Research on the design and implementation of a financial data analysis system based on the background of big data

Xinchen Shi*

Faculty of Customs and Public Economics, Shanghai Customs College, Shanghai, China

{*18757211867@163.com}

Abstract. This paper focuses on the system design and research of the financial system based on the new big data context. The system design and implementation research is carried out on the basis of the development of the current situation of the company combined with certain system optimisation exit functions. Financial analysis is an important piece of information for decision makers and investors in companies. Traditional methods of financial analysis include quantitative modelling and textual analysis. Quantitative model analysis is the analysis of financial data through the use of statistical analysis tools or artificial intelligence techniques. The models rely on selected key indicators such as financial ratios, technical indicators and macroeconomic indicators. Text analysis is the contextual analysis of the content of financial reports using text mining methods that rely on the identification of predefined sets of keywords. Although there are commonalities in the indicators and keywords chosen by different researchers in their studies, the choice of indicators and keywords differs due to personal preferences and subjective and objective circumstances, so the results of the analysis of these two types of methods are often subjective.

Keywords: Big data; The financial system; Method of implementation; Optimisation studies

1 Introduction

In reality, financial analysis indicators are often high-dimensional, often with a lot of redundant information, but in order to ensure the integrity of the information, analysts are often difficult to choose, these complicated indicators often obscure or disrupt the presentation of the essential characteristics of the financial situation [1]. Quantitative Structure-Property Relationship (QSPR) is an important research method in the natural sciences. The core idea of QSPR is that the microstructure of a substance determines its macroscopic properties [2]. This led to the inspiration to further discover the overall characteristics of the market by identifying the intrinsic relationships between microscopic financial data points. Stream shape learning aims to discover the low-dimensional stream shape structure embedded in high-dimensional data, which is exactly the approach needed for our research. In the field of financial analysis, the information characteristics of data, such as probability distributions, are more important and

© The Author(s) 2023
C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_2

meaningful than their geometric characteristics. However, existing streamform learning algorithms basically consider the spatial geometric structure of the data. If the formation of a sub-flow shape is limited by its probability density function, then it is inaccurate to describe the flow shape distance by the shortest spatial distance between data points. For the financial dataset, each data point represents a listed company, and the distance between each data point characterizes the degree of dissimilarity between the financial positions of two listed companies. If this degree of dissimilarity is expressed only by the geometric spatial distance between data points, it may not only be unsuitable for the practical meaning of financial analysis, but also lead to large errors in the subsequent analysis. Therefore, we use the information distance to measure the relationship between listed companies, and thus obtain the relationship metric model. It has been shown that the relationship between financial data points is nonlinear, and the existing stream learning algorithms capture the nonlinear relationship in high-dimensional space and then obtain the low-dimensional coordinate representation through linear mapping, which is tantamount to destroying the eigenrelationships in the original data set again. Some scholars have proposed Fisher's information embedding flow learning algorithm (FINE), information preserving principal component analysis (IPCA), and information maximizing principal component analysis (IMCA). Although these flow learning algorithms maintain the information metric between data points, they all use linear mapping functions when mapping from high-dimensional space to low-dimensional Euclidean space, which cannot correctly output the topology of the original data logical relationships of the original data. Qiao Hong et al. proposed a nonlinear and explicit mapping method, but the method is computationally intensive.

2 Kernel entropy-based streaming learning algorithm for financial data (KEML)

After applying the information distance to measure the difference between financial data points, we map the data points into a linear feature space by means of a kernel function, and finally use a linear mapping function to obtain a low-dimensional coordinate representation. The above algorithm is a KernelEntropy Manifold Learning (KEML) algorithm proposed in this paper, which uses the method of manifold learning to obtain low-dimensional inlays of high-dimensional financial data [3]. To prove the effectiveness of the algorithm, we selected financial data from the financial annual reports of manufacturing SMEs in the Chinese A-share stock market from 2006 to 2010 and conducted simulation experiments on the algorithm. The experiments were divided into two parts: financial status analysis and stock market volatility analysis, and the empirical analysis verified the effectiveness of the algorithm and provided a new objective analysis basis for financial practice analysis [4]. The technical route of the research in this paper is shown in Figure 1:

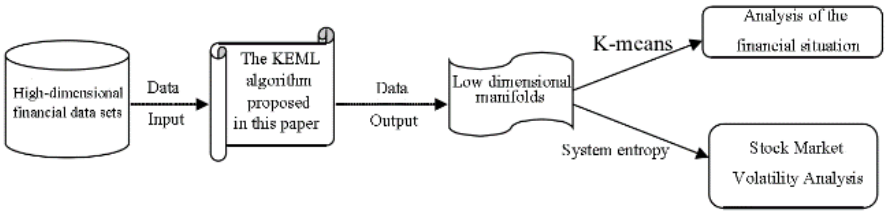


Fig. 1. Technology roadmap for text research

In this paper, we propose a kernel entropy-based manifold learning method (Kenerl Entropy ManifoldLearning, KEML) for extracting low-dimensional manifolds embedded in high-dimensional financial datasets [5]. We first construct a mathematical model of the state space of financial data points; use the information distance as the distance metric between data points in the mathematical model; map the financial data points represented by the information metric to the feature space by means of a kernel function; and finally use a linear mapping to obtain a low-dimensional coordinate representation [6].

Our study is based on a dataset of n listed companies $(X_1, X_2, \dots, X_n)$, with D financial indicators for each listed company X_i $(X_i^D, X_i^D, \dots, X_i^D)$. Thus each X_i is a dataset containing financial indicators, i.e. $X_i = (X_i^1, X_i^2, \dots, X_i^D)$. We define χ to be a cluster of data sets, whereupon we have:

$$\chi = \{X_1, \dots, X_i, \dots, X_n\}(i = 1, \dots n) \ \ and \ \ X_i = (X_i^1, X_i^2, \dots, X_i^D) \tag{1}$$

It is assumed that for each data set X_i there exists a potential probability distribution function p_i determined by the financial indicators. Can be defined:

$$S = \{p_i(x{:}\,\theta)|\theta \in \Theta, \mathrm{i} = 1, \dots, \mathrm{n}\} \tag{2}$$

where x is each data point in the state space X of financial data points, $p(x{:}\,\theta)$is the probability density function of x, and θ is a D-dimensional vector representing D indicators $\theta = (\theta^1, \theta^2, \dots, \theta^D)$, Θ for which D is the open set of the real space R^D. We thus obtain a statistical manifold S on which the set $P = \{p_1, p_2, \dots, p_n\}$ of probability distribution functions lies. In a statistical manifold S, each element is a probability distribution function p_1, and our task is to reconstruct a manifold π in density space using valid probabilistic information. The problem then translates into finding an embed:

$$A = p(x) \to y(y \in R^m, m < D) \tag{3}$$

Unlike traditional stream learning algorithms that are constructed in Euclidean space, our proposed approach discovers a low-dimensional embedding in density space.

3 Empirical studies

In this paper, we select Chinese A-share market data and examine the performance of the KEML algorithm through three simulation experiments. The experiments were conducted on the MATLAB 2010R platform [7].

3.1 Selection of data sets

We selected the annual financial data of manufacturing SMEs in China's A-share market from 2006 to 2010, excluding those with missing data and those that went bankrupt and delisted during this period, to obtain a total sample of 205 companies. Data from Wind Information Database [8].

According to modern financial theory, the above financial indicators have the following correlation: firstly, solvency and cash flow are positively correlated. This indicates that when a listed company has insufficient cash flow, its debt level is high and its solvency is weak. Secondly, there is a significant positive correlation between operating capacity, profitability and growth capacity [9].

3.2 Experimental design

The experiment consisted of three parts:

(1) The KEML algorithm was compared with six commonly used classical stream shape learning algorithms, namely KPCA, LTSA, LMDS, ISOMAP, LLE and PCA, and the quantitative metric Procrustes Measure (PM) was used to evaluate the feature extraction results obtained by these seven algorithms. PM is a quantitative metric of non-linearity. The smaller the PM value, the higher the accuracy of the obtained low-dimensional embedding, and PM is now widely used in the evaluation of embedding quality. A package on PM is available in MATLAB [10].

(2) The low-dimensional embedding obtained in experiment (1) is applied to the experiment on financial situation analysis. The clustering results are evaluated by F1-scores indicators and risk expectations respectively.

(3) The final part of the experiment was to apply the KEML algorithm to the characterisation of the overall stock market movements. Financial markets can be viewed as a highly complex evolutionary system, and each experimental data set constitutes a sub-system of that complex system. We have explored the embedded structure of this subsystem through KEML and will use relevant analytical methods to further investigate the operation of the system [11].

3.3 Experimental results

Experiment: Financial Position Early Warning Experiment

In the Chinese A-share market, the symbol "ST" or "*ST" means that the listed company is in an abnormal financial condition and is at risk of investment or delisting.

The results of low-dimensional embedding of a high-dimensional financial dataset were obtained in the experiments. In this paper, we apply the K-means clustering algorithm low-dimensional embedding for cluster analysis and identify companies with abnormal financial status. Here is a brief description of the K-means clustering method: assume that there are n data points$\{x_1, x_2, \dots x_n\}$,for $\{a_1, a_2, \dots a_k\}$ of the K clusters, the distance squared W_n from each data point to the nearest centroid is required to be minimized, and using this distance squared as the objective function, there are:

$$W_n = \sum_{i=1}^{n} min_{i \le i \le k} |x_i - a_j|^2 \tag{4}$$

The K-means algorithm is simple and runs efficiently for clustering of large data sets. According to the existing literature on the determination of the number of clusters K_{max},$K_{max} \le \sqrt{n}$, the best clustering effect and the number of clusters were taken separately for the experimental algorithm.

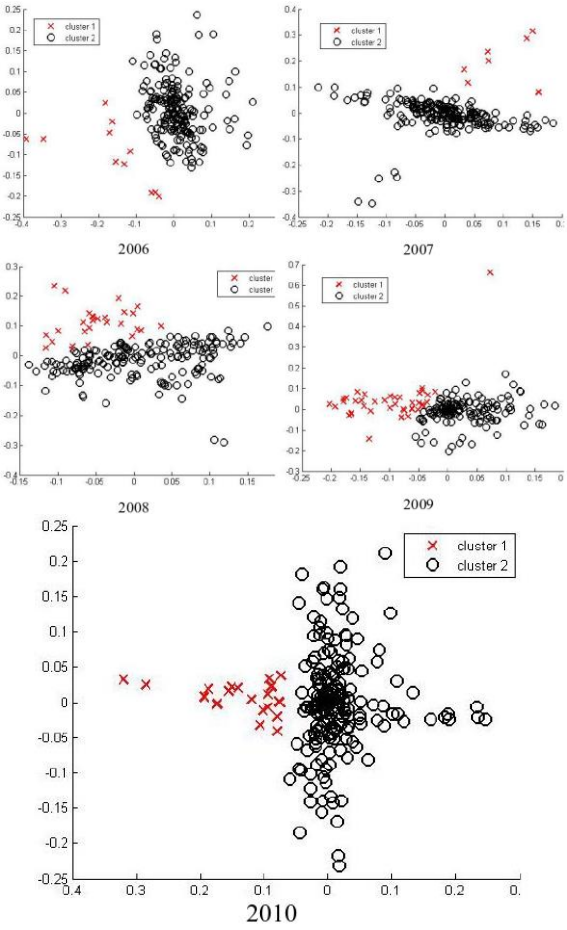


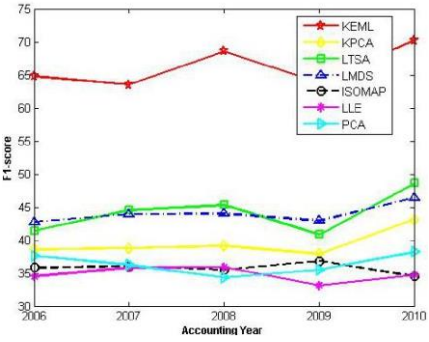
Fig. 2. Results for KEML+(K-means)

In Figure 2, "x" represents a dysfunctional company and "○" represents a normal company. In 2006 and 2007, the number of companies with "irregularities" in the data set was relatively small. From 2008 onwards, this number began to increase sharply, so much so that it peaked in 2009, but declined again in 2010. In 2006 and 2007, a large number of companies were in a superficially prosperous position due to the inflation of the bubble economy in the Chinese stock market. In 2008, the financial crisis reached China and a large number of SMEs were in financial crisis and even went bankrupt. In 2009, the effects of the financial crisis continued in China, with many companies still in financial crisis. In 2010, the impact of the crisis was mitigated by the efforts of the Chinese government. the clustering results of KEML (K-means) reflect the actual situation, which indicates that KEML helps in the early warning identification of financial crises.

We used F1-scores to evaluate each clustering result. F1-scores is a quantitative indicator made up of a search completion rate (r) and a search accuracy rate (p), as expressed in the following equation:

$$F_1(r,p) = \frac{2rp}{r+p} \tag{5}$$

The checking accuracy (p) is the ratio of the number of correct categories classified by the classifier to the overall sample size. The check-completion rate (r) is the ratio of the number of correct categories classified by the classifier to the total number of correct categories. The higher the value of F1-scores, the better the performance of the algorithm. Figure 3 shows the evolution of the F1-scores obtained after clustering the low-dimensional embeddings of the seven algorithms:



T-test:t_{D_1}=36.4082, t_{D_2}=29.1003, t_{D_3}=26.5499
t_{D_4}=19.7146, t_{D_5}=26.4080, t_{D_6}=22.9986;α=0.01

Fig. 3. F1-scores of clustering results for low-dimensional embeddings obtained by seven algorithms

The graph shows that in this experiment KEML outperforms several other algorithms by a wide margin. Similarly, we perform a t-test on the F1-scores above. The test results show that the H_1-set can be rejected at the significance level of 0.01, which means that the F1-scores of KEML are higher than those of several other algorithms.

F1-scores are microscopic averages that are often influenced by the performance of the classifier itself. Here we also consider the cost of misclassification and the a priori probability that a firm's financial position is "abnormal", and Altman proposes to assess the effectiveness of financial warnings in terms of risk expectations, which are defined as follows:

$$c_1\pi_1\frac{n_1}{|A_1|}+c_0(1-\pi_1)\frac{n_0}{|A_0|} \tag{6}$$

$A_1 and A_0$ are the "abnormal" and "normal" firms in the experimental data set, $n_1 and n_0$ are the misclassified abnormal and normal firms, π_1 are the a priori probabilities of abnormal firms, and $c_1 and c_0$ are the costs incurred by misclassifying "abnormal" and "normal" firms. Altman notes that when $\pi_1 = 0.016: 0.03$, $32 \le c_1/c_0 \le 62$ in this experiment, the rate of "abnormal" firms is close to $\pi_1 = 0.1: 0.2$ so we can get $4 \le c_1/c_0 \le 9$. We obtain $c_1/c_0 \cong 6$,so that the total risk expectation is:

$$0.51\frac{n_1}{|A_1|}+0.49\frac{n_0}{|A_0|} \tag{7}$$

The experimental results of the seven algorithms above were used to calculate risk expectations as shown in Table 1:

Table 1. Risk expectation of clustering results (K-means) (%)

Dataset	KEML	KPCA	LTSA	LMDS
2006 dataset	4.25	9.53	9.43	9.48
2007 dataset	5.13	9.51	9.15	9.28
2008 dataset	3.82	8.98	8.54	887
2009 dataset	6.51	9.95	9.58	9.69
2010 dataset	2.05	7.83	6.93	7.44
T-test	t_{D_1}=-12.8766, t_{D_2}=-12.6904, t_{D_3}=-12.1460			

Continued from Table 1

ISOMAP	LLE	PCA
11.36	15.13	9.68
10.54	12.03	10.34
12.47	11.47	15.74
10.08	15.81	10.11
13.92	14.65	8.79
t_{D_4}=-5.6556, t_{D_5}=-9.9479, t_{D_6}=-4.9928;α =0.01		

Table 1 shows the risk expectation for all clustering results, with smaller values indicating higher accuracy of the clustering results and better performance of the algorithm. The last row of Table 1 is about the results of the t-test, and it is clear that the H_0

hypothesis can be rejected, which means that the error of the K-means clustering results after using the KEML algorithm for feature extraction is lower than the error of the clustering after using other feature extraction algorithms. As the KEML algorithm more realistically reflects the logical relationships between financial data points, the resulting low-dimensional coordinates are more accurately represented and therefore the clustering results obtained with the same clustering algorithm are more accurate and the risk expectation is smaller.

4 Conclusions

Based on the assumption that high-dimensional data lie in a low-dimensional embedded manifold, we attempt to discover intrinsic manifolds in a high-dimensional financial dataset, thus proposing a manifold learning algorithm for financial datasets - KEML. Unlike traditional stream learning algorithms, KEML uses information distance to measure the relationships between financial data points and obtains suitable and accurate low-dimensional embeddings in high-dimensional financial datasets. The results of the empirical study show that KEML is able to obtain more accurate low-dimensional embeddings than the other six traditional streamlining learning algorithms. In subsequent experiments on financial situation analysis, the clustering results generated by KEML had lower error rates than the other comparison algorithms. The experiments further derive the K-entropy, which describes the nature of the system dynamics in the original data space. The K-entropy derived from KEML is able to explain and predict the volatility of the stock market when compared to the CSI 300 index.

References

1. Cai J. J., Wang A. Q., Zou J. Y., Wang Y. J., Qi Y. M. Design and implementation of financial analysis and decision-making system based on big data and artificial intelligence [J]. Modern Industrial Economics and Informatization,2016,6(11):86-88+97.DOI:10.16525/j.cnki.14-1362/n.2016.11.36.
2. Chen Meng. The dilemma facing the development of big data finance and ideas for coping with it [J]. Fiscal Science,2020(12):65-76. doi:10.19477/j.cnki.10-1368/f.2020.12.008.
3. Chen T. Research on the authentication model of financial big data based on blockchain technology[J]. Modern Electronic Technology,2020,43(06):171-174. doi:10.16652/j.issn.1004-373x.2020.06.042.
4. Ding Lianye. Big data finance: innovation and reflection of financial services for small and micro enterprises[J]. Southwest Finance,2021(07):62-73.
5. Liu Xinyuan. Research on the risks and challenges of big data-based finance[J]. Times Finance,2021(15):5-7.
6. Lai SQ, Yang L. Legal investigation on the security of financial data in the context of big data finance[J]. Economic Law Series, 2022,39(01):25-40.
7. Liu Hui. Legal regulation of big data financial algorithms [J]. Financial Theory and Practice,2021,42(02):148-154.DOI:10.16339/j.cnkihdxbcjb.2021.02.020.

8. Wang Yilang. Construction of a national anti-money laundering computer network system based on financial big data[J]. Financial Development Research,2017(06):62-66. doi:10.19647/j.cnki.37-1462/f.2017.06.009.
9. Zhao Jieru. An analysis of the risks and challenges of the development of big data finance[J]. Business Development Economics,2021(14):62-65.
10. Stven kelly. Research on the strategy of Bank A Yunnan Branch in serving small and micro enterprises' financing products in the context of big data finance[D]. Kunming University of Technology, 2021. DOI:10.27200/d.cnki.gkmlu.2021.000487.
11. Mhannod D. The social value and strategic choice of big data [D]. Party School of the Central Committee of the Communist Party of China, 2014.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.