (10), it follows that

$$D_{IPD}(A^T, \bar{A}^T) \leqslant D_{euc}(S, \bar{S}). \qquad \square$$

Theorem 2. Let T, $\hat{T}$ be two time series, and A^T, $\overset{\frown}{A^T}$ be the corresponding vectors of IP coefficients. Then:

$$D_{IPD}(A^T \overset{\frown}{A^T}) \leqslant D_{euc}(T, \hat{T})$$

Proof. Let S and $\hat{S}$ be star coordinate series transformed from T and $\hat{T}$ respectively. Then from Lemma 1, it follows that $D_{euc}(S, \hat{S}) \leqslant D_{euc}(T, \hat{T})$. Furthermore, from Lemma 3, it follows that $D_{IPD}(A^T A^T) \leqslant D_{euc}(S, \hat{S})$, and consequently $D_{IPD}(A^T A^T) \leqslant D_{euc}(T, \hat{T})$. □

5. Experimental evaluation

In this section, we illustrated the effectiveness of IPA through measuring its ability and competing it with other methods in supporting time series classification. Furthermore, we demonstrated the relation between error rate of classification and compression rate for IPA, and compared it with two state-of-the-art reduction techniques, PLA and CHEBY.

We conducted the experimental evaluation on many real data sets. Table 1 provided a summary of those reported here. In particular, Adiac, Beef, Coffee, ECG200, Trace, Diatom, FaceAll, ECGFive, GunPoint, Haptics and SwedishLeaf are available at http://www.cs.ucr.edu/∼eamonn/time_series_data/, whereas Mixed-Bag-Shapes can be found at http://www.cs.ucr.edu/∼eamonn/shape

This work focused on assessing the impact of the proposed time series representation model in similarity detection rather than finding the best strategy of time series classification. Hence, we conceived classification frameworks for time series data. Specifically, we used nearest neighbor classification to evaluate time series representation model. Ding and Trajcevsk, (2008) discuss the advantages with this approach.

5.1. Assessment criteria

In order to valuate the efficacy of dimensionality reduction methods, we used one nearest classifier (1NN) (Tan et al., 2006) on labeled time series data, each of which has a correct class label, and the classifier tries to predict the label as that of its nearest neighbor. In addition, Error rate, F-measure and accuracy are selected as assessment criteria. Error rate is the number of fault results of classification divided by the number of all test data. Clearly, the less error rate is, the better the result of classification. F-measure is a measure of a test's accuracy. It considers both the precision(P) and the recall(R) of the test to compute the measure. That is, F-measure is harmonic mean of precision and recall

$$F = \frac{2PR}{P+R}$$

Given a set T of time series of size L, let the expected organization of the series in T be $\{\alpha_1, ...\alpha_k\}$ and the output of a classification algorithm be $\{\beta_1, ...\beta_k\}$. The precision of β_j with respect to α_i is defined by $P_{ij} = |\beta_j \cap \alpha_i|/|\beta_j|$. The recall of β_j with respect to α_i is defined by $P_{ij} = |\beta_j \cap \alpha_i|/|\alpha_j|$. The overall of precision and recall are defined as

$$P = \frac{1}{k}\sum_{i=1}^{k} P_{ii}, R = \frac{1}{k}\sum_{i=1}^{k} R_{ii}$$

The F-measure can be interpreted as a weighted average of the precision and recall, which reaches its best value at 1 and worst score at 0. Accuracy (A) is the proportion of true results (both true positives and true negatives) in the population

$$A = 1 - \frac{1}{2L}\sum_{i=1}^{k} (|\beta_i \setminus \alpha_i| + |\alpha_i \setminus \beta_i|)$$

Note that all measures above range with zero and one. Unlike the error rate, higher values of F-measure and accuracy indicate better quality.

In order to access the stability of representation, we explored the relation between the error rate and compression rate (C). Specifically, compression rate is the dimension after dimensionality reduction (s) divided by the length of original time series (N), so that $C = s/N$. For example, we can use PAA to represent time series with the various dimensions respectively, such as 10, 20, 30. Clearly, higher dimensions can capture more detailed information time series representation can capture. Theoretically, as the compression rate (the dimensions after dimensionality reduction) increases, the error rate of classification decreases. In general, the error rates of classification with the most powerful representation method should be the lowest under the same compression rates among the various representation methods. Then, we consider this representation method as the most stable one.

5.2. Setup of the competing methods

To make comparative evaluation possible in term of error rate, F-measure and accuracy, we performed the three methods at levels of data compression that is close as possible. For instance, if we used IP curve of degree 6 (the number of its coefficients is 28) to represent time series dataset, we should set the parameters of PAA, APCA, DWT, DFT, PLA and CHEBY so that dimensions of their representations are also 28. For example, for PAA algorithm, we should choose 28 segments to represent the time series.

In stability experiment, for each time series dataset and algorithm, we varied the setting of each parameter of these methods in such way that it achieved the same compression (i.e. the number of segments). That means, for IPA, the numbers of coefficients of IP curve are 14, 28 and 44 corresponding to degree 4, 5 and 8 respectively. Then the parameters of competing methods should be adjusted so that their dimensions are 14, 28 and 44 respectively too.

5.3. Accuracy in time series classification

We compared IPA against state-of-the-art methods for modeling and comparing time series data, which include Euclidean distance (L2) and DTW as distance measures, and APCA, PAA, PLA, CHEBY, DWT and DFT as dimensionality reduction methods.

We assessed the performance of IPA and the competing methods using the 1-NN classification algorithm. Error rates, F-measure and accuracy obtained by the various methods are shown in Table 2 and Table 3 respectively. For the sake of fair comparison, the various methods used the same number of segments or coefficients of various datasets representation, which of Adiac, Beef, Coffee, ECG200, Mixed-BagShapes, Trace, Diatom, FaceAll, ECGFive, GunPoint, Haptics and SwedishLeaf are 28, 6, 28, 14, 14, 28, 14, 28, 28, 28, 28 and 28 respectively. From Table 2 and Table 3, L2 on DFT performed better than other methods on Adiac, Mixed-BagShapes, ECGFive, but not on Beef, Coffee, FaceAll, Trace and other datasets. Using L2 on CHEBY led to most results on Coffee, but not on the others. Among PAA, APCA and PLA, we can see that PLA is the best method because PLA represents datasets most accurately among these three methods. However, comparing PLA with IPA, we can find that the error rates of L2 on IPA in all datasets are less than ones of L2 on PLA. Specially, IPA led to quality improvements up to about 50% with respect to PLA, and up to about 60% with respect to DFT and DWT, which meant IPA was superior to piecewise approximation for representation of time series.

From the above discussion, We can see that L2 on IPA is the first ranked method in all the datasets, which illustrated that IPA was the best methods among the competing ones for representation of times series. We can reach the same conclusion about DTW on IPA. In addition, L2 on IPA and DTW on IPA always led to better results than L2 and DTW alone.

5.4. Stability of representation comparison

Fig 5 illustrated the stability of representation of PLA, CHEBY and IPA over the above twelve real datasets, where the *x*-axis is compression rate, and *y*-axis is error rate. In the experiments, For the three methods, as the compression rates increased, most of error rates decreased except PLA and IPA on Beef, PLA and CHEBY on Trace, but we can see that among all the competing methods, our method has the lowest error rate of classification at each datasets. For example, the error rates of IPA on Trace, GunPoint, ECGFive were far lower than ones of PLA and CHEBY. All the experiment results indicated that the proposed method had significantly higher performance for representation of time series than all the other competing methods.

From Fig 5, we observed the error rates of Beef for PLA and IPA, Trace for PLA and CHEBY, SwedishLeaf for PLA, CHEBY and IPA also increased as the compression rates increased. That is, increasing dimensions of various datasets representation may not have good result for representation of time series. We called the problem overfitting. The possibility of overfitting exists because representation methods of higher compression rate capture too much local information of time series. Hence, it is of importance to choose adaptive compression rate to represent time series. The determination of degree of IP curve discussed in Section 3.3 provided a method to solve the problem. Again, from most datasets in Fig 5 except Adiac, we can see the IP curve of degree 6 is good enough to represent the time series, and IP curve of higher degree cannot further improve the representation of time series. The fact illustrated the IPA was the most stable in competing methods.

6. Conclusion

In this paper, we explored how to apply IP curve to represent time series. IP curve representation enjoys the property that they approximate time series with least square, by which, they are easy to compute. In order for IP curve representation to be used for indexing, classification etc., we proved the Lower Bounding Lemma, and gave the definition of a distance function between two vectors of IP coefficients. We experimentally evaluated IPA in classification frameworks and compared it with state-of-the-art dimensionality reduction methods. Experiments show the efficiency of IPA in time series representation.

Generally, star coordinate is used to visualize to multidimensional data in data mining. However, in our work, star coordinate series is regarded as boundary points dataset of planar region. Therefore, the IP curve can represent it. Actually, there are many methods in the field of image processing to study boundary points dataset. We think analysis of star coordinate series is new direction in the study of time series.

We plan to extend the Lower Bounding Lemma to other distance functions, such as DTW, LCSS, Mahalanobis distance etc. In addition, we would like to study the feasibility of applying implicit polynomial surface to multivariate time series.

Acknowledgments

This work was supported in part by Natural Science Foundation of Jiangsu Province China under Grant BK2009349, Grant BK2009348 and the Natural Science Foundation of the Jiangsu Higher Education Institutions under Grant 09KJB520005.

References

Ahmet Yasin Yazicioglu, BerkCalli, Mustafa Unel, 2009. Image based visual servoing using algebraic curves applied to shape alignment. Proc. IEEE Int'l Conf. On Intell. Robots and Systems, pp. 5444–5449.

Blane, M.M., Lei, Z., 2000. The 3L algorithm for fitting implicit polynomial curves and surfaces to data. IEEE Trans. Pattern Anal. Machine Intell. 22 (3), 298–313.

Bo Zheng, Ryo Ishikawa Oishi T., Takamatsu J. Ikeuchi K., 2009. A fast registration method using IP and its application to ultrasound image registration. IPSJ Trans. On Comput. Vision and Application, 1, pp. 209–219.

Cai Y., Ng R.T., 2004. Indexing spatio-temporal trajectories with chebyshev polynomials. In SIGMOD Conf., pp. 599–610.

Chen L., Ng R.T., 2004. On the marriage of Lp-norms and edit distance. In Proc. of the Thirtieth Internat. Conf. on VLDB, pp. 792–802.

Chen L., 2005. Robust and fast similarity search for moving object trajectories. In Proc. of the ACM SIGMOD Internat. Conf. on Management of Data, pp. 491–502.

Chen Q., Chen L., 2007. Indexable PLA for efficient similarity search. In Proc. of the Thirty Third Internat. Conf. on VLDB, pp. 435–446.

Chen Y., Nascimento M.A., 2007. SpADe: On shape based pattern detection in streaming time series. In IEEE Twenty Third Internat. Conf. on Data, Eng., pp. 786–795.

Fu, T.C., Chung, F.L., 2008. Representing financial time series based on data point importance. Eng. Appl. Artif. Intell. 21 (2), 277–300.

Fu, T.C., Chung, F.L., 2011. A review on time series data mining. Eng. Appl. Artif. Intell. 24, 164–181.

Faloutsos C., Ranganahan M., 1994. Fast subsequence matching in time series databases. In SIGMOD Conf., pp. 419–429.

Gullo, F., Ponti, G., 2009. A time series representation model for accurate and fast similarity detection. Pattern Recognition 42 (11), 2998–3014.

Hui Ding, Goce Trajcevsk., 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. In Proc. of the VLDB Endowment, pp. 1542–1551.

Helzer A., Bar Zohar M., Malah D., 2000. Using implicit polynomials for image compression. In Proc. of Twenty First IEEE Convention of the Electrical and Electronic Eng., pp. 384–388.

Heizer, A., Barzohar, M., malah, D., 2004. Stable fitting of 2D curves and 3D surfaces by implicit polynomials. IEEE Trans. Pattern Anal. Machine Intell. 26 (10), 1283–1294.

Aßfalg, Johannes., Kriegel, Hans.Peter., 2006. Similarity search on time series based on threshold queries. Adv. Database Technol. 3869, 276–294.

Kautsky, Jaroslav., Flusser, Jan., 2007. Implicit invariants and object recognition. Digital Image Comput. Techn. Appl. 3, 462–469.

Jiang Xiaoqian, Xu Wanhong, 2007. 2D image database indexing: a coefficient-based approach. In proc. of IEEE ICME, pp. 2210–2213.

Korn F., Jagadish H. V., Faloutsos., 1997. Efficiently supporting Ad Hoc queries in large datasets of time sequences. In Proc. of ACM SIGMOD on management of data, pp. 510–535.

Keogh, E., Chakrabarti, K., 2001. Dimensionality reduction for fast similarity search in large time series databases. Knowl. Inf. System 3 (3), 263–286.

Keogh, E., Chakrabarti, K., 2002. Locally adaptive dimensionality reduction for indexing large time. ACM Trans. Database Systems 27 (2), 188–228.

Keogh, E., Ratanamahatana, C.A., 2005. Exact indexing of dynamic time warping. Knowl. Inf. System 7 (3), 358–386.

Lin, J., Keogh, E., 2007. Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Disc. 15 (2), 107–144.

LI Dao lun, ZHA Wen shu, Lu De tang, 2011. Implicit interpolation of time vector series. in proc. of 2011 Seventh Internat. Conf. on Natural Computation, pp. 151–155.

Marola, G., 2005. A technique for finding the symmetry axes of implicit polynomial curves under perspective projection. IEEE Trans. Pattern Anal. Machine Intell. 27 (3), 465–470.

Morse M.D., Patel J.M., 2007. An efficient and accurate method for evaluating time series similarity. In Proc. of ACM SIGMOD Internat. Conf. on Management of Data, pp. 569–580.

Oden, C., Ercil, A., Yildiz, V.T., Kirmiztia, H., Buke, B., 2001. Hand recognition using implicit polynomials and geometric features. Springer Lecture Notes Comput. Sci. 2091, 336–341.

Pong Chan K., Fu A.W., 1999. Efficient time series matching by wavelets. Fifteenth Internat. Conf. on Data Eng., pp. 126–133.

Pang Ning Tan, Michael Steinbach, Vipin Kumar., 2006. Introduction to data mining, Addison Wesley.

Lebmeir, Peter., Richter Gebert, Jurgen, 2008. Rotations translations and symmetry detection for complexified curves. Comput. Aided Geometric Design 25, 707–719.

Subrahmonia, J., Cooper, D., Keren, D., 1996. Practical reliable Bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants. IEEE Trans. Pattern Anal. Machine Intell. 18, 505–519.

Taubin, G., Cukirman, F., Sullivan, S., 1994. Parameterized families of polynomials for bounded algebraic curve and surface fitting. IEEE Trans. Pattern Anal. Machine Intell. 16 (3), 286–303.

Tarel, J.P., Cooper, D.B., 2000. The complex representation of algebraic curves and its simple exploitation for pose estimation and invariant recognition. IEEE Trans. Pattern Anal. Machine Intell. 22 (7), 663–674.

Tasdizen T., Cooper D.B., 2000. Boundary estimation from intensity color images with algebraic curve models. In Proc. of Fifteenth Internat. Conf. on, Pattern Recognition, pp. 225–228.

Tasdizen, T., Tarel, J.P., 2000. Improving the stability of algebraic curves for application. IEEE Trans. Image Process. 9 (3), 405–416.

Vlachos M., Gunopulos D., Kollios G., 2002. Discovering similar multidimensional trajectories. In Eighteenth Internat. Conf. on Data, Eng., pp. 673–684.

Wu, Gang, Li, Daolun, 2002. Object representation and symmetry detection based on implicit polynomial curves. J. Comput. Res. Dev. 39 (10), 1337–1344.

Wu Gang, Li Dao lun., 2004. Object recognition based on affined invariants in implicit polynomial curves. Acta Electronica Sinica., 32 (12), 1987–1991.

Wu, Gang, 2007. Research on degree of fitting implicit polynomial curves and surfaces. J. Comput. Res. Dev. 44 (1), 148–153.