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

- *Liang Zhao is with the School of Software Technology, Dalian University of Technology, Dalian 116600, China. E-mail: matthew1988zhao@mail.dlut.edu.cn.*
- *Zhikui Chen is with the School of Software Technology, Dalian University of Technology, and the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116600, China. E-mail: zkchen@dlut.edu.cn.*
- *Yueming Hu is with the College of Natural Resources and Environment, South China Agricultural University, Guangzhou 510642, China. E-mail: ymhu163@163.com.*
- *Geyong Min is with the College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, U.K. E-mail: g.min@exeter.ac.uk.*
- *Zhaohua Jiang is with the School of Public Administration and Law, Dalian University of Technology, Dalian 116024, China. E-mail: jiang_zhaohua@163.com.*

Manuscript received April 19, 2005; revised September 17, 2014.

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 indictors 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_1SVM [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 $Economic$

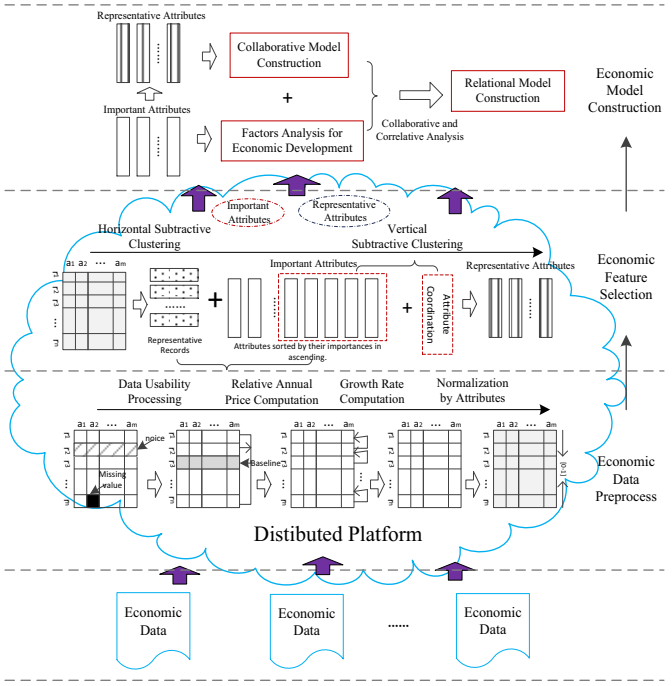

Fig. 1. The proposed framework for economic big data analysis. It includes three components: (1) Economic Data Preprocess; (2) Economic Feature Selection; and (3) Economic Model Construction.

Feature Selection are deployed in distributed platform [36].

- *Economic Data Preprocess*

The raw data always contains the most important information. However, it is difficult to mine useful information from the mass as it is mixing with incomplete, incorrect and nonstandard items. Thus the methods that can improve the data quality should be developed for economic big data analysis. We propose to exploit the methods of noise elimination [28] and missing value imputation [32] to enhance the data usability. For the influence of inflation or deflation, the currency prices corresponding to economic indicators in different years cannot be measured directly. In this paper, we project the economic data to the same domain with baseline data in 2012 based on corresponding price indexes, so that the data in different year can be processed fairly. As a rule of thumb, the growth rate of economic indicators can better reflect economic development than their raw forms. Hence, we compute the relative growth rates of a year to its previous year for all numerical indicators. Moreover, to avoid the influence of absolute values on the analytic results, the min-max normalization technique for all numerical attributes is approached to unify all attribute values to the same metric space.

- *Economic Feature Selection*

The preprocessed data obtained from the first phase is unsuitable for econometric analysis due to its high-dimensionality. Therefore, it is essential to select the representative economic indicators and their related important ones for econometric model construction. To tackle this problem, we propose a two-stage distributed subtractive clustering based feature selection method. Firstly, the important attributes that are more relevant to economic development are selected by the horizontal distributed subtractive clustering. Secondly, by approaching the improved attribute coordination based distributed subtractive clustering on the selected attributes vertically, we can gain the representative attributes.

- *Economic Model Construction*

With the combination of the selected indicators, we can construct the economic prediction models. However, a weakness of most traditional econometric methods for constructing models is that they take no consideration of the indirect relations between response indicators and economic factors. For example, many existing methods combine the representative factors with urbanization to establish the relational models between urbanization and economic development [6, 7]. Obviously, they ignore the indirect effects of urbanization on the important factors that are related to the representative ones. Hence, we integrate correlative and collaborative analysis simultaneously in this work to construct novel economic models.

In sum, our proposed framework outperforms the existing econometric methods for economic big data analysis. The economic big data usually has the characteristics of low-quality, high-dimensionality and huge-volume, which pose great challenges to existing econometric methods. To tackle these problems, we propose a three-layer model to embrace all related data for efficient economic analysis. Firstly, the low-quality and huge-volume economic data is cleaned to improve the data usability and transformed to consist with economic rules. After that, the attributes that can represent the high-dimensionality and huge-volume economic data are selected by the distributed feature selection method, which can fully consider the relationships among attributes yet reduce the influences of past experience in indicator selection for economic analysis. Finally, the correlative and collaborative analysis are approached to distill the direct and indirect corrections among the selected indicators, thus to construct the distinctive economic models.

4 A DISTRIBUTED FEATURE SELECTION MODEL

This paper aims to reduce the potentially huge set of candidate attributes produced by the preprocess layer to a small set of possible attributes, which are diverse and similar to the attributes in the original data set. However, there is no universal method for all problem settings, so we design a novel, systematic attribute selection approach for economic analysis. 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.

4.1 Important attribute selection

For economic analysis, some records may be related to other records and some indicators can be represented by the combination of other indicators. Therefore, by approaching

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 attribute, the contribution of the attribute to preserve the relationships can be achieved. According to this idea, we use SC to identify the important indicators for economic analysis.

As shown in Fig. 2, the economic data contains decades 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:

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

Herein, the data set $\{x_1, x_2, ..., x_n\}$ is denoted as matrix $X \in \Re^{n\times m}$ with n records and m economic factors, and r^* is a positive constant representing a normalized radius defining the neighborhood. According to Eq. (1), a high density 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_{c_k} is obtained with density D_{c_k}, the density of each data point is updated by:

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

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_{c_1} of the first cluster center. The stopping criterion usually adopted is $D_{c_k}/D_{c_1} < \varepsilon$. 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:

$$I(i,a)_k = \sum_{p=1}^{n} \left|\frac{\partial D_i}{\partial x_{pa}}\right|, \quad (3)$$

which means how much information there is to cluster the i-th sample with the a-th feature. According to Eq. (3), $\frac{\partial D_i}{\partial x_{pa}}$

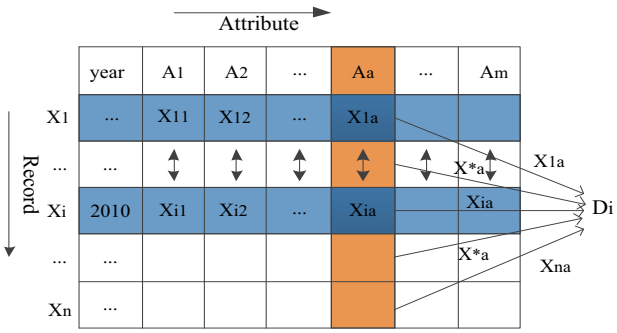

Fig. 2. The contribution of a feature to a record's density. When calculating density D_i for record x_i, all attribute values in A_a corresponding to records $x_1, x_2, ..., x_n$ are approached.

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 $i = p, j = p$,

$$\frac{\partial D_i}{\partial x_{pa}} = \frac{4(x_{pa} - x_{c_k a})}{(r^*/2)^2} \exp\left(\frac{-2\sum_{r=1}^{m}(x_{pr} - x_{c_k r})^2}{(r^*/2)^2}\right), \quad (4)$$

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

$$\frac{\partial D_i}{\partial x_{pt}} = \begin{pmatrix} \frac{-2(x_{pa}-x_{ja})}{(r^*/2)^2} \exp\left(\frac{-\sum_{r=1}^{m}(x_{ir}-x_{jr})^2}{(r^*/2)^2}\right) + \\ \frac{2(x_{pa}-x_{c_k a})}{(r^*/2)^2} \exp\left(\frac{-\sum_{r=1}^{m}(x_{pr}-x_{c_k r})^2}{(r^*/2)^2}\right) \\ \exp\left(\frac{-\sum_{r=1}^{m}(x_{c_k r}-x_{jr})^2}{(r^*/2)^2}\right) \end{pmatrix}, \quad (5)$$

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

$$\frac{\partial D_i}{\partial x_{pa}} = \begin{pmatrix} \frac{2(x_{ia}-x_{pa})}{(r^*/2)^2} \exp\left(\frac{-\sum_{r=1}^{m}(x_{ir}-x_{pr})^2}{(r^*/2)^2}\right) - \\ \frac{2(x_{c_k a}-x_{pa})}{(r^*/2)^2} \exp\left(\frac{-\sum_{r=1}^{m}(x_{c_k r}-x_{pr})^2}{(r^*/2)^2}\right) \\ \exp\left(\frac{-\sum_{r=1}^{m}(x_{ir}-x_{c_k r})^2}{(r^*/2)^2}\right) \end{pmatrix}, \quad (6)$$

(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

$$I(a)_k = \sum_{i=1}^{n} I(i,a)_k. \quad (7)$$

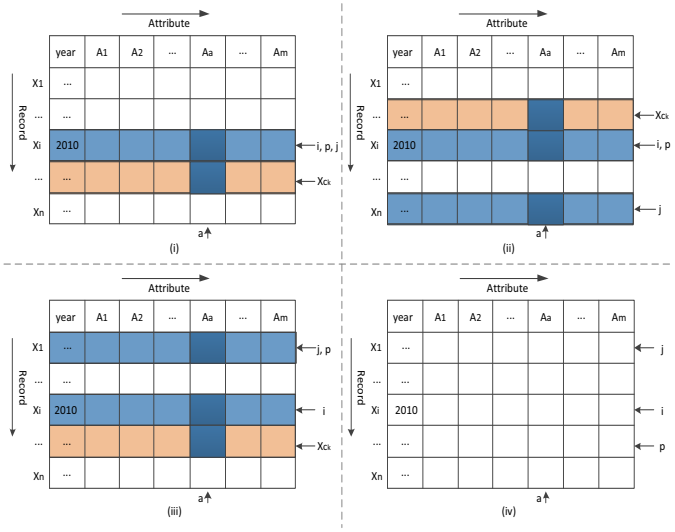

Fig. 3. Records selection for the contribution calculation of attribute a to record i. (i) Records x_i and x_{ck} are selected. (ii) Records x_i, x_j and x_{ck} are selected. (ii) Records x_i, x_p and x_{ck} are selected. (iv) No record is selected.

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

Algorithm 1 : Important attribute selection

Input: Data matrix $X \in \Re^{n\times m}$, and parameter ε , σ.
Output: Important attributes and cluster centers for X.

1: Initialize the neighborhood radius $r^* = \sqrt{\sum_{j=1}^{n}\sum_{i=1}^{n}||x_i - x_j||^2 \Big/ n(n-1)}$, 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** $D_{c_k}/D_{c_1} > \varepsilon$ **do**
7: **for** each point $x_i \in X$, $i = 1...n-k$, 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. (3);
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 to select the k-th representative record is obtained.
16: **end for**
17: **end while**
18: The attributes with $I = \sum_{j=2}^{k} I(a)_j \Big/ (k-1) > \sigma$ are selected; // I is the importance of the a-th attribute for clustering.

$O(mn^2/q) + M(q)$, 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 processes of the important attribute selection algorithm. More importantly, by partitioning the matrix G and the initial dataset X appropriately, we can achieve 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 2 : Parallelization of distance matrix calculation

Input: Two copies of dataset $X \in \Re^{n\times m}$, named as Y, Z.
Output: Distance Matrix $G \in \Re^{n\times n}$.

1: The diagonal elements of G are set to be '0'; $//G$ is symmetric, and its upper triangular matrix is calculated.
2: Function **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 **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: $dist_a = (Y[i,a] - Z[j,a])^2$;
13: **end for**
14: Sum the $dist_a$ for all 'a'. We can get the distance G_{ij} between records i and j;
15: Function **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

$$attr\_sum(a,b) = ||a+b||^2 = \sum_{i=1}^{n}(a_i+b_i)^2. \quad (8)$$

Definition 2 (Attribute Product) The attribute product between them is defined as

$$attr\_pro(a,b) = ||ab||^2 = \sum_{i=1}^{n}(a_i b_i)^2. \quad (9)$$

Definition 3 (Attribute Coordination) Given the attribute summation and attribute product, the attribute coordination between a and b is define as

$$attr\_coo(a,b) = \frac{attr\_sum(a,b)}{attr\_pro(a,b)} = \frac{||a+b||^2}{||ab||^2}. \quad (10)$$

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).

$$D_i = \sum_{j=1}^{m} \exp\left[\frac{-||A_i + A_j||^2}{||A_i A_j||^2 (0.5R^*)^2}\right] \quad (11)$$

$$D_i = D_i - D_{c_k} \exp\left[\frac{-||A_i + A_{c_k}||^2}{||A_i A_{c_k}||^2 (0.5R^*)^2}\right] \quad (12)$$

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 ε.
Output: Representative attributes.

1: Set the records of attributes as sample data, and initialize the neighborhood radius $R^* = \sqrt{\sum_{j=1}^{m}\sum_{i=1}^{m}(||A_i+A_j||^2/||A_iA_j||^2)\Big/m(m-1)}$;
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** $D_{c_k}/D_{c_1} > \varepsilon$ **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 matrixes A and B have n records and m attributes, $A\Delta B \in \Re^{m\times m}$ can be defined as an attribute coordination matrix

$$(A\Delta B)_{ij} = \frac{\sum_{r=1}^{n}(a_{ri}+b_{rj})^2}{\sum_{r=1}^{n}(a_{ri}b_{rj})^2} = \frac{(a_{1i}+b_{1j})^2+(a_{2i}+b_{2j})^2+...+(a_{ni}+b_{nj})^2}{(a_{1i}b_{1j})^2+(a_{2i}b_{2j})^2+...+(a_{in}b_{nj})^2}. \quad (13)$$

$(A\Delta B)_{ij}$, 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.

Algorithm 4 : Parallelization of coordination matrix calculation

Input: A, B are two copies of $AS = X^{\mathrm{T}} \in \Re^{m \times n}$.
Output: Attribute Coordination Matrix $A\Delta B \in \Re^{m \times m}$.

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 $< (i,j), A\Delta B_{ij} >$, 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 coordinations between different attributes. After that, the coordination items are matched and sorted to calculate the corresponding attribute coordinations (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.

5 Economic Analysis of a Real-Word Case Study

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 relationships 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 abnormity, 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

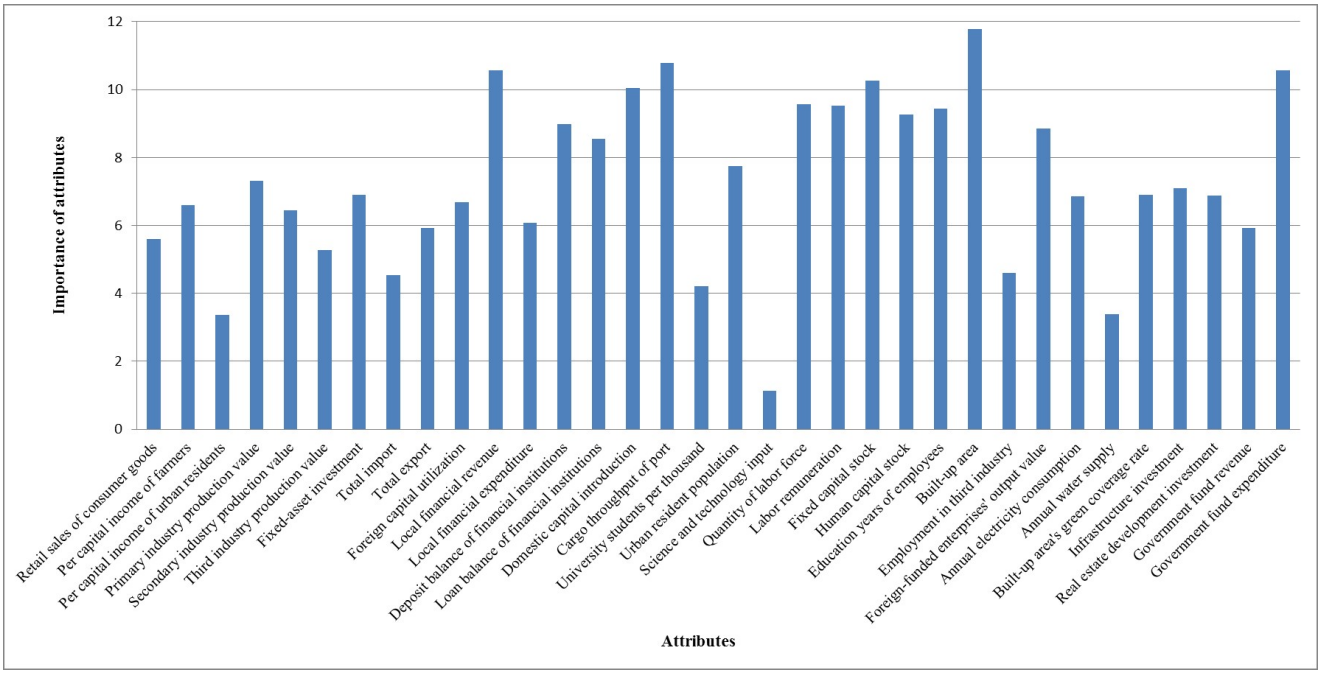

Fig. 4. The importance of attributes for economic development in Dalian, China.

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.

Important attributes	Representative attributes	Research perspectives
Per capital income of farmers	Per capital income of farmers	Analysis of economic growth factors
Labor remuneration		
Urban resident population		
Quantity of labor force	Quantity of labor force	
Human capital stock	Human capital stock	
Education years of employees		
Fixed-asset investment	Fixed-asset investment	
Foreign capital utilization		
Domestic capital introduction		
Deposit balance of financial institutions		
Loan balance of financial institutions		
Real estate development investment		
Infrastructure investment		
Fixed capital stock	Fixed capital stock	
Primary industry production value		Analysis of the industrial structure
Secondary industry production value	Secondary industry production value	
Local financial revenue	Local financial revenue	Institutional functions for economic growth
Local financial expenditure	Local financial expenditure	
Government fund expenditure		
Cargo throughput of port	Cargo throughput of port	Environment externality of economic growth
Output value of foreign-funded enterprise		
Annual electricity consumption		
Built-up area	Built-up area	
Built-up area's green coverage rate		

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 Externality".

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

$$\begin{array}{l} GDP = Labor\ Remuneration + Capital\ Gain + \\ Collaborative\ Interests + Others \end{array}, \quad (14)$$

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:

$$Y = aL^{\alpha}H^{\beta}S^{\gamma}D^{\delta} + bK + cSD/K + u, \quad (15)$$

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:

$$\begin{array}{l} y = \frac{bK - cSD/K}{Y}k + \frac{cSD/K + a\delta L^{\alpha}H^{\beta}S^{\gamma}D^{\delta-1}}{Y}d \\ + \frac{cSD + a\gamma L^{\alpha}H^{\beta}S^{\gamma-1}D^{\delta}}{Y}s + \frac{\alpha\beta L^{\alpha}H^{\beta-1}S^{\gamma}D^{\delta}}{Y}h \\ + \frac{a\alpha L^{\alpha-1}H^{\beta}S^{\gamma}D^{\delta}}{Y}l + i + e \end{array}, \quad (16)$$

in which y, k, d, s, h, and 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).

$$\eta_K = \frac{bK - cSD/K}{Y} \times \frac{k}{y}, \quad (17)$$

$$\eta_D = \frac{cSD/K + a\delta L^{\alpha}H^{\beta}S^{\gamma}D^{\delta-1}}{Y} \times \frac{d}{y}, \quad (18)$$

$$\eta_S = \frac{cSD + a\gamma L^{\alpha}H^{\beta}S^{\gamma-1}D^{\delta}}{Y} \times \frac{s}{y}, \quad (19)$$

$$\eta_H = \frac{\alpha\beta L^{\alpha}H^{\beta-1}S^{\gamma}D^{\delta}}{Y} \times \frac{h}{y}, \quad (20)$$

$$\eta_L = \frac{a\alpha L^{\alpha-1}H^{\beta}S^{\gamma}D^{\delta}}{Y} \times \frac{l}{y}, \quad (21)$$

$$\eta_I = {}^{i}\!/_{y}, \quad (22)$$

$$\eta_E = {}^{e}\!/_{y}. \quad (23)$$

Herein, η_K, η_D, η_S, η_H, η_L, η_I and η_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.

TABLE 2
Analysis of driving factors of economic growth in Dalian-the contribution rates of them to economic growth (1997-2013, %).

Factors	1997-2008	2009-2013
Fixed capital stock	12	23
Fixed-asset investment	41	37
Innovation input	32	33
Human capital	10	18
Quantity of labor force	5	7
Institutional innovation	5	0
Environment externality	-5	-17

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.

$$\begin{array}{l} Y = 0.247(HL)^{0.509}(SD/L)^{0.282} + 0.178K + \\ 45.5SDH/K^2 + 1.08L \end{array}. \quad (24)$$

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.

$$H = 13.64U + 0.21, \tag{25}$$

$$L = 0.874UP, \tag{26}$$

$$D = 0.86UV - 380.5, \tag{27}$$

$$S = 0.222HLU + 0.012D - 22.8. \tag{28}$$

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 3
The marginal incomes (Billion Yuan) of 1 percent urbanization rate improvement on factors "Labor force", "Human capital", "Fixed-asset investment" and "Innovation input".

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

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).

$$\begin{aligned} Y = {} & 0.247(13.64U + 0.21)^{0.509}(0.874UP)^{0.227} \\ & (0.86UV - 380.5)^{0.282}(2.647U^3P + 0.01UV + \\ & 0.041U^2P - 27.366)^{0.282} + 0.178(0.9K_{t-1} + \\ & 0.86UV - 380.5) + 45.5(2.647U^3P + 0.01UV + \\ & 0.041U^2P - 27.366)(0.86UV - 380.5)(13.64U + \\ & 0.21)/(0.9K_{t-1} + 0.86UV - 380.5)^2 + 0.944\text{UP} \end{aligned} \tag{29}$$

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

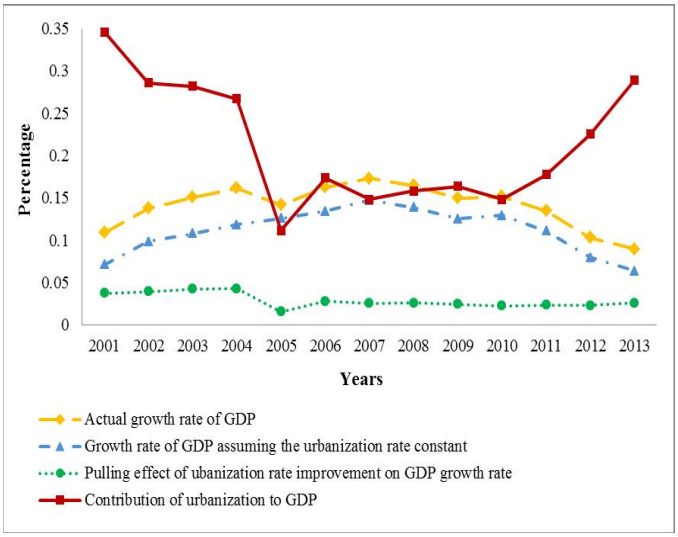

Fig. 5. The results of correlation analysis between urbanization and economic development in Dalian.

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.

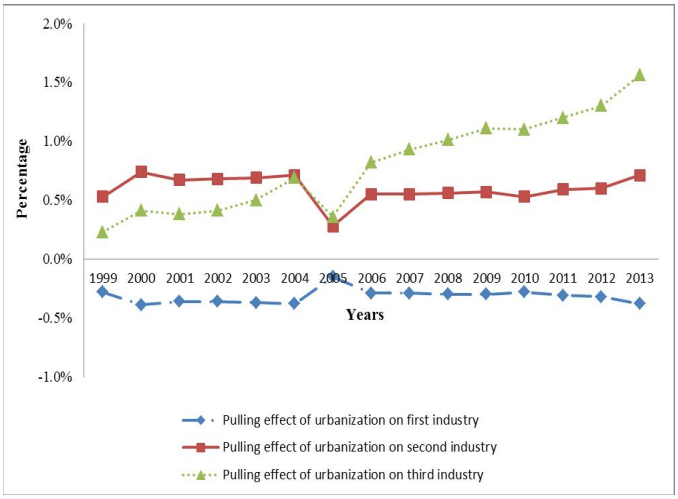

Fig. 6. The pulling effects of urbanization on the first, second and third industry in Dalian.

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 yet 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 accords 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

[1] A. Sheth, "Transforming Big Data into Smart Data: Deriving Value via Harnessing Volume, Variety, and Velocity Using Semantic Techniques and Technologies," in *Proc. 30th IEEE Int. Conf. on Data Engineering*, 2014, pp.2.

[2] World Economic Forum, "Big Data, Big ImpactNew Possibilities for International Development," *http : //www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_*2012.*pdf*, 2012.

[3] "Big Data across the Federal Government," *http : //www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_*1.*pdf*, 2014.

[4] H. Giersch, "Urban Agglomeration and Economic Growth," *Springer Science & Business Media*, 2012.

[5] R. B. Ekelund Jr and R. F. Hbert, "A History of Economic Theory and Method," *Waveland Press*, 2013.

[6] B. Liddle, "The Energy, Economic Growth, Urbanization Nexus across Development: Evidence from Heterogeneous Panel Estimates Robust to Cross-sectional Dependence," *The Energy Journal*, vol.34, no.2, pp.223-244, 2013.

[7] S. Ghosh and K. Kanjilal, "Long-term Equilibrium Relationship between Urbanization, Energy Consumption and Economic Activity: Empirical Evidence from India," *Energy*, vol.66, no.3, pp.24-331, 2014.

[8] S. H. Law and N. Singh, "Does Too Much Finance Harm Economic Growth?," *Journal of Banking & Finance*, vol.41, no.4, pp.36-44, 2014.

[9] D. Baglan and E. Yoldas, "Non-linearity in the Inflation-growth Relationship in Developing Economies: Evidence from a Semiparametric Panel Model," *Economics Letters*, vol.125, no.1, pp.93-96, 2014.

[10] Q. Ashraf and O. Galor, "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *The American Economic Review*, vol.103, no.1, pp.1-46, 2013.

[11] V. Boln-Canedo, N. Snchez-Marono and A. Alonso-Betanzos, "A Review of Feature Selection Methods on Synthetic Data," *Knowledge and Information Systems*, vol.34, no.3, pp.483-519, 2013.

[12] S. Alelyani, J. Tang and H. Liu, "Feature Selection for Clustering: A Review," *Data Clustering: Algorithms and Applications*, vol.29, 2013.

[13] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," *The University of Waikato*, 1999.

[14] M. Dash and H. Liu, "Consistency-based Search in Feature Selection," *Artificial Intelligence*, vol.151, no.1, pp.155-176, 2003.

[15] M. A. Hall and L. A. Smith, "Practical Feature Subset Selection for Machine Learning," in *Proc. 21st Australian Computer Science Conf.*, 1998, pp.181-191.

[16] L. Beretta and A. Santaniello, "Implementing ReliefF Filters to Extract Meaningful Features from Genetic Lifetime Datasets," *Journal of Biomedical Informatics*, vol.44, no.2, pp.361-369, 2011.

[17] Q. Gu, Z. Li and J. Han, "Generalized Fisher Score for Feature Selection," *arXiv preprint arXiv*:1202.3725, 2012.

[18] H. Peng, F. Long and C. Ding, "Feature Selection based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, no.8, pp.1226-1238, 2005.

[19] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," *Morgan Kaufmann*, 2005.

[20] J. G. Dy and C. E. Brodley, "Feature Subset Selection and Order Identification for Unsupervised learning," in *Proc. International Conference on Machine Learning*, 2000, pp.247-254.

[21] P. S. Bradley and O. L. Mangasarian, "Feature Selection via Concave Minimization and Support Vector machines," in *Proc. International Conference on Machine Learning*, 1988, pp.82-90.

[22] A. Rakotomamonjy, "Variable Selection Using SVM based Criteria," *The Journal of Machine Learning Research*, vol.3, no.3, pp.1357-1370, 2003.

[23] M. Mejła-Lavalle, E. Sucar and G. Arroyo, "Feature Selection with a Perceptron Neural Net," in *Proc. of the International Workshop on Feature Selection for Data Mining*, 2006, pp.131-135.

[24] G. C. Cawley, N. L. C. Talbot and M. Girolami, "Sparse Multinomial Logistic Regression via Bayesian L1 Regularisation," in *Proc. Advances in Neural Information Processing Systems*, 2007, pp.209-216.

[25] T. Prasartvit, A. Banharnsakun and B. Kaewkamnerdpong, "Reducing Bioinformatics Data Dimension with ABC-kNN," *Neurocomputing*, vol.116, no.9, pp.367-381, 2013.

[26] P. Ghamisi, M. S. Couceiro and J. A. Benediktsson, "A Novel Feature Selection Approach Based on FODPSO and SVM," *IEEE Transactions on Geoscience and Remote Sensing*, vol.53, no.5, pp.2935-2947, 2015.

[27] H. Uguz, "A Two-stage Feature Selection Method for Text Categorization by Using Information Gain, Principal Component Analysis and Genetic Algorithm," *Knowledge-Based Systems*, vol.24, no.7, pp.1024-1032, 2011.

[28] M. Jamjoom, and K. E. Hindi, "Partial Instance Reduction for Noise Elimination," *Pattern Recognition Letters*, vol.74, no.4, pp.30-37, 2016.

[29] B. Xue, M. Zhang and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-objective Approach," *IEEE Transactions on Cybernetics*, vol.43, no.6, pp.1656-1671, 2013

[30] J. Liang, F. Wang, C. Dang and Y. Qian, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique," *IEEE Transactions on Knowledge and Data Engineering*, vol.26, no.2, pp.294-308, 2014.

[31] Q. Song, J. Ni and G. Wang, "A Fast Clustering-based Feature Subset Selection Algorithm for High-dimensional Data," *IEEE Transactions on Knowledge and Data Engineering*, vol.25, no.1, pp.1-14, 2013.

[32] L. Zhao, Z. Chen, Z. Yang, and Y. Hu , "A Hybrid Method for Incomplete Data Imputation," in *Proc. 17th IEEE International Conference on High Performance Computing and Communications*, 2015, pp.1725-1730.
[33] Y. Yang, Z. Ma, A. G. Hauptmann and N. Sebe, "Feature Selection for Multimedia Analysis by Sharing Information Among Multiple Tasks," *IEEE Transactions on Multimedia*, vol.15, no.3, pp.661-669, 2013.
[34] G. Casalino, N. Del Buono and C. Mencar, "Subtractive Clustering for Seeding Non-negative Matrix Factorizations," *Information Sciences*, vol.257, no.2, pp.369-387, 2014.
[35] J. Liu and Z. Jiang, "Innovation-driven and Investment-supportive Strategies for China's Economic Transformation based on Collaborative Theory," *Science of Science and Management of S.&T.*, vol.36, no.2, pp.25-33, 2015.
[36] E. P. Xing, Q. Ho, W. Dai, J. K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar and Y. Yu, "Petuum: A New Platform for Distributed Machine Learning on Big Data," *IEEE Transactions on Big Data*,vol.1, no.2, pp.49-67, 2015.
[37] X. Hu, L. Tang and H. Liu, "Embracing Information Explosion without Choking: Clustering and Labeling in Microblogging," *IEEE Transactions on Big Data*, vol.1, no.1, pp.35-46, 2015.

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.