ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Exploration of value density enhancement and clustering application of financial big data based on deep feature learning

Cong Xie[1,2,3], *Wanzhao Zhao*[4,*] *and Yi Zeng*[3]

[1] Guangxi Key Laboratory of Big Data in Finance and Economics (Guangxi University of Finance and Economics), Nanning, Guangxi, 530007, China
[2] School of Judicial Application, Guangxi Police College, Nanning, Guangxi, 530028, China
[3] School of Information Engineering, Guangxi Vocational University of Agriculture, Nanning, Guangxi, 530007, China
[4] School of Intelligent Equipment Engineering, Guangxi Vocational University of Agriculture, Nanning, Guangxi, 530007, China

SUMMARY: *In this paper, based on the characteristics of real-time and time characteristics of financial data, time series is used as an entry point to mine financial data. Using the trigonometric polynomial graph method, the visualization of financial and economic cube data and the definition of clustering distance are carried out. Combined with wavelet analysis theory to decompose the periodic, trend and random terms of financial data time series, so as to predict the development of financial data with different scale components. In addition, BIRCH method is utilized to construct CF tree through data clustering features for micro-clustering of financial data. And K-meadiods method is chosen to fill the gap of BIRCH method in the application of large datasets to realize the macro-clustering of financial and economic data, and to establish BIRCH.K-meadiods clustering method. Compared with similar models, the prediction and clustering model of financial data based on time series has an overlap rate of >99.00% between the predicted and actual values of financial data and an error in the interval of (0,300), which can assist in the in-depth mining and clustering analysis of the value of financial data with high-precision prediction, and expand the level of decision-making considerations for the internal management of the enterprise and even for the external construction.*

KEYWORDS: *trigonometric polynomial graph; BIRCH.K-meadiods; financial data prediction; time series; wavelet analysis*

1 Introduction

Due to the constant development of the social economy, the financial market has become one of the essential elements of the national economic system and one of the significant indicators of a nation-wide economic performance [1]. Over the past few years, the Chinese financial market has been developing at a very fast pace, starting with a comparatively weak base. This is accompanied by the extreme growth of financial data which places significantly greater demands on organization, storage, and management of data. The ability to quickly extract useful information of large volumes of financial data, and subsequently apply this information to perform effective analysis and prediction has turned into an acute problem that has received

*13978605116@163.com
https://doi.org/10.65102/is2026382
Published: 30 April, 2026

considerable attention in the academic sphere and industry [2, 3]. That is why investigating the nature of the financial markets and studying the underlying developmental tendencies that are hidden in the data will be able not only to offer consistent technical assistance to financial management and investment practices, but also contribute greatly to the successful and sound functioning of the financial market.

As a huge and complex system, financial market is easily affected by social, economic, political, investor tendency, financial activities and other factors [4]. Financial big data can be viewed as a mirror image of market conditions and is commonly used to discover the intrinsic value and pattern of market operations. Financial market analysis is effectively the analysis of time-series data since large amounts of such data are typically used in chronological form in practice, from the analysis methods of financial market data, there are mainly fundamental analysis, technical analysis, mathematical and statistical analysis, data mining and deep learning analysis [5-8]. Among them, fundamental analysis and technical analysis, using the surface information data, analyze the time series data, but can not be deep mining data between the internal hidden laws [9]. Compared with the fundamental and technical analysis methods, the traditional mathematical and statistical analysis methods, for the analysis of small data volume effect and accuracy is more obvious, but for the financial data containing a large amount of complex information its analysis efficiency is low [10]. And financial big data as a kind of time series data, both with the commonality of time series, at the same time with the nonlinear data between the chaotic, non-smooth and high-dimensional characteristics [11].

Although conventional approaches to analyzing data may be effective in cases where the amount of data is small, these methods tend to fail in the case of large-scale non-stationary data, and it is hard to find hidden relationships in the data. Nevertheless, deep feature learning can discover natural patterns of large data and can be useful to classify the data and predict them with some confidence [12-14]. In addition to the fast development of artificial intelligence, deep feature learning has slowly been applied to a wide range of research and applications.First of all, a large amount of historical data through the integration of data, cleaning, transformation, mining, model evaluation, the use of deep learning methods to analyze and predict the time series data, to extract the implied laws, for the discovery of the market's potential laws and information, to assist investors in avoiding the market risk has a very important significance, and at the same time provide important technical support for financial decision-making [15].

Efficient extraction of key information out of large volumes of text data has come to be a fundamental problem of natural language processing. The problem is particularly acute in finances and economics whereby unstructured texts, including financial news, corporate announcements and research reports, are filled with event information, and such extracted information is crucial in the process of revealing the mechanisms behind economic activities [16]. Event extraction, as one of the significant areas of research on information extraction, is to automatically find and categorize event-related information in financial texts, which can offer significant assistance to investment decision-making and market prediction [17]. For example, Yang, H et al [18] customized relevant event templates for equity changes, a financial type of event, and although this approach can achieve better results on specific text data, it requires preset templates for each event type, which undoubtedly increases the complexity of the task and restricts the scope of its application. Cheng, W et al [19] noted the dramatic increase in financial news, social media content, and similar data, underlined the fact that this kind of data can help to explain the unexplainable volatility in financial markets and enhance the accuracy of forecasts, and discussed semantic-based, sentiment-based, event-based and hybrid extraction of the stock prediction. Lu, Y et al [20] used a special

decoding strategy to transform event extraction into a sequence generation problem by targeting the problems of traditional news text categorization methods that require a large number of fine-grained annotations for different sub-tasks, are data inefficient, and designing the optimal combinatorial structure for different sub-tasks is challenging and often leads to error propagation. Deng, H et al [21] used an open model to define the event as a ternary structure including subject, predicate and object.This approach is suitable for news headline datasets with relatively simple structure.Since the syntactic structure of financial news headlines is usually simple, the ternary structure can cover the event content in a more complete way.

With the rapid development of machine learning models and artificial intelligence technologies, event extraction methods have undergone an evolutionary process from the initial basic machine learning models to reading comprehension models to the generative-based models that are widely used today. For example, Wan, Q et al [22] constructed Bidirectional Dependency Analysis Graph (Bi-DPG) by adjusting the structure of syntactic dependency tree and captured the semantic information of the graph structure through an improved Graph Attention Network (GAT), which effectively enhanced the model's extraction of open domain events. Liu, X et al [23]investigated the identification of open-pattern events from clustered news texts. They developed a novel latent-variable neural architecture for inducing event patterns and extracting events, and they also released a large-scale dataset, GNBusiness, which covers multiple event categories and interpretable event patterns. Zheng, S et al [24] transformed the filling problem of event table into the path extension problem of directed acyclic graph, proposed the Doc2EDAG model, and constructed a large financial text dataset, which provides new ideas and new resources for the subsequent research. Du, X and Cardie, C [25] reformulated event extraction as question answering by following an end-to-end model of event argument extraction, where no preprocessing steps such as entity recognition are used. The approach is able to identify semantic similarities between various argument roles and may retrieve those roles that are infrequent or not present in the training set. To solve the problem of lack of data in the current classification-based methods of event extraction. Liu, J et al [26] addressed the issue of insufficient data in existing classification-based event extraction approaches by treating event extraction as a machine reading comprehension task and introducing an unsupervised question-generation algorithm that produces topic-related and context-aware questions.

Financial information on the Internet, especially news reports on major financial websites, is the main way of information transmission, which prompts many investors to pay more attention to key information, such as bank interest rate hike policy, strong US dollar, Internet finance, etc [27]. By mining the keywords of financial news, finding the main variables that represent the heat of investors' attention, and analyzing the correlation between these variables and the rise and fall of stocks, it is of great significance to mine the latent value in financial big data and formulate investment strategies [28, 29]. In the mining and clustering research of financial big data, Guo, Y et al [30] proposed an improved clustering framework based on a deep sparse autoencoder. Their method first uses a text vector model together with cosine distance to build a similarity matrix, then applies an unsupervised deep sparse autoencoder to reduce dimensionality and derive structural features from complex text vectors, and finally performs clustering through the K-Means algorithm. Carta, S et al [31] proposed an event detection method based on real-time domain specific clustering for identifying relevant news stories published by globally recognized news sources for clustering, the experiments showed that the method is capable of extracting valuable information from financial texts as well as its effectiveness in discovering topical events in the financial domain. Wang, J [32] proposed a graph clustering framework (FinGC) for extracting summaries of

financial news, which performs clustering in an unsupervised manner and enhances the representation of financial news by combining graph structures containing cross-sentence relationships into text embeddings, extracting valuable information from financial news. de Oliveira, A et al [33] proposed a technique to predict price movements based on a collection of similar stocks based on historical stock data and stock information from Google financial news, and used a K-Means clustering algorithm to identify similar stock portfolios. Sidorov, S et al [34] proposed a procedure to identify economic and financial events by constructing data-similarity graphs at specific scales and detecting both types of events using a common graph-based clustering algorithm.

In this paper, the trigonometric polynomial graph is chosen as the data visualization method in the framework of time series analysis, and the distance definition for data clustering under this method is projected. Based on the characteristics of financial data time series, we analyze the different decomposition components of time series under wavelet variations, and establish a time series decomposition and reconstruction model based on wavelet analysis. Then design the prediction steps and propose the wavelet analysis model to form the financial data time series decomposition and prediction scheme based on wavelet analysis. The BIRCH method is used to develop the microclustering of the data, and the K-meadiods method is introduced to improve the shortcomings of the BIRCH method, and a combined hierarchical clustering method combining the BIRCH and the K-meadiods is constructed, so as to form the prediction and clustering model of financial data based on time series. The experimental dataset and the control model are selected to draw and compare the prediction effect and prediction error of the time series-based financial data prediction and clustering model, and initially verify the overall feasibility of the model. The model is used to denoise the experimental data set, set the financial and economic evaluation indexes, and carry out the visualization and effect analysis of the experimental data. And through the form of comparison, the optimal visualization parameter settings are determined.

2 Visualization, clustering based on financial data mining

2.1 Time series based financial data mining

Let X_{ktij} be a financial data cube, where $k=1,2,\cdots,p$, p denotes the number of members of the financial variety of the class, for example, if we study the A shares, then p denotes the number of all the A shares; $t=1,2,\cdots,\tau$, τ is a definite point in time; $i=1,2,\cdots,m$, m denotes the number of significant time measures of a single species, such as the closing price, opening price, high price, low price, etc. of a particular stock; $j=1,2,\cdots,n$, n denotes the number of sequence values within a cell, such as the number of trading days per week, the number of trading hours per day, etc.

In general, noting the matrix $X_{kt}=(X_{ktij})_{m\times n}$, the vector $X_{kt}\eta_n$ obtained by right-multiplying it by a variable η_n is known as a weighted average, e.g., by taking $\eta_n=(1/n,1/n,\ldots,1/n)^{\mathrm{T}}$, then $X_{kt}\eta_n$ is the n-day average. If a vector ξ_m^{T} is left-multiplied, the corresponding result $\xi_m^{\mathrm{T}}X_{kt}$ is called the mid-price. Let equation (1):

$$y_{kt}=\xi_m^T X_{kt}\eta_n \tag{1}$$

Thus, k sequences $\{y_{kt}\}$ are obtained.

The difference with the traditional time series analysis is that the data mining of the above sequences needs to process a large set of sequences automatically and quickly at the same time, the above financial sequences $\{y_{kt}\}$ exist in the massive dynamic data warehouse, and need to mine the useful information that exists in many sequences in order to support the decision-making and rapid reflection, then, the study of the operation law of the individual sequences is less important, so that the The methodology is heavily simplified to accommodate the requirement for algorithmic simplicity and speed.

2.2 Visualization of data and definition of distances used for clustering

Data warehousing and OLAP systems should be based on a multidimensional modeling framework because they need to handle information that is structured as data cubes. To graph such data with cube structure, the proposed solution of this research is the implementation of trigonometric polynomial maps as an efficient method.Take the orthogonal function system $\{\sin t, \cos t, \sin 2t, \cos 2t, \cdots\}$ on $[-\pi, \pi]$ and build the mapping as in equation (2):

$$\begin{aligned} V &= (v_1, v_2, \cdots, v_m) \to \hat{f}_i(t) \\ &= v_1 \sin t + v_2 \cos t + v_3 \sin 2t + v_4 \cos 2t + \cdots - \pi \le t \le \pi \end{aligned} \tag{2}$$

The mapping has the following excellent properties:

(1) Linearity-preserving relations.

(2) Euclidean distance is preserved.

(3) A one-to-one mapping on R^m to L^m.

(4) If the components of V are independently homoskedastic σ^2, then there is equation (3) when m is even:

$$\operatorname{var}\left(\hat{f}_i(t)\right) = \frac{m\sigma^2}{2}, -\pi \le t \le \pi \tag{3}$$

When m is odd, there is equation (4):

$$\frac{(m-1)\sigma^2}{2} \le \operatorname{var}\left(\hat{f}_i(t)\right) \le \frac{m\sigma^2}{2}, -\pi \le t \le \pi \tag{4}$$

(5) If $V \sim N_p(\mu, \sigma^2 I_p)$, then the probability of being in $1-\alpha$ has equation (5):

$$\left| f_V(t) - f_\mu(t) \right|^2 \le \frac{\sigma^2 (m+1)}{2} \chi_m^2(\alpha), -\pi \le t \le \pi \tag{5}$$

where $\chi_m^2(\alpha)$ is the α quantile of the cartesian distribution.

Let $V, W \in R^m$, then the Euclidean distance between V, W is equation (6):

$$d_{VW}^2 = (V-W)'(V-W) \tag{6}$$

Since both $f_V(t)$ and $f_W(t)$ are square-integrable functions on $[-\pi, \pi]$, the Euclidean distance between them can be defined as Eq. (7):

$$d_{f_V f_W}^2 = \int_{-\pi}^{\pi} \left| f_V(t) - f_W(t) \right|^2 dt \tag{7}$$

and there is equation (8):

$$d_{f_V f_W}^2 = \pi d_{VW}^2 \tag{8}$$

This strict equation solves the problem of the complexity of this graphical method of distance calculation, i.e., the complexity of $d_{f_V f_W}^2$ is at most $O\left(n^2\right)$. However, it is very appropriate to use the area between two curves as the distance measure between them, so consider the following definition of distance as in equation (9):

$$\tilde{d}_{f_V f_W} = \int_{-\pi}^{\pi} \left| f_V(t) - f_W(t) \right| dt \tag{9}$$

It not only has computational advantages, but is also more robust than the Euclidean distance, and does not exaggerate the differences in a particular local space in the case of a large number of dimensions. It can be utilized as a distance definition for clustering to achieve fast clustering.

Two approximation algorithms can be used in practical calculations, both of which have an algorithmic complexity of $O\left(n^1\right)$!, and with equation (10):

$$\tilde{d}_{f_V f_W} \approx \sqrt{\pi} \sum_{i=1}^{m} \left| v_i - w_i \right| \tag{10}$$

If there is no requirement for accuracy you can use this formula, the current clustering algorithms in data mining widely used distances such as minimum distance, maximum distance, average distance, etc. and it is similar to it, there is equation (11):

$$\tilde{d}_{f_V f_W} \approx \frac{2\pi}{s} \sum_{i=1}^{s} \left| \sum_{j=1}^{m} c_{ij} \left(v_j - w_j \right) \right| \tag{11}$$

where c_{ij} are all constants and have equation (12):

$$\begin{aligned} \left(c_{i1}, c_{i2}, ..., c_{im}\right) = \big(&\sin\left(-\pi + 2\pi i / s\right), \cos\left(-\pi + 2\pi i / s\right) \\ &\sin 2\left(-\pi + 2\pi i / s\right), \cos 2\left(-\pi + 2\pi i / s\right), \cdots\big) \end{aligned} \tag{12}$$

This formula can be controlled for accuracy, and since $\tilde{d}_{f_V f_W}$ denotes the area between two curves f_V and f_W, which is not only robust, but also more geometrically obvious compared to the Euclidean distance between f_V and f_W, a commonly used approximation algorithm is used, which first divides the The interval $[-\pi, \pi]$ is equidivided into s small intervals, resulting in equation (13):

$$\tilde{d}^{*}_{f_V f_W} = \frac{2\pi}{s}\sum_{i=1}^{s}\left| f_V\left(-\pi+\frac{2\pi}{s}i\right) - f_W\left(-\pi+\frac{2\pi}{s}i\right)\right| \xrightarrow{n\to\infty} \tilde{d}_{f_V f_W} \tag{13}$$

From this, we can obtain equation (14):

$$\tilde{d}^{*}_{f_V f_W} = \frac{2\pi}{s}\sum_{i=1}^{s}\left|\sum_{j=1}^{m} c_{ij}\left(v_j - w_j\right)\right| \tag{14}$$

By choosing different s, it is possible to adapt to different requirements on computational accuracy.

3 Forecasting and clustering model for financial data based on time series

3.1 Time series decomposition based on wavelet analysis

Based on wavelet theory, a signal or a time series may be divided up level by level into components at various frequency bands. Each of the decomposed components is typically far less rough compared to the original signal in terms of frequency content and wavelet decomposition hence offers increased smoothness over the raw series. Practically, using different wavelet transforms it is possible to divide a time series with trend terms, periodic terms, and random fluctuations into numerous scales thus separating these components into distinct elements and treating them according to their scale, making it simpler to analyze a complex phenomenon.

3.1.1 Wavelet Decomposition and Reconstruction Models

Let the time series $x(n)=\{x_1, x_2, \cdots, x_L\}$ be denoted as c_0, and wavelet decomposition of the original series is done using tower decomposition algorithm as in equation (15):

$$\begin{cases} c_{j+1} = Hc_j \\ d_{j+1} = Gd_j \end{cases} \quad j = 0,1,2\cdots J \tag{15}$$

That is, c_0 is decomposed into $d_1, d_2 \cdots d_J, c_J$, and the number of points of each decomposed detail signal d_{j+1}, and approximation signal c_{j+1} is reduced by a factor of two compared with that of the pre-decomposition signal c_j.

In order not to affect the prediction accuracy, the decomposed signal is reconstructed as in equation (16):

$$c_j = H_1 c_{j+1} + G_1 d_{j+1},\ j = J-1, J-2, \cdots, 0 \tag{16}$$

where H^*, G^* is the dual operator of H, G, and the original signal can be reconstructed exactly when $H^*H + G^*G = I$ is satisfied.

In case the wavelet basis used for decomposition is not canonically orthogonal, the reconstruction formula is equation (17):

$$c_j = H_1 c_{j+1} + G_1 d_{j+1}, j = J-1, J-2, \cdots, 1, 0 \tag{17}$$

where H, G, H_1, G_1 are obtained from $\phi_{j,k}, \varphi_{j,k}$ and their duals respectively, the original signal can be reconstructed exactly when $H_1 H + G_1 G = I$ is satisfied.

In this way, the number of points of each reconstructed approximation signal c_j is again doubled compared to the signal before reconstruction to have the same length as the original signal, yielding $D_1, D_2, \cdots D_J, C_J$, i.e., Eq. (18):

$$X(n) = D_1 + D_2 + \cdots D_J + C_J \tag{18}$$

It is shown after test:

(1) If the selected wavelet has certain regularity, after several times (generally 3-7 times is appropriate) of wavelet transform, the random and periodic terms of the original time series are gradually separated, i.e., the anomalies or random fluctuations in the sequence are gradually eliminated, and the edges and corners are smoothed. The time series tends to be smoothed until the last low frequency that is the long-term trend part.

(2) In the wavelet analysis, the layers other than the long-term trend are the periodic and random terms, and the random factor often corresponds to the small-scale components. The high half-frequency D_1 of the wavelet decomposition in the first layer contains the most random terms, while the following layers have more periodic terms. For some layers where the change is not obvious, wavelet decomposition can be carried out again until the periodic and random terms are well separated.

(3) The long-term trend part of the time series can be observed through the graph of C_J, while the random fluctuations and periodicity of the time series can be observed in each detail layer $D_1, D_2 \cdots D_J$.

3.1.2 Wavelet analysis models

Let the non-smooth time series be $x(n) = \{x_1, x_2 \cdots x_N\}$, the wavelet prediction steps are as follows:

(1) Perform J-layer wavelet decomposition to obtain $d_1, d_2 \cdots d_J, c_J$, then reconstruct each layer to obtain $D_1, D_2 \cdots D_J, C_J$, and $D_1 = \{d_{11}, d_{12} \cdots d_{1N}\}, \cdots D_J = \{d_{J1}, d_{J2} \cdots d_{JN}\}$, $C_J = \{c_{J1}, c_{J2} \cdots c_{JN}\}$, and so there is equation (19):

$$X(n) = D_1 + D_2 + \cdots D_J + C_J \tag{19}$$

(2) Plot C_J, if the trend is obvious, then modeling prediction; if the trend is not obvious, increase J to the first step. The size of J is generally taken as 3~7. The trend term models are generally: linear trend, polynomial trend, exponential trend and logarithmic trend, which can be predicted by least squares fitting and regression.

(3) For each layer of the periodical term, the prediction is modeled by the periodogram method. Let the data series on a layer also be $x_1, x_2 \cdots x_N$, which is approximated to minimize the mean squared error by a family of sine functions, i.e., $x_t = \sum_{i=1}^{k} c_i \cos\left(2\pi f_i t + \varphi_i\right) + \varepsilon_t$, $k, c_i, f_i \left(i = 1, 2 \cdots k\right)$ are constants, $\varphi_i \in \left(-pi, \pi\right)$, and ε_t is a stochastic process with mean

0 and variance σ_ε^2, then we have equation (20):

$$\hat{x}_t = \sum_{i=1}^{k} \left(a_i \cos 2\pi f_i t + b_i \sin 2\pi f_i t \right) + \varepsilon_t \tag{20}$$

where $a_i = c_i \cos \varphi_i$ and $b_i = -c_i \sin \varphi_i$. Equation (21) is solved by the least squares method:

$$\hat{a}_i = \frac{2}{N} \sum_{t=1}^{N} x_t \cos 2\pi f_i t \tag{21}$$

Thus, future values can be predicted by $\hat{x}_t = \sum_{i=1}^{k} \left(a_i \cos 2\pi f_i t + b_i \sin 2\pi f_i t \right)$.

(4) For the smallest scale random layer, it can be considered as white noise. Since the value is very small, it can be ignored in general prediction without leading to large errors. After removing the random layer with large fluctuations, each random term is regarded as a smooth process that can be predicted by AR, MA or ARMA models. The random term after the trend term, the periodic term and the partial random term is denoted as $y_1, y_2 \cdots y_k$, where $y_t = \{d_{1t}, d_{2t} \cdots d_{Mt}\}$ for which the $AR(n)$ model is used for prediction. Generally $n = 2 \sim 3$ can be taken, i.e., $y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \alpha_t$, and model parameters are estimated and model tests are performed with the known $d_{i,j}$. Finally, the best prediction formula is used to predict the state values y_{M+l} after l steps for a sequence of known first M values. The optimal prediction formula is equation (22):

$$\hat{x}_t(l) = \begin{cases} \sum_{i=1}^{n} \varphi_i \hat{x}_{t-1,l} & l = 1 \\ \sum_{i=1}^{n} \varphi_i \hat{x}_t(l-i) + \sum_{i=1}^{n} \varphi_i \hat{x}_{t,l-l} & l \in (1, n) \\ \sum_{i=1}^{n} \hat{x}_t(l-i) & l > n \end{cases} \tag{22}$$

$y_{M+l} = dd_{1,M+l} + dd_{2,M+l} + \cdots dd_{F,M+l}$, where $dd_{i,M+l}$ is the predicted value of the i th random term.

If we set the predicted value of the trend term to be $\hat{c}_{J,M+l}$, the predicted value of each layer of the periodic term to be $\hat{d}_{1,M+l}, \hat{d}_{2,M+l} \cdots \hat{d}_{F,M+l}$, and the predicted value of each layer of the stochastic term to be $\hat{dd}_{1,M+l}, \hat{dd}_{2,M+l} \cdots \hat{dd}_{E,M+l}$, where $E + F <= J$, the predicted value of the original sequence from the moment M to the step l backward is Eq. (23):

$$\hat{x}_{M+l} = \hat{c}_{J,M+l} + \sum_{i=1}^{F} \hat{d}_{i,M+l} + \sum_{i=1}^{E} \hat{dd}_{i,M+l} \tag{23}$$

3.2 BIRCH.K-meadiods approach

BIRCH method in particular, Balanced Iterative Reduction and Clustering using Hierarchies, is an integrated multistage hierarchical clustering algorithm which can be divided into two steps, micro-clustering and macro-clustering. In the micro-clustering step, BIRCH uses CFs (clustering features) to represent and generalize clusters and subsequently creates a CF tree that is a compressed representation of hierarchical organization of clusters but retains much of the information. In the macro-clustering step, the leaf nodes of the CF tree are clustered again with other clustering methods, sparse clusters are considered as anomalies and eliminated, and dense clusters are combined into bigger ones. The macro stage clustering results are based on the input data objects identified in the microclustering stage which thus finalizes the complete overall clustering analysis. Figure 1 illustrates the hierarchical structure of the CF tree.

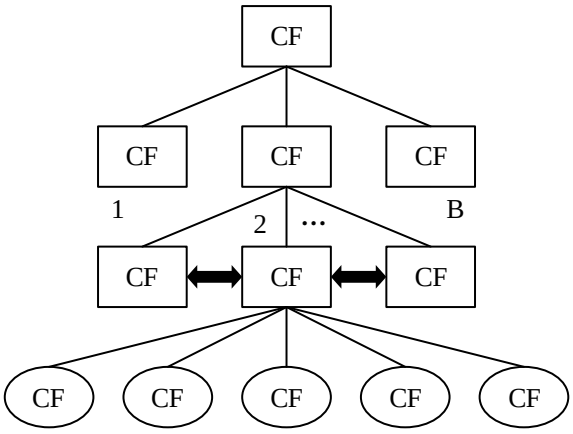


Figure 1: CF (Clustering Feature) tree hierarchical structure

This paper presents an improved version of the BIRCH framework along with the use of K-medoids clustering. Macroclustering stage involves adoption of K-medoids as a way of enhancing the entire BIRCH scheme. The K-medoids algorithm is more robust than the traditional K-means algorithm and it also works well on small datasets, but the substitution and traverse mechanisms used in the algorithm restrain its performance on extremely large datasets. Herein, the combination of K-medoids and BIRCH is done in such a way that the disadvantage of K-medoids on large data sets is addressed and the hybrid approach enables the two techniques to serve each other.

The CF tree is built in the course of microclustering, and the assumption is made on minimum loss of information, so that it would be possible to derive the information about the data objects. The hybrid approach is more efficient on large and streaming datasets, it scales well, allows incremental clustering and consumes little space, which is an important property of BIRCH clustering. In the macroclustering stage, it is compensated by combining K-medoids due to its low applicability to large datasets. To be more precise, once the microclustering has been performed, K-medoids are not used on the original data objects but are used on the leaf nodes of the CF tree constructed. With the assumption that the size of the dataset is minimized and the level of information loss is kept low, anomalies may be eliminated using the leaf nodes, thus the combined approach is less susceptible to abnormal data and provides a higher level of resistance to anomalous interference.

3.2.1 Microclustering phase: construction of the CF tree hierarchy

The main component of BIRCH in the microclustering stage is the creation of the CF tree. Suppose the data objects are a set of n-objects in a p-dimensional space. CF tree is denoted

by a three-dimensional vector consisting of information on the data objects as given by equation (24):

$$CF = (n, LS, SS) \tag{24}$$

where LS is a linear sum of n points (i.e., $LS = \sum_{i=1}^{n} x_i$), and SS is a linear sum of n point squares (i.e., $SS = \sum_{i=1}^{n} x_i^2$).

CF (Clustering Feature) is essentially a summary of statistical information about a given cluster. Using CF (Clustering Feature) a portion of the cluster's statistics can be derived: the cluster's form center x_0, radius R, and diameter d, in order as in Eqs. (25)-(27):

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{LS}{n} \tag{25}$$

$$R = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_0)^2}{n}} = \sqrt{\frac{nSS - 2LS^2 + nLS}{n^2}} \tag{26}$$

$$d = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} (x_i - x_j)^2}{n(n-1)}} = \sqrt{\frac{2nSS - 2LS^2}{n(n-1)}} \tag{27}$$

In this case, both the radius R and the diameter d can be used to define the tightness of the clustering around the center. The radius R indicates the mean separation between the member objects and cluster center, while the diameter d indicates the mean separation between any two members of the cluster as given in equation (28):

$$d^* = \sqrt{\frac{SS_1}{N_1} + \frac{SS_2}{N_2} - \frac{2LS_1 LS_2}{N_1 N_2}} \tag{28}$$

where d^* reflects the diameter of the two clusters after merging, with equation (29):

$$dd = \left(\frac{\sum_{i=1}^{N_1} \sum_{j=N_1+1}^{N_1+N_2} (x_i - x_j)^2}{N_1 N_2} \right)^{\frac{1}{2}} = \sqrt{\frac{2N^* * SS^* - 2LS^{*2}}{N^*(N^* - 1)}} \tag{29}$$

where dd reflects the inter-cluster distance, and N^*, SS^*, and LS^* reflect the N, SS, and LS of the larger clusters of the two clusters merged, respectively.

The key to the effectiveness of BIRCH clustering in space is in its storage space fixation.

Because the use of CF (clustering features) to represent clusters avoids storing detailed information about individual objects, its only necessary to determine the space to fix the storage of CF.

In addition, the rapidity of BIRCH clustering is also reflected in the additivity of CF (clustering features). It is defined as follows:

For two disjoint clusters C_1 and C_2, the CFs of both are $CF_1 = (n_1, LS_1, SS_1)$ and $CF_2 = (n_2, LS_2, SS_2)$, and the CF of the merged clusters is as in equation (30):

$$CF_1 + CF_2 = (n_1 + n_2, LS_1 + LS_2, SS_1 + SS_2) \tag{30}$$

Suppose there are three objects $(3,8)$, $(4,7)$ and $(5,6)$ on the cluster C_1, and the CF (clustering feature) of C_1 is Eq. (31):

$$\begin{aligned} CF_1 &= \left(3,\left(3+4+5,8+7+6\right),\left(3^2+4^2+5^2,8^2+7^2+6^2\right)\right) \\ &= \left(3,\left(12,21\right),\left(50,149\right)\right) \end{aligned} \tag{31}$$

Assume that C_1 is disjoint from another cluster C_2, where the CF (clustering feature) of C_2 is Eq. (32):

$$CF_2 = \left(3,\left(9,10\right),\left(29,38\right)\right) \tag{32}$$

Then, C_1 and C_2 are merged to form a new cluster C_3 whose CF (clustering feature) is the sum of CF_1 and CF_2, i.e., equation (33):

$$\begin{aligned} CF_3 &= \left(3+3,\left(12+9,21+10\right),\left(50+29,149+38\right)\right) \\ &= \left(6,\left(21,31\right),\left(79,187\right)\right) \end{aligned} \tag{33}$$

Each node in the CF tree has the same CF value, which is the sum of the CF values of all its child nodes, and any subtree with any node as its root may also be considered as a cluster. Furthermore, once two clusters are combined to form a larger cluster, the diameter of this new cluster is also achievable using the diameters of the two initial clusters respectively (refer to d^*).

The creation of the CF-tree hierarchy depends on two key parameters: the largest number of objects that may be included in a subcluster, which is indicated by B and the largest diameter permissible to objects in a subcluster, which is indicated by T.

3.2.2 Macroclustering phase: k-mediods clustering based on CF tree leaf nodes

The CF values of every leaf node can be computed using the given algorithm, and outliers are eliminated, which are the sparse clusters, i.e., those clusters with a small number of objects. This is followed by K-medoids clustering of the rest of the leaf nodes. Initial leaf nodes are chosen randomly to serve as representative candidates in the K-medoids process. Real leaf nodes are seen as clusters and one representative leaf node is chosen per cluster. The algorithm considers all possible exchanges and determines whether quality of clusters

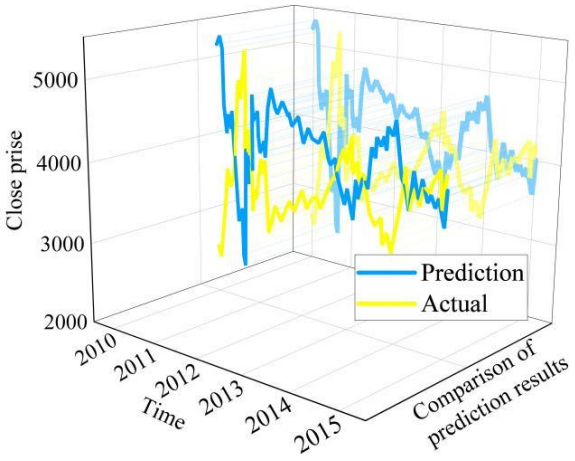
improves or reduces as the non-representative leaf node is substituted by a representative one. Stopping occurs when there is no additional improvement in clustering quality.

In the traditional K-medoids algorithm, the dataset consists of n data objects in a p -dimensional space. By contrast, after the microclustering phase, the dataset becomes q three-dimensional data objects, where q is the number of leaf nodes in the CF tree. This transformation clearly improves the efficiency of the K-medoids algorithm. Since K-medoids is not well suited to large datasets, the CF tree plays the role of compressing the dataset effectively.

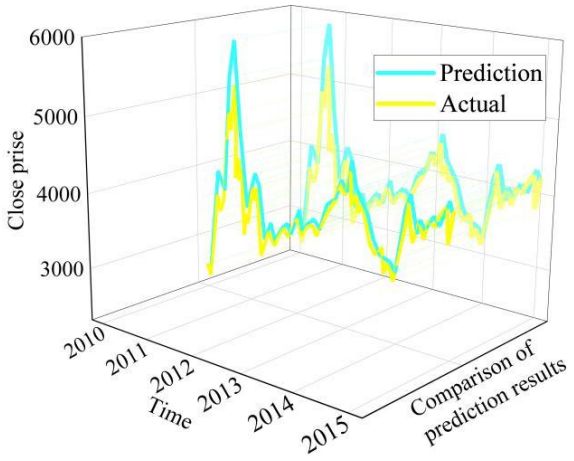
4 Density Enhancement and Clustering Visualization and Analysis of Financial Data

4.1 Predictive performance of the model

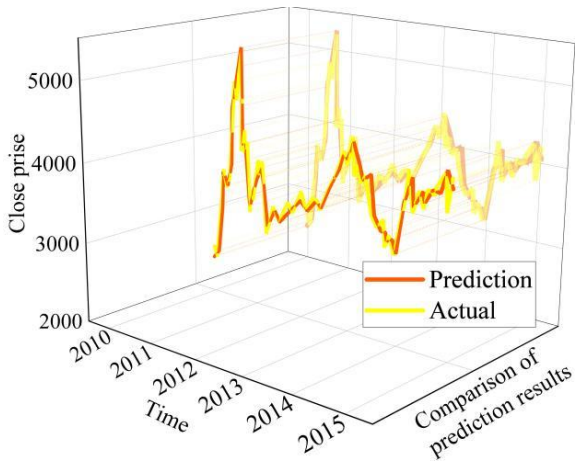
The financial and economic indices from January 2010 to March 2015 are selected to form the experimental dataset, and the acquired financial and economic index data are preprocessed using the methods proposed in Chapter II. A commonly used model in financial time series forecasting (K1) LSTM and a classical model in generative adversarial network model (K2) WGAN-gp are chosen to compare with (K3) model in this paper.The experimental results of the three models are shown in Fig. 2(a)-(c), and the error results are shown in Fig. 3(a)-(c). The predictive ability of the (K1) LSTM model is a little bit worse, its predicted data trend is slightly different from the real data trend, and the overlap rate is low. The predicted data of (K2)WGAN-gp and (K3)this paper's model are basically the same as the real data, but the details of (K2)WGAN-gp still can't match the real data completely, and the overlap rate is 87.23%. The predicted data of (K3)WGAN-gp and the real data achieve almost full-phase coincidence with the real data, with the coincidence rate of >99.00%, which preliminarily verifies that (K3)WGAN-gp is able to better capture the characteristics of the development of the financial data and better fit the real data. In terms of error performance, the prediction errors of (K1)LSTM and (K2)WGAN-gp models are in the range of (0,500), which are not only large, but also fluctuate frequently, with large fluctuations and weak stability. On the other hand, the prediction error of the (K3)WGAN-gp model is in the range of (0,300), with high prediction accuracy and superior prediction performance for financial data development.



(a) (K1) LSTM

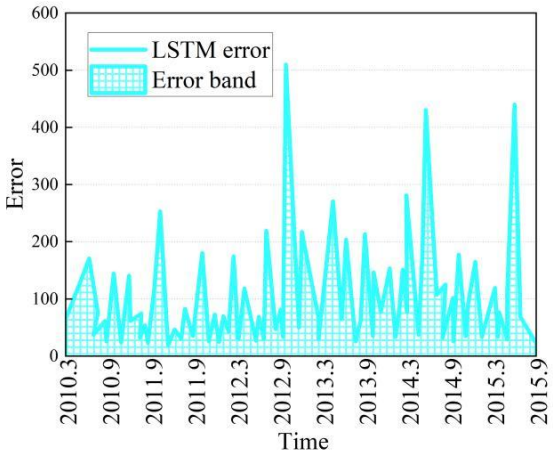


(b) (K2) WGAN-gp

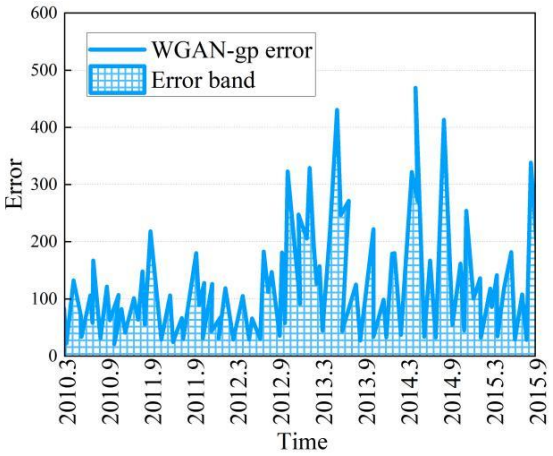


(c) (K3) Textual model

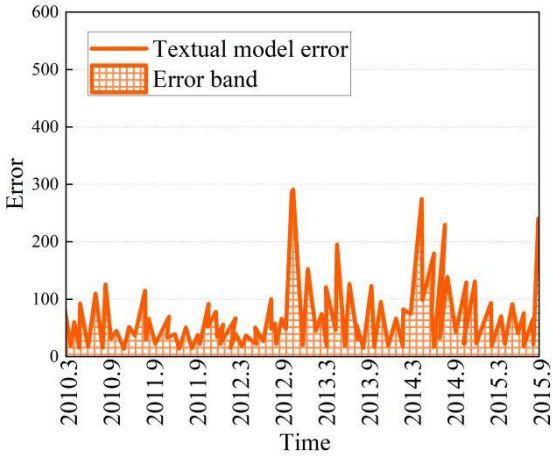
Figure 2: Comparison of prediction results from different methods



(a) (K1) LSTM



(b) (K2) WGAN-gp



(c) (K3) Textual model

Figure 3: Comparison of prediction errors of different methods

4.2 Denoising of data

Time-series decomposition using wavelets is utilized on the experimental data as a method of reducing noise. To achieve this, the db3 wavelet is chosen because with this wavelet basis, one gets good orthogonality, N-order vanishing moments, and excellent digital support, and its support length is 2N-1, which is not too large. Regarding the level of decomposition, the wavelet decomposition is performed at five levels. Such setting allows analyzing the data in sufficient detail without losing significant signal information due to the high number of decomposition layers. Figure 4 illustrates the variation trend of raw experimental data.

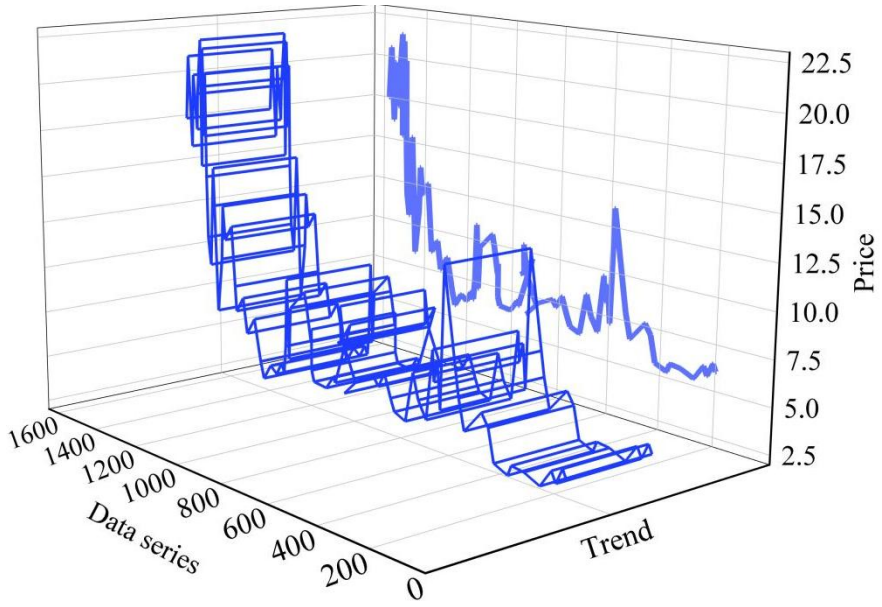
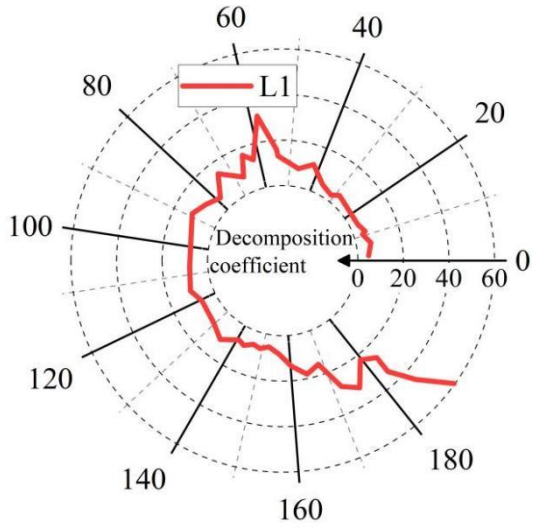
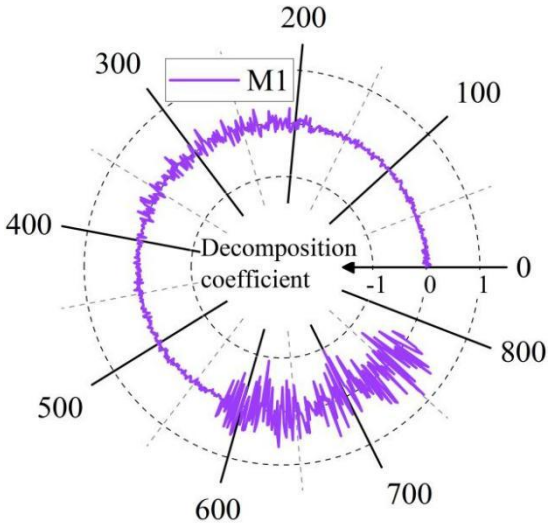


Figure 4: Original data trend

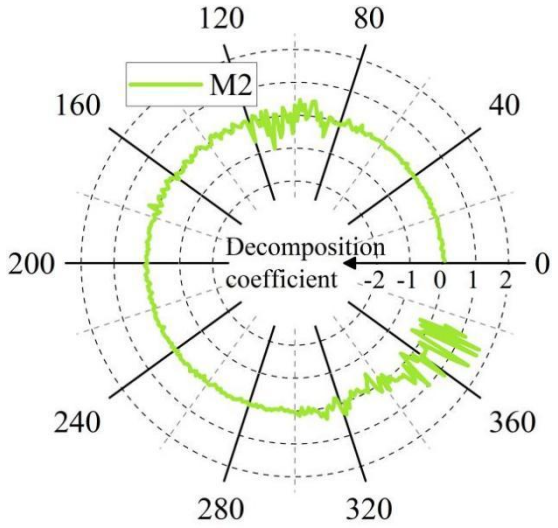
The original data are wavelet decomposed using the model in this paper, and the low-frequency coefficients L1 and high-frequency coefficients M1, M2 and M3 are obtained after decomposing the three layers, which are shown in Fig. 5(a)-(d) in turn.



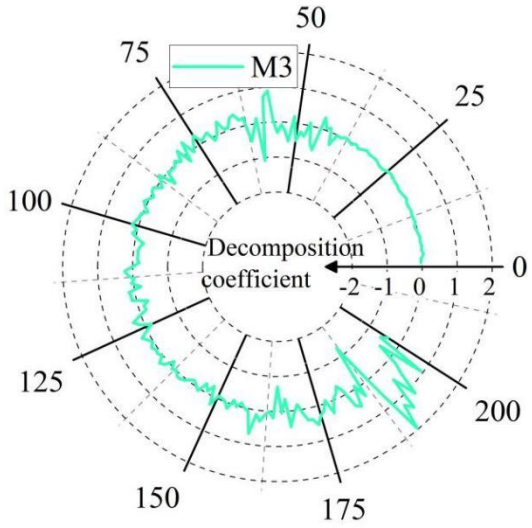
(a) Low-frequency coefficient L1



(b) High-frequency coefficient M1



(c) High-frequency coefficient M2



(d) High-frequency coefficient M3

Figure 5: Wavelet decomposition coefficient

A combination of fixed threshold and hard threshold denoising method is chosen to denoise the three high frequency coefficients obtained, and wavelet reconstruction is utilized to produce the new denoised data trend is shown in Fig. 6.

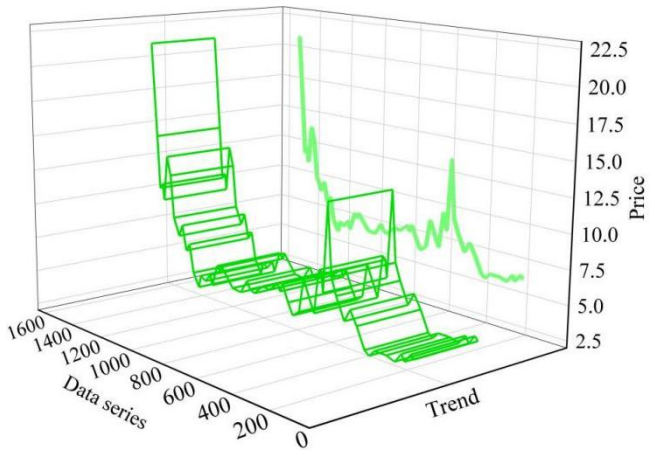


Figure 6: The trend after wavelet threshold denoising

Comparing Fig. 4 and Fig. 6, it can be clearly seen that the wavelet denoising processed data is smoother and also preserves the main fluctuation characteristics of the data, indicating that the model in this paper can achieve the purpose of denoising while retaining the useful information in the original signal. Therefore, in the next financial data analysis process, denoising the data can effectively improve the accuracy of modeling.

4.3 Visual presentation and analysis of data

4.3.1 Visualization process and effects

The annual reports of 127 listed companies in the experimental dataset are randomly selected, and (N1) current ratio, (N2) gearing ratio, (N3) accounts receivable turnover ratio, (N4) total asset turnover ratio, (N5) return on net assets, (N6) total asset compensation ratio, (N7) net profit growth rate, and (N8) total asset growth rate are used as the eight indicators for

assessing the financial status of the enterprise, which reflect the enterprise's The company's solvency, operating ability, profitability and growth ability. Meanwhile, in order to facilitate the user's investment decision, the following weights are set for each indicator: 0.1228, 0.1563, 0.1118, 0.1345, 0.1102, 0.1113, 0.1388, 0.1143.

After completing the pre-processing of the data, the K-means clustering algorithm is used for its clustering process, based on which the algorithm of this paper is further used for processing, the financial data processed by the algorithm of this paper is added with the transfer function Cs (0.05) and opacity, and the four phases of the changes in the financial data in the period are shown in Fig. 7. It can be seen that the original financial data is more cluttered and can not be analyzed and evaluated, and the clustered data is slightly more cluttered after the clustering process, which is not analyzed. After the clustering process, the data are slightly differentiated, but still difficult to recognize. The algorithm in this paper is able to clearly present the intrinsic connection between the financial data and assist the user in finding and searching the data. After setting the transfer function and opacity, the user can quickly and accurately find the region of interest, for example, the orange curve is mostly located in the upper position of the image, indicating that the overall financial situation is better, which is suitable for further investment decision analysis.

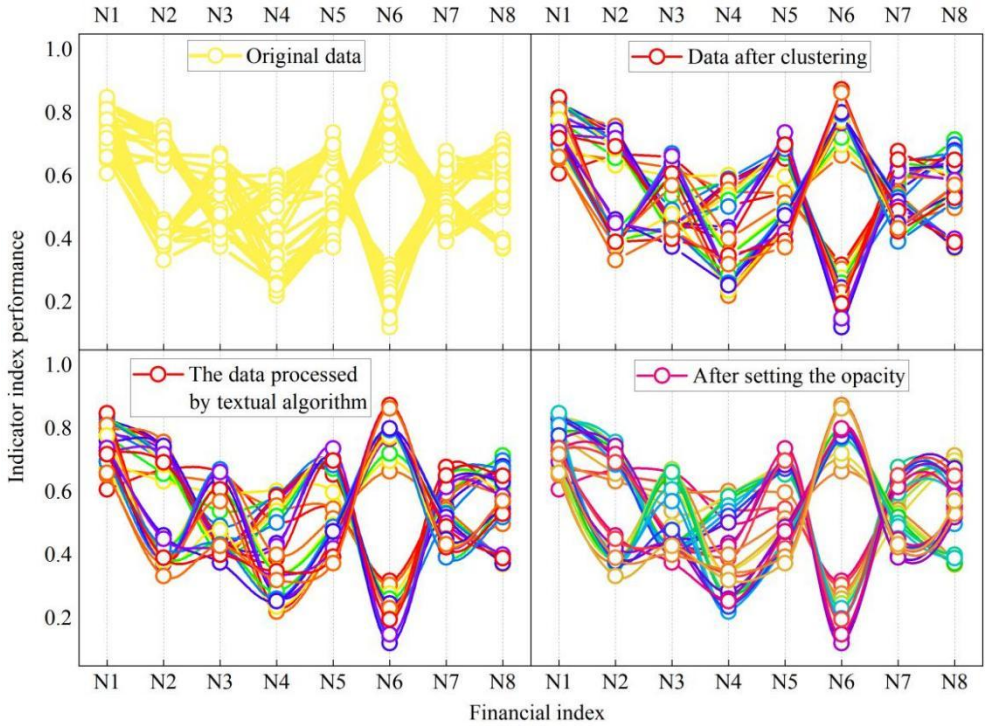
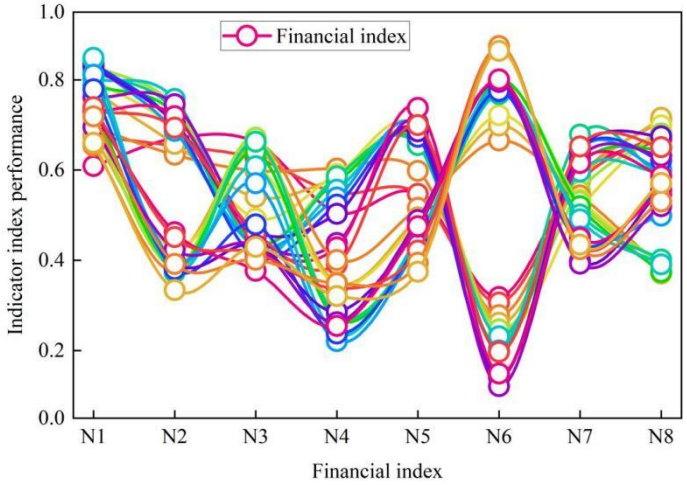


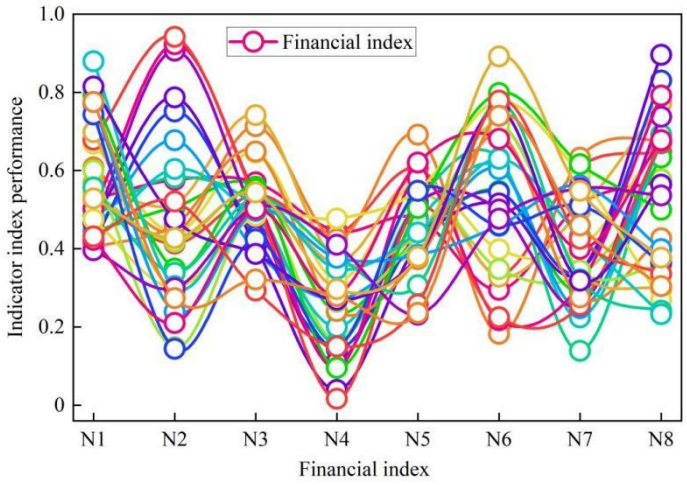
Figure 7: The display of financial data

4.3.2 Visualization enhancements

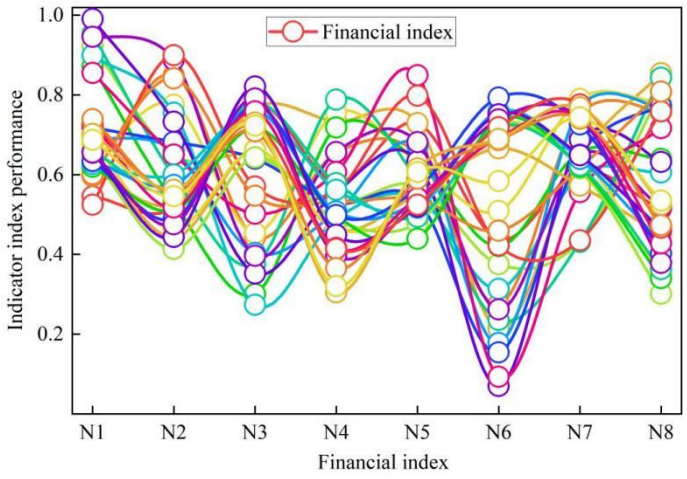
Appropriate settings of the transfer function C_s are able to cluster the interclass line segments, and the transfer functions of 0.05, 0.1, and 0.25 are selected, and the outputs of their corresponding effects are shown in Fig. 8(a)-(c). On the whole, when the transfer function C_s is 0.05, the effect of inter-class line segments clustering is better, when it is 0.1 or 0.25, the financial line segments between the varying degrees of dispersion, and the overlap between the class and the class is serious, and it is difficult to assist in the financial data of the discrimination, evaluation and even investment decision-making, so the selection of the optimal clustering 0.05 for the optimal parameter settings of the transfer function.



(a) C_s=0.05



(b) C_s=0.1



(c) C_s=0.25

Figure 8: The effects of different transfer functions

5 Conclusion

The present research uses trigonometric polynomial graph as one of the visualization methods to visualize financial data with the time-series feature. Also, the time-series-based prediction and clustering model of financial data is developed through the combination of wavelet analysis and the BIRCH-K-medoids algorithm. Comparative experiments on similar models indicate that the agreement rate between predicted values and actual values is greater than 99.00% and the prediction error is limited to the range of (0, 0.300) indicating much higher predictive performance compared to similar models. Following the denoising, the experimental financial data are smoother and their significant features remain intact. Additionally, in the visualization application, the best transfer function of the model is 0.05, which allows exact reproduction and explicit representation of the inherent relations found in financial data. Hence, the proposed strategy offers a viable way to enhance the level of value density in financial data and visual clustering in the big-data age.

Funding

This work was supported by Support from Guangxi Key Laboratory of Big Data in Finance and Economics (Grant No. FEDOP2022B03).

About the Author

Cong Xie was born in Luchuan County, Guangxi, China, in 1982. He received his undergraduate degree from Jiangxi Normal University, a Master’s degree from Guangxi University, and a PhD from Jose Rizal University. Currently, he works at Guangxi Police College. His research interests include intelligent algorithms, cybersecurity, and big data.

Wanzhao Zhao was born in Chongzuo County, Guangxi, China, in 1985. He received his undergraduate degree and a master’s degree from Nanchang Hangkong University.He works at Guangxi Vocational University of Agriculture. His research interests include artificial intelligence, Intelligent Applications and computer graphics.

Yi Zeng was born in Hezhou County, Guangxi, China, in 1982. She received her undergraduate degree from Guangxi Normal University and a Master’s degree from Uttar Pradesh Technical University. Currently, she works at Guangxi Vocational University of Agriculture. Her research interests include computer application, algorithms design and artificial intelligent.

References

[1] Slepov, V. A., Kosov, M. E., Chalova, A. Y., Gromova, E. I., & Voronkova, E. K. (2019). Integration of the financial market sectors: factors, risks and management approaches. International Journal of Civil Engineering and Technology, 10(2), 1243.

[2] Shen, D., & Chen, S. H. (2018). Big data finance and financial markets. In Big data in computational social science and humanities (pp. 235-248). Cham: Springer International Publishing.

[3] Yang, J., Zhao, Y., Han, C., Liu, Y., & Yang, M. (2022). Big data, big challenges: risk management of financial market in the digital economy. Journal of Enterprise

Information Management, 35(4/5), 1288-1304.

[4] Chen, H. (2021). Analysis of influencing factors of financial market volatility based on cluster analysis. Mobile Information Systems, 2021(1), 2313259.

[5] Zhuang, E., Small, M., & Feng, G. (2014). Time series analysis of the developed financial markets' integration using visibility graphs. Physica A: Statistical Mechanics and its Applications, 410, 483-495.

[6] Idrees, S. M., Alam, M. A., & Agarwal, P. (2019). A prediction approach for stock market volatility based on time series data. Ieee Access, 7, 17287-17298.

[7] Gidea, M., & Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. Physica A: Statistical mechanics and its applications, 491, 820-834.

[8] Lahmiri, S. (2016). A variational mode decompoisition approach for analysis and forecasting of economic and financial time series. Expert Systems with Applications, 55, 268-273.

[9] Chen, J. F., Chen, W. L., Huang, C. P., Huang, S. H., & Chen, A. P. (2016, November). Financial time-series data analysis using deep convolutional neural networks. In 2016 7th International conference on cloud computing and big data (CCBD) (pp. 87-92). IEEE.

[10] Zhang, P., Shi, X., & Khan, S. U. (2018). QuantCloud: enabling big data complex event processing for quantitative finance through a data-driven execution. IEEE Transactions on Big Data, 5(4), 564-575.

[11] Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2022). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. Journal of Artificial Intelligence and Big Data, 2(1), 153-164.

[12] Li, Q., Li, J., Sheng, J., Cui, S., Wu, J., Hei, Y., ... & Yu, P. S. (2022). A survey on deep learning event extraction: Approaches and applications. IEEE Transactions on Neural Networks and Learning Systems, 35(5), 6301-6321.

[13] Phyu, S., Li, W., Liu, Q., & Zhu, H. (2024, July). A Deep Learning Approach for Document-level Chinese Financial Event Extraction. In Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering (pp. 88-95).

[14] Wujec, M. (2021). Analysis of the financial information contained in the texts of current reports: A deep learning approach. Journal of Risk and Financial Management, 14(12), 582.

[15] Bach, M. P., Krstić, Ž., & Seljan, S. (2019). Big data text mining in the financial sector. In Expert systems in finance (pp. 80-96). Routledge.

[16] Tharaniya, B., Liyanapathirana, C., Rupasinghe, L., & Sampath, K. K. (2018).

Extracting Unstructured Data and Analysis and Prediction of Financial Event Modeling. In In Conference proceedings of the Annual Conference IET (pp. 6-11).

[17] Shang, C., Panangadan, A., & Prasanna, V. K. (2015, August). Event extraction from unstructured text data. In International Conference on Data Management in Cloud, Grid and P2P Systems (pp. 543-557). Cham: Springer International Publishing.

[18] Yang, H., Chen, Y., Liu, K., Xiao, Y., & Zhao, J. (2018, July). Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In Proceedings of ACL 2018, System Demonstrations (pp. 50-55).

[19] Cheng, W. K., Bea, K. T., Leow, S. M. H., Chan, J. Y. L., Hong, Z. W., & Chen, Y. L. (2022). A review of sentiment, semantic and event-extraction-based approaches in stock forecasting. Mathematics, 10(14), 2437.

[20] Lu, Y., Lin, H., Xu, J., Han, X., Tang, J., Li, A., ... & Chen, S. (2021). Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. arXiv preprint arXiv:2106.09232.

[21] Deng, H., Zhang, Y., Zhang, Y., Ying, W., Yu, C., Gao, J., ... & Zhou, T. (2022). 2event: Benchmarking open event extraction with a large-scale chinese title dataset. arXiv preprint arXiv:2211.00869.

[22] Wan, Q., Wan, C., Xiao, K., Hu, R., & Liu, D. (2023). A multi-channel hierarchical graph attention network for open event extraction. ACM Transactions on Information Systems, 41(1), 1-27.

[23] Liu, X., Huang, H. Y., & Zhang, Y. (2019, July). Open domain event extraction using neural latent variable models. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 2860-2871).

[24] Zheng, S., Cao, W., Xu, W., & Bian, J. (2019). Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. arXiv preprint arXiv:1904.07535.

[25] Du, X., & Cardie, C. (2020). Event extraction by answering (almost) natural questions. arXiv preprint arXiv:2004.13625.

[26] Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020, November). Event extraction as machine reading comprehension. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 1641-1651).

[27] Naveed, M., Ali, S., Iqbal, K., & Sohail, M. K. (2020). Role of financial and non-financial information in determining individual investor investment decision: a signaling perspective. South Asian Journal of Business Studies, 9(2), 261-278.

[28] Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. Sustainability, 11(5), 1277.

[29] Yang, N. (2022). Financial big data management and control and artificial intelligence analysis method based on data mining technology. Wireless Communications and

Mobile Computing, 2022(1), 7596094.

[30] Guo, Y., Fei, R., Zhang, K., Tang, Y., & Hu, B. (2020, December). Developing a clustering structure with consideration of cross-domain text classification based on deep sparse auto-encoder. In 2020 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 2477-2483). IEEE.

[31] Carta, S., Consoli, S., Piras, L., Podda, A. S., & Recupero, D. R. (2021). Event detection in finance using hierarchical clustering algorithms on news and tweets. Peerj computer science, 7, e438.

[32] Wang, J., Tan, J., Jin, H., & Qi, S. (2021, December). Unsupervised graph-clustering learning framework for financial news summarization. In 2021 International Conference on Data Mining Workshops (ICDMW) (pp. 719-726). IEEE.

[33] de Oliveira, A. D., Pinto, P. F., & Colcher, S. (2020, October). Stocks Clustering Based on Textual Embeddings for Price Forecasting. In Brazilian Conference on Intelligent Systems (pp. 665-678). Cham: Springer International Publishing.

[34] Sidorov, S. P., Faizliev, A. R., Levshunov, M., Chekmareva, A., Gudkov, A., & Korobov, E. (2018, September). Graph-based clustering approach for economic and financial event detection using news analytics data. In International Conference on Social Informatics (pp. 271-280). Cham: Springer International Publishing.