



Time series clustering based on sparse subspace clustering algorithm and its application to daily box-office data analysis

Yan Wang¹ · Yunian Ru² · Jianping Chai¹

Received: 4 June 2018 / Accepted: 10 September 2018 / Published online: 21 September 2018
© The Natural Computing Applications Forum 2018

Abstract

Movie box-office research is an important work for the rapid development of the film industry, and it is also a challenging task. Our study focuses on finding the regular box-office revenue patterns. Clustering algorithm is unsupervised machine learning algorithm which classifies the data in the absence of early knowledge of the classes. Unlike static data, the time series data vary with time. The work focused on time series clustering analysis is relatively less than those focused on static data. In this paper, the sparse subspace clustering (SSC) algorithm is introduced to analyze the time series data. The SSC algorithm has a better performance both on the artificial data set and the daily box-office data than recently developed well-known clustering algorithm such as K-means and spectral clustering algorithm. On the artificial data set, SSC is more suitable for time series, whether from the angle of clustering error or visualization. On the actual data, movies are divided into five clusters by SSC algorithm, and each cluster represents a distinct type of distribution pattern. And these patterns can be used in movie recommendation, film evaluation and can guide theater exhibitors and distributors. In addition, this is the first time to apply SSC to deal with time series clustering problem and get a pleasant effect.

Keywords Movie box office · Sparse subspace clustering · Time series · Data mining

1 Introduction

With the increasing demand for cultural consumption and the rapid growth of theaters and screens, Chinese film industry continues to show a boom. However, the film industry is a high-investment, high-risk industry. Most movies are not able to recover the cost of production by box-office revenues. Therefore, the research of the film box office plays a significant role in avoiding risks.

The box-office study began in the last century, and most previous studies focused on the motion picture box-office forecasting [1–3], the influencing factors of the box-office revenue [4, 5] and the box-office prediction based on social media [6–8]. These studies focused mainly on forecasting the total revenue of a film or the movies' opening weekend revenues and do not pay attention to the daily box-office trend of each movie. With the advent of the big data era, more detailed data are being stored, and people in the film industry want to know the pattern of daily box-office trends. Therefore, studying the trend of daily box office has become a new problem: whether there are some regular patterns of box-office trend.

There have been several prior studies aimed at finding the regular box-office revenue patterns. Jedidi, Krider and Weinberg used a finite mixture regression method to find the existence of the regular sales pattern in weekly box-office revenues of 102 successful movies. Based on an exponential decay model, four clusters of movies, varying in opening strength and decay rate, were found [9]. Li Bo and Lu Fengbin established the Gamma demand model

✉ Yan Wang
wy@cuc.edu.cn

Yunian Ru
ruyn2016@cuc.edu.cn

Jianping Chai
jp_chai@cuc.edu.cn

¹ School of Data Science and Intelligent Media, Communication University of China, No. 1 East Street, Dingfu Village, Chaoyang District, Beijing 100024, China

² Department of Information Engineering College, Communication University of China, No. 1 East Street, Dingfu Village, Chaoyang District, Beijing, China

with seasonal factor and some other important factors, to analyze the movie life cycles in Chinese movie market [10]. Inbal Yahav suggested that films which attract the same audience similar film audiences are thought to have similar demand pattern, and therefore, he offered a model to forecast the entire weekly per-screen demand of a film, by using information on movie similarity network which is an important step to understanding consumers' choice in the film industry [11]. However, with the rapid increase of Chinese film production, the release time of films is shortened dramatically. The life cycles of the film are generally three weeks, and only a few films can last for more than five weeks. Therefore, weekly box-office trend research is no longer suitable to the film market analysis, and the daily box-office trend research, with finer particle size, is more in line with the current situation of Chinese film.

In this paper, our study focuses on finding the regular box-office revenue patterns by clustering algorithm to understand box office as a process which changes with time. Daily box-office data are a time series, whereas the traditional methods which are used to describe similarity, such as Euclidean distance, will not be able to provide an effective description of the similarity of time series [12]. A suitable clustering algorithm should be able to identify the trend similarity of time series [13]. Traditional clustering algorithms based on Euclidean distance, such as K-means, ignore time information when distance is calculated [14].

To address the above challenges, we try to use the SSC model to cluster the data of daily box-office revenue to find the regular box-office revenue patterns. Sparse subspace clustering can reveal the real subspace structure of high-dimensional data, and it can also deal with noise data and missing items [15]. So, it would be a better choice to apply SSC algorithm to cluster time series of movie box office. The spectral clustering and K-means algorithm model are also used to compare with SSC algorithm. As clustering data sets are not labeled, it is an open problem to measure the effect of clustering. Visualization and some computational indexes are often used to measure the effectiveness of clustering. We first used an artificial labeled data to analyze the performance of different clustering algorithms. Then, we use SSC to analysis the patterns in box-office data. In addition, for each cluster, the distributions of other information of movies such as genre and origin country are analyzed to find characteristics of each cluster.

The main contributions of this paper are as follows:

- It is the first time to apply SSC to deal with time series clustering problem and get a pleasant effect.
- This paper analyzes the applicability of SSC in time series clustering problem by artificial data sets. The result is that SSC is more suitable for time series than well-known clustering algorithm such as K-means and

spectral clustering algorithm, whether from the angle of clustering error or visualization.

- This paper successfully extracts five distinct potential trend patterns from box-office data using SSC algorithm and analyzes the characteristics of films with these patterns. These patterns are interesting and valuable for movie recommendation, cinema management and movie box-office forecasting.

This paper is divided into the following sections: In the second chapter, the SSC, spectral clustering and K-means algorithm model are introduced; in the third chapter, the data source is introduced; in the fourth chapter, we use SSC algorithm to analyze the data, and compare the clustering results of several common clustering algorithms. And the box-office revenue shape patterns and characteristics of each cluster are discussed. In the fifth chapter, we summarize the main conclusions of the study and analyze the realistic significance of the result of sparse subspace clustering.

2 Model

2.1 K-means algorithm

K-means is a commonly used and simple clustering algorithm. Given an integer K and data set $X = \{x_1, x_2, x_3, \dots, x_n\}$, $x_i \in R^d$, the algorithm cluster the data into K categories, $C = \{c_k, i = 1, 2, 3, \dots, K\}$. The Euclidean distance is taken as the criterion of distance, and the sum of square distances between each point and its center of the cluster is calculated as [16]

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - u_k\|^2. \quad (1)$$

Clustering goal is to minimize the sum of squared for all clusters [17],

$$J(c) = \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - u_k\|^2. \quad (2)$$

The K-means clustering algorithm starts with an initial K class partition and then assigns the data points to each category to reduce the total distance square sum. As the $J(c)$ of K-means clustering algorithm tends to decrease with the increase of the number of categories K (when $K = n$, $J(c) = 0$), the sum of squares of distances can only be obtained under the K of a certain number of categories [18]. Its key steps are as follows:

Step 1 Arbitrarily select K initial clusters centers $C = \{c_k, i = 1, 2, 3, \dots, K\}$.

- Step2* Obtain a new classification by distributing points to its closest cluster center.
- Step3* Calculate new cluster center for each cluster.
- Step4* Repeat steps 2 and 3 until C stop to changes.

2.2 Spectral clustering

Compared to the common algorithms such as K-means, spectral clustering algorithm based on spectrum theory often has better performance [19]. The scheme the algorithm generally includes two steps. First, similarity graph is constructed to describe the similarity of all the data points, and the Laplacian matrices are solved. Then, the graph is divided into some sub-graphs which are as disjoint as possible, according to some optimization goal. The point set contained in each sub-graph is considered as a cluster [20].

Given a set of n data points x_1, \dots, x_n , similarity graph $G = (V, E)$ is a useful way to represent the data. G is an undirected graph and its vertex set is V , and vertex v_i in the graph represents a data point x_i , and edge e_i is formed between every pair of nodes. $w_{(i,j)}$ is the weight of each edge, and it represents the similarity between nodes i and j [21]. Generally, the data points are in the Euclidean space, so the Gaussian similarity function is chosen to represent the similar relationship. Then, the definition of similar matrix W matrix is:

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right). \tag{3}$$

The parameter σ in (3) also needs to be chosen, and the degree d_i of a vertex $v_i \in V$ is defined as

$$d_i = \sum_{j=1}^n w_{ij}. \tag{4}$$

The degree matrix D is the diagonal matrix with the degrees d_1, \dots, d_n on the diagonal.

The import tools for spectral clustering are Laplacian matrices. Laplacian matrices are divided into unnormalized graph Laplacian and the normalized graph Laplacians. The unnormalized graph Laplacian is defined as $L = D - W$, and the normalized graph Laplacians have two forms: L_{sym} and L_{rw} . They are defined in (5) and (6).

$$L_{sym} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \tag{5}$$

$$L_{rw} = D^{-1}L = I - D^{-1}W. \tag{6}$$

L_{sym} is a symmetric matrix, and the L_{rw} is a random walk matrix [20].

According to different similarity functions and graph partitioning methods, the spectral clustering algorithm has

a lot of different implementation methods, but they can be summarized as the following three main steps:

- Step1* Construct a Laplacian matrix L , representing the data set.
- Step2* Figure out the eigenvalues and eigenvectors of the Laplacian matrix.
- Step3* Use K-means or other classical clustering algorithms to cluster the eigenvectors in eigenvector space.

2.3 Sparse subspace clustering

Sparse subspace clustering has been widely used in machine learning, image processing and pattern recognition, such as face recognition, image segmentation, the detection of closely related gene clusters in the genome and detection of epileptic seizures in patients with EEG data [22].

For data sets $x_n, n = 1 \dots N$, assuming that each data point is one of the elements in the K subspace (usually K needs to be given), then the subspace clustering is to solve the problem for which data points belong to which subspace. And the scheme of sparse subspace clustering is shown in Fig. 1. The sparse representation coefficient matrix is obtained. Then, the similarity matrix is constructed by sparse representation coefficient matrix. Finally, we use the spectral clustering framework to calculate the clustering result of the data set [15].

Sparse representation (SR) is a hot topic in the field of image processing and applied mathematics. Sparse representation is designed to represent the data more effectively or to reveal the essential structure of the data. Sparsity refers to the use of a smallest group of basis to linear combination of data. For the given data, the position of nonzero element indicates that the point belongs to the subspace, which is spanned by the corresponding base, and the number of nonzero element indicates the intrinsic dimension of the data. The sparsity of a vector is usually characterized by the number of nonzero elements in a vector, that is, the l_0 norm which corresponds to the general NP-hard problem. So, under mild assumptions, the SR can be obtained by using l_1 optimization [23].

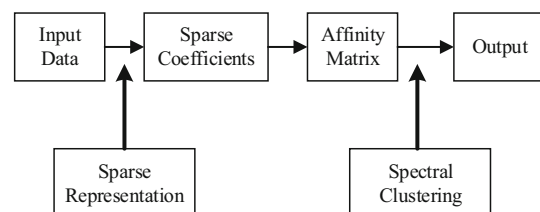


Fig. 1 Scheme of sparse subspace clustering

More precisely, each data point $x_i \in U_{l=1}^k S_l$ can be expressed as

$$x_i = Xz_i, z_{ii} = 0. \tag{7}$$

And the purpose of constraint $z_{ii} = 0$ is to avoid the trivial solution. The coefficient z_i is arranged in a matrix in a certain way, and then (7) is equivalent to

$$X = XZ, \text{diag}(Z) = 0. \tag{8}$$

Here, X is the data matrix, and Z is the matrix of coefficients. In order to find the sparse representation, (9) can be written as

$$\min Z_1 \text{ s.t. } X = XZ, \text{diag}(Z) = 0 \tag{9}$$

where $\text{diag}(Z) \in R^N$ is the vector of the diagonal elements of Z .

If the subspace structure of the data is known, and the data are clustered by class, then the coefficient matrix Z can be made with block diagonal structure under certain conditions [24].

$$Z = \begin{bmatrix} z_1 & \dots & 0 \\ & z_2 & \\ \vdots & & \vdots \\ 0 & \dots & z_k \end{bmatrix}$$

However, as is often the case, data are often corrupted by noise. In such cases, we change (9) to (10).

$$\min \|Z_1\| + \lambda_1 \|E_1\| + \lambda_2 \|R_F^2\| \text{ s.t. } X = XZ \tag{10}$$

where E is a matrix of noise, and R is a matrix of sparse outlying entries. λ_1 and λ_2 are used to balance these three terms of the objective function. Then, (10) can be solved efficiently by common convex programming tools, and then the similarity matrix can be obtained by (11).

$$W = |Z| + |Z|^T \tag{11}$$

In summary, sparse subspace clustering can be divided into two steps:

- (1) Construct a similarity matrix of all points by using sparse representation, which can be solved using convex optimization tools.
- (2) Use spectral clustering on the similarity matrix to obtain the data clustering results, which can be solved by using existing spectral clustering methods. What is more, the spectral clustering step can get rid of the noise in the similarity matrix, which can be regarded as a process of information gain [25].

Algorithm: Sparse Subspace Clustering

Input: A set of points $\{x_i\}$ lying in a union of n linear subspaces $\{S_l\}, l = 1, \dots, k$.

1: Solve the sparse optimization program (9) if the data is uncorrupted, or optimization program (10) if the data is corrupted.

2: Normalize the columns of Z .

3: Make a similarity graph with N nodes to represent the data set. Set the weights on the edges between the nodes by (11).

4: Use spectral clustering to the similarity graph to get the ultimate clustering result.

Output: Segmentation of the data: $X_1; X_2; \dots; X_k$.

3 Data source

As clustering data sets are not labeled, it is an open problem to measure the effect of clustering. Visualization and some computational indexes are often used to measure the effectiveness of clustering. We first use the labeled data to analyze the effect of different clustering algorithms on the clustering of time series. Then, we analyze the box-office data using clustering algorithm.

3.1 Artificial data set

This paper first uses the artificial time series data set with label information constructed in the paper [14]. Four groups of five vectors were structured, and each group has the same parameters of linear transformation between time points. For the group $j, 1 \leq j < 4, x_{n(t)} = w_{jt}x_{n(t-1)} + b_{jt}$, with $0 \leq t < (nt - 1)$ and $1 \leq n < 5$. The values of W and b were obtained randomly for each group. A group of sequence with similar shape can be gained using different initial value, which is shown in Table 1.

3.2 Movie box-office data set

From the Maoyan (<http://piaofang.maoyan.com/>), this article climbed box-office data in Chinese movies' market, from January 1, 2015, to March 10, 2016. These data are basically the same with the data from The State Press and Publication Administration. So, the knowledge mined from

Table 1 Artificial data set

Time points	Value
x_0	Initial value
x_1	$w_1x_0 + b_1$
x_2	$w_2x_1 + b_2$
...	...
x_{nt}	$w_{(nt)}x_{(nt-1)} + b_{(nt)}$

these data sets can reflect the real regular pattern of Chinese film market at the box office. In order to analyze the characteristics of each cluster of the film, we got the “Watercress Movie Ratings” for each film and the Comments number from the Internet from (<https://movie.douban.com/>). In addition, From the Chinese box-office (<http://www.cbooo.cn/>), we got the basic information of the film such as the origin country and genre (see Table 2 for detailed information).

4 Results and analysis

4.1 Experiment on artificial data set

The resulting artificial data set, shown in Fig. 2a, was clustered using SSC, K-means and spectral clustering algorithms, respectively. We can see that from the perspective of visualization, SSC algorithm can accurately identify patterns in time series and successfully cluster the similar shape sequences into one class.

From Fig. 2, we can see that from the perspective of visualization, SSC algorithm can accurately identify patterns in time series and successfully cluster the similar shape sequences into one class. From Table 3, we can find that SSC can classify data without error.

Therefore, in view of this problem, the SSC algorithm is superior to the traditional clustering algorithm in terms of classification accuracy and subjective feeling.

4.2 Experiment on box-office data set

The movies with low box-office and short release period have little research value, and they have complex shape patterns which make them not easy to study and use. Therefore, this paper selects 68 movies from Chinese

movies’ market, from January 1, 2015, to March 10, 2016. These films were released on Friday, and their release period is more than 22 days. Time series of movie box office over release period for 68 movies are shown in Fig. 3. In Fig. 3, it clearly illustrates the heterogeneity of daily box-office shape patterns, in which some movies’ decay over time, with different decay rates and shapes; some movies have one peak point and some have several peak points. And the peak time is also different.

4.2.1 Comparison of the SSC and classical algorithm

In the following sections, this article uses the classical K-means and spectral clustering algorithm to analyze the 68 films’ box-office data. In the experimental analysis, in order to facilitate the comparison with the SSC algorithm, the parameter (number of cluster) in the K-means and spectral clustering algorithm are set to 5. The maximum number of iterations of the K-means will be set to 1000. We select canonical Laplacian matrix in spectral clustering. The experimental results are shown in Fig. 4a, b.

It can be found through Fig. 4a, b, either the K-means or the spectral clustering algorithm can be used to interpret the clustering results of the box-office data which have time information. They cannot be used to analyze the different aspects of daily box-office patterns, such as decay rate, time of first box-office peak, revenue value at peak time, existence of other peaks and their time. The K-means clustering results are related to the total box-office revenue, and it clustered high box-office films into one category, low box-office film into another class. Using spectral clustering algorithm, the majority of the film is divided into a class. Therefore, the above two methods are not like the SSC algorithm, which can find regular patterns from the complex data.

Table 2 Movies’ raw characteristics

Variable name	Description	Value range
Genre	Movie genre, or combination of genres (e.g., adventure/drama)	23 distinct values
Origin country	Film producing country	7 countries and regions
Distributor	The name of the film distribution company	28 distributors
Runtime	Film runtime	[100 min, 137 min]
Release date	Date of movie has released to screens	From January 1, 2015, to March 10, 2016
Watercress movie ratings	To investigate the powerful influence of electronic word of mouth (eWOM), we select China’s famous film critic website “Watercress”	[0,10]
Comments number	Number of movie comments for a movie	[247, 299584]
Box office	Total box-office revenues	[2.192 million, 1.467 billion]

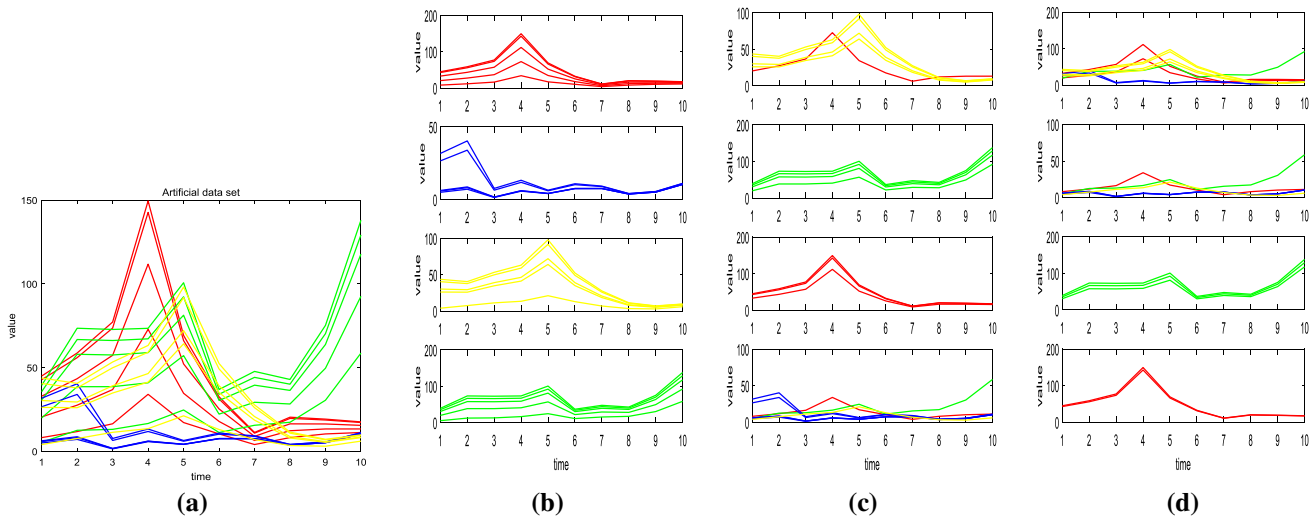


Fig. 2 K-means clustering and spectral clustering result

Table 3 Classification error

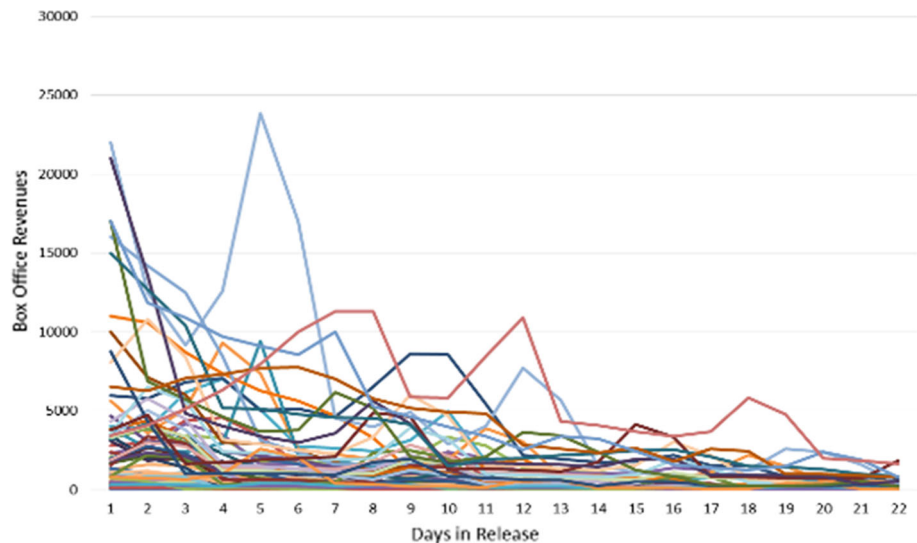
Clustering algorithm	SSC	K-means	Spectral clustering
error	0%	20%	40%

4.2.2 Data analyzed by SSC

In order to achieve the best clustering results, SSC code in [19] needs to optimize the parameters. This paper chooses “outliers” as “true,” so the program will remove outliers, and chooses “affine” as “false.” And this paper chooses the number of the clusters as 5.

Clustered by SSC, the data are divided into five clusters, and the results are shown in Fig. 5a–e. Figure 5f is an average of every day’s box office for the movie of each

Fig. 3 Time series of movie box office over release period for 68 movies



cluster, and it can be used to represent the pattern of the box office for each cluster.

Therefore, 68 movies can be clustered into five categories, according to the different patterns hidden in time series of each movie box office as is shown in Table 4. The following section will add the basic information of the film such as origin country and genre of the film and other information, to analyze the characteristics of each cluster.

4.3 Clustering result analysis

Figure 6 reflects the distribution of different genres of films in these five clusters of films.

In order to find characteristics of five clusters, we analyze the country of origin, “Watercress” Movie Ratings, comments and box office of each movie, and the results are written in Table 5.

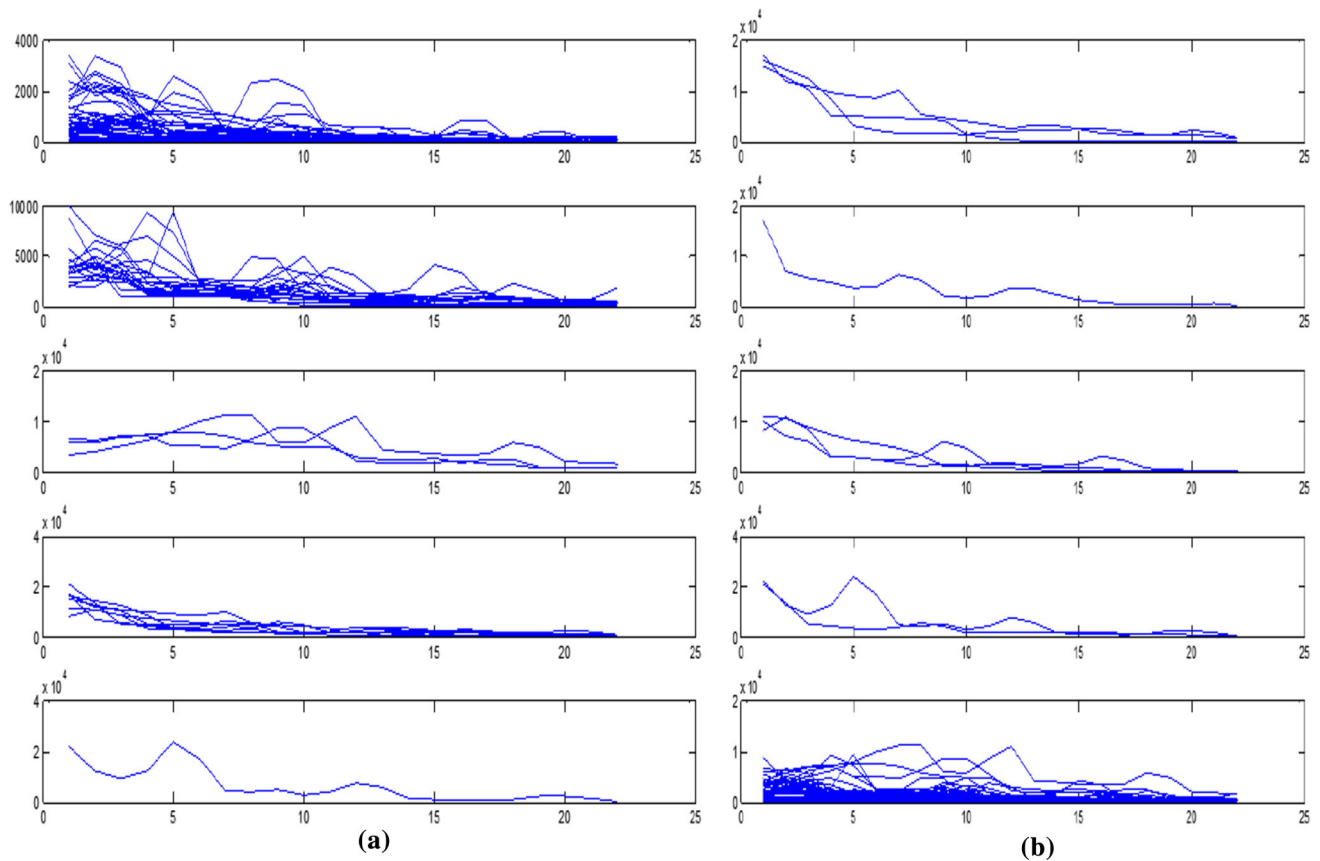


Fig. 4 K-means clustering and spectral clustering result

The following section mainly analyzes each cluster, to explore whether these features are related to cluster membership, combining with the clustering results and the basic data of the movie.

Cluster 1 The typical representatives of these films were *Ant Man* and *Taken3*. This kind of film was mainly imported films which accounted for as high as 70.59%. And they also had good scores. Interestingly, they did not have a great advantage at the box office and the number of comments. Adventure and animation films accounted for a large proportion of these films. Adventure movies accounted for 23.81% and animation films accounted for 14.29%.

Cluster 2 The typical representatives of these films were *Wolf Warriors* and *Mr. Six*. This kind of films was mainly domestic films, which are good at the box office and word of mouth. In this cluster, films imported from USA accounted for only 14.29%. The average “Watercress” Movie Ratings of this cluster was the highest of these five clusters and the average number of comments was also the largest one. From the time series of the movies’ box office, it could be seen that the box-office peak does not appear in the opening day, and these movies still have strong competitive advantages in the second weeks. So, their good

word of mouth could gradually open up. In these movies, romance and nostalgia movies had a larger proportion.

Cluster 3 The typical representatives of these films were *Devil and Angle*, *The Assassin* and *Chappie*, and this kind of film was mainly domestic films. The average score in “Watercress” Movie Ratings was only 5.4, which is the lowest score in these five clusters. In this cluster, films imported from USA accounted for only 16.67%. What is more, the number of comments and the box office of these movies was also the worst in the five categories. So, this type of film mainly represents a class of films with poor competitiveness. In this kind of film, the action and comedy movies accounted for a large proportion, and they reached 25% and 21.43%, respectively.

Cluster 4 The typical representatives of these films were *Chronicles of the Ghostly Tribe*, *Where Are We Going?* *Dad 2* and *Pixels*, and domestic films accounted for a large proportion. It is interesting that though these films have a lot of the number of reviews, their score is low and the box-office performance is poor. The feature of this type of film was that they are popular IP on the Internet or had a strong star lineup. So, the audience had a great expectation on them before they are released, but the bad reputation hindered the growth of their box-office revenues. In this kind

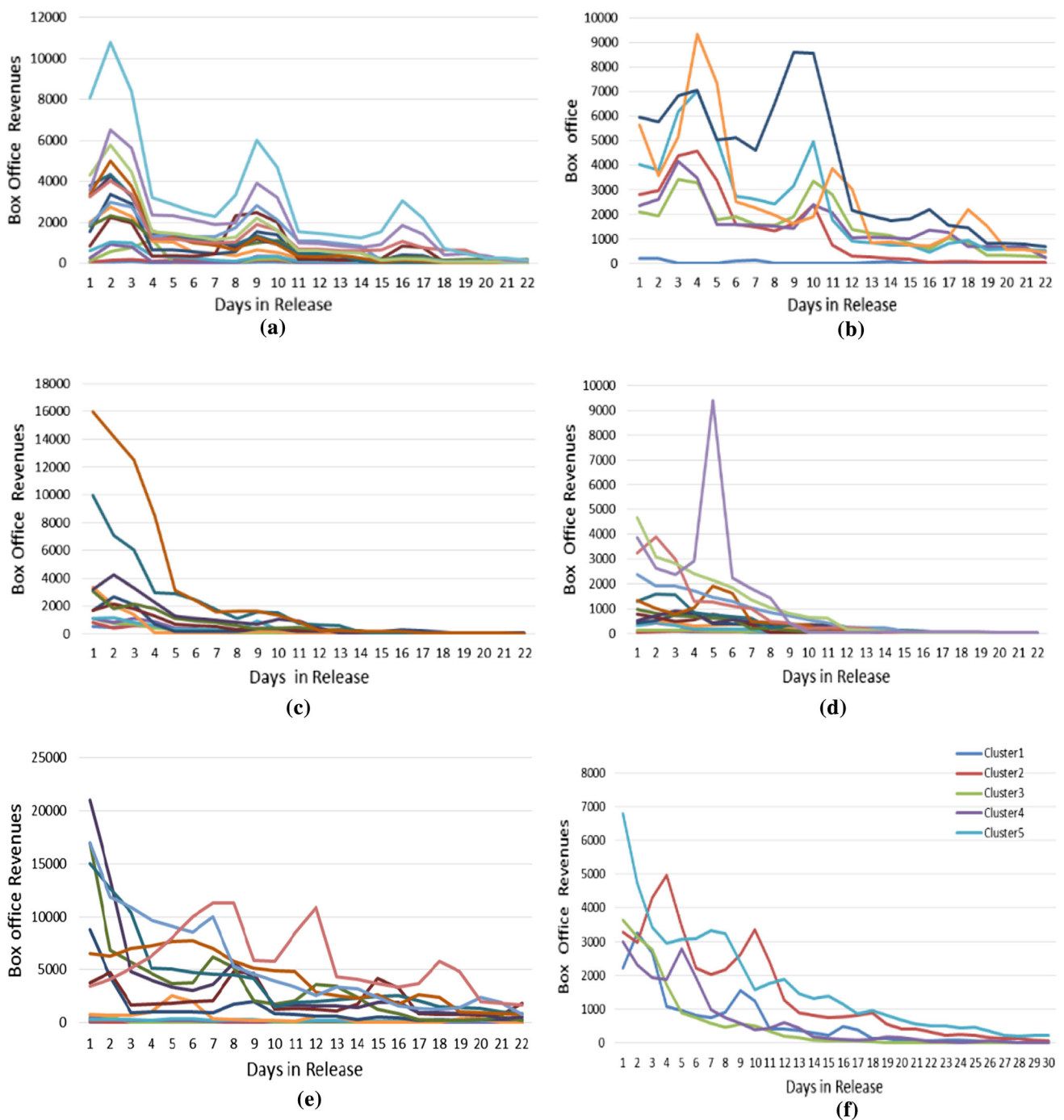


Fig. 5 Clustering results for SSC. **a** Cluster 1, **b** Cluster 2, **c** Cluster 3, **d** Cluster 4, **e** Cluster 5, **f** means of each cluster

of film, motion film took up a large proportion, and the types of these movies were more diversified.

Cluster 5 The typical representatives of these films were *Kung Fu Panda 3*, *The Man From Macau II* and *The Monkey King 2*, and they were mostly blockbusters with high quality. Most of them were the sequels of the successful movies. This kind of film had the best box-office performance and the largest number of comments, and they

represented the most successful movies. In this kind of film, comedy and action films accounted for a larger proportion, and they, respectively, reached 20.59% and 17.65%.

Table 4 Description of clusters

Cluster	Proportion	Trend description
Cluster1 Figure 4a	10.29%	The time series of box office has (have) obvious periodicity, and the peak appears on Saturday in every week
Cluster2 Figure 4b	26.47%	Box office performs well in the first two weeks, but it decays rapidly in the third week. Compared with the first cluster, it has a short life cycle, and the weekly box-office peak has delays
Cluster3 Figure 4c	17.65%	Time series of box office correspond to a rapid decay pattern, and box office in second week and third week is far lower than it in the first week
Cluster4 Figure 4d	25.00%	Compared to third cluster, the peak of its box office is not on the first day of release
Cluster5 Figure 4e	20.59%	Box office of this cluster performs well, and the peak does not appear in the weekend

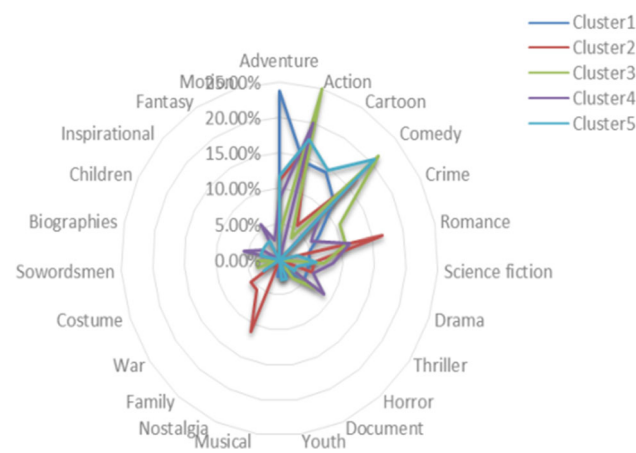


Fig. 6 Genre distribution of the movies in each of the five clusters

5 Conclusion

Sparse subspace clustering has been widely used in image processing and pattern recognition. It is a useful and effective exploration to use SSC to cluster the time series. It effectively solves the problem how to measure the

similarity of time series. Compared with the traditional clustering algorithm, the SSC clustering algorithm has the best results of these three methods to express the correlation of time series, in terms of classification accuracy and subjective feeling. The SSC is also superior to K-means and spectral clustering algorithms in pattern discovery. Through clustering analysis, we find that the movie box-office data can be clustered into five clusters, and each cluster has its own unique pattern.

The first cluster movies are mainly imported films, and their life cycle can last for three weeks. But it is interesting that their box-office revenues and the number of comments are very ordinary. The second cluster is mainly domestic films, with good word of mouth and box-office revenues. There are two obvious peaks in their time series of the box office. So, good reputations play an important role in the high box office. The third cluster is mainly domestic films, with low score, their time series of box-office correspond to a rapid decay pattern. They represent a kind of film with poor competitiveness. The fourth cluster of films has many comments, but their score and the box office are poor. The

Table 5 Characteristics of five clusters

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
<i>Country of origin</i>					
China	29.41%	57.14%	66.67%	50.00%	50.00%
Europe and America	70.59%	14.29%	16.67%	27.78%	42.86%
China–Hong Kong	0.00%	14.29%	16.67%	16.67%	7.14%
China–Taiwan	0.00%	14.29%	0.00%	0.00%	0.00%
Korea	0.00%	0.00%	0.00%	5.56%	0.00%
<i>“Watercress” Movie Ratings:</i>					
Average score	6.79	7.07	5.40	5.92	6.65
Average comments number	48,313.44	145,692.57	43,996.42	52,678.50	89,134.64
<i>Box office (million)</i>					
Mean	188.94	429.80	161.41	193.68	506.82

fifth cluster is mainly the sequel of the successful movies, with high box office and good word of mouth.

The clustering analysis of the time series of movie box office based on SSC algorithm can effectively cluster the films based on their time information, and it can find the unique pattern of the box-office trend. These patterns can be used in the field of box-office prediction, movie recommendation, film evaluation and so on, in future research.

Acknowledgements This paper is financially supported by the Fundamental Research Funds for the Central Universities, the Outstanding Young Teacher Training Project of Communication University of China (YXJS201527), Engineering Planning Project of Communication University of China (3132018XNG1823) and the Research of Key Technology in modeling digital movie service management intelligent data repository (2015-56).

Compliance with ethical standards

Conflict of interest No conflict of this work and authors.

References

- Barry RL, Kohl LS (1989) Predicting financial success of motion pictures: the '80 s experience. *J Media Econ* 2(2):35–50
- Sochay S (1994) Predicting the performance of motion pictures. *J Media Econ* 7(4):1–20
- Park J, Chung Y, Cho Y (2015) Using the hierarchical linear model to forecast movie box-office performance. the effect of online word of mouth. *Asia Pac J Inf Syst* 25(3):563–578
- Karniouchina EV (2010) Impact of star and movie buzz on motion picture distribution and box-office revenue. *Int J Res Mark* 28(1):62–74
- Chakravarty A, Liu Y, Mazumdar T (2009) The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation. *Soc Sci Electron Publ* 24(24):185–197
- Asur S, Huberman BA (2010) Predicting the future with social media. *Proc of Wiat* 7(2):492–499
- Koehler-Derrick G (2013) Quantifying anecdotes: google search data and political developments in Egypt. *Ps Political Sci Politics* 46(2):291–298
- Wang X (2015) The brief introduction of the history of box-office research. Doctorate dissertation, Chongqing University
- Jedidi K, Krider R, Weinberg C (1998) Clustering at the Movies. *Market Lett* 9(4):393–405
- Li B, Lu F, Zhao X, Wang Q, Wang S (2010) Chinese movies' life cycle model and empirical analysis. *Syst Eng Theory Prac* 30(10):1790–1797
- Yahav Inbal (2016) Network analysis: understanding consumers' choice in the film industry and predicting pre-released weekly box-office revenue. *Appl Stoch Models Bus Ind* 32(4):409–422
- Lu Y et al (2013) Implementation of the fuzzy C-means clustering algorithm in meteorological data. *Int J Database Theory Appl* 6(6):1–18
- Yang L et al (2015) A novel combination forecasting algorithm based on time series. *Int J Database Theory Appl* 8(2):157–170
- Möller-Levet CS, Klawonn F, Cho KH, Wolkenhauer O (2003) Fuzzy clustering of short time-series and unevenly distributed sampling points. In: Berthold MR, Lenz HJ, Bradley E, Kruse R, Borgelt C (eds) *Advances in intelligent data analysis V. IDA 2003. Lecture notes in computer science*, vol 2810. Springer, Berlin, Heidelberg, pp 330–340
- Elhamifar E, Vidal R (2013) Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans Pattern Anal Mach Intell* 35(11):2765–2781
- Macqueen J (1967) Some methods for classification and analysis of multi variate observations. In: *Proceedings of, Berkeley symposium on mathematical statistics and probability*, pp 281–297
- Zhang T, Chen L, Ma F (2014) An improved algorithm of rough K-means clustering based on variable weighted distance measure. *Int J Database Theory Appl* 7(6):163–174
- Zhang Z, Wang C (2015) Research on the AE signal de-noising based on k-means clustering and the wavelet transform. *Int J Multimed Ubiquitous Eng* 10(7):223–228
- von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17:395–416
- Bhissy KE, Faleet FE, Ashour W (2014) Spectral clustering using optimized Gaussian Kernel function. *Int J Artif Intell Appl Smart Device* 2(1):41–56
- Malik J, Shi J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905
- Vidal R (2011) Subspace Cluster Application in motion segmentation and face clustering. In: *IEEE signal processing magazine* 2011, March
- Wang WW et al (2015) A survey on sparse subspace clustering. *Zidonghua Xuebao/acta Automatica Sinica* 41(8):1373–1384
- Elhamifar E, Vidal R (2009) Sparse subspace clustering. In: *IEEE conference on computer vision and pattern recognition, 2009, CVPR 2009. IEEE*, pp 2790–2797
- Li CG, Vidal R (2015) Structured sparse subspace clustering: a unified optimization framework. In: *IEEE conference on computer vision and pattern recognition*, pp 277–286