International Journal of Database Theory and Application
Vol.9, No.4 (2016), pp. 161-172
http://dx.doi.org/10.14257/ijdta.2016.9.4.15

Identifying University Cash Flow Pattern Recognition: A Data Clustering Approach

Yixuan Ma and Zhenji Zhang

Economics and Management, Beijing Jiaotong University
Beijing 100044, China
mayixuan@bjtu.edu.cn, zhjzhang@bjtu.edu.cn.

Abstract

The fast-developing Chinese higher education and research institutions are faced with more and more serious challenges in their financial regulation. Traditional accounting approaches may fail to distill useful decision-making suggestions confronted with the multi-species and huge-quantity financial "big data". To reveal valuable information, this research focuses on the pattern recognition of the funding flows in 76 universities under Chinese Ministry of Education. Given the trend feature of the data series detected by the Mann-Kendall Non-Parameter Ranking Test, the low-frequency parts of the funding flows are distilled with wavelet transform to represent their basic features. Results show a clear hierarchy structure in the investments and expenses of universities according to their "titles" and advantage disciplines. Specifically, the comprehensive universities fall to different categories according to their "titles", while the professional universities are classified according to their disciplines. The scientific and technical focused universities show larger variance among different categories than other specific discipline focused universities.

Keywords*: Higher Education, Funding Flow, Pattern Recognition, Wavelet Clustering, Agglomerative Hierarchical Clustering*

1. Introduction

Higher education in China is continuously growing, changing and developing, which is playing an important part in the country's economic growth, scientific progress and social development [1]. Contrary to the trend of funding decreasing in higher education around the world, the expenditure and government spending on education services in China have been dramatically increasing and flowing to emerging new disciplines in its 2,000 universities and colleges. Among all these institutions, the universities directly under Chinese Ministry of Education are of most significance for their academic contribution and reputation, whose expenditures are accordingly higher and diverse. The government has published series of regulations in detecting and managing the funding flow among these universities. However, most of the ordinances are qualitative, focusing on the moral conscience and power of the managers. The traditional accounting approaches are failing to reveal useful information for decision making when the data set is growing too big. We call for quantitative and executable managing approaches. An initial step toward this goal is to distill useful information from the existed funding flow data.

To distill information from data is the process of finding pattern, or in the other words, finding the entity in chaos [2]. The blowing data set scale together with the fast-development of computing techniques formed the specific domain of pattern recognition, which uses machine learning algorithms to automatically discover regularities in data. Originated from artificial intelligence and statistics, pattern

ISSN: 2005-4270 IJDTA

recognition has found wide application in scientific and engineering areas. It is also gaining popularity in the domain of financial and management, such as financial fraud detection [3-5], financial invoice [6], financial forecasting [7-12], bankruptcy prediction [13-14], corporations financial conditions [15-18] and financial risk analysis [19-20]. However, there is little research in funding flow analysis area. Here we apply pattern recognition algorithm in distilling patterns in the university funding flows to provide information for research funding decision and management.

The choice of pattern recognition algorithms should be decided by the research purpose and data characteristics. In this paper, we aim at revealing the similarities and differences of funding flows between universities, which belongs to the domain of "clustering algorithm". Clustering algorithms use unsupervised algorithms to predict categorical labels so that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Specific clustering algorithms include connectivity-based clustering (also known as hierarchical clustering), centroid-based clustering, distribution-based clustering and density-based clustering [21]. Given the significant increasing trend of the original data series and the calculating complexity, we apply the wavelet-based clustering method, which combines the advantages of connectivity-based clustering and density-based clustering methods through the wavelet transform of the original data. The traditional hierarchical clustering method is implemented as control group.

The rest of paper is organized as follows. In Section 2, the methodology for recognize university funding flow is introduced. The method is based on data features by wavelet transformation. Each parts of the algorithm are depict in detail. In Section 3, we describe the university funding flow time series data and its trend feature. Then we adopt the algorithm in the experiment. The clustering results and evaluation results of the experiment are shown in Section 4. In the Section 5, we make discussion about the results of the experiment. In Section 6, we draw the conclusion of the paper.

2. Methodology

2.1. Trend Test

Given the background of higher education reform and development, the funding flow of universities may have significant decreasing or increasing trend, which bring significant influence on the choice of clustering algorithms. For example, the inconsistent data set puts more weight on those larger values when applying distance measurement between data series. The ignorance of the small values may distort the results and cause unreasonable inferences.

Here we apply a non-parameter ranktest, namely, Mann-Kendall Test [22-23], to determine the significance of trend in funding flows. The Mann-Kendall Method has been widely used in statistics for its sound theoretical foundation, non-parameter property and loose constrains of original data set.

To implement the trend test, the Mann-Kendall Statistic (S) is calculated:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} Sgn(T_j - T_k) \tag{1}$$

$$Sgn(T_j - T_k) = \begin{cases} +1, (T_j - T_k) > 0 \\ 0, (T_j - T_k) = 0 \\ -1, (T_j - T_k) < 0 \end{cases} \tag{2}$$

Here T_j and T_k are elements in a time series. Given the hypothesis of non-trend, S should be normal distributed with zero mean and variance $V(S) = n(n-1)(2n+5)/18$, where

n is the length of the time series. When n>10, the normalized statistics is calculated as follows:

$$Z = \begin{cases} (S-1)/\sqrt{V(s)}, S < 0 \\ 0, S = 0 \\ (S-1)/\sqrt{V(s)}, S < 0 \end{cases} \quad (3)$$

In this trend test, if $|Z| \geq Z_{1-\alpha/2}$, non-trend hypothesis is unacceptable, which means at confidence level α, time series exists significant trend.

$$Z = \begin{cases} Z > 0, Up \\ Z < 0, Down \end{cases} \quad (4)$$

2.2. Data Feature Extraction

There are many approaches for time series trend extraction. The wavelet transform is among the most well-known method suited for financial time series. In wavelet analysis, a signal is split into an approximation and details. The approximation is then itself split into a second-level approximation and details, and the process is repeated. Figure1 shows the three levels of discrete wavelet transform.

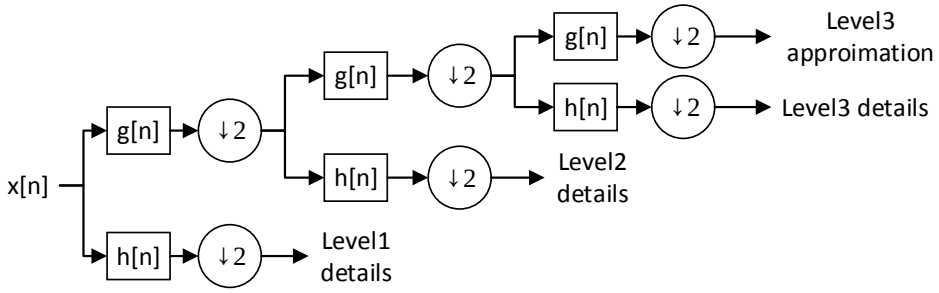


Figure 1. Three Levels of Discrete Wavelet Transform

In this paper, Haar wavelet transform is adopted. Haar wavelet is a sequence of square-shaped functions. Because of the orthogonal property of the Haar wavelet function, the frequency components of input data can be analyzed. Besides, Haar transform is one of the oldest and simplest transform functions. In this paper, we use Haar transform to extract trend feature of these time series data. Leave alone the detail (high-frequency), the approximation (low-frequency) of the split signal is trend feature of the object.

2.3. Similarity Measurement

With the trend feature extracted, the distances between different time series should be measured to implement the clustering algorithm. There are many different distance metrics, such as 1-norm distance, 2-norm distance (Euclidean distance), p-norm distance and infinite norm distance. Here we adapt Euclidean distance for its simplicity and non-parameter property [24].

For two time series $X=\{x_1, x_2,\ldots,x_n\}$, $Y=\{y_1, y_2,\ldots, y_n\}$, their Euclidean distance is calculated with the following equation:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \forall x_i \in X, y_i \in Y \quad (5)$$

Table 1. Agglomerative Hierarchical Clustering Based on Trend Feature Extraction Algorithm Process

Algorithm	Agglomerative Hierarchical Clustering Based on Trend Feature Extraction
Input:	Time Series Data Set Di, Di={T1, T2, …Tm}i=1,2,…,76
Output:	Clusters binary tree
1	Extract data trend feature by Haar wavelet transform
2	Assign each trend feature of data to a cluster;
3	Repeat: Evaluate all pair-wise distance between clusters;
4	Look for the pair of clusters with the shortest distance;
5	Remove the pair from the matrix and merge them;
6	Evaluate all distances from this new cluster to all other clusters, and update the matrix;
7	Until the distance matrix is reduced to a single element.

2.3. Clustering

With the basic features extracted and their distances calculated, the closest time series are clustered to form the bottom layer of the agglomerative hierarchy. The developing strategy of the hierarchy structure follows the rule of smallest distance within same cluster in a Bottom-Up manner [25]. Firstly, each observation is considered as a single cluster. Then in each successive iteration, the closest pair of clusters which satisfying certain similarity criteria are merged during moving up the hierarchy, until all of the data are in one cluster. The process of the algorithm is as follow.

Given the descriptions above, the flow chart of agglomerative hierarchical clustering based on trend feature extraction algorithm is as shown in Figure 2.

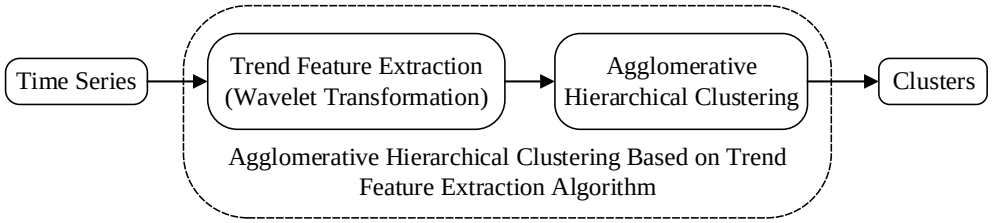


Figure 2. Agglomerative Hierarchical Clustering Based on Trend Feature Extraction Algorithm

2.4 Evaluation Indicators

In this paper, two evaluation indicators, namely Root Mean Square Standard Deviation(RMSSTD) are R Square Indicator are introduced to validate the approach brought forward above and the traditional agglomerative hierarchical clustering methods [26].

RMSSTD is the square root of the variance of all the variables. This indicator measures the homogeneity of the formed clusters at each steps of agglomerative hierarchical clustering. The smaller the indicator value indicates a higher degree of similarity of objects within the class and the better clustering results. The definition of RMSSTD is as follows:

$$RMSSTD = \sqrt{\sum_{i=1}^{K}\sum_{k=1}^{n_i}(X_k - \overline{X_k})^2 / \sum_{i=1}^{K}(n_i - 1)} \quad (6)$$

n refers to the number of data in data set.
K refers to the number of clusters.
ni refers to the number of data in ith class.

RS is the metric of differences between clusters. The greater the differences between groups, the more homogenous each group is. The value of RS ranges between 0 and 1. RS equals zero indicate that there is no difference among clusters. RS equals 1 means significant difference among groups. The definition of RS is as follows:

$$RS = \sqrt{(\sum_{k=1}^{n}(X_k - \overline{X_k})^2 - \sum_{i=1}^{K}\sum_{k=1}^{n_i}(X_k - \overline{X_k})^2) / \sum_{k=1}^{n}(X_k - \overline{X_k})^2} \quad (7)$$

n refers to the length of time series.
K refers to the number of clusters.
n_i refers to the number of data in i^{th} cluster.

3. Experiment

3.1 Data Description

There are seventy-six universities directly under the Ministry of Education of China. For confidentiality, the names are represented by 1, 2, 3…, 76. Their monthly income and outcome cash flows (17 items for each university) are stored in the SQL Server 2008 database as the original data set. Each kind of funding flow of seventy-six universities is regarded as one sub dataset. There are seventeen datasets overall. The temporal coverage of these funding flows ranges from May 2003 to December 2011.Their attribute fields, descriptions and data types are listed in Table 2 and Table 3.

Table 2. Information of Universities Incomes under Ministry of Education

Attribute Field	Data Type
TIME_ID	int
UNIT_ID	int
Total_Income	float
Financial_Allocation	float
Central_Financial_Allocation	float
Local_Financial_Allocation	float
Self_Funding	float
Education_Income	float
Research_Income	float
Education_Funding	float
Research_Funding	float

The 76 universities could be classified according to their “titles” and advantageous disciplines. For example, nine universities in this list hold the C9 League Membership, which is referred to as the Chinese equivalent of lvy league universities in the United States [27], that bears the largest expectation of the government and citizens. 39 universities in this list are under the title of 985 Project, which is a constructive project for founding world-class universities in the 21st century [28]. And almost universities are under project 211, which includes one hundred institutions of higher education and key disciplinary areas as a national priority for the 21st century [29]. Some of universities

focus more on scientific and technical disciplines while some focus on financial, art, normal education or agriculture. Their funding flows also vary accordingly.

Table 3. Information of Universities Outcomes under Ministry of Education

Attribute Field	Data Type
TIME_ID	int
UNIT_ID	int
Total_Outcome	float
Out_Funding	float
Main_Outcome	float
Education_Outcome	float
Research_Outcome	float
Operational_Outcome	float
Subunit_Outcome	float
Self_Funding_Outcome	float
Total_Outcome	float

3.2 Data Trend Test

The figure above shows the distribution of normalized Mann-Kendall Statistics of all 1292 (17*76) time series funding flows. It could be detected that more than 70% time series hold significant increasing trend. Given this, the changing characteristics of the funding flow should firstly be distracted before implementing clustering algorithms.

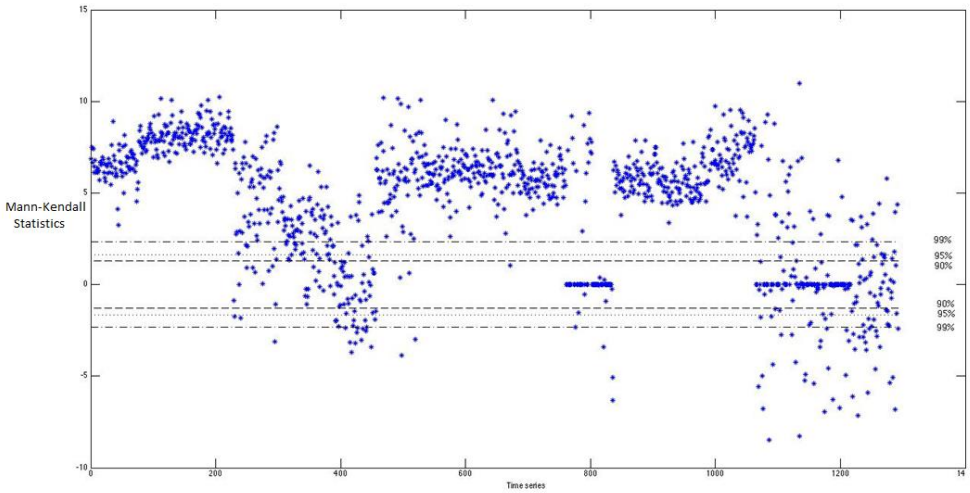


Figure 3. Time Series Mann-Kendall Trend Test Statistic Z Results

3.3 Clustering Based on Data Features Extraction

A typical decomposition of the funding flow by Harr Wavelet transform is shown in Figure 4. The a5 layer could best represent the trend feature in most of the data series.

Then agglomerative hierarchical clustering is adopted based on trend feature extraction.

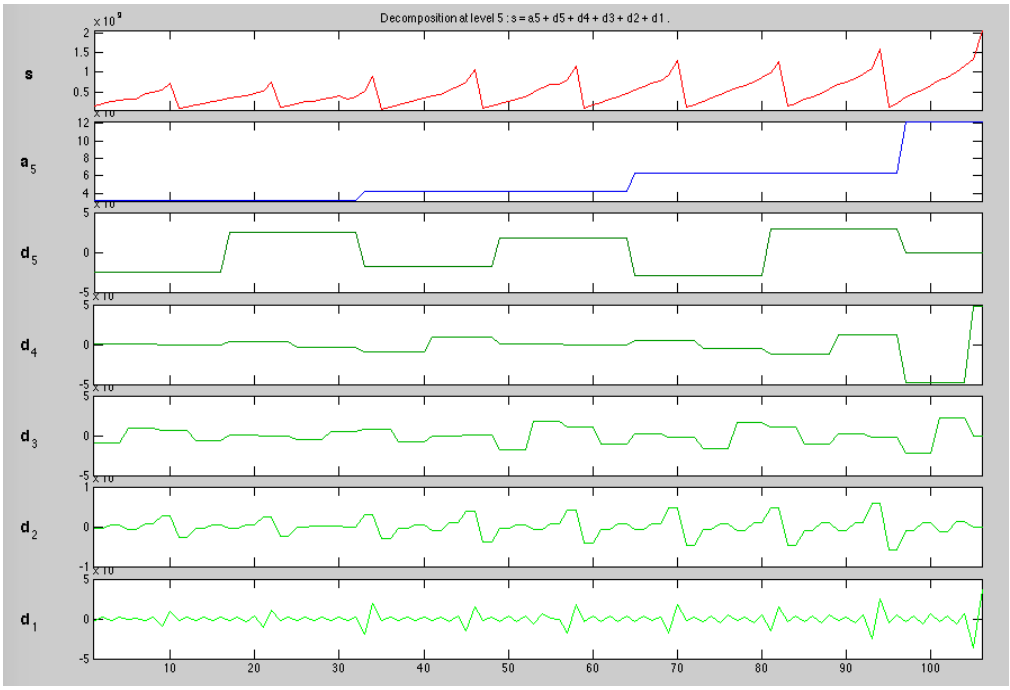
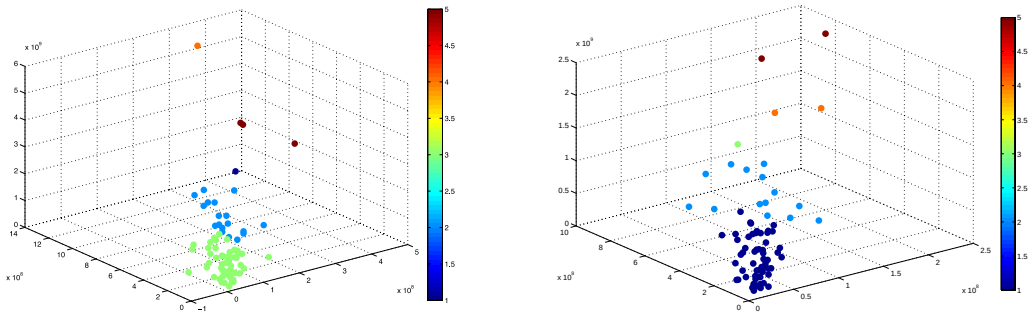


Figure 4. Time Series Wavelet Transform Trend Feature Extraction Process

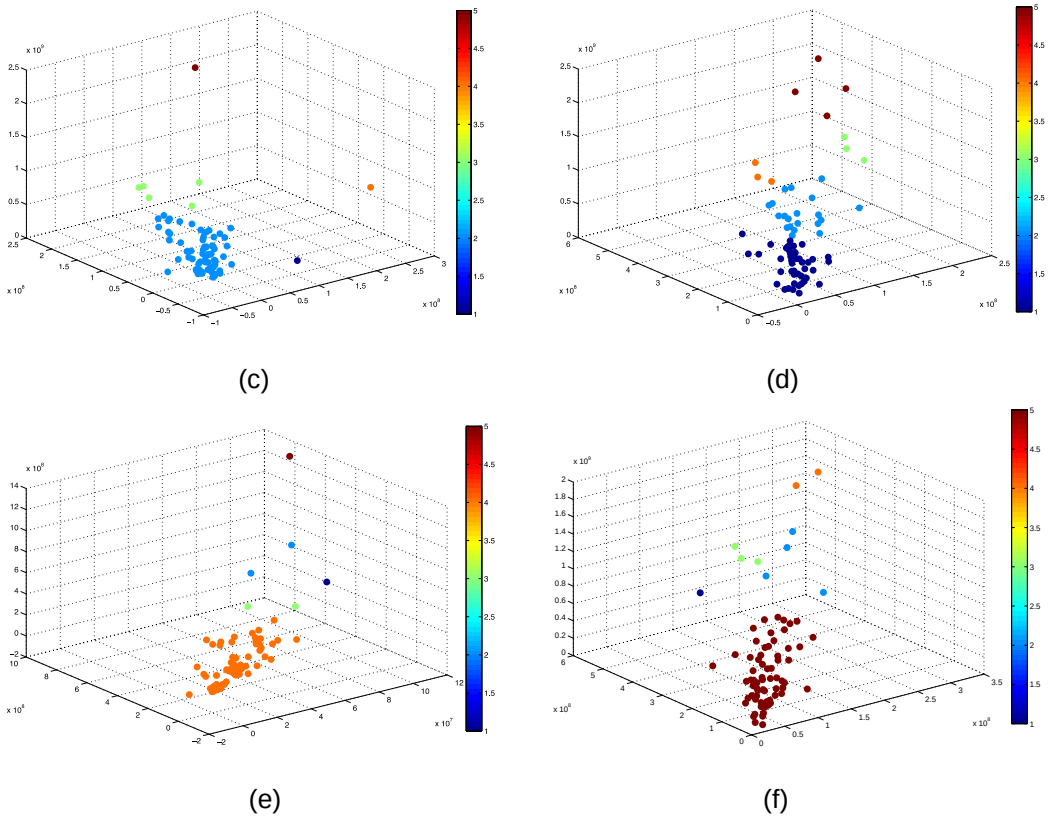
4. Results

4.1. Clustering Results

The agglomerative hierarchical clustering is performed to analysis on extracted features of the date set. All seventy-six funding series eventually agglomerate to a hierarchical tree. The distances between the objects are the height of the upside-down U-shaped. In this experiment, five clusters are chosen in each dataset clustering. Figure 5 (a) to (f) respectively show the part of clustering results of Total Income, Central Financial Allocation, Self-Funding, Education Funding, Research Funding and Education Outcome.



(a) (b)



(c) (d)

(e) (f)

Figure 5. (a) Clustering Result of Total Income; (b) Clustering Result of Central Financial Allocation; (c) Clustering Result of Self-Funding; (d) Clustering Result of Education Funding; (e) Clustering Result of Research Funding and (f) Clustering Result of Education Outcome

4.2. Evaluation Results

Traditional agglomerative hierarchical clustering is carried out as a control experiment of agglomerative hierarchical clustering based on trend feature extraction in this paper. RMSSTD and R Square are used to evaluate the performance of these two algorithms.

When it comes to RMSSTD, the smaller the value is the higher degree of similarity one cluster within and the better the clustering results are. Because RMSSRD calculate standard deviation of all the variables, the results is decreasing. It means that distance between data and central point within each cluster decrease while the number of clusters increase. RMSSTD value of both algorithms shows decreasing trend. However, when the number of clusters is the same, the value of A2 is less than the value of A1. Thus agglomerative hierarchical clustering based on trend feature extraction gets better results.

On the country, another evaluation indicator RS indicates the difference among clusters. The higher the value is, the better the results of clustering are. Unlike RMSSTD, RS shows increasing trend, which means the value increase while the number of clusters increase at the same time. The results in Table IV show that value of RS of both algorithms increasing from one cluster to ten clusters. However, when the number of cluster is equal, the value of A2 is smaller than the value of A1. It indicates that agglomerative hierarchical clustering based on trend feature extraction better than traditional hierarchical clustering.

Table 4. Information of Universities Incomes under Ministry of Education

K	RMSSTD		RS	
	A1	A2	A1	A2
2	2.5392	1.0071	0.2013	0.2271
3	2.2891	0.9884	0.3596	0.2655
4	2.2684	0.8117	0.3797	0.5114
5	2.239	0.7994	0.4041	0.5327
6	2.2193	0.7864	0.4228	0.5541
7	2.1952	0.7508	0.4434	0.5995
8	1.9917	0.7463	0.5484	0.61
9	1.9687	0.6956	0.5653	0.6661
10	1.9476	0.6907	0.5809	0.6757

A1: Traditional Agglomerative Hierarchical Clustering
A2: Agglomerative Hierarchical Clustering Based on Feature Extraction

The effectiveness of clustering algorithm is determined by how well patterns from different classes can be separated and how well within same class can be similar. From the evaluation indicators, agglomerative hierarchical clustering based on trend feature extraction is more effective than traditional ones.

5. Discussion

This part takes the research funding series as an example to explain the practical significance of the clustering results. The classification of research funding series in seventy-six universities under Chinese Ministry of Education is listed in the following table. The distributions of university types among the 5 clusters are shown as follow:

Table 5. Universities Category Analysis of Research Funding Clustering Results

Cluster	Objects	
	Category	Amount(Total)
1	Comprehensive University	8 (21)
	University of Science and Technology	17 (28)
	Normal University	5 (5)
	Foreign Language University	3 (3)
	University of Agriculture	6 (7)
	University of Finance, politics and law	6 (6)
	Medical University	2 (2)
	University of Art	4 (4)
2	Comprehensive University	7 (21)
	University of Science and Technology	11 (28)
	University of Agriculture	1 (7)
3	Comprehensive University	3 (21)
4	Comprehensive University	2 (21)
5	Comprehensive University	1 (21)

It could be depicted that for specific domain universities, the research funding sare largely determined by their specialized disciplines. However, this conclusion cannot be generalized to science and technology focused universities, which are clearly classified into two clusters given their detailed dominants. This shows a less differentiation pattern

in specific disciplines except science and technology in the most significant universities in China.

For comprehensive universities, their research funding flows are more homogeneously distributed into 5 categories, which are closely related to the labels or privileges entitled by the government. For example, University 29 monopolizes cluster 5.It is recognized as Top 1 University in China. It receives most rewards in the Major National Science and Technology Ceremony in 2012. University No. 3 and No. 4 are also known as C9 members, which distinguish them from other universities.

6. Conclusion

The significant development of higher education in China brings serious challenge its financial regulation. Given the significant increasing trend of the funding flow series, this research applies wavelet transform to distill trend features in the original data and clusters these series accordingly. Results show that the specialized disciplines of universities could largely determine its funding flows, which reveal a homogeneous development of certain subjects. The science and technology universities are more differentiated. The comprehensive universities fall to different categories, which are clearly determined by their abilities and reputation. More detailed pattern needs to be recognized with specified data.

Acknowledgment

This work is supported by Natural Science Foundation of China under Grant No. 713900334 and Grant No. 7132008.

References

[1] "Higher education in China", Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. **(2015)**.
[2] G. Zhang and M. Y. Hu, "Neural network forecasting of the British pound/US dollar exchange rate", Omega, vol. 26, no. 4, **(1998)**, pp. 495-506.
[3] E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements", Expert Systems with Applications, vol. 32, no. 4, **(2007)**, pp. 995-1003.
[4] E. W. T. Ngai, H. Hong, Y. H. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", Decision Support Systems, vol. 50, no. 3, **(2011)**, pp. 559-569.
[5] P. Ravisankar, V. Ravi, G. R. Rao and I. Bose. "Detection of financial statement fraud and feature selection using data mining techniques", Decision Support Systems, vol. 50, no. 2, **(2011)**, pp. 491-500.
[6] D. Ming, J. Liu and J. Tian, "Research on Chinese financial invoice recognition technology", Pattern recognition letters, vol. 24, no. 1, **(2003)**, pp. 489-497.
[7] A. Sfetsos and C. Siriopoulos, "Time series forecasting with a hybrid clustering scheme and pattern recognition", Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 34, no. 3, **(2004)**, pp. 399-405.
[8] K. Kim, "Artificial neural networks with evolutionary instance selection for financial forecasting", Expert Systems with Applications, vol. 30, no. 3, **(2006)**, pp. 519-526.
[9] A. Bagheri, H. M. Peyhani and M. Akbari, "Financial forecasting using ANFIS networks with quantum-behaved particle swarm optimization", Expert Systems with Applications, vol. 41, no. 14, **(2014)**, pp. 6235-6250.
[10] L. J. Cao and F. E. Tay, "Support vector machine with adaptive parameters in financial time series forecasting", Neural Networks, IEEE Transactions on, vol. 14, no. 6, **(2003)**, pp. 1506-1518.
[11] K. J. Kim, "Artificial neural networks with evolutionary instance selection for financial forecasting", Expert Systems with Applications, vol. 30, no.3, **(2006)**, pp. 519-526.
[12] G. Zhiqiang, Guo, W. Huaiqing and Q. Liu, "Financial time series forecasting using LPP and SVM optimized by PSO", Soft Computing, vol. 17, no. 5, **(2013)**, pp. 805-818.
[13] K. S. Shin, S. L. Taik and K. Hyunjung, "Anapplication of support vector machines in bankruptcy prediction model", Expert Systems with Applications, vol. 28, no. 1, **(2005)**, pp. 127-135.
[14] D. L. Olson, D. Delen and Y. Meng, "Comparative analysis of data mining methods for bankruptcy prediction", Decision Support Systems, vol. 52, no. 2, **(2012)**, pp. 464-473.

[15] L. Liang and D. Wu, "An application of pattern recognition on scoring Chinese corporations financial conditions based on back propagation neural network", Computers & Operations Research, vol. 32, no. 5, **(2005)**, pp. 1115-1129.
[16] J. Sun and H. Li, "Data mining method for listed companies' financial distress prediction", Knowledge Based Systems, vol. 21, no. 1, **(2008)**, pp. 1-5.
[17] J. Sun and H. Li, "Financial distress prediction using support vector machines: Ensemble vs. individual", Applied Soft Computing, vol. 12, no. 8, **(2012)**, pp. 2254-2265.
[18] L. Liang and D. Wu, "An application of pattern recognition on scoring Chinese corporations financial conditions based on back propagation neural network", Computers and Operations Research, vol. 32, no.5, **(2005)**, pp. 1115-1129.
[19] K. Gang, Y. Peng and G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods", Information Sciences, vol. 275, **(2014)**, pp. 1-12.
[20] C. A. Basch, B. J. Bruesewitz, K. Siegel and P. Faith, "Visa Internation Service Association", Financial risk prediction systems and methods therefor. U.S. Patent 6,658,393, **(2003)**.
[21] P. Berkhin, "A survey of clustering data mining techniques", In Grouping multidimensional data, Springer Berlin Heidelberg, **(2006)**, pp. 25-71.
[22] H. B. Mann, "Nonparametric tests against trend", Econometrica: Journal of the Econometric Society, **(1945)**, pp. 245-259.
[23] M. G. Kendall, "Rank correlation methods", Griffin, London, **(1948)**.
[24] R. Agrawal, C. Faloutsos and A. Swami, "Efficient similarity search in sequence databases", Springer Berlin Heidelberg, **(1993)**.
[25] J. W. Han, K. Micheline and P. Jian, "Data mining: concepts and techniques", Elsevier, **(2011)**.
[26] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On Clustering validation techniques", Journal of Intelligent Information Systems, vol. 17, no. 2-3, **(2001)**, pp. 107-145.
[27] "C9 League", Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. **(2015)**.
[28] "Program 985", http://www.chinaeducenter.com/en/cedu/. China Education Center Ltd. **(2015)**.
[29] "Program 211", http://www.chinaeducenter.com/en/cedu/. China Education Center Ltd. **(2015)**.

