Applied Mathematics and Nonlinear Sciences

<https://www.sciendo.com>

Application of mathematical probabilistic statistical model of base – FFCA financial data processing

Zhengqing Li^{1†}, Jiliang Mu², Mohammed Basher³, Hafnida Hasan⁴

¹University of Sanya Hainan, Sanya, 570200, China

²North University of China, School of Instrument and Electronics, Shanxi, Taiyuan, 03000, China

³Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

⁴Applied Science University, Al Eker, Kingdom of Bahrain

Submission Info

Communicated by Juan Luis García Guirao

Received June 16th 2021

Accepted September 24th 2021

Available online December 15th 2021

Abstract

In order to improve the detection and filtering ability for financial data, a data-filtering method based on mathematical probability statistical model, a descriptive statistical analysis model of big data filtering, probability density characteristic statistical design data filtering analysis combined with fuzzy mathematical reasoning, regression analysis according to probability density of financial data distribution, and threshold test and threshold judgment are conducted to realize data filtering. The test results show that the big data filtering and the reliability and convergence of the mathematical model are optimal.

Keywords: mathematical probability model, data filtering, financial data, inspection statistics.

AMS 2010 codes: 91-04

1 Introduction

As the size of the big data information is increasing, big data containing different types of information require effective filtration and separation of different big data, improvement in the storage and management capabilities of data, optimization of the storage space of data, and improvement in large data information identification. Ability research on information classification methods and filtration methods for big data is of great significance in large database construction and cloud storage model design.

[†]Corresponding author.

Email address: lzqsy000@126.com

The filter model for data is based on mathematical statistical analysis and applies the statistical regression analysis model of data filtration. Using probability distribution and adaptive fuzzy clustering analysis methods, we optimize the filtration of data. Typical large data filtration models mainly include quantitative regression analysis model, autoregressive moving average (ARMA) model, correlation statistical analysis model, Backlund model etc. Based on the design of the probability mathematical model, combined with the pattern identification and the characteristic clustering method, data filtering is done; but the traditional method has a large interference with mode identification during data filtration. In response to the above problem, this paper proposes a data filtration method based on the probability mathematical model, and the statistical characteristic analysis model for large data filtration is constructed, combined with mathematical modeling and statistical analysis methods, and the data filter probability mathematical model is optimized. Finally, data analysis is performed to obtain conclusions regarding the effectiveness of the model [1, 2].

Financial data are typical time series data, which contain financial data, exchange rate interest rates, futures prices, etc. and various forms of one-dimensional or multidimensional data, which are characterized by nonlinearity of dynamics. Analysis of financial data is not only a hot problem of social hotspots but also a subject of academic community research. For decades, a large number of scholars have made great contributions to the development of the study, whether economics, statistics, or computer science. There are many kinds of financial data, such as stock, futures, exchange rates, etc., characterized by nonlinear dynamic changes, caused also because of the social development and economic policies of various countries. Correspondingly, a variety of research methods are more likely. Data are endless, and when they are financial, statistical or computer scholars hope to establish a relatively reliable model through a large amount of financial data to predict and analyze financial markets, thereby reducing investors' mistakes in transaction decision-making, avoiding operational risks.

1.1 More economic benefits

At the same time, these models explore the internal law of the financial market, providing an efficient solution for managers and decision-makers in the financial market to improve the market economy system. The statistical model of real-world data usually involves continuous probability distribution, such as normal, Laplace, or index distribution. This distribution is supported by many probabilistic modeling forms, including probability database systems. However, the traditional theoretical framework of the probability database is completely concerned with the limited probability database [3]. Yue et al. [4] developed mathematical theories in unlimited probability databases. In a previous paper (GROHE, LINDNER; ICDT 2020), a very general framework is proposed for the probability database that may involve a continuous probability distribution; the query has a clear definition in this framework. Extending the probability programming language to generate data records (PODS 2020) to achieve a continuous probability distribution and displaying such programs generates a continuous probability database generation model [4]. Zhang [5] has developed a new way to have a nonprogrammetry of boundaries to accurately surround all uncertain parameter data extracted in the actual project. This method is based on conventional statistical methods and related analysis techniques. First, the average value and the correlation coefficient of the uncertain parameter are calculated by using all given data.

Then, first, a simple but effective optimization process is introduced in the mathematical modeling process to induce the uncertain parameters to obtain their exact boundaries. This process operates on all of the given data by optimizing the convex mode. Thus, an effective mathematical expression of the final formulation of the convex mode is obtained. In order to test the predictive capacity and generalization ability of the proposed convex mode, the evaluation criteria used are volume ratio, standard volume ratio, and prediction accuracy. The performance of the proposed method is studied by testing standards and comparing with other existing competition methods. The results prove the effectiveness and efficiency of the method [5]. Herrera-Vega et al. [6] propose a probabilistic hesitant financial data (PHFD)-based portfolio selection and risk portfolio selection model, i.e., probabilistic hesitant portfolio selection (PHPS) model and risk PHPS (RPHPS) model, of risk probability hesitation.

In addition, investment decision methods are provided to show their practical applications in the financial

market. It is expected that the PHPS model of general investors is built based on the highest score or minimum deviation principle to obtain the best investment ratio, and the RPHPS model provides the best investment ratio for the best return or to undertake risk. Finally, an empirical study based on the actual data of China's stock market is shown in detail. The results verify the effectiveness and practicality of the proposed method [6]. Li et al. used mathematical statistical methods to study the effectiveness of single-stage and multistage hydraulic fracturing technology. These studies are based on the geological and physical properties of the circulation of the prime wall, which makes it possible to demonstrate the choice of hydraulic fracturing technology.

Based on the probability and statistical analysis of the field efficiency data, quantitative standards selected by the crack parameters of propanes by length, height, and the injection of propane have been established. They can be used to predict the use of multistage and single-stage hydraulic fracturing methods. High-efficiency horizon AS12-3 has complex geological and physical structures and low filtering and storage performance, making it complicated, and contributes to the active formation of the reserves that are difficult to recover. Single-stage hydraulic fracturing technology is only 3–4 years old. Hydraulic fracturing technology with single-stage hydraulic fracturing technology is more effective [7].

2 Research method

2.1 Mathematical probability statistics model – preparatory knowledge

The mathematical modeling analysis of data filtration is performed by the probability mathematical model. First, the sample descriptive statistical analysis method is used to analyze large data samples, the sampled major households are screened, and the statistical feature of the characteristic attribute that can effectively reflect the data filtration is determined. Sample regression analysis and adaptive processing on the sampling data are performed; the combined information fusion method implements the pattern identification of data filtration, and regression analysis is conducted according to the probability density of the detection statistic distribution in massive data environments; thus, data filtration according to the mode identification results. According to the above analysis, the overall structure of the probability mathematical model of data filtration is shown in Figure 1.

According to the probability mathematical model shown in Figure 1, the mathematical information flow analysis of data filtration is performed, and the sample time series of the sampled data information stream is assumed. $Y_1, Y_2, \dots, Y_N$. Let $A_j(L)$ for the fuzzy cluster center of the data probability statistics model, $j = 1, 2, \dots, k$; the statistical regression analysis method was used, and the descriptive statistical analysis feature distribution matrix of massive data filtration was obtained from $K = [K_1^T K_2^T K_3^T K_4^T K_5^T]^T$, $L = [L_1^T L_2^T L_3^T L_4^T L_5^T]^T$, $M = [M_1^T M_2^T M_3^T M_4^T M_5^T]^T$, and $W = [W_1^T W_2^T W_3^T W_4^T W_5^T]^T$, according to the above-described overall model design, using a descriptive statistical analysis method to construct a large data filtration statistical analysis model.

2.2 Descriptive statistical analysis

Data filtration statistical analysis in Leslie–Gower space is performed. The associated property rule analysis method is used, and the algorithm for data filtration is designed. The feature extraction algorithm for data filtration is described as follows.

Algorithm 1

Enter: construct data filtering table with minimum multiplining method $S = (U, A, V, f)$, $U = \{u_1 u_2 \dots, u_n\}$, $A = \{a_1, a_2, \dots, a_m\}$. Output: $S = (U, A, V, f)$, indicating the information entropy of data filtration, the descriptive statistic $\lim_{n \rightarrow \infty} \|X_n - X^*\| = 0$ of the output, where $\|\cdot\| = 0$ indicates a certain number. The characteristic sampling steps of data filtration are as follows:

Step 1: $n \leftarrow |U|$, $matrix \leftarrow \Phi$, $m \leftarrow |A|$;

Step 2: $s \leftarrow 1$, $d(s) \leftarrow \Phi$;

Fig. 1 Overall design of the probability mathematical model of data filtration.

Step 3: $i \leftarrow 1$;

Step 4: $j \leftarrow i + 1$;

Step 5: $k \leftarrow 1$;

Step 6: *if* $f(i, a_k) \neq f(j, a_k)$ *then* $matrix(i, j) \leftarrow matrix(i, j) + a_k$;

Step 7: $k \leftarrow k + 1$, using probability density feature extraction methods, high-order differential equations are constructed, and when statistical feature amount $k < m$, move to Step 6. In the limited weight Morrey–Herz convex space, the global regular fuzzy decision vector set of data filtered $d(s) \leftarrow matrix(i, j)$, $s \leftarrow s + 1$, move to Step 8;

Step 8: $j \leftarrow j + 1$, if $j \leq n$, move to Step 5, correct the verification statistic, convert the data filtering problem into the probability density function, and go to Step 9;

Step 9: $i \leftarrow i + 1$, if $i \leq n - 1$, turn to Step 4, through multilevel distribution decision, data feature extraction, if convergence conditions are met, turn to Step 10;

Step 10: Fusion decisions are performed within a limited field $U = \{x_1, x_2, \dots, x_m\}$ of the data feature distribution space, and the conditional attribute element is spaced;

Step 11: Descriptive statistics matrix for output data filtration [8].

2.3 Hidden Markov model (HMM) prediction method for mathematical probability statistics model

2.3.1 Choice of raw data

Looking for a series of factors affecting financial data change as an input of the HMM, i.e., observed sequences. The observed value belongs to the mixed Gaussian distribution; then:

$$b_j(o_i) = \sum_{k=1}^M w_{jk} b_{jk}(o_i), j = 1, 2, \dots, N, 1 \leq k \leq M \quad (1)$$

where m represents the number of mixed components; w_{jk} indicates the weight of each mixed component under each hidden state and is represented as follows:

$$w_{jk} \geq 0, \sum_{k=1}^M w_{jk} = 1 \quad (2)$$

For $b_{jk}(o_t)$, the common Gaussian distributions are as follows:

$$b_{jk}(o_t) = N(o_t, \mu_{jk}, \sigma_{jk}^2) \quad (3)$$

where μ_{jk} indicates the mean of the Gaussian distribution, and σ_{jk}^2 indicates the difference between the Gaussian distributions. Gaussian distribution has a wide range of applications, such as medical statistics, height statistics, quality testing, etc., but there is no sufficient theoretical basis and scientific guidance for financial data. Because financial data are not naturally formed, they are affected by various factors, state policies, bank deposit interest rates, industry factors, etc. Gaussian distribution cannot accurately simulate financial data probability distribution. Although many experiments use Gaussian distribution as a probability density function of continuous observations and achieve better experimental results, it is also limited to further improve prediction accuracy. In addition, the mixed Gaussian model also increases the difficulty of constructing HMM, and the parameters m and w_{jk} need to be estimated by expectation–maximization (EM) algorithm, increasing the workload. In addition to direct prices as the observations, the stock price-related indicator function can be used as the input of HMM. Select the average real range, random momentum index closing price position value, smooth and rich indexes, average line, parabolic stop loss and volatility indicator, average tendency index, and cloth belt. Cai jin wave indicator, amu simple fluctuation indicator, fund flow index and other technical indicators are screened using random forests. Select a number of factors such as interest rate, fund flow average, moving average, and basic morphology and find the most suitable factor combination by multiple factors [9].

2.4 Probability mathematical model construction – optimization of data filtration

On the basis of constructing a statistical characteristic analysis model for large data filtration, the optimization design of data filtering model is performed, and this paper proposes a data filtration method based on the probability mathematical model and gives data filtration probability density feature distribution mapping $f: U \times A \rightarrow V$ in order to build data filtration in the phase space; the information function $f(x, a)$ is the value on the element A in the resolution matrix, and the global registered gradient diversity method is used, and the core function of the data filter is $X \subset U$, the probability density is domain $POS_A^*(X)$, the negative domain is $NEG_A^*(X)$; in the boundary arrest, the feature distribution set $BDN_A^*(X)$ of massive data is obtained, whereby the fuzzy mathematical reasoning attribute set of data filtration is represented as follows:

$$POS_A^*(X) = \cup \{E | P(X|E) \rangle P(X), E \in U/A\} \quad (4)$$

$$NEG_A^*(X) = \cup \{E | P(X|E) \langle P(X), E \in U/A\} \quad (5)$$

$$BDN_A^*(X) = \cup \{E | P(X|E) = P(X), E \in U/A\} \quad (6)$$

In the formulas $P(X|E) = \frac{P(X)}{P(E)} = \frac{card(X \cap E)}{card(E)}$, $P(X) = \frac{card(X)}{card(U)}$, $card(X)$, indicating the probability density feature statistic in line data filtration [10].

The information entropy feature extraction is performed according to the association of the attribute. The scalar data time sequence attribute of data filtering is $D_x = 2$, and the regression analysis of the probability density distribution of the detection statistics in mass data is $\xi_{c_1}^{d_2} = \frac{3}{5}$, $\xi_{c_2}^{d_2} = \frac{2}{5}$, $\max g_{c_2}(d_2) = \frac{3}{8}$, $\max g_{c_3}(d_2) = \frac{1}{10}$.

Assuming that the decision threshold of X 's data filtering is $D_x = 3$, the reference component of the feature statistic of the data filtered is $\xi_{c_1}^{d_3} = 1$, $\xi_{c_2}^{d_3} = 1$, $\xi_{c_3}^{d_3} = 1$, $\max g_{c_1}(d_3) = \frac{7}{4}$, $\max g_{c_2}(d_2) = \frac{3}{8}$, $\max g_{c_3}(d_2) = \frac{7}{4}$.

According to the results of the decision statistical regression analysis, data filtration is achieved by the decision attribute identification method, and the filter output is as follows:

$$\begin{aligned}
 m_1^{(1)}(d_1) &= 0.2027, m_1^{(1)}(d_2) = 0.1622, m_1^{(1)}(d_3) = 0.2703; \\
 m_1^{(1)}(\Theta) &= 0.3649; \\
 m_2^{(1)}(d_1) &= 0.1786, m_2^{(1)}(d_2) = 0.1429, m_2^{(1)}(d_3) = 0.3571; \\
 m_2^{(1)}(\Theta) &= 0.3214; \\
 m_3^{(1)}(d_1) &= 0.2000, m_3^{(1)}(d_2) = 0.1333, m_3^{(1)}(d_3) = 0.3333, m_3^{(1)}(\Theta) = 0.3333; \\
 m_1^{(2)}(d_1) &= 0.1512, m_1^{(2)}(d_2) = 0.1708, m_1^{(2)}(d_3) = 0.2491; \\
 m_1^{(2)}(\Theta) &= 0.4288; \\
 m_2^{(2)}(d_1) &= 0.3000, m_2^{(2)}(d_2) = 0.3000, m_2^{(2)}(d_3) = 0.3000; \\
 m_1^{(2)}(\Theta) &= 0.1000; \\
 m_3^{(2)}(d_1) &= 0.1625, m_3^{(2)}(d_2) = 0.0250, m_3^{(2)}(d_3) = 0.4375; \\
 m_3^{(2)}(\Theta) &= 0.3750.
 \end{aligned}$$

The probability mathematical model of data filtration is constructed using a fuzzy decision method, using a threshold test and threshold decision method, to achieve large data filtration, and the implementation process of the algorithm is as follows:

Algorithm 2

1. //Set the primary value of the probability mathematical model of data filtering

2.

$$X(1) = X, Y(1) = Y, N(1) = n, i = 1$$

3.

$$\text{while } N(i) \geq L$$

4. //Calculate data characteristics positioning matrix to get a fuzzy decision power index $p(i)$

5.

$$\text{if } N(i) \bmod L \neq 0$$

6.

$$\text{then } p(i) = N(i) / L$$

7. else

8.

$$p(i) = N(i) / L + 1$$

9. //With an independent main component analysis method, the fuzzy decision and probability statistics of data filtering are obtained, and the fuzzy identification matrix of data filtration output is obtained $X(i)$ and $Y(i)$.

10.

$$X_{ij} = X(i) [(j-1+1 : L^*j)]$$

11.

$$Y_{ij} = Y(i) [L^*(J-1) + 1 : L^*j]$$

12. //use SVD breakout operation method, power index statistics, execute the following cycle.

13. *for* $j = 1$ *to* $p(i)$

14. $X_{ij} = U_{ij} \Sigma_{ij} V_{ij}^T$

15. End

16. Feature statistic for data filtering output $X(i+1), Y(i+1)$

17. $i = i + 1$

18. //End of iteration, realize the statistical regression analysis of large data filtration.

19. $k = i - 1$

20. $X(k) = U(k) \Sigma(k) V(k)^T$

21. Seeking the main feature β^* , determine whether the convergence condition is met, if not satisfied, go to Step 10.

22. If the convergence condition is met, the algorithm ends.

3 Analysis of results

In order to test the application performance in data filtration, the data set of data sample distribution is database2016; the test data set is CSLOGS SET2, and the data sample number is $N = 1024$, the data set of the test sample is 100. The time interval of the original data sampling is 0.02 s; the classification attribute set distribution of data filtration is shown in Table 1.

Table 1 Classification attribute set of data filtration

Discourse domain	u	c_1	c_2	c_3	$\mathbf{D}$
x_1	53	2	2		
3	6	3	3		
3	4	4	3		
x_4	53	6	3		
3	2	5	4		
x_6	25	4	5		
6	6	5	4		
6	5	2	3		
7	3	1	4		
5	4	1	5		

According to the statistical rules of the large data sheet described above, the data are filtered, and data filtration is obtained as shown in Table 2.

Table 2 Detection statistics rules table for data filtration

Regular	D			
1	4	*	2	1
2	3	*	3	5
3	6	*	2	4
4	*	4	6	4
5	3	*	5	5
6	4	4	*	4
7	5	*	6	5
8	3	4	6	6

According to the statistical rule table, based on the above data filtration, data filtration is achieved, and the error rate of filtration is obtained, e.g., as shown in Figure 2; Figure 2 is analyzed and the accuracy of data filtration in this method is derived; the error is low.

Fig. 2 Performance comparison of data filtration.

4 Conclusion

Effective filtration and separation of different types of big data improve data storage and management capabilities, optimize data storage space, and improve large data information recognition capability. This paper proposes a data filtration method based on the probability mathematical model, using descriptive statistical analysis method, the statistical feature analysis model of large data filtration, combined with mathematical modeling and statistical analysis, implementation of data filter probability mathematical model, using threshold test and threshold decision methods to implement data filtering. Studies show that the method of data filtration is better and the error is low.

Acknowledgment: Hainan Provincial Natural Science Foundation of China (Research on financial data processing based on FPGA; grant no. 619QN244).

References

- [1] Zeng M, Hao H, Zhou H. Super parametric convex model and its application for non-probabilistic reliability-based design optimization[J]. *Applied Mathematical Modelling*, 2017, 55(MAR.):354-370.
- [2] T, VAN, HECKE. PROBABILISTIC MODEL OF CRYPTOGRAPHIC KEY COLLISIONS IN RELATION TO THE KEY LENGTH[J]. *The Mathematical Scientist*, 2016, 41(1):21-24.
- [3] Chudoba R, Li Y, Rypl R, et al. Probabilistic multiple cracking model of brittle-matrix composite based on a one-by-one crack tracing algorithm[J]. *Applied Mathematical Modelling*, 2021, 92(3):315-332.
- [4] Yue N, Xie J, Chen S. Some new basic operations of probabilistic linguistic term sets and their application in multi-criteria decision making[J]. *Soft Computing*, 2020, 24(16):12131-12148.
- [5] Zhang M, Yao Q, Sun S, et al. An efficient strategy for reliability-based multidisciplinary design optimization of twin-web disk with non-probabilistic model[J]. *Applied Mathematical Modelling*, 2020, 82(Jun.):546-572.
- [6] Herrera-Vega J, F Orihuela-Espina, PH Ibargüengoytia, et al. A local multiscale probabilistic graphical model for data validation and reconstruction, and its application in industry[J]. *Engineering Applications of Artificial Intelligence*, 2018, 70(APR.):1-15.
- [7] Li Y, Wang X, Wang C, et al. Non-probabilistic Bayesian Update Method for Model Validation[J]. *Applied Mathematical Modelling*, 2018, 58(JUN.):388-403.
- [8] Rosas-Carbajal M, Lin De N, Peacock J, et al. Probabilistic 3-D time-lapse inversion of magnetotelluric data: Application to an enhanced geothermal system[J]. *Geophysical Journal International*, 2015, 203(3):1946-1960.
- [9] Ziane D, Cherif M H, Cattani C, et al. Yang-Laplace Decomposition Method for Nonlinear System of Local Fractional Partial Differential Equations[J]. *Applied Mathematics and Nonlinear Sciences*, 2019, 4(2):489-502.
- [10] Ahin R, Yagci O. Fractional Calculus of the Extended Hypergeometric Function[J]. *Applied Mathematics and Nonlinear Sciences*, 2020, 5(1):369-384.

This page is intentionally left blank