An FDA-based Stock Exchange Price Curve Feature Recognition and Analysis Method

Yan Xue[1, a], Tong Ge[2, b], Hongxia Bie[2, c]

[1]Cui Huang Kou High School, Wuqing District, Tianjin, China

[2]Beijing University of Posts and Telecommunications, Haidian District, Beijing, China

Email: [a]1005901738@qq.com, [b]byrgtclaire@126.com, [c]biehx@bupt.edu.cn

Keywords: functional data analysis, curve feature extraction, K-means clustering, curve classification

Abstract. The classification and trend prediction of stock exchange price via trading history becomes the crucial part of intelligent stock analysis software nowadays. To figure out the variation pattern of stock price better, the curve-based method is proved to be efficient when applied to large discrete dataset. This paper proposed a FDA-based stock price curve recognition method in order to provide support for stock price prediction. On the basis of fitting function, extract segmented variation trend, segmented variation rate and segmented Root Mean Square as features which reflect the information of curve shape. And give weight to the three features to form the feature vector of the curve. Then conduct K-means clustering on these feature vectors. The result is the same to the subjective classification, so that in this way obtaining the class label of each curve. Finally classify the unknown curve with neural network. On the test set, the correct recognition rate reaches 80%.

Introduction

Curve is an important representation of discrete data. In the era of big data, the analysis of historical data, especially economic data such as large amounts of stock trading prices, provides great support for price prediction, economic law discovery and data comparison.

Reasonable feature extraction is the basis of implementing intelligent analysis. After feature extraction, we represent one curve with finite dimensional features. There are many features of curves. The most common features include first-order features, high-order features, transform domain features and model parameter based features.

Traditional multivariate analysis doesn’t make full use of the procedural characteristic of the curves, namely integrity, randomness and intuition. By contrast, the Functional Data Analysis (FDA) doesn’t rely too much on assumed condition and reflects the procedural characteristic better.

The FDA based curve recognition methods are widely used in economic field. In 2013, Tiantian Yu has fitted the discrete data of group purchase on the internet based on FDA and has conducted statistics analysis and principal component analysis to compare different kinds of group purchase and group purchase in different area [1]. In 2015, Yu Zhao has analyzed the application prospect of FDA in eco-economic system [2]. In 2015, Chunyi Liu has fitted the economic growth data with sine function base, consequently data periodicity is reflected well. This provides support for doing research on the law of economic growth [3].

In recent years, stock analysis software is hot. This paper proposed a stock exchange price curve analysis method based on FDA to figure out its variation pattern. Mainly do research on how to extract effective feature vector for curve recognition.

FDA based stock price curve feature extraction

This paper collects samples via Straight Flush stock analysis software, 92 samples in all.

There are two kinds of FDA based features, one is to get the coefficients of the fitting function as feature vector, the other is to regard curve as a fitting function and get features out of the function.

The functional fitting of stock exchange price data. Firstly, fit the discrete data to smooth function. To eliminate the error, we adopt smoothing method. The smoothing method includes linear smoothing, base function smoothing and kernel function smoothing. This paper chooses to use base function smoothing. Base function is a collection of some independent functions. Giving different weight to each base function can fit any curve [4]. Base function includes Fourier basis, B-spline basis, Polynomial basis, wavelet basis, Bernstein basis [5].

The choice of base function is important to whether the model will work well. Choosing proper base function can get a good approximation to the original data, reduce computational complexity and be convenient for processing fitting function such as derivation.

Fourier basis is mainly used for periodic curve, B-spline basis used for non-periodic curve. Most of the functional data is non-periodic, so we don't choose Fourier basis as base function. B-spline basis is the expansion of Polynomial basis and can remedy the defeat of the implicit local characteristic of Polynomial basis and Fourier basis. However, the expression of spline basis is complex and its computation of function fitting and derivation is large. According to the need of analyzing stock price variation pattern, we compromise precision and computational complexity to choose Polynomial basis as base function.

Polynomial basis is composed of 0 to P polynomial in variable t, namely:

$$\vec{b}(t)=\{(t-q)^0,(t-q)^1,\mathbf{L},(t-q)^p\} \tag{1}$$

the number of polynomial is p+1.

Fig. 1 is the 10 and 60 polynomial fitting of stock price curve.

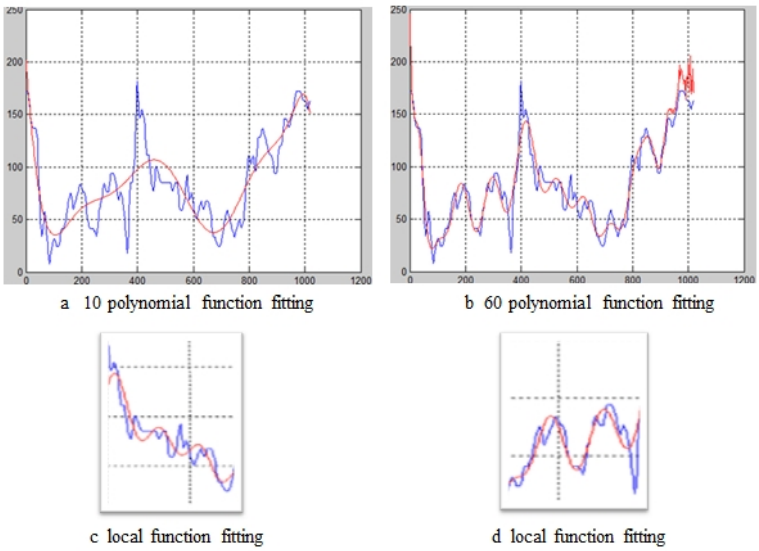

Fig. 1 10 and 60 polynomial fitting of stock price curve

As Fig. 1 shows, a can only plot the trend of the curve in general, but can not reflect the variation of local part of the curve. b approximates the original curve to a very similar extent and also plots the variation of local part. c fluctuates intensely, the fitting function ignore the fluctuation and only

focus on the entire trend. As for d, the subtle fluctuation is relatively less and the trend is obvious, so the fitting function reflects the trend better. This paper cares about the entire shape of the curve and ignore the local variation. Experiments indicate that 10 to 20 polynomial meet the requirements.

FDA based feature extraction. After functional fitting of the curve, the object we extract feature from is no longer the discrete data, instead the fitting function. As for the function, we can extract coefficients as feature vector or the first-order features, high-order features, transform domain features and model parameter based features of the function. Getting coefficients of the function as feature vector can not reflect the shape of the curve. If we conduct clustering on the coefficients, the curve in the same cluster only have parameter similarity rather than variation pattern similarity. Researches demonstrate that milestone is an important factor to plot the shape of curve. Milestone means extreme point and it belongs to first-order feature. Extreme point is used to divide the domain of function definition, consequently obtain the monotonicity of the curve and then the variation pattern.

Segmented variation trend. At first, divide the curve with milestone. Then get the variation trend on each segment as feature. We conduct FDA on the stock exchange price data and obtain the segmented variation trend of its 10 polynomial fitting function. The categories of variation trend are shown in Fig. 2.

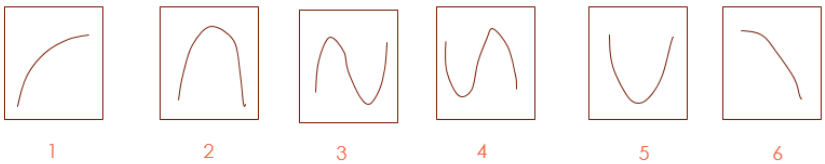

Fig. 2 10 polynomial fitting function variation trend categories of stock price curve

The algorithm of extracting segmented variation trend:

(1) Take the first derivative of the fitting function, get the extreme points and store the milestones in ascending order.

(2) Take the second derivative of the fitting function and determine that each extreme point is maximum or minimum.

(3) Divide the domain of the function definition into 10 equal segments. Determine the number of extreme points in each segments. The segment is denoted by [xmin,xmax]. The extreme point which is smaller than xmin is denoted by nmin. The extreme point which is greater than xmax is denoted by nmax.

(4) Distinguish the segments who have the same number of extreme points. The difference between 1 and 6 (2 and 5,3 and 4) is that the nmax (or nmin) is maximum or minimum.

According to the algorithms proposed above, we extract the segmented variation trend of the curve as feature vector. As the curve is divided into 10 segments, the feature vector has 10 dimensions.

Segmented variation rate. The segmented variation trend can only reflect the rough trend of the curve. The segmented variation rate can describe the speed of the variation further. We can distinguish the curves which have the same variation trend but have different variation speed using segmented variation rate.

The formula to calculate the variation rate depends on the category of variation trend. As for 1 and 6, we calculate the absolute value of the segment midpoint slope. For 2 and 5, the extreme point divide the segment into 2 parts, we calculate the average of the absolute value of midpoint slope in

the two parts. For 3 and 4, the extreme points divide the segment into 3 parts, so we calculate the average of the absolute value of the midpoint slope in the three parts. When the number of extreme point in one segment is greater than 2, we still process it fuzzily with the formula of 3 and 4.

Segmented root mean square. To distinguish curves which have the same variation trend and variation rate but have different value domain. We use segmented root mean square feature to reflect the difference. In each of the 10 segment, sample 100 points of the fitting function with equal interval and calculate the root mean square via the formula below.

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^{\,2}} \qquad N = 100 \tag{2}$$

10 dimensions of segmented variation trend, segmented variation rate and segmented root mean square form the 30 dimension feature vector. We conduct clustering on the vectors using Euclidean distance as similarity metric and give 5, 10, 250 as weight to the three kinds of features with respect to their value domain and parameter significance. To weight the features is equivalent to normalization to some extent.

FDA based curve analysis

curve classification. If we are going to classify a dataset containing n samples into k classes. The specific steps of K-means algorithm are as follows.

(1) Choose samples at random for k clusters. Each cluster represents a class, namely C={c_i, i=1,2…k}.

(2) Define the center of a cluster as the average of the samples in this cluster. Calculate the distance between samples and all the cluster centers and set the sample to the cluster which the Euclidean distance is the minimum.

$$J(c_i) = \sum_{x_j \in c_i} \| x_j - u_i \|^2 \tag{3}$$

(3) Calculate the new centers of the cluster and the sum of squared distance of different clusters.

$$J(C) = \sum_{i=1}^{k} j(c_i) = \sum_{i=1}^{k} \sum_{x_i \in c_i} \| x_i - u_i \|^2 \tag{4}$$

We conduct K-means on the FDA based feature vector to classify the stock trading prices curves according to their shape so that the curves with the different variation pattern can be distinguished. To test the algorithms, we classify 10 curve samples into 4 classes based on subject shape judgement and then classify the samples with the algorithm we proposed. The result is showed in Fig. 3.

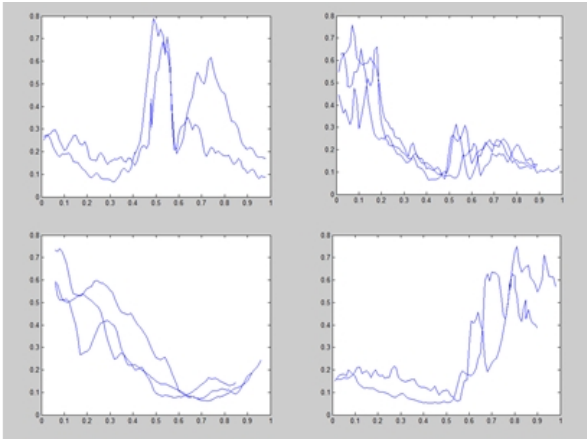

Fig. 3 10 stock price curves after FDA based K-means clustering

Neural network based curve classification. This paper extracts FDA based features to help to recognize the curve whose label is unknown. We sample 92 curves using Straight Flush stock analysis software to establish the dataset. Then process the raw data to get the 30 dimension feature vector. The feature vectors of the 92 curves are showed in Fig. 4.

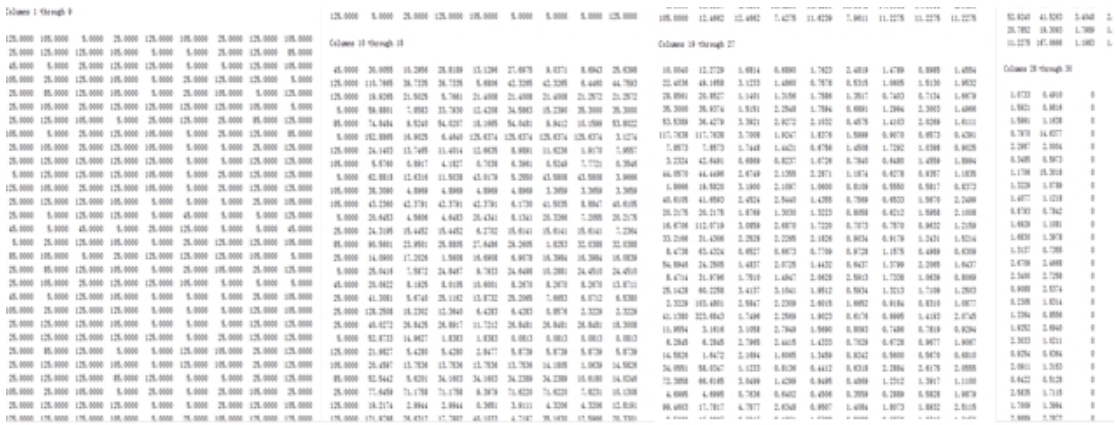

Fig. 4 30 dimension feature vector of 92 stock trading price curves

One row contains a 30 dimension feature vector of one curve.1 to 10 column is the segmented variation trend feature.11 to 20 column is the segmented variation rate feature. 21 to 30 column is the segmented root mean square feature. The significance of feature influence to clustering is sorted as follows, segmented variation trend > segmented variation rate > segmented root mean square.

Firstly, we give the 92 curve labels using K-means clustering. Then divide the 92 samples into training set and testing set, containing 72 and 20 samples respectively. Finally training the neural network based classifier and compare the output with the class label given by K-means clustering. The accuracy reaches 80%.

Conclusions

To implement stock exchange price curve classification and recognize the category of curves according to the shape intelligently, this paper proposed a FDA based feature extracting algorithm. The FDA based features include weighted segmented variation trend, weighted segmented variation rate and weighted segmented root mean square. The weight given to each features vary with respect to their value domain and influence significance. Conduct K-means clustering on the feature vectors,

the classification result agrees with subject classification. In the end, we use neural network to classify the curves, the accuracy is 80%.

References

[1]. Tiantian Yu, Xiaolin Lv. Analyze the structure and market development of group purchase on the internet via functional data analysis [J]. statistics and information forum,2013.28(3): 68-75.DOl:10.3069/j.issn.1007-3116.2013.03.012(In Chinese)

[2]. Yu Zhao, Zengju Qin. The prospect of functional data analysis in eco-economic system [J]. Technology of Gansu, 2015.31(16):66-68.DOl:10.3969/j.issn.1000-0952.2015.16.023. (In Chinese)

[3]. Chunyi Liu, Liming Liu, Shaoguo Wang. New perspective to measure the economic cycle – functional data analysis based method [J]. the world of survey and research, 2015, (6):4 2-4 6.DO l:10.13778/j.cnki.11-3705/c.2015.06.009(In Chinese)

[4]. Yuyu Zeng, Jinzhong Wen. Research on the functional data clustering methods [J]. statistics and information forum,2007,22(5):10-14.DOI:10.3969/j.issn.1007-3116.2007.05.002. (In Chinese)

[5]. Wenping Wang. Research on classification of employers based on functional data theories [D]. Northeast Normal University.2010(In Chinese)

[6]. Ramsay J O. When the data are functions [J]. Psychometrika, 1982 (47): 379-396

[7]. Ramsay J O, Dalzell C J. Some tools for functional data analysis [J]. Journal of the Royal Statistical Society, 1991(3):539-572