Hongyi Cao,
Junhui Yang,
Li Wang

Xi'an Medical University, Xi'an 710021, Shaanxi, China

FAST TIME SEQUENCE DATA MINING ALGORITHM BASED ON GREY SYSTEM THEORY

Хуні Цао,
Цзюньхой Ян,
Лі Ван

Сіаньський медичний університет, м. Сіань, Шеньсі, Китай

ШВИДКИЙ АЛГОРИТМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ЧАСОВИХ РЯДІВ НА ОСНОВІ ТЕОРІЇ СІРИХ СИСТЕМ

Purpose. With the development of the big data technology, the time sequence data mining has become a hot spot that attracts the attention of the public. Based on the correlation and cooperativity of the time sequence data, we propose the fast time sequence data mining model based on the grey system theory.

Methodology. The correlation determination method that is based on the features of the relevant coefficient of the time shift sequence is obtained. As a result, a kind of fast time sequence data mining model based on the grey system theory is proposed.

Findings. The correlation determination methodology proposed in this paper is more effective than the Pearson linear correlation coefficient, Spearman rank correlation coefficient, Kendall rank correlation coefficient and Granger causality test.

Originality. In this paper, the double sequence fast correlation determination method and curve alignment method are provided. So far, we have not found other literature on the related research.

Practical value. The research results can provide theoretical basis for the determination of the correlation of regression analysis and the time alignment.

Keywords: *grey system theory, time warping, correlation, time sequence data, data mining, curve alignment, causality test*

Introduction. Time series data is one of the most commonly seen types of data in the data mining, which is applied in many fields such as the monthly volume of runoff of a certain river, the local mean monthly temperature and precipitation, our country's consumer price index (CPI) and the gross domestic product (GDP), the earthquake wave sequence acquired at multiple observation points at the time of the occurrence of the earthquake, etc. [1]. Through the analysis on the time sequence data, some useful conclusions can be drawn, for example: Through studying the history of flow volume of the river, the temperature, precipitation and other characteristics, the forecast level for flood can be effectively improved; through the application of CPI and GDP, the degree of inflation and the momentum of economic development of the country or region can be analysed; according to the multiple seismic wave sequence, the seismic source and seismic magnitude, etc. can be accurately located [2]. Some of the non-time sequence data can also be handled through the conversion and application of the time sequence data mining method for processing, for example: Through the application of the distance from the edge of leaf to the center of the mass to describe its characteristics from different angles, a column of data can be acquired, which then can further distinguish the type of the leaf [3].

In the process of time sequence data mining, if the time difference of the data is not taken into consideration, it is easy to be influenced by intuition or prejudice, and thus cause wrong determination on the correlation; however, it does not make sense if the time difference of the related time sequence is not taken into account [4]. That is to say, in the assessment of the correlation of the sequence, it does not only require considering the time difference, but also requires that the data is correlated, therefore, the correlation and time difference between the sequences are mutually restrained [5]. At present, the correlation analysis on the time sequence data is faced with some problems, for example, the data relationship is relatively complex, the data contains noise, there is missing data or abnormal data, etc. [6]. Homogeneous data (data from the same source or with the same attributes, such as the earthquake seismic data of the same earthquake acquired in multiple locations) has natural similarity, which does not require determining the correlation, and there is no correlation or time difference constraint problem either, hence mostly applied for the classification or clustering. For the heterogeneous data (data from different sources or with different attributes, such as the amount of precipitation and river runoff volume, CPI and GDP), its correlation shall be determined, and its relevance and regression analysis etc. shall be conducted [7–9]. Therefore, the main object of the time sequence data correlation analysis is the heterogeneous data.

Similar to the binary classification problems or the two types of errors in hypothesis testing, in correlation with heterogeneous data mining or data regression to two types of errors: 1) it is believed that the related data does not have correlation; 2) it is believed that the irrelevant data has correlation and regression analysis is conducted [10]. The former often appears in actual application, such as the elevation of the sun and the ground temperature, precipitation and river runoff, if in accordance with the time control study the correlation of two groups of data may not be able to get relevant conclusions, but in fact, if both will be in time for translation, there will be close correlation. The above two types of errors can occur, for example, in the past 20 years China's GDP and the height of a person can grow before the age of 20 certainlyhas significant positive correlation, and this is meaningless, illogical regression, which is called nonsense regression or spurious regression. So, prior to the analysis of the data, if not considering the correlation of data, class 1 correlation error will cause the waste of potential information, data category 2 correlation errors may cause misleading for subsequent analysis. The correlation of multiple sets of data can be provided through the correlation of two groups of data.

Homogeneous data has natural similarity and a curve line can be developed. Constrained for heterogeneous time sequence data correlation and the time difference problem, based on the theory of grey system, fixed time determine the move sequence correlation, on the basis of serial correlation, and then through the curve line refining time function. When making correlation judgment on heterogeneous data, on the one hand, there is a deviation due to the sample correlation coefficient and the overall correlation coefficient; it studies the upper and lower bounds of the overall correlation coefficient. On the other hand, in order to prevent the two kinds of correlation mistakes, starting from the main causes, the study of the characteristics of two kinds of correlation error and corresponding correlation judgment method is put forward. Applicable to the curve of the heterogeneous data, line method is also applicable to homogeneous data, but what applies to homogeneous data (such as AISE) does not apply to heterogeneous data (dimension is not unified, and negative correlation, etc.). Therefore, the maximum correlation coefficient (absolute value) curve standard mainly based on the characteristics of heterogeneous data is put forward, and the GQS algorithm is applied for the solution.

Basic concept. ***Grey Theory Class Set.*** Customer confidence in businesses depends on many factors, "complete trust", "somewhat trust", "general trust" and "distrust" and other information can exactly describe the state. Therefore, we introduce a grey element, the concept of grey number, ash content and ash. Ash element refers to the incomplete information elements, grey number refers to the volume of incomplete information, grey variables refer to incomplete information, and grey variables of a specific value constitute a grey class. For example, users who evaluate the quality of products constitute a grey element; product quality assessment values, such as about 0.40 points, constitutes a grey number. All the sets of gray classes are called grey class set, noted as $G = \{g_k | k = 1, 2, ..., r\}$, in which g_k is the k-th grey class.

For example, assume that the product quality is the grey variable, which can be "good", "general" and "poor", etc. Let the grey class set $G = \{g_1, g_2, g_3\}$, which, respectively, represents the first, second, third class, in turn, showing good, general, poor product quality in turn.

Grey class whitening function, weight matrix. Let us set clustering entity set $D = \{d_i | i = 1, 2, ..., m\}$, the key attributes $A = \{a_h | h = 1, 2, ..., e\}$, grey class set $G = \{g_k | k = 1, 2, ..., r\}$, $T_i(a_h)$ represents the entity d_i's key attribute a_h's score value, the monotonic function f_{hk} is defined as the grey class whitening function, as shown in Fig. 1, $E(\lambda_{hk}, 1)$ is the turning point, $f_{hk}(T_i(a_h)) = 0.81$ represents the rating value. $T_i(a_h)$ belongs to grey class with the possibility of 0.81.

Weight matrix is defined as W, as shown in Fig. 2, the matrix elements

$$W_{jk} = \lambda_{hk}/(\lambda_{1k} + \lambda_{2k} + ... + \lambda_{hk} + ... + \lambda_{ek});$$

$$1 \le i \le m, \quad 1 \le h \le e, \quad 1 \le k \le r;$$

$$W = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1k} & \cdots & w_{1r} \\ w_{21} & w_{22} & \cdots & w_{2k} & \cdots & w_{2r} \\ \vdots & \vdots & & \vdots & & \vdots \\ w_{h1} & w_{h2} & \cdots & w_{hk} & \cdots & w_{hr} \\ \vdots & \vdots & & \vdots & & \vdots \\ w_{e1} & w_{e2} & \cdots & w_{ek} & \cdots & w_{er} \end{pmatrix}.$$

Clustering vector. W represents the weight matrix, the whitening matrix of the entity d_i is F_i, F_i and W are noted as column matrix respectively

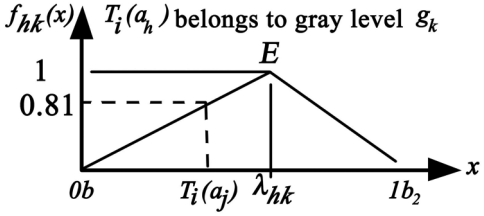


Fig. 1. Grey class whitening function

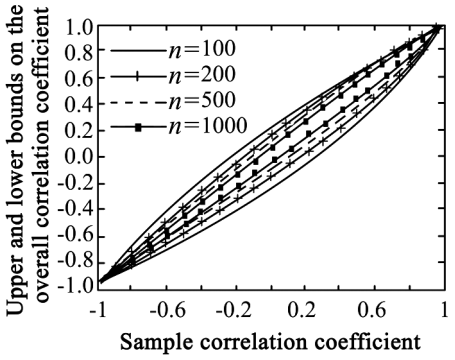


Fig. 2. Overall Correlation Coefficient Bounds (*significance level* $\alpha = 0.05$)

$$F_i = (\alpha_1, \alpha_2, ..., \alpha_k, ..., \alpha_r),$$

in which

$$a_k = (f_{1k}(T_i(a_1)), f_{2k}(T_i(a_2)), ..., f_{nk}(T_i(a_h)), ...;$$
$$f_{ek}(T_i(a_e))); \quad W_i = (\beta_1, \beta_2, ..., \beta_k, ..., \beta_r),$$

in which

$$\beta_k = (W_{1k}, W_{2k}, ..., W_{hk}, ..., W_{ek}).$$

The clustering vector of the entity d_i is defined as σ_i,

$$\sigma_i = (\sigma_{i1}, \sigma_{i2}, ..., \sigma_{ik}, ..., \sigma_{ir}) =$$
$$= (\alpha_1 \times \beta_1, \alpha_2 \times \beta_2, ..., \alpha_k \times \beta_k, ..., \alpha_r \times \beta_r),$$

in which

$$\sigma_{ik} = \alpha_k \times \beta_k = (f_{1k}(T_i(a_1)), f_{2k}(T_i(a_2)), ..., f_{hk}(T_i(a_h)), ...;$$
$$f_{ek}(T_i(a_e))) \times (w_{1k}, w_{2k}, ..., w_{hk}, ..., w_{ek}), \quad k = 1, 2, ..., r.$$

Curve alignment relevant mining method. As in the solution for the actual problem, only sample data can be obtained. When using the sample, the overall estimation may be biased, as a result, this paper uses the sample correlation coefficient to infer the overall correlation coefficient with certain level of significance on the boundary. At the same time, in order to prevent the two types of error, this paper studies the correlation coefficient of two kinds of error under the move sequence characteristics to rule out two types of error correlation accordingly. From the above two aspects, the correlation determination method of the two sets of time sequence data can be obtained.

Correlation determination to the related sequence with the time warping. In order to determine the serial correlation, it is necessary to obtain the upper and lower bounds of the general correlation coefficient. In this paper, through the two asymptotic distributions of the sample correlation coefficients, the upper and lower bounds of the overall correlation coefficient at a certain significance level are obtained, and combined with the characteristics of the correlation errors of the first type, the method of the correlation determination of the correlated sequence with the time warping is achieved.

Bounds of the correlation coefficient. Pearson correlation coefficient is the most commonly used when measuring serial correlation measure. If there are two sets of corresponding data $\{(x_i, y_i), i = 1, 2, ..., n\}$ (n is the quantity of the samples) which is from the bivariate normal overall sample $(x, y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, the sample correlation coefficient is as follows

$$\hat{\rho}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$
$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}. \quad (1)$$

In which, $\bar{x}$, $\bar{y}$ are the sample mean of X, Y respectively.

Sample correlation coefficient $\hat{\rho}(X,Y)$ can be used as the correlation coefficient ρ of two normal population (X, Y) of unbiased estimator and consistent, but the correlation coefficient has an obvious shortcoming, namely the degree that it is close to 1 is related to the number of n sets of data, it is easy to send a kind of illusion. When n is small, the absolute value of the correlation coefficient of some samples is close to 1, and when $n = 2$, the absolute value of correlation coefficient is 1. When n is bigger, the absolute value of correlation coefficient is smaller. There are many scholars who have obtained the distribution results about the sample correlation coefficient, sample size and bivariate normal of overall correlation coefficient.

For binary normal population (X, Y) and under the assumption $\rho = 0$, the following distribution is obtained

$$T = \frac{\sqrt{n-2}\hat{\rho}}{\sqrt{n-\hat{\rho}^2}} \sim t(n-2). \quad (2)$$

When $\rho = \rho_0$, Fisher gives a relatively complex $\hat{\rho}$ probability density function, after proper transform, the asymptotic distribution can be obtained

$$z = \frac{\varphi(\hat{\rho}) - \varphi(\rho)^{n\to\infty}}{2\sqrt{n-3}} \sim N(0.1), \quad (3)$$

where $\varphi(x) = \ln\frac{1+x}{1-x}$. When the sample size is large, the overall correlation coefficient can be estimated from the sample correlation coefficient.

The literature proves that in binary normal population, the extraction of n samples and the asymptotic distribution can be obtained.

$$\sqrt{n}(\hat{\rho} - \rho)^{n\to\infty} \sim N\left(0, (1-\rho^2)^2\right),$$
$$\text{that is } \frac{\sqrt{n}(\hat{\rho} - \rho)^{n\to\infty}}{(1-\rho^2)} \sim N(0.1). \quad (4)$$

This paper has estimated the overall correlation coefficient based on the above two asymptotic distributions.

As the $\varphi(x)$ in Formula (3) is monotonically increasing, it can be known that:

- When $\rho \geq \hat{\rho}$

$$P\left\{\rho \leq \varphi^{-1}\left[\varphi(\hat{\rho}) + 2z_{1-\frac{a}{2}} \cdot \sqrt{n-3}\right]\right\} = 1 - a. \quad (5)$$

- When $\rho \leq \hat{\rho}$

$$P\left\{\rho \geq \varphi^{-1}\left[\varphi(\hat{\rho}) - 2z_{1-\frac{a}{2}} \cdot \sqrt{n-3}\right]\right\} = 1 - a, \quad (6)$$

in which, $\varphi^{-1}(x)=\ln\frac{e^x-1}{e^x+1}$; Z_α is the α quantile of standard normal distribution, namely, $P(x \le Z_\alpha) = \alpha$; Random variable $x \sim N(0, 1)$.

In this paper, on the basis of Formula (4) further inference of the bounds of the overall correlation coefficient can be obtained, namely:

- When $\rho \ge \hat{\rho}$

$$P\left\{\frac{-\sqrt{n}-\sqrt{n+4\sqrt{n}z_{1-\frac{a}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{a}{2}}}}{2z_{1-\frac{a}{2}}}\le\rho\le\right.$$
$$\left.\le\frac{-\sqrt{n}+\sqrt{n+4\sqrt{n}z_{1-\frac{a}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{a}{2}}}}{2z_{1-\frac{a}{2}}}\right\}=1-a. \quad (7)$$

- When $\rho \le \hat{\rho}$

$$P\left\{\frac{\sqrt{n}-\sqrt{n-4\sqrt{n}z_{1-\frac{a}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{a}{2}}}}{2z_{1-\frac{a}{2}}}\le\rho\le\right.$$
$$\left.\le\frac{\sqrt{n}+\sqrt{n-4\sqrt{n}z_{1-\frac{a}{2}}\cdot\hat{\rho}+4z^2_{1-\frac{a}{2}}}}{2z_{1-\frac{a}{2}}}\right\}=1-a. \quad (8)$$

Integrate Formulas (5–8), when $\alpha = 0.05$, the approximation can be obtained:

- When $\rho \ge \hat{\rho}$

$$\inf_{\alpha=0.05}\rho=\max\left\{\frac{-\sqrt{n}-\sqrt{n+8\sqrt{n}\cdot\hat{\rho}+16}}{4},\hat{\rho},-1\right\}; \quad (9)$$

$$\sup_{\alpha=0.05}\rho=$$
$$=\min\left\{\frac{-\sqrt{n}-\sqrt{n+8\sqrt{n}\cdot\hat{\rho}+16}}{4},\varphi^{-1}\left[\varphi\left(\hat{p}\right)+4\sqrt{n-3}\right],1\right\}. \quad (10)$$

- When $\rho \le \hat{\rho}$

$$\inf_{\alpha=0.05}\rho=\max\left\{\frac{\sqrt{n}-\sqrt{n-8\sqrt{n}\cdot\hat{\rho}+16}}{4},\right.$$
$$\left.\varphi^{-1}\left[\varphi\left(\hat{p}\right)-4\sqrt{n-3}\right],-1\right\}; \quad (11)$$

$$\sup_{\alpha=0.05}\rho=\min\left\{\frac{\sqrt{n}+\sqrt{n-8\sqrt{n}\cdot\hat{\rho}+16}}{4},\hat{\rho},1\right\}. \quad (12)$$

Fig. 2 provides the upper and lower bounds of the general correlation coefficient under different sample size and sample correlation coefficient. As can be seen from the figure, in this paper, the upper and lower bounds of the curve provided are of the following features:

1. The larger the sample size is, the more compact the upper and lower bounds are.

2. With the same sample size, the upper and lower curve centres are in symmetry.

3. The larger the absolute value of the correlation coefficient is, the more compact the upper and lower bounds are.

The above features can easily be proved by formulas (9–12).

Correlation determination method. For the convenience of the description of the relevant characteristics of the sequence, the definition of the time shift sequence (time-lag series) is first given.

Assume two sequences $(X, Y) = \{(x_i, y_i), i = 1, 2, ..., n\}$, the following time shift sequences are defined

$$\begin{gathered}(X_t,Y_{t+m})=\left\{\left(x_i,\mathrm{y}_{i+m}\right),i=1,2,...,n-m\right\},\\ 1\le m<n,m\in N^+;\\ (X_t,Y_{t-m})=\left\{\left(x_i,\mathrm{y}_{i-m}\right),i=m+1,2,...,n\right\},\\ 1\le m<n,m\in N^+.\end{gathered} \quad (13)$$

On the first type of regression errors, start the consideration of the initial sequence directly, and their relevance must be relatively small. If considering the moving time sequence correlation, there must be $m_0(1 \le |m_0| \ll n, |m_0| \in N^+)$, which makes the correlation coefficient relatively large.

The resulting time sequence correlation methods are: the sequence correlation coefficient changes with m, and achieves the maximum at m_0, namely, the graph $\rho=\hat{\rho}(m)$ presents obvious convex phenomenon, according to the formulas (9–12) it can estimate the range of overall correlation coefficient. If $|\rho(m_0)| > \rho_0$ (i. e., more than the given threshold value, such as 0.6), it is believed that the time shift sequence $\left(X_t,Y_{t+m_0}\right)$ has correlation, and the curve alignment and regression analysis, etc. can be conducted.

Curve alignment method based on grey theory. According to the correlation coefficient analysis of the time-lag series, we can determine whether there is correlation existing between the series. In the presence of correlation between two series but with the time bias, through the curve alignment method, they can be aligned so as to eliminate the difference on the phase time axis. For heterogeneous data, when AISE criterion is used, the result will change with the dimensional change and, therefore, it is necessary to propose a dimensionless criterion to align the heterogeneous data.

Curve alignment method based on grey theory. Pearson correlation coefficient is actually a dimensionless measure, with the inner product to represent the connectivity of the continuity function, at the same time, to achieve the inner product value in line with the norm; it is then divided by two functions of the norm. The alignment criteria of curves composed of heterogeneous data, in fact, can be constructed through the function correlation coefficient

$$\max_{h(t)}\left|\rho\left(x_1^*, x_2\right)\right| = \max_{h(t)}\left|\frac{\int_T x_1^*(s)x_2(s)ds}{\sqrt{\int_T\left[x_1^*(s)\right]^2 ds}\sqrt{\int_T\left[x_2(s)\right]^2 ds}}\right|, \quad (14)$$

where $x_1^*(t) = x_1\left[h(t)\right]$ represents the function after alignment. As the representation is too complicated, this paper gives the corresponding discretization state.

Let us assume that the two functional data $x_1(t)$ and $x_2(t)$ at the sampling time point $T = (t_1, t_2, ..., t_n)$ have the sample sequence

$$x_1(T) = [x_1(t_1), x_1(t_2), ..., x_1(t_n)],$$

and

$$x_2(T) = [x_2(t_1), x_2(t_2), ..., x_2(t_n)],$$

the alignment of function $x_1(t)$ in contrast to curve $x_2(t)$ is to be performed.

Let $\Delta = (\delta_1, \delta_2, ..., \delta_n)$ be the offset of $x_1(t)$ at the time point T relative to $x_2(t)$, namely, the time curve function shall meet $h(T) = T + \Delta$, then the temporal samples are transformed into $x_1(T + \Delta) = [x_1(t_1 + \delta_1), ..., x_1(t_n + \delta_n)]$ after alignment, the two groups of functional data sequence of samples will have high correlation. The curve alignment problem can be converted into solving the following

$$\max_{\Delta}\left|\rho\left[x_1(T+\Delta)x_2(T)\right]\right|. \quad (15)$$

General time warping function features the consistent monotonicity, namely, to meet $t_{i-1} + \delta_{i-1} < t_i + \delta_i < t_{i+1} + \delta_{i+1}$. However, the offset vector which is chaotic at bending function does not meet the same time and will make $bndl_i = t_{i-1} + \delta_{i-1}^k - t_i$, $bndr_i = t_{i+1} + \delta_{i+1}^k - t_i$, δ_i^k represents the value of δ_i at the k-th iteration. In specific implementation, the search interval of δ_i^{k+1} can be narrowed down as the closed interval $\left[bndl_i + p\cdot(bndr_i - bndl_i), bndr_i - p\cdot(bndr_i - bndl_i)\right]$, in which, p is the constant within (0.05).

Finally, the problem of curve alignment is transformed into solving the following constrained optimization problem

$$\begin{cases} \Delta^* = \arg\max_{\Delta}\left|\rho\left[x_1(T+\Delta), x_2(T)\right]\right| \\ s.t.\ \delta_i \in \left[\begin{array}{l} bndl_i + p\cdot(bndr_i - bndl_i), \\ bndr_i - p\cdot(bndr_i - bndl_i) \end{array}\right]. \end{cases} \quad (16)$$

At last, the time offset vector Δ^* is transformed into functional form, and the time offset function $d(t)$ is obtained, the corresponding time for bending function is $h(t) = d(t) + t$.

Model solution — gqs algorithm. When Parameter has high dimension, it is hard to solve the maximization problem of function Q. In order to overcome this problem, in this paper, the problem of the objective functionis taken as function Q (i.e., the expectation of the logarithm likelihood function in the EM algorithm), the extended EM algorithm (generalized maximum expectation maximization algorithm (GEM)) is applied to solve the problem. Due to the good time-smoothness of the time warping function, every time the update on Δ^* is performed, spline smoothing is performed on P for one time, thus the smoothness of the time difference vector can be increased, but the spline has regular term, which can prevent the problem of time difference function instability caused by overoptimization. Therefore, the smooth generalized expectation maximization method (S-GEM) for solving the model is obtained. The solving steps are as follows:

Input: Two sets of correlated time sequence data TS_1 and TS_2 with the time warping on time $T_0 = (t_{01}, t_{02}, ..., t_{0m})$;

Step 1: Initialization time vector $\Delta_0 = zeros(1, n)$, error tolerance for iteration is eps.

Step 2: Time sequence data functions TS_1, TS_2 can be converted into function type data $x_1(t)$ and $x_2(t)$, take n points $T = (t_1, t_2, ..., t_n)$ in T_0 uniformly, the smooth sequence $\{x_1(t_i)\}$ and $\{x_2(t_i)\}$ $(i = 1, 2, ..., n)$ can be obtained, in which $t_1 = t_{01}$, $t_n = t_{0m}$.

Step 3: Using generalized expectation maximization for the time difference vector. Record the k-th iteration time difference vector as $\Delta^k = \left(\delta_1^k, \delta_2^k, ..., \delta_n^k\right)$, perform $n - 2$ times of conditional maximization (assuming starting point without time difference, namely $\delta_1^k = \delta_n^k = 0$), $\Delta^{k+1} = \left(\delta_1^{k+1}, \delta_2^{k+1}, ..., \delta_n^{k+1}\right)$ can be obtained.

Step 4: Smooth processing of the time difference vector: Using P spline function fitting on sequences Δ^{k+1} of function $d^{k+1}(t)$, and use the fitting value to replace the original value, namely: $\Delta^{k+1} = d^{k+1}(t_i)$ $i = 1, 2, ..., n$.

Step 5: Repeat Step 3 and Step 4 until convergence $(|\Delta^{k+1} - \Delta_k| < eps)$.

Output: Time difference function $d(t) = d^K(t)$ or time warping function $(|\Delta^{k+1} - \Delta_k| < eps)$.

Similar to many functional data mining methods, the proposed algorithm can deal with large volume of data line problems, even if there is missing data or abnormal data, the current information can still be fully used; in addition, through smooth processing, the algorithm can quickly converge to the extreme. It is important that the running time or time complexity of the algorithm mainly depends on the number of sampling, and has nothing to do with the number of the original sample. GQS algorithm complexity analysis is shown in Table 1, where m is the initial sample size, n is the uniform sampling volume, d is functional order (highest number minus 1), *fm* is the condition of maximizing the average time complexity in *Step 3*, k is the total number of iterations. As *fm* and k is related to the accuracy, parameters, and the problem itself, it is set out separately. P spline is applied when the data is functional or smooth, and the most complicated part in the coefficient vector estimation of the spline function obtains the inversion in the spline basis function matrix, and the time complexity and space complexity are 3 and 2 times of the number of the base function,

Table 1

GQS Algorithm Complexity Analysis

Steps	Time complexity	Space complexity
Step 1	$O(n)$	$O(n)$
Step 2	Function: $O[(m+d-2)^3]$ Value: $O(nd^2)$	Function: $O[(m+d-2)^2]$ Value: $O(n)$
Step 3	$O[n \times fm]$	$O(n)$
Step 4	Smooth: $O[(m+d-2)^3]$ Value: $O(nd^2)$	Smooth: $O[(m+d-2)^2]$ Value: $O(n)$
Overall	$O[(m+d-2)^3 + k((n+d-2)^3 + n \times fm)]$	$O[(m+d-2)^2 + (n+d-2)^2]$

respectively. In general, m^3 is a limited amount of calculation, relative to the iterative process it can be left out; when d takes smaller values, in the experiment of this paper, $d = 4$, at this point the time complexity and space complexity of the algorithm is $O[k(n^3 + n \times fm)]$ and $O[m^2 + n^2]$ respectively.

Experimental result and analysis. This paper has validated the proposed correlation determination method and the curve alignment method based on the simulated data. And the sensitivity of the parameters is compared with the existing methods for the time warping data collected by the curve alignment method and analytical method as well.

Time warping sequence correlation determination experiment. Select the Sinc function $\mathrm{Sinc}(x) = \sin \pi x / \pi x$, $\xi \in [-6, 6]$ with volatility as the research object for the correlation of simulation data, and make the following two time functions: $d_1(t) = 0.01t^2 - 0.36$, $d^2(t) = 0.005t(t-6)(t+6)$. In both cases, the trend of variation with the standard Sinc function and the time shift sequence correlation coefficient are shown in Fig. 4.

As it can be seen from Fig. 3, *b*: when two sequence correlation coefficient curve is on the convex, and the bounds of the two correlation coefficients are [0.991, 0.996] and [0.914, 0.962] respectively. Thus, it can be determined that based on the sequence correlation between the two groups $d_1(t)$ and $d_2(t)$, with the standard Sinc function sequence, the average amount of lag is 0 and 3, respectively.

Curve alignment experiment. This section mainly tests on the GQS algorithm performance, and makes comparison with the classical CMRM algorithm, maximum likelihood registration (hereinafter referred to as MLR) and the self-modelling registration (referred to as SMR) are compared. To be fair, the results of CMRM algorithm and MLR are 5 times the average results of operation. The experimental machine is configured as the following: Intel quad-core CPU (frequency of 2.83 GHz), 3G memory.

Align the Sinc function containing noise with the time difference functions $d_1(t) = 0.01t^2 - 0.36$ and $d_2(t) = 0.005t(t-6)(t+6)$ with the standard Sinc function curve. And the trend of variation with the standard Sinc function and the time-lag series correlation coefficient is shown in Fig. 4.

As shown in Fig. 4: For the time difference function, MLR alignment effect is poorer, the other three kinds are close to the actual value; for the time difference function, the MLR and CMRM alignment effect is poorer, SMR and S-GEM are very close to the time difference function.

Tables 2–5 are the alignment results of the Sinc function, respectively, using CMRM, MLR, SMR, S-GEM under two kinds of time functions.

As can be seen from Table 5, GQS algorithm accuracy is closely related to the sampling points, and for more complicated time function, it requires more sampling points; the run time of GQS algorithm changes mainly with the increase of the sampling points.

It can be seen from the results of the comparison of Tables 2–5: when time difference function is $d_1(t)$,

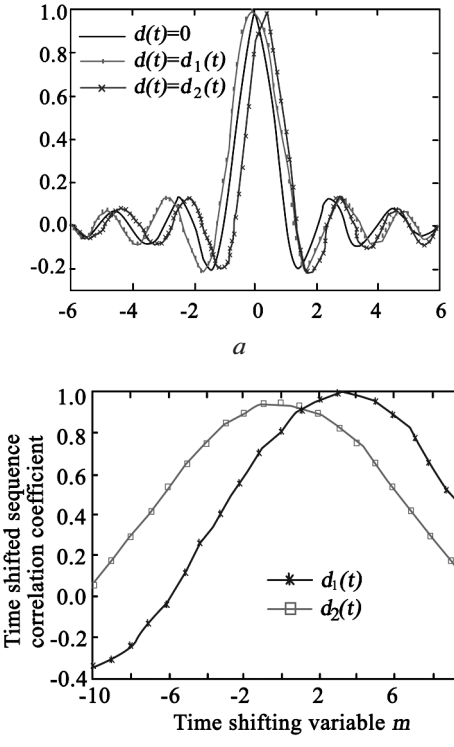
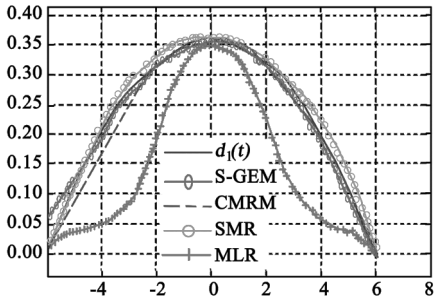
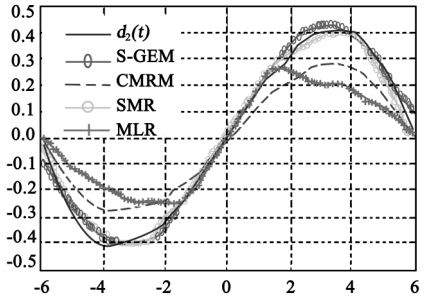


Fig. 3. Sinc Function and Time Shift Sequence Correlation Coefficient Variation Diagram



a



b

Fig. 4. Results of 4 Methods of Alignment

CMRM has the highest precision, but also the lowest efficiency; SMR and S-GEM's accuracy and efficiency are about the same, but the SMR results are not as stable as S-GEM; the precision of MLR is the worst, and the efficiency is not high; when time difference function takes $d_2(t)$, the precision of SMR and S-GEM is relatively high, but the SMR efficiency is far worse than S-GEM, but not as stable as S-GEM.

The above experiments show that, when the time difference function is simple, the precision of CMRM,

Table 2

Sinc Function CMRM Alignment Results
(mean of 5 Times)

Time difference function	Run duration (*s*)	RMSE
$d_1(t)$	20.32	0.0038
$d_2(t)$	30.74	0.0891

Table 3

Sinc Function MLR Alignment Results
(mean of 5 times)

Time difference function	Parameter		Run duration (*s*)	RMSE
	Scale parameter	Maximum number of iterations		
$d_1(t)$	0.1	50	14.60	0.1086
	0.1	100	22.24	0.1462
$d_2(t)$	0.1	50	14.51	0.1504
	0.1	100	21.83	0.1337

Table 4

Sinc Function SMR Alignment Results

Time difference function	Experiment serial number	Parameter		Run duration (*s*)	RMSE
		Number of components	Number of primary functions		
$d_1(t)$	1	4	7	5.49	0.0606
	2	3	8	6.87	0.0118
	3	3	8	6.05	0.0133
	4	4	7	5.65	0.0632
	5	3	7	6.27	0.0512
	Mean	3.4	7.4	6.06	0.0400
$d_2(t)$	1	4	12	21.99	0.0208
	2	4	11	25.10	0.0474
	3	5	13	24.56	0.0674
	4	4	12	23.82	0.0611
	5	4	15	23.00	0.0888
	Mean	4.2	12.6	23.69	0.0571

Table 5

Sinc Function GQS Algorithm Alignment Results

Time difference function	Iteration error *eps*	Sampling points *n*	Run duration (*s*)	RMSE
$d_1(t)$	0.01	10	0.15	0.2618
		30	4.29	0.0293
		50	7.76	0.0119
		80	16.52	0.0135
		100	25.47	0.0169
	0.05	10	1.13	0.0835
		30	3.2	0.0180
		50	7.96	0.0120
		80	17.15	0.0239
		100	24.87	0.0245
	0.1	10	0.15	0.2618
		30	3.13	0.0301
		50	7.33	0.0191
		80	15.19	0.0232
		100	19.72	0.0298
$d_2(t)$	0.01	10	0.15	0.2969
		30	4.74	0.0356
		50	8.04	0.0297
		80	16.61	0.0269
		100	27.82	0.0340
	0.05	10	0.15	0.2969
		30	4.09	0.0337
		50	8.15	0.0224
		80	16.78	0.0261
		100	22.75	0.0288
	0.1	10	0.15	0.2969
		30	3.31	0.0379
		50	7.09	0.0324
		80	14.73	0.0265
		100	20.60	0.0374

SMR and S-GEM is all relatively higher, but the effect of CMRM is relatively poor; when the time difference function is relatively complicated, the alignment results of SMR and S-GEM are good, but S-GEM is better in the efficiency and stability.

In conclusion, it is noteworthy that in the alignment of the leading index contrast consistent process, compared with CMRM, SMR and MLR, GQS algorithm is superior in alignment effect (the correlation coefficient and intuitive graphics) to the other methods, and the computation efficiency is higher than CMRM and SMR.

Summary and outlook. At a certain significance level, this paper has proposed that the upper and lower bounds of the overall correlation coefficient shall be applied to determine the correlation; there are multiple causes of spurious regression problem, at present a strict and accurate identification method has not yet been identified. For the correlation error, this paper can determine its correlation from the characteristics of the relevant coefficient of the grey theory time shift sequence. For the relevant sequences with time warping, the maximum model based on the grey theory correlation coefficient and the improved algorithm is established, with the applicable scope wider than the AISE standards. The experimental results show that the correlation determination method proposed in this paper is more effective than Pearson linear correlation coefficient, Spearman rank correlation coefficient, Kendall rank correlation coefficient and the Granger causality test. In most cases the proposed GQS algorithm is superior to CMRM, SMR and MLR. This paper has taken the double sequence linear correlation issue and function type curve alignment method into consideration, and the results can be used to provide the theoretical basis for the correlation determination of the regression analysis and the time alignment, and can also offer the reference direction for the multiple sequential correlation mining and curve alignment.

References / Список літератури

1. Yin, M. S., 2013. Fifteen years of grey system theory research: a historical review and bibliometric analysis. *Expert systems with Applications*, Vol. 40, No. 7, pp. 2767–2775.

2. Low, Y., Bickson, D., Gonzalez, J., Guestrin, C., Kyrola, A. and Joseph M. Hellerstein, 2012. Distributed GraphLab: a framework for machine learning and data mining in the cloud' *Proceedings of the VLDB Endowment*, Vol. 5, No. 8, pp. 716–727.

3. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M. and Zupan, B., 2013. Orange: data mining toolbox in Python. *The Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 2349–2353.

4. Nguyen, P. H., Sheu, T. W., Nguyen, P. T., et al., 2014. Taylor Approximation Method in Grey System Theory and Its Application to Predict the Number of Foreign Students Studying in Taiwan, *International Journal of Innovation and Scientific Research*, Vol. 10, No. 2, pp. 409–420.

5. Romero, C. and Ventura, S., 2013. Data mining in education, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 3, No. 1, pp. 12–27.

6. Tserng, H. P., Ngo, T. L. and Chen, P. C., 2015. A Grey System Theory Based Default Prediction Model for Construction Firms. *Computer Aided Civil and Infrastructure Engineering*, Vol. 30, No. 2, pp. 120–134.

7. Wei, M.C., 2014. The Influence Factor Analysis for Sexual Harassment on Campus in Taiwan via Grey System Theory. *Journal of Grey System*, Vol. 17, No. 4, pp. 207–213.

8. Ghodrati Amiri, G., Zare Hosseinzadeh, A. and Jafarian Abyaneh, M., 2016. A new two-stage method for damage identification in linear-shaped structures via Grey System Theory and optimization algorithm, *Journal of Rehabilitation in Civil Engineering*, Vol. 3, No. 2, pp. 36–50.

9. Raju, P. S., Bai, D. V. R. and Chaitanya, G. K., 2014. Data mining: Techniques for enhancing customer relationship management in banking and retail industries. *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 1, pp. 2650–2657.

10. Liao, S. H., Chu, P. H. and Hsiao, P. Y., 2012. Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications*, Vol. 39, No. 12, pp. 11303–11311.

Мета. Із розвитком передачі великих об'ємів даних, інтелектуальний аналіз часових рядів став важливою темою, що привертає до себе увагу суспільства. На основі кореляції та спільності часових рядів, розглянута швидка модель інтелектуального аналізу часових послідовностей.

Методика. Запропоновано метод визначення кореляції, заснований на особливостях відповідного коефіцієнта зсунутої часової послідовності. У кінцевому рахунку, запропонована швидка модель інтелектуального аналізу часових послідовностей, заснована на теорії сірих систем.

Результати. Методологія визначення кореляції, запропонована в даній роботі, є більш ефективною, ніж коефіцієнт лінійної кореляції Пірсона, коефіцієнт рангової кореляції Спірмена, коефіцієнт рангової кореляції Кендалла та тест Гренджера на причинність.

Наукова новизна. У роботі запропоноване об'єднання швидкого метода визначення кореляції послідовностей і метода вирівнювання по кривій.

Практична значимість. Результати можуть забезпечити теоретичну базу для визначення кореляції регресійного аналізу та часового вирівнювання.

Ключові слова: *теорія сірих систем, часове вирівнювання, кореляція, часова послідовність даних, інтелектуальний аналіз даних, вирівнювання по кривій, тест на причинність*

Цель. С развитием технологии передачи больших объемов, интеллектуальный анализ временных рядов стал важной темой, которая привлекает к себе внимание общественности. На основании корреляции и общности временных рядов, рассмотрена быстрая модель интеллектуального анализа временных последовательностей.

Методика. Предложен метод определения корреляции, основанный на особенностях соответствующего коэффициента сдвинутой временной последовательности. В конечном счете, предложена быстрая модель интеллектуального анализа временных последовательностей, основанная на теории серых систем.

Результаты. Методология определения корреляции, предложенная в данной работе, является более эффективной, чем коэффициент линейной корреляции Пирсона, коэффициент ранговой корреляции Спирмена, коэффициент ранговой корреляции Кендалла и тест Гренджера на причинность.

Научная новизна. В работе предложено объединение быстрого метода определения корреляции последовательностей и метода выравнивания по кривой.

Практическая значимость. Результаты могут обеспечить теоретическую базу для определения корреляции регрессионного анализа и временного выравнивания.

Ключевые слова: *теория серых систем, временное выравнивание, корреляция, временная последовательность данных, интеллектуальный анализ данных, выравнивание по кривой, тест на причинность*

Рекомендовано до публікації докт. техн. наук В. В. Гнатушенком. Дата надходження рукопису 20.10.15.

Fei Hu[1,2], Changjiu Pu[2], Haowei Gao[3], Mengzi Tang[1], Li Li[1]

1 – School of Computer and Information Science, Southwest University, Chongqing, China
2 – Network Centre, Chongqing University of Education, Chongqing, China
3 – The Webb Schools, 1175 West Baseline Road Claremont, CA 91711, USA

IMAGE COMPRESSION AND ENCRYPTION SCHEME BASED ON DEEP LEARNING

Фей Ху[1,2], Чанцзю Пу[2], Хаовей Гао[3], Менцзи Тан[1], Лі Лі[1]

1 – Школа комп'ютерних та інформаційних наук, Південно-Західний університет, Чунцін, Китай
2 – Мережевий центр, Чунцінський університет освіти, Чунцін, Китай
3 – Школа Уебб, Клермонт, США

СХЕМА СТИСНЕННЯ ТА ШИФРУВАННЯ ЗОБРАЖЕНЬ НА ОСНОВІ ГЛИБИННОГО НАВЧАННЯ

Purpose. With the growing demands of image processing on the Internet, image compression and encryption have been playing an important role in image protection and transfering. In this paper we will investigate deep learning technology in image compression, and chaotic logistic map in image encryption, to obtain a scheme in image compression and encryption. We have evaluated this scheme with some performance measures and results show it is effective.

Methodology. We formulate the scheme using deep learning and chaos. With the deep learning technology, levels of features are extracted from an image and a certain level of features can be used as a compressed representation of the image. Chaos is used to encrypt the compressed image.

Findings. We first introduceda five-layer Stacked Auto-Encoder model, which is trained by the Back Propogration method, and then we obtained the compressed representation of an image. By using the logistic map method, a pseudo-stochastic sequence is generated to encrypt the compressed image.

Originality. We conducted a study of image compression and encryption. Image characteristics are extracted from an arbitrary level of our deep learning model, and they are used as the compressed representation of the image. The research on this aspect has not been found at present.

Practical value. We have evaluated this scheme on several randomly selected images. And results show it is robust and can be widely used for most images.

Keywords: *stacked auto-encode, deep learning, image protection, image feature, image compression, image encryption*