Distributed Feature Selection for Efficient Economic Big Data Analysis

Liang Zhao, Zhikui Chen, Senior Member, IEEE, Yueming Hu, Geyong Min, Senior Member, IEEE, and Zhaohua Jiang

Abstract—With the rapidly increasing popularity of economic activities, a large amount of economic data is being collected. Although such data offers super opportunities for economic analysis, its low-quality, high-dimensionality and huge-volume pose great challenges on efficient analysis of economic big data. The existing methods have primarily analyzed economic data from the perspective of econometrics, which involves limited indicators and demands prior knowledge of economists. When embracing large varieties of economic factors, these methods tend to yield unsatisfactory performance. To address the challenges, this paper presents a new framework for efficient analysis of high-dimensional economic big data based on innovative distributed feature selection. Specifically, the framework combines the methods of economic feature selection and econometric model construction to reveal the hidden patterns for economic development. The functionality rests on three pillars: (i) novel data pre-processing techniques to prepare high-quality economic data, (ii) an innovative distributed feature identification solution to locate important and representative economic indicators from multidimensional data sets, and (iii) new econometric models to capture the hidden patterns for economic development. The experimental results on the economic data collected in Dalian, China, demonstrate that our proposed framework and methods have superior performance in analyzing enormous economic data.

Index Terms—feature selection, big data, subtractive clustering, collaborative theory, economy, urbanization

1 INTRODUCTION

Big data, as a term often defined around four V’s: Volume, Velocity, Variety, and Veracity has attracted many interests in solving social and economic problems, with anticipation of efficient organizations and decision-making [1]. For example, the World Economic Forum claimed that big data had significant and would provide new opportunities for international development in 2012 [2]. The White House also published the white paper in May 2014, stating that big data offered a marvelous opportunity for the economy, people’s health and education, national security, and energy efficiency of the United States [3]. However, only having massive data is inadequate, because our interests are focused on the valuable information, that is usually characterized by ‘Value’ instead of the four V’s, buried in the mass [37]. Therefore, to support social and economic development, the key is to capture valuable information, meanings, and insights hidden in big data.

With the increasing popularity of economic activities, a large number of factors and records are involved in economic development. At present, the volume of data in many financial institutions is more than 100TB in China. Meanwhile, an average of about 820GB data is produced continuously for 1 million dollars in revenue for a bank. In addition, the electronic commerce and other economic activities also produce enormous data for economic analysis constantly. For example, in double eleven shopping festival of Alibaba 2014, there were a total of 240 million Internet users visiting Taobao, making the trading volume peak at 2.85 million in one minute. The total turnover reached at 57.1 billion yuan, resulting in 278.5 million package deliveries. While all of these provide sufficient information for economic analysis, the issues of dimension and volume overload pose great challenges: (1) The collected huge volume data usually contains incomplete, incorrect and nonstandard items, which are difficult for processing. (2) The high-dimensionality of economic indicators makes manual factors selection for economic model construction impossible. (3) Statistical analysis software (e.g. Statistical Product and Service Solutions, SPSS) often generates runtime errors when dealing with the high-dimensionality and huge-volume economic data. Hence, it is necessary to provide an efficient way to extract the useful features contained in the massive data. Then the extracted features can be used to identify valuable information through economic models analysis. Such valuable information extraction process calls for novel economic big data analysis frameworks and advanced mining techniques.

Unfortunately, there are few intelligent schemas that can be used to gain actionable knowledge and valuable insights from the large amount of economic data. For economic development, most of the existing methods are
involved with econometric analysis [4-6], including basic element method, cost saving method, elements and internal associations method, and retarded economy method. They exploit econometric models, such as cointegration model [7], regression model [8], semi-parametric model [9], hypothesis model [10] and hybrid model, to quantitatively analyze the relations between response indicators and economic development. Thus the effects of them on economic development can be obtained. However, most existing methods identify the response factors related to economic development based on past experience and directly embody them into production function to build the correlations with economic growth, overlooking the indirect effects caused by other factors related to them. Besides, the existing methods rely too much on the knowledge of economists and embrace limited indicators and records for analysis, without fully considering the intrinsic characteristics of high-dimensional economic data. Therefore, they cannot effectively reveal the impacts of response indicators on economic development.

To address these challenges, we explore the hidden relations between economy and its response indicators from a new angle and extract the meaningful knowledge from economic big data in order to derive right insights and conclusions based on an innovative distributed feature selection framework that integrates advanced feature selection techniques and econometric methods. First, in order to reduce the noise yet promote the data quality, we propose to use usability preprocessing, relative annual price computation, growth rate computation and normalization techniques to clean and transform the collected economic big data. Then, to distill the features related to economic development from high-dimensional economic data, distributed feature selection methods are proposed to quickly partition the importance of given economic indicators. After that, the relations between response indicators and economic growth can be established by conducting correlative and collaborative analysis. Our main contributions are summarized as follows:

- We present a novel framework combining distributed feature selection methods and econometric models for efficient economic analysis, which can reveal the valuable insights from the low-quality, high-dimensionality, and huge-volume economic big data.

- We develop a subtractive clustering based feature selection algorithm and an attribute coordination based clustering algorithm to select and identify the important features of data in horizontally and vertically. Also, we extend these two methods to distributed platform for economic big data analysis.

- We conduct correlative and collaborative analysis simultaneously to explore the direct and indirect relations between economy and its response indicators based on the identified economic features.

- We evaluate the proposed framework and algorithms on the economic development data in Dalian, a fast developing city in China, over the past 30 years. Extensive experiments and analysis demonstrate that the designed framework and algorithms can distill the hidden patterns of economic development efficiently and the achieved results accord with the actual development situation in Dalian city.

The rest of this paper is organized as follows. Section 2 reviews related works on feature selection and econometric analysis methods. Section 3 formulates the problem to be addressed and introduces our proposed framework for economic big data analysis. The subtractive clustering based feature selection method and attribute coordination based clustering method, as well as their parallel methods are described in Section 4. Section 5 presents the processes of constructing economic models and demonstrates the efficiency of the proposed methods through a case study. Section 6 concludes the paper and directs future work.

2 RELATED WORK

This section reviews related works on feature selection and econometric methods.

2.1 The feature selection methods

Feature selection aims to process multidimensional data by detecting the relevant features and discarding the irrelevant ones. Effective feature selection can lead to reduction of measurement costs yet generate a better understanding of the original domain [11, 12, 30, 31, 33]. With respect to different selection strategies, feature selection algorithms can be categorized into four groups, namely the filter, wrapper, embedded, and hybrid methods.

The filter methods present the feature selection process independent of any classifier and evaluate the relevance of a feature by studying the characteristics of training data using certain statistical criteria. The correlation-based feature selection [13], consistency-based filter [14], information gain [15], relief [16], fisher score [17], and minimum redundancy maximum relevance [18] are the most representative filter techniques.

The wrapper methods integrate a classifier, such as SVM [21], KNN [25], and LDA [12], to select a set of features that have the most discriminative power. Representative wrapper feature selection methods include: wrapperC4.5 [19], wrapperSVM, FSSEM [20], and $l_1$SVM [21]. Other examples of the wrapper method could be any combination of a preferred search strategy and given classifiers.

The embedded methods perform feature selection in the process of training and achieve model fitting to a given learning mechanism simultaneously. For example, SVM-RFE [22] trains the current features of the given data set by a SVM classifier and removes the least important features indicated by the SVM iteratively to achieve feature selection. Other embedded methods include FS-P [23], BlogReg and SBMLR [24].

In summary, the filter methods, independent of any classifier, have lower computational complexity than wrapper methods yet with favorable generalization ability. Unlike filters, the wrapper methods are superior to filters in terms of classification accuracy, whereas they take more time due to the cost of expensive computation. The embedded methods, with lower computational cost than wrappers, are also integrated with classifiers, leading the risk of over-fitting.
Due to the shortcomings in each method, the hybrid methods [26, 27, 29] are proposed to bridge the gaps between them. However, the existing feature selection methods are incapable of being adapted to economic analysis. Since they analyze the data through its inherent knowledge characteristics, they cannot identify the feature cointegration and intrinsic association between economic indicators. Besides, the low-quality and huge-volume characteristics of economic big data present great challenges when the existing feature selection methods are directly applied to process inductive analysis. 

2.2 The econometric methods

Econometric analysis, based on economic theory and data, uses mathematical and statistical methods to study the quantitative relations and rules of economy [4,5]. The existing econometric studies on economic development and its response factors address the following aspects:

First, basic elements are applied to describe the mechanism of economic growth. The economic growth can be promoted by increasing consumption and investment, as well as affecting related decisive factors. When approaching economic analysis, the contributing factors are selected to identify the relations between them and economic development. Second, from the perspective of cost saving, urbanization can bring more workforces into city, which reduces the economic costs and boosts facilities sharing to cut down transaction costs. Meanwhile, through the agglomeration and diffusion effects, the economic growth can be accelerated. Third, elements and internal associations are involved to comprehensively explain the correlations between economy and its decisive factors. For example, Brant integrates two aggregate production function models, one with urbanization as a shift factor and the other that combines energy consumption and physical capital, to estimate the internal relevance among urbanization, energy consumption, and economic growth [6]. In addition, some researchers pose retarded economy theory to argue the restraining factors for economic development.

Moreover, there are an army of quantitative studies concentrating on this thesis [6-10], such as cointegration analysis, regression analysis, semi-parametric methods, hypothesis methods and hybrid methods. Sajal et al. approach threshold cointegration method to examine the cointegrating relationship between energy consumption, urbanization and economic activity for India [7]. In [8], the authors use a regression model, that allows the relationship between finance and economic growth to be piecewise linear, based on the concept of threshold effects to reveal the effects of finance on economic growth. By approaching data on developing economies, the semi-parametric method can estimate the potentially nonlinear effects of inflation on economic growth [9]. Moreover, in [10], the hypothesis is established that variation in migratory distance has a long-lasting effect on genetic diversity and the pattern of economic development. Based on this, the effects of genetic diversity on economic development can be obtained by approaching regression analysis. Although all the methods mentioned above can shed light on the patterns of economic development, they rely too much on the past experience and the knowledge of economists. Besides, they involve limited indicators and records for analysis, which will yield unsatisfactory results when approaching high-dimensional economic data.

3 DISTRIBUTED ECONOMIC BIG DATA ANALYSIS

In this section, we define the problem of statement of economic big data analysis, and then present a framework based on distributed feature selection.

3.1 Problem statement

The increasing economy related activities provide a wide range of indicators and records for economic analysis. Facing such large amount of data, how to detect useful information from it has drawn extensive attention in academia and industry. Traditional econometric methods cannot embrace the high-dimensionality data since they only involve limited economic factors for model construction based on past experiences. For example, some economists analyze economic development from the perspective of industrial structure. They select three indicators, namely the added value of primary industry, secondary industry and tertiary industry, to establish the production function for predicting GDP growth. Obviously, the obtained result is not persuasive because many other indicators also have impact on the economy. Besides, the existing statistical analysis software (e.g. SPSS) would generate runtime errors when dealing with the high-dimensionality and huge-volume economic data. While some methods are able to process the massive data, their computation costs are expensive [26-28]. Therefore, we aim to provide an efficient way to bridge the gap between data analysis methods and economic big data in real word. Specifically, it consists of two major tasks.

Task 1: Feature Selection. Let \( A = \{a_1, a_2, ..., a_m\} \) be a corpus of \( m \) economic indicators. Among these \( m \) indicators, there are \( m' \) features more relevant to economic development than others. And they can be grouped into \( k \) clusters according to their internal relevances. We aim to select the \( m' \) features and partition them to \( k \) groups \( \{c_1, c_2, ..., c_k\} \) with the representative features as centroids.

Task 2: Econometric Model Construction. For each cluster \( c_i \), we aim to conduct correlative analysis between the representative feature and other related ones to generate relational model. By combining all the models based on collaborative analysis, we can establish the economic prediction model.

Economic big data analysis is important and challenging in many ways. In the next subsection, we present a novel framework by combing distributed feature selection and econometric analysis to achieve the task of predicting economic development.

3.2 The framework of economic big data analysis

Our proposed framework consists of three phases, 1) Economic Data Preprocess, 2) Economic Feature Selection, and 3) Economic Model Construction, as shown in Fig. 1. Specially, to speed up the process of data analysis, the Economic Data Preprocess and Economic
Our objectives of such an ideal approach are two-fold: (i) the parallel subtractive clustering is generalized to select important attributes, and (ii) the attribute coordination based parallel clustering is designed to identify representative ones. Thus, we can make full use of the representative factors and their related important factors to mine the direct and indirect effects on economic development.
correlation analysis on economic data, the important and representative records and indicators can be identified. Subtractive clustering (SC) [34], a density-based clustering algorithm, is a favorable method to investigate the correlations between data samples. It assumes that each data point is a potential cluster center and calculates a measure of the likelihood based on the density of surrounding data points. In this way, it can construct the relationships among all the data points. When decomposing the relationships to a same center, it can construct the relationships among all the data records of the most representative for economic development. According to this idea, we use SC to identify the important indicators for economic analysis.

As shown in Fig. 2, the economic data contains dozens of records, with a range of indicators, sorted by year. For each record, its density value contributed by other records can be calculated as follows:

\[ D_i = \sum_{j=1}^{n} \exp \left[ -\frac{||x_i - x_j||^2}{(0.5r^*)^2} \right]. \]  

(1)

Herein, the data set \( \{x_1, x_2, ..., x_n\} \) is denoted as matrix \( X \in \mathbb{R}^{n \times m} \) with \( n \) records and \( m \) economic factors, and \( r^* \) is a positive constant value corresponds to a data point with many neighborhood data points. Hence, the point with the highest density is selected as the first cluster center. In other words, the annual data record of the most representative for economic development is found. Then, in order to avoid the points near the first cluster center being selected as other centers of clusters, an amount of density proportional is subtracted from each point to its distance from the first cluster center. After the reduction, the data point with the highest remaining density is selected as the second cluster center and the density of each data point is further reduced according to its distance to the second cluster center. Generally, after the \( k \)-th cluster center \( x_{ck} \) is obtained with density \( D_{ck} \), the density of each data point is updated by:

\[ D_i = D_i - D_{ck} \exp \left[ -\frac{||x_i - x_{ck}||^2}{(0.5r^*)^2} \right]. \]  

(2)

The processes of finding new cluster centers and reducing the density for each data point iterate until the remaining densities of all data points are bounded by some fraction of the density \( D_{ck} \) of the first cluster center. The stopping criterion usually adopted is \( D_{ck}/D_{ci} < \varepsilon \). Until now, all the typical annual data records for economic development are selected for the past decades. They can provide references for further development of economy.

Fortunately, we find that the density measure implies the relations between the features and the clusters. For example, in Eq. (2) the density value \( D_i \) is affected by the sample feature values \( A_a \) (see Fig. 2). So we define the contribution of attribute \( a \) to sample \( i \) as:

\[ I(i, a)_k = \sum_{p=1}^{n} \left| \frac{\partial D_i}{\partial x_{pa}} \right|, \]  

(3)

which means how much information there is to cluster the \( i \)-th sample with the \( a \)-th feature. According to Eq. (3), \( \frac{\partial D_i}{\partial x_{pa}} \) has the following four forms, in which part of the records are selected to calculate the feature contribution, as shown in Fig. 3:

(i) when \( i = p, j = p \),

\[ \frac{\partial D_i}{\partial x_{pa}} = 4(x_{pa} - x_{ca}) \exp \left( -\frac{2 \sum_{r=1}^{m} (x_{pr} - x_{cr})^2}{(r^*/2)^2} \right), \]  

(4)

(ii) when \( i = p, j \neq p \),

\[ \frac{\partial D_i}{\partial x_{pt}} = \frac{-2(x_{pa} - x_{ca})}{(r^*/2)^2} \exp \left( -\sum_{r=1}^{m} (x_{pr} - x_{cr})^2/(r^*/2)^2 \right) - \]

\[ \exp \left( -\sum_{r=1}^{m} (x_{cr} - x_{ir})^2/(r^*/2)^2 \right) \]  

(5)

(iii) when \( i \neq p, j = p \),

\[ \frac{\partial D_i}{\partial x_{pa}} = \frac{2(x_{ca} - x_{pa})}{(r^*/2)^2} \exp \left( -\sum_{r=1}^{m} (x_{pr} - x_{cr})^2/(r^*/2)^2 \right) - \]

\[ \exp \left( -\sum_{r=1}^{m} (x_{cr} - x_{ir})^2/(r^*/2)^2 \right) \]  

(6)

(iv) and when \( i \neq p, j \neq p, \frac{\partial D_i}{\partial x_{pa}} = 0 \).

Here, \( c_k \) is the \( k \)-th cluster center chosen by SC. Hence, the importance of the \( a \)-th attribute to select the \( k \)-th representative economic record can be defined as:

\[ I(a)_k = \sum_{i=1}^{n} I(i, a)_k. \]  

(7)
The attributes with the higher-ranking value contain more information of clusters than others, namely they have powerful impacts on typical economic phenomena analysis. The detailed algorithm to select the important attributes from a data set is summarized in Algorithm 1. In Algorithm 1, the neighborhood radius is initialized in Step 1. Then the representative economic records named as cluster centers are selected from Step 2 to Step 10. Steps 11 to 16 calculate the importance of attributes for preserving the correlations among records in clustering.

4.2 Distributed important attribute selection

It can be seen from Fig. 2 and Fig. 3 that the main procedure of important attribute selection is to calculate the distances between samples. For a large volume of data, computing the distance matrix $G$ itself becomes comparatively expensive. Hence, we investigate a strategy to reduce this cost via optimization, in particular via the MapReduce parallelization.

In the parallelization of calculating distance matrix $G$, the idea of matrix multiplication is adopted. Two copies of the original data set are generated at first, naming as $Y$, and $Z$. For each record in $Y$, the distances between it and the records in $Z$ need to be calculated respectively. Obviously, the records in $Y$ (or $Z$) are independent of each other, thus $Y, Z$ can be divided into equal blocks by row. Then the data blocks are sent to different data processing nodes in the distributed system. By decomposing, mapping, reducing and integrating operations, the distance matrix $G$ is achieved. The detailed algorithm to parallel this process is summarized in Algorithm 2.

In Algorithm 2, the diagonal elements of distance matrix $G$ are set to be ‘0’ in Step 1. Steps 3 to 8 generate the distance items for calculating distances between different data points. After that, the distance items are matched and sorted to calculate the corresponding distances (see Steps 9 to 14). For Algorithm 2, the complexity of computing the distances between samples is cut from $O(mn^2)$ to $O(mn^2/q) + M(q)$, where $q$ and $M(q)$ are the number of computing nodes and the additional overhead of the distributed system. Considering the distance matrix $G$, the distances between sample points are no longer computed in the parallelization of Algorithm 1 based on MapReduce.

4.3 Representative attribute identification

Traditional subtractive clustering approaches the Euclidean distance to partition sample data points, which takes no consideration of the relevance and dependency among them. For some economic indicators, there are cointegration relations among them. Thus, a method which makes better use of the indicators’ relations is necessary for representative attributes identification. In this subsection, we propose to utilize an improved subtractive clustering algorithm to classify the attributes. By combining the attribute coordination with SC, it can identify the representative attributes.

Assuming that $a = \{a_1, a_2, ..., a_n\}^T, b = \{b_1, b_2, ..., b_n\}^T$ are two attributes:

Algorithm 1: Important attribute selection

**Input:** Data matrix $X \in \mathbb{R}^{n \times m}$, and parameter $\varepsilon, \sigma$.

**Output:** Important attributes and cluster centers for $X$.

1. Initialize the neighborhood radius $r^* = \sqrt{\frac{\sum_{j=1}^{n} \sum_{i=1}^{n} ||x_i - x_j||^2}{n(n-1)}}$, and Euclidean distance matrix $G$ between data points;
2. for each data point $x_i \in X, i = 1...n$ do
3. Calculate the density $D_i$ as Eq. (1);
4. end for
5. The point with the highest density $D_{c_1}$ is selected as the first center. Set $k = 1$;
6. while $\frac{D_{c_k}}{D_{c_{k-1}}} > \varepsilon$ do
7. for each point $x_i \in X, i = 1...n$, except the chosen centers do
8. Update the density $D_i$ as Eq. (2);
9. end for
10. The new center with the highest density $D_{c_k}$ is selected. Set $k = k + 1$;
11. for attribute $a \in A, A$ is the attribute set of $X$ do
12. for each data point $x_i \in X, i = 1...n$ do
13. Calculate the effect of attribute $a$ to sample $i$ in clustering as Eq. (7); // The importance of the $a$-th attribute to select the $k$-th representative record is obtained.
14. end for
15. Sum the effects of attribute $a$ to all samples in clustering as Eq. (7); // The importance of the $a$-th attribute for clustering.
16. end for
17. end while
18. The attributes with $I = \sum_{j=2}^{k} I(a)_j \frac{1}{(k-1)} > \sigma$ are selected; // $I$ is the importance of the $a$-th attribute for clustering.
Algorithm 2: Parallelization of distance matrix calculation  
Input: Two copies of dataset \(X \in \mathbb{R}^{n \times m}\), named as \(Y, Z\).  
Output: Distance Matrix \(G \in \mathbb{R}^{n \times n}\).  
1. The diagonal elements of \(G\) are set to ‘0’; //\(G\) is symmetric, and its upper triangular matrix is calculated.  
2. Function \texttt{Map} reads the records of matrix \(Y\), and \(Z\);  
3. for each item \(Y[i,a], i = 1...n, a = 1...m\) do  
4. Generate a series of key-value pairs \(<(i,t),(Y,a,Y[i,a])>\), \(n-i > 0\), \(t\) decreases progressively from \(n\) to \(n-i\);  
5. end for  
6. for each item \(Z[j,a], j = 1...n, a = 1...m\) do  
7. Generate a series of key-value pairs \(<(t,j),(Z,a,Z[j,a])>\), \(j-1 > 0\), \(t\) increases progressively from \(1\) to \(j-1\);  
8. end for  
9. Function \texttt{Reduce} collects the items \((Y,a,Y[i,a])\) and \((Z,a,Z[j,a])\) that have the same key, and puts them in different tables according to their matrix value (\(Y\) or \(Z\)).  
10. In each table, the items are sorted by ‘\(a\)’ in ascending order;  
11. for item \(Y[i,a],Z[j,a], a = 1...m\) in different tables do  
12. \(\text{dist}_a = \|Y[i,a] - Z[j,a]\|^2\);  
13. end for  
14. Sum the \(\text{dist}_a\) for all ‘\(a\)’. We can get the distance \(G_{ij}\) between records \(i\) and \(j\);  
15. Function \texttt{Reduce} outputs the key-value pairs \(<(i,j),G_{ij}>\);  
16. The Distance Matrix \(G\) containing \(n \times n\) items is achieved;  

Definition 1 (Attribute Summation) The attribute summation between them is defined as  
\[
\text{attr}_\text{sum}(a,b) = \|a + b\|^2 = \sum_{i=1}^{n} (a_i + b_i)^2. \tag{8}
\]

Definition 2 (Attribute Product) The attribute product between them is defined as  
\[
\text{attr}_\text{pro}(a,b) = \|ab\|^2 = \sum_{i=1}^{n} (a_i b_i)^2. \tag{9}
\]

Definition 3 (Attribute Coordination) Given the attribute summation and attribute product, the attribute coordination between \(a\) and \(b\) is defined as  
\[
\text{attr}_\text{coo}(a,b) = \frac{\text{attr}_\text{sum}(a,b)}{\text{attr}_\text{pro}(a,b)} = \frac{\|a + b\|^2}{\|ab\|^2}. \tag{10}
\]

In order to make attribute coordination well suited for subtractive clustering, it requires the following properties.  

Properties of Attribute Coordination:  
Property 1: Attribute coordination is not less than zero.  
Property 2: The smaller the attribute coordination value is, the better coordination two attributes get, namely the more correlative they are.  
Property 3: The attribute coordination between one attribute and itself is zero.  
By combining the attribute coordination with subtractive clustering, the Eq. (2) and Eq. (3) can be written as Eq. (11) and Eq. (12).  
\[
D_i = \sum_{j=1}^{m} \exp \left[ -\frac{|A_i + A_j|^2}{|A_i + A_j|^2(0.5R^2)^2} \right] \tag{11}
\]
\[
D_i = D_i - D_{ck} \exp \left[ -\frac{|A_i + A_{ck}|^2}{|A_i + A_{ck}|^2(0.5R^2)^2} \right] \tag{12}
\]

Until now, we can cluster the important attributes and identify the representative attributes more persuasively. Algorithm 3 presents the processes in details. In Algorithm 3, the neighborhood radius is initialized in Step 1. Then the representative economic attributes are identified by the attribute coordination based subtractive clustering (see Steps 2 to 11).  

Algorithm 3: Representative attribute identification  
Input: Attribute set \(AS = \{A_1, A_2, ..., A_m\}\), parameter \(\varepsilon\).  
Output: Representative attributes.  
1. Set the records of attributes as sample data, and initialize the neighborhood radius \(R^* = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} \frac{|A_i + A_j|^2}{|A_i + A_j|^2}} / m(m-1)\);  
2. for each data point \(A_i \in AS, i = 1...m\) do  
3. Calculate the density as Eq. (11);  
4. end for  
5. The attribute with the highest density is selected as the first representative attribute. Set centers number \(k = 1\);  
6. while \(\frac{D_{ck}}{D_i} > \varepsilon\) do  
7. for attributes (except the chosen ones) \(A_i \in AS, i = 1...m - k\) do  
8. Update the density as Eq. (12);  
9. end for  
10. The new representative attribute with the highest density is selected. Set \(k = k + 1\);  
11. end while  
12. Output all the representative attributes;  

4.4 Distributed representative attribute identification  
Similar to Algorithm 1, the major procedure of Algorithm 3 is to calculate the coordination between attributes. We define an attribute coordination matrix to optimize it.  

Definition 4 (Attribute Coordination Matrix) Given matrices \(A\) and \(B\) have \(n\) records and \(m\) attributes, \(A \Delta B \in \mathbb{R}^m \times m\) can be defined as an attribute coordination matrix  
\[
(A \Delta B)_{ij} = \frac{\sum_{i=1}^{n} (a_{ri} + b_{rij})^2}{\sum_{i=1}^{n} (a_{ri} + b_{rij})^2} \tag{13}
\]
\[
(A \Delta B)_{ij} = \frac{(a_{1i} + b_{r1j})^2 + (a_{2i} + b_{r2j})^2 + ... + (a_{ni} + b_{rnj})^2}{(a_{1i} + b_{r1j})^2 + (a_{2i} + b_{r2j})^2 + ... + (a_{ni} + b_{rnj})^2}
\]

\((A \Delta B)_{ij}\), the item of row \(i\) and column \(j\) in the matrix, is the coordination between attributes \(i\) and \(j\).  
Further, the attribute set is decomposed and the parallel model-MapReduce is applied to speed up solving coordination matrix. More details are in Algorithm 4.
The aim of this paper is to establish the analytical models for economic development so that the hidden patterns of economy and the correlations between economy and its response indicators can be captured. In this section, we describe the construction of economic models in details based on the important and representative attributes identified by the proposed feature selection method at first. Then the dependencies between economic growth and its response indicators are discussed. To present some concreteness to our discussion, the economic data in Dalian, China, is exploited to construct the model of factor analysis of economic growth. After that, the correlation between urbanization and economic growth is obtained based on the constructed economic models.

### 5.1 Selection of representative attributes

The focus of model construction in this case study is to reveal the contribution of decisive factors to economic growth and the hidden relationship between urbanization and economic growth from enormous economic data. So 52 economic indicators of more than 300 towns and streets over the past 30 years are collected in Dalian. For meaningful information is often buried in the mass, we extract the indicators that actually make sense to establish correlative models. After eliminating redundancy and abnormality, 32 factors are selected for important attributes analysis. Fig. 4 shows the importance of these attributes, which are obtained by the algorithms in Section 4.

It can be seen from Fig. 4 that the indicators “Built-up Area” and “Science and Technology Input” own the highest and lowest importance respectively. The “Built-up Area” reflects the functions of urbanization for economic growth in Dalian. Over the past decade, urbanization has been steadily advanced with a large percentage of land converted to built-up area in Dalian. It promotes the economic growth through increasing consumption and investment, as well as affecting the related decisive factors. The “Science and Technology Input” has had less impact than other factors on the economic growth in the past few years in Dalian, which means that the capacity for innovation is not enough. It is in line with the actual situation. So the investment in science and technology should be increased to promote the innovation of economic growth.

Fig. 4 reveals that different economic attributes have different contributions to economic development, which can provide some references for economists in construction of economic models, and thus to reduce the task of manual analysis. In this section, the important attributes whose attribute weights are not less than $\sigma = 6$ are selected for analysis. Afterwards, the proposed algorithms in Sections 4.3 and 4.4 are approached to extract the representative ones (see Table 1) for models construction.

Table 1 shows that 10 representative attributes are extracted and all the important attributes are partitioned into four groups corresponding to different economic research perspectives.

### 5.2 Construction of economic models

We conduct the collaborative theory [35] to analyze the contributions of economic growth factors based on the selected indicators in this subsection. Collaborative theory proposes the method of income value decomposition to research on economic growth. It holds the view that the output value of

---

**Algorithm 4**: Parallelization of coordination matrix calculation

**Input**: $A, B$ are two copies of $AS = X^T \in \mathbb{R}^{m \times n}$.

**Output**: Attribute Coordination Matrix $A\Delta B \in \mathbb{R}^{m \times m}$.

1. The diagonal elements of $A\Delta B$ are all set to be ‘0’; //The upper triangular matrix of $A\Delta B$ is calculated.
2. Function Map reads records of matrix $A, B$ simultaneously;
3. for each item $A[i, c], i = 1...m, c = 1...n$ do
   4. Generate a series of key-value pairs $(i, t), (A[c, A[i, c]])$, $m - i > 0$, $t$ decreases progressively from $m$ to $m - i$;
5. end for  
6. for each item $B[j, c], j = 1...m, c = 1...n$ do
   7. Generate a series of key-value pairs $(t, j), (B[c, B[j, c]])$, $j - 1 > 0$, $t$ increases progressively from 1 to $j - 1$;
8. end for  
9. Function Reduce collects the items $(A[c, A[i, c])$ and $(B[c, B[j, c)])$ that have the same key, and puts them in different tables according to their matrix value ($A$ or $B$);
10. In each table, the items are sorted by ‘$c$’ in ascending order;
11. for item $A[i, c], B[j, c], c = 1...n$ in different tables do
   12. $sum_1 = sum_1 + (A[i, c] + B[j, c])^2$, $sum_2 = sum_2 + (A[i, c] \times B[j, c])^2$;
13. end for  
14. We can get $A\Delta B_{ij} = sum_1/sum_2$, the coordination between attributes $i$ and $j$;
15. Function Reduce outputs the key-value pairs $(i, j), A\Delta B_{ij} > 0$, and $A\Delta B$ is obtained.

In Algorithm 4, the diagonal elements of attribute coordination matrix $A\Delta B$ are set to be ‘0’ in Step 1. Steps 3 to 8 generate the attribute coordination items for calculating attribute coordinates between different attributes. After that, the coordination items are matched and sorted to calculate the corresponding attribute coordinates (see Steps 9 to 14). It can be seen from Algorithm 4 that the time complexity of solving the attribute coordination matrix is also cut from $O(nm^2)$ to $O(nm^2/q) + M(q)$, here, $q$ is the number of computing nodes and $M(q)$ is the additional overhead of distributed systems.

According to the attribute coordination matrix $A\Delta B$, the coordination distances between attribute samples are no longer calculated in Algorithm 3. Moreover, by partitioning $A\Delta B$ and the attribute set, the parallelization of Algorithm 3 based on the MapReduce can be realized. In addition, when grouping attribute set with enormous records, the gained coordination matrix can make the execution time reduced considerably without computing every record’s items repeatedly.
TABLE 1
Results of representative attributes identification. The representative attributes are extracted from the selected important ones based on our proposed methods. Meanwhile, all the important attributes are partitioned into four groups corresponding to different economic research perspectives.

<table>
<thead>
<tr>
<th>Important attributes</th>
<th>Representative attributes</th>
<th>Research perspectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per capital income of farmers</td>
<td>Per capital income of farmers</td>
<td></td>
</tr>
<tr>
<td>Labor remuneration</td>
<td>Quantity of labor force</td>
<td>Analysis of economic growth factors</td>
</tr>
<tr>
<td>Urban resident population</td>
<td>Human capital stock</td>
<td></td>
</tr>
<tr>
<td>Quantity of labor force</td>
<td>Education years of employees</td>
<td></td>
</tr>
<tr>
<td>Human capital stock</td>
<td>Fixed-asset investment</td>
<td></td>
</tr>
<tr>
<td>Education years of employees</td>
<td>Foreign capital utilization</td>
<td></td>
</tr>
<tr>
<td>Fixed-asset investment</td>
<td>Domestic capital introduction</td>
<td></td>
</tr>
<tr>
<td>Foreign capital utilization</td>
<td>Deposit balance of financial institutions</td>
<td></td>
</tr>
<tr>
<td>Domestic capital introduction</td>
<td>Loan balance of financial institutions</td>
<td></td>
</tr>
<tr>
<td>Deposit balance of financial institutions</td>
<td>Real estate development investment</td>
<td></td>
</tr>
<tr>
<td>Loan balance of financial institutions</td>
<td>Infrastructure investment</td>
<td></td>
</tr>
<tr>
<td>Real estate development investment</td>
<td>Fixed capital stock</td>
<td></td>
</tr>
<tr>
<td>Infrastructure investment</td>
<td>Primary industry production value</td>
<td>Analysis of the industrial structure</td>
</tr>
<tr>
<td>Fixed capital stock</td>
<td>Secondary industry production value</td>
<td></td>
</tr>
<tr>
<td>Primary industry production value</td>
<td>Local financial revenue</td>
<td>Institutional functions for economic growth</td>
</tr>
<tr>
<td>Secondary industry production value</td>
<td>Local financial expenditure</td>
<td></td>
</tr>
<tr>
<td>Local financial revenue</td>
<td>Government fund expenditure</td>
<td></td>
</tr>
<tr>
<td>Local financial expenditure</td>
<td>Cargo throughput of port</td>
<td>Environment externality of economic growth</td>
</tr>
<tr>
<td>Government fund expenditure</td>
<td>Cargo throughput of port</td>
<td></td>
</tr>
<tr>
<td>Cargo throughput of port</td>
<td>Output value of foreign-funded enterprise</td>
<td></td>
</tr>
<tr>
<td>Output value of foreign-funded enterprise</td>
<td>Annual electricity consumption</td>
<td></td>
</tr>
<tr>
<td>Annual electricity consumption</td>
<td>Built-up area</td>
<td></td>
</tr>
<tr>
<td>Built-up area</td>
<td>Built-up area’s green coverage rate</td>
<td></td>
</tr>
</tbody>
</table>
the economic system includes not only the labor remuneration of the employees and the capital gain of the investors, but also the collaborative interests of them. Its form depends on the interaction of "Labor Force", "Fixed Capital", "Science and Technology", "Human Capital", "Institutional Factors" and "Environment Externalities".

In view of this thought, the economic growth model can be designed as

\[ GDP = Labor\ Remuneration + Capital\ Gain + Collaborative\ Interests + Others, \tag{14} \]

where "Collaborative Interests" promotes the mutual support, mutual benefit and common prosperity of laborers, investors and other stakeholders. "Collaborative Interests", called "item three", is independent of "Labor Remuneration" and "Capital Gain" and is usually regarded as the sources to improve innovation ability.

Coincidentally, the selected important indicators, the identified representative indicators and their corresponding economic research perspectives are in accord with collaborative theory for economic analysis (see Table 1). Notably, the importance of "Science and Technology Input" reflects that the capacity for innovation is not enough in Dalian, but we also want to verify how it affects the economic growth. Therefore, when combined with the extracted attributes in Section 5.1 and the innovation indicator, Eq. (14) can be written in quantitative form:

\[ Y = aL^\alpha H^\beta S^\gamma D^\delta + bK + cSD/K + u, \tag{15} \]

where \(aL^\alpha H^\beta S^\gamma D^\delta\) is "Labor Remuneration", \(bK\) is "Capital Gain", \(cSD/K\) is "Collaborative Interests" and \(u\) is the item "Others". \(Y, L, K, D, S, H\) stand for "GDP", "Quantity of labor force", "Fixed capital stock", "Fixed-asset investment", "Innovation input" and "Human capital" respectively. Parameters \(\alpha, \beta, \gamma, \delta, a, b, c\) are decided by item "Others", usually institutions and external environments.

Based on Eq. (15), we can deduce the following economic growth rate decomposition model:

\[ y = \frac{bK - cSD/K}{Y} \times \frac{k}{y} + \frac{cSD/K + a\delta L^\alpha H^\beta S^\gamma D^\delta d}{Y} \times \frac{d}{y} + \frac{cSD + a\beta L^\alpha H^\beta S^\gamma D^\delta s}{Y} \times \frac{s}{y} + \frac{a\alpha L^\alpha H^\beta S^\gamma D^\delta h}{Y} \times \frac{h}{y} + \frac{a\alpha L^\alpha H^\beta S^\gamma D^\delta l}{Y} \times \frac{l}{y} + i + e, \tag{16} \]

in which \(y, k, d, s, h, l\) are the rates of change of \(Y, K, D, S, H\), and \(L\) respectively. \(i\) is the effect of institutional innovation on economic growth, and \(e\) is the effect of environment externality on economic growth. As a result, the contribution of each factor to economic growth can be gained based on Eq. (16).

\[ \eta_K = \frac{bK - cSD/K}{Y} \times \frac{k}{y}, \tag{17} \]

\[ \eta_D = \frac{cSD/K + a\delta L^\alpha H^\beta S^\gamma D^\delta d}{Y} \times \frac{d}{y}, \tag{18} \]

\[ \eta_S = \frac{cSD + a\beta L^\alpha H^\beta S^\gamma D^\delta s}{Y} \times \frac{s}{y}, \tag{19} \]

\[ \eta_H = \frac{a\alpha L^\alpha H^\beta S^\gamma D^\delta h}{Y} \times \frac{h}{y}, \tag{20} \]

\[ \eta_L = \frac{a\alpha L^\alpha H^\beta S^\gamma D^\delta l}{Y} \times \frac{l}{y}. \tag{21} \]

Herein, \(\eta_K, \eta_D, \eta_S, \eta_H, \eta_L, \eta_I\) and \(\eta_E\) stand for the contributions of "Fixed capital stock", "Fixed-asset investment", "Innovation input", "Human capital", "Quantity of labor force", "Institutional innovation" and "Environment externality" to economic growth respectively.

Through regression analysis and data envelopment analysis based on the preprocessed data in Dalian, the contributions of driving factors (see Table 2) to economic growth and the model (see Eq. (24)) for economic development are obtained.

\[
\begin{align*}
\eta_I &= \frac{i}{y}, \tag{22} \\
\eta_E &= \frac{e}{y}. \tag{23}
\end{align*}
\]

As can be seen in Table 2, the contribution rate of "Fixed capital", including "Fixed capital stock" and "Fixed-asset investment", is over 50%, which indicates that the economic growth mainly depends on investment in Dalian. As for "Innovation input", also called "Science and Technology input", its contribution rate to economic growth is about 30%. To accelerate economic growth, the science and technology input should be further increased. It can also been seen from Table 2 that the "Institutional innovation" has neither hindered nor pulled economic growth over the period 2009-2013. Besides, the investment environment is favorable in Dalian, with its contribution rate negative. Hence, it is important to take full advantage of the existing investment environment and strengthen the institutional innovation to promote the economic development in Dalian.

\[
Y = 0.247(1.1)^{0.509}(SD/L)^{0.282} + 0.178K + 45.5SDH/K^2 + 1.08L. \tag{24}
\]

In Section 5.3, the obtained economic growth model as shown in Eq. (24), will be used to establish the correlation between economy and urbanization.

### 5.3 Correlation analysis between urbanization and economic development

As described in Section 5.1, urbanization has great effect on the economic development in Dalian. In this subsection, we study the correlation between urbanization and economic development. Based on the relevance analysis between urbanization and the constituent factors of the collaborative economic model in Section 5.2, we can establish the relationship between economic growth and urbanization.

As shown in Eq. (24), "Quantity of labor force", "Fixed capital stock", "Fixed-asset investment", "Innovation input"
and “Human capital” are the representative factors to determine the economic growth in Dalian. Hence, the correlative models (see Eqs. (25)-(28)) between urbanization and them can be established through regression analysis.

\[
H = 13.64U + 0.21, \quad (25)
\]
\[
L = 0.874UP, \quad (26)
\]
\[
D = 0.86UV - 380.5, \quad (27)
\]
\[
S = 0.222HLU + 0.012D - 22.8. \quad (28)
\]

Herein, \(U\) is denoted as the urbanization rate, which is the ratio of the urban resident population to the total population. \(P\) is the total resident population and \(V\) is the sum of the foreign capital utilization, domestic capital introduction, and loan balance of financial institutions.

When combining Eqs. (25)-(28) with Eq. (24) and calculating the derivative of \(Y\) on \(U\), the marginal income of 1 percent urbanization rate improvement on each factor can be gained in Table 3.

<table>
<thead>
<tr>
<th>Years</th>
<th>“Labor force”</th>
<th>“Human capital”</th>
<th>“Fixed-asset investment”</th>
<th>“Innovation input”</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>0.747</td>
<td>0.691</td>
<td>1.337</td>
<td>0.151</td>
</tr>
<tr>
<td>2005</td>
<td>1.382</td>
<td>1.625</td>
<td>1.753</td>
<td>0.238</td>
</tr>
<tr>
<td>2006</td>
<td>1.461</td>
<td>1.723</td>
<td>1.828</td>
<td>0.283</td>
</tr>
<tr>
<td>2007</td>
<td>1.542</td>
<td>1.845</td>
<td>1.878</td>
<td>0.316</td>
</tr>
<tr>
<td>2008</td>
<td>1.737</td>
<td>2.274</td>
<td>2.229</td>
<td>0.360</td>
</tr>
<tr>
<td>2009</td>
<td>1.845</td>
<td>2.481</td>
<td>2.549</td>
<td>0.447</td>
</tr>
<tr>
<td>2010</td>
<td>2.017</td>
<td>2.837</td>
<td>3.180</td>
<td>0.563</td>
</tr>
<tr>
<td>2011</td>
<td>2.078</td>
<td>3.103</td>
<td>3.704</td>
<td>0.644</td>
</tr>
<tr>
<td>2012</td>
<td>2.222</td>
<td>3.416</td>
<td>4.004</td>
<td>0.728</td>
</tr>
<tr>
<td>2013</td>
<td>2.326</td>
<td>3.571</td>
<td>4.125</td>
<td>0.753</td>
</tr>
</tbody>
</table>

Table 3 illustrates that the global urbanization in Dalian makes the major constituent factors of economic development grow steadily in recent years. As observed, the income of "Labor force" increased from 0.747 billion Yuan in 2001 to 2.326 billion in 2013, with urbanization rate being improved by 1 percent. The other factors saw the similar trend as well.

After integrating all equations, we can obtain the correlation between GDP(\(Y\)) and urbanization rate (\(U\)).

\[
Y = 0.247(13.64U + 0.21)0.509(0.874UP)^{0.227}(0.86UV - 380.5)^{0.282}(2.647U^3P + 0.01UV + 0.041U^2P - 27.366)^{0.282} + 0.178(0.9K_{t-1} + 0.86UV - 380.5) + 45.5(2.647U^3P + 0.01UV + 0.041U^2P - 27.366)(0.86UV - 380.5)(13.64U + 0.21)/(0.9K_{t-1} + 0.86UV - 380.5)^2 + 0.944UP \quad (29)
\]

Here, \(K_{t-1}\) stands for the fixed capital stock of the previous year. When calculating the derivative of \(Y\) on \(U\), the marginal income of urbanization on economic growth can be obtained, namely the pulling effect that raises one unit of urbanization rate on economic growth. Fig. 5 depicts the detailed information.

It can be seen from Fig. 5 that the contribution of urbanization to GDP has been back to positive territory since the implementation of the global urbanization in 2009, rising from 16% to 29%. That is to say, urbanization has become an important factor to stabilize the GDP growth rate in Dalian. For instance, the improvement of urbanization rate had made the GDP growth rate increase 2.6% in 2013, as shown by the circle-dotted line.

Moreover, we discuss the impact of urbanization on economic development from the perspective of industrial structure in Dalian.

Fig. 6 shows the pulling effects of urbanization on different industries in Dalian. In general, the urbanization makes the proportions of the second and third industry added value accounted for GDP increase continuously. Specially, the pulling effects of urbanization on the two industries climbed dramatically between 2005 and 2013. In contrast, the urbanization has negative impact on the first industry, with the proportion of the first industry added value in the total GDP added value decreasing gradually.
6 Conclusion

In this paper, we have proposed a novel feature selection based framework, aiming at effective and efficient analyzing the economic big data. In particular, it tries to learn the important features from the high-dimensionality, huge-volume, and low-quality economic data for economic model construction. Firstly, in order to reduce the noise and promote the data quality, the usability preprocessing, relative annual price computation, growth rate computation and normalization techniques are approached to clean and transform the collected economic big data. After that, a distributed subtractive clustering algorithm and its improved algorithm are proposed to construct a two-layer feature selection model, which selects the important features and identifies the representative ones of economic big data in horizontally and vertically. With the representative economic factors extracted by the feature selection model, we construct the collaborative model for driving factors analysis of economic development. Based on the collaborative model, collaborative analysis and correlative analysis are integrated to explore the direct and indirect relationships between response indicators and economy. The proposed framework and algorithms are evaluated on the economic development data in Dalian over the past 30 years. All experimental results demonstrate that our work not only agrees with the actual development situation in Dalian, but also distills the hidden relations between economy and urbanization efficiently.

In the future work, we plan to establish a platform of algorithm library based on the proposed framework. By analyzing and matching various of methods, it can provide more efficient solutions for economic and social development.

Acknowledgments

This work is supported by National Natural Science Foundation Project of China (U1301253), Science and Technology Planning Key Project of Guangdong Province, China (2015B010110006) and Research Office of Dalian Government in China.

References


Liang Zhao received his B.S. and M.S. degrees in Software Engineering from Dalian University of Technology, China, in 2011 and 2014, respectively. He is a doctoral student in the School of Software Technology, Dalian University of Technology. His research interests are usability of big data and cloud computing.

Zhikui Chen received his Ph.D. degree in Digital Signal Processing and M.S. degree in Mechanics from Chongqing University, China, in 1998 and 1993, respectively. He obtained his B.S. degree in the Department of Mathematics and Computer Science from Chongqing Normal University, China. Zhikui Chen is working as a full professor at Dalian University of Technology, China. He is leading the Institute of Ubiquitous Network and Computing of Dalian University of Technology. He was a general chair of IEEE Ithings2011 and IEEE Smartdata2015, advisor chair of IEEE Ithings2012-2015, and program chair of IEEE ICDH2014. His research interests are big data processing, mobile cloud computing, ubiquitous network and its computing. He is a senior member of IEEE.

Yueming Hu received his Ph.D. degree in Soil Science from Zhejiang Agricultural University, China, in 1997, and M.S. degree in Soil Science from Northwestern Agricultural University, China, in 1990. He is currently a professor in the College of Natural Resources and Environment, South China Agricultural University, Guangzhou, China. His research interests are land resource management, geographic information system application, and agricultural information.

Geyong Min is a Professor of High-Performance Computing and Networking with the Department of Mathematics and Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, U.K. He received the Ph.D. degree in Computing Science from the University of Glasgow, Glasgow, U.K., in 2003, and the B.Sc. degree in Computer Science from the Huazhong University of Science and Technology, Wuhan, China, in 1995. His research interests include next-generation internet, wireless communications, multimedia systems, information security, high-performance computing, ubiquitous computing, modeling, and performance engineering.

Zhaohua Jiang received his MS degree in Philosophy of Science from Northeast University, China in 1988. He is currently a professor in the the School of Public Administration and Law, Dalian University of Technology. His research interests are scientific & technological progress and economic growth, and economics of technological innovation.