Recognition and Prediction of Chinese Energy Efficiency Influence Factors Based on Data Mining Algorithm

Fansheng Meng[1]; Bin Li[1]; Jiao Liu[2]; Zenglei Yue [3]; Donghui Yang[4]; Zhi Liu [5]; Jie Wan [2,*]

1 *School of Economics and Management,* Harbin Engineering University, Harbin, Heilongjiang, China
2 *School of Energy Science and Engineering*, Harbin Institute of Technology, Harbin, Heilongjiang, China
3 *Heilongjiang Science and Technology Information Research Institute*, Harbin, Heilongjiang, China
4 *School of Economics and Management*, Southeast University, Nanjing, Jiangsu, China
5 Power Horizon Information Technology Co., Ltd, Nanjing, Jiangsu, China

***Abstract* — Rise of green industrial revolution' objected to low-carbon and energy conservation has made it an important research direction to measure energy efficiency and its influence factors. In this paper, character identification method has been proposed to determine influence factors of energy efficiency, and energy efficiency of 24 provinces is analyzed and evaluated by the method of data mining algorithm. In the research, some common classification algorithms are utilized to build three classification models with collected data, with the accuracy of over 90%. Then energy efficiency of other six provinces are predicted with this model. Furthermore, provinces with high and low energy efficiency are distinguished by cluster algorithm, and the trend of a fall in energy efficiency of the whole country is discovered. In the end, on the basis of above analysis, a strategy is put forward to improve Chinese energy efficiency.**

Keywords - energy efficiency; data mining; feature selection; classification; cluster

I. INTRODUCTION

As the condition of energy and environment is being more and more severe, worldwide attention has been paid to the field of energy efficiency. Many scholars have studied the improvement of energy utilization efficiency and its energy-saving potential from different aspects, and have drawn the conclusion that to improve energy utilization efficiency is an important method to reduce energy consumption, mitigate supply and demand contradiction as well as keep economy growth sustained and steady[1-3]. Some policies in China also state that it is essential to accelerate the progress in converting the way economy develops, enforcing energy resource saving and ecological environment protection, and enhancing sustainable development ability[4]. It cannot be denied that China is still under the condition of high resource and energy consumption, low energy utilization and severe pollution[5]. Energy utilization efficiency is still low among the world, which has become challenge to sustainable development[6]. To release the mentioned energy contradiction, it is of emergency to save energy and increase energy utilization efficiency. To conclude, it is necessary to find out the key influence factors of energy efficiency, and the degree of influences should be quantitatively analyzed.

Energy utilization efficiency is mostly calculated by the method of DEA. Hu and Wang raised the concept of total factor energy efficiency with EDA[7]. Afterwards, researches were following this thought, for example, make GDP the output index, and make energy consumption, capital, labor, energy consumption per unit of GDP the input index[8,9]. Some scholars also studied the effects of the industrial structure, technological progress, openness and other factors on energy efficiency, on the basis of estimating the total factor energy efficiency[10]. Due to the complexity of Chinese regions and imbalance of spatial development, many scholars use energy panel data of different provinces to analyze energy efficiency of different regions or provinces. Some effective calculation method and evaluation methods were given out in this way[11-13]. However, as different scholars would use different metrics to calculate the energy efficiency, it is difficult to figure out the actual factors that affect energy efficiency. In this paper, data mining methods are used to take existing literature factors comprehensively into account, and a gather of energy efficiency feature is formed. According to the energy efficiency published in former papers, an experiential learning sample is formed, and key factors affecting energy efficiency are identified with the method of feature selection. So influence degrees of various factors can be quantitatively measured. Then an energy efficiency classification model for prediction is formed by classification algorithm. In this way, it is unnecessary to calculate energy efficiency value. Instead, a scientific and targeted method, data mining, can be applied to solve the issue of energy efficiency evaluation.

The key innovation point of this paper is taking charge of data mining method. 8 indicators of 24 provinces from 2005 to 2013 is treated as feature space, and the feature contribution rate is evaluated by feature selection method, which helps identifying key factors that affect energy efficiency. The paper uses classification algorithm to acquire energy efficiency categories models for each province in China, and predicts energy efficiency categories of the other four provinces as well as those in the future. At the same time, energy efficiency cluster results were obtained by clustering algorithm. The results avoids influence of existing results, and distributes trends for each province, which would be reference for energy efficiency development of each province.

II. DATA COLLECTION AND FEATURE SELECTION

Energy efficiency can be defined as measurement of the largest extent that practical output capacity can make up under the current fixed energy input, or can be defined as the smallest input under fixed output condition. It is a dimensionless value between 0 and 1, and is an important part of a regional economic efficiency. Due to the level of energy efficiency is influenced by many factors, the measurement of energy efficiency should consider multiple indicators. On this basis, key factors are identified, and prediction for future energy efficiency of the regions can be obtained,According to China Statistical Yearbook, 'China Energy Statistical Yearbook' and the regional statistical yearbook, this paper collects panel data from 24 provinces and autonomous regions (excluding Tibet, Hong Kong, Macao and Taiwan, Jilin, Heilongjiang, Guizhou, Yunnan Gansu, Qinghai) from 2005 to 2013. The identification of energy efficiency factors considers selection of existing researches, meaning input elements generally consider energy consumption, labor, energy, industrial investment and capital investment; output elements include GDP, environmental pollution. etc[14]. Since the study only needs to classify the levels of energy efficiency, there is no need to calculate the value of the specific energy efficiency. Considering the availability of data, this paper selects the primary energy production, the total energy consumption, energy consumption elasticity coefficient, energy and industrial investment capital stock as an indicator of inputs, the provinces GDP, unit GDP energy consumption and sulfur dioxide emission coefficient as the output indicators. Therefore, the selected feature space includes primary energy production (F1), the total energy consumption (F2), energy consumption elasticity coefficient (F3), GDP (F4), the energy industry investment (F5), unit production capacity consumption (F6), capital stock (F7), and sulfur dioxide emission factors (F8). Among them, the primary energy production (F1) is qualified products that primary energy enterprises (companies) produce from nature existing energy during the reporting period, such as crude oil, coal, gas and hydropower electricity. Total energy consumption (F2) refers to the real energy consumption that companies consume during statistics reporting period, the value is calculated by taking a predetermined sum and converting to the measurement unit. Energy consumption elasticity coefficient (F3) is an indicator of the relationship between the proportion of gross domestic product growth rate of energy consumption and economic growth rate. GDP (F4), is market price of all final goods and services of all resident units of a country (within national borders) in a given period. GDP is a key indicator of national accounting, as well as an important indicator of the overall economic situation of a country or region. Energy Industry Investment (F5) is the total funds invested in the energy industry. Unit production capacity consumption (F6) refers to energy that per unit of gross domestic product consumes in a country or region during a period of time. Capital stock (F7) is all existing capital resources of an enterprise. It is sum of all types of resources that put into the enterprises. It exists in the form of assets and is also known as the stock of assets. According to its status in the production process, it can be divided into two categories: stock of assets that involved in the reproduction and the stock of idle assets, including idle plant, machinery and equipment. Sulfur dioxide emission factor (F8) is the volume of sulfur dioxide emissions from energy combustion or usage (per unit).

III. ENERGY EFFICIENCY OF DATA MINING ALGORITHM

This paper use the classification algorithm, the collected 216 energy efficiency influencing factors of instance data for analysis. Due to the range between the eight characteristics are different. For data preprocessing, the characteristic value standardization that correct for feature selection. On this basis, the comprehensive documented different size of the provincial energy efficiency calculation, Category annotations to the data collection, the class label for learning to get training set classification algorithm. By comparing the advantages and disadvantages of different classification algorithm, this article can use the classification model, so that they can use in the prediction. In order to compare the energy efficiency analysis under the category labels. This paper adopts two kinds of clustering algorithms, the collected data classified with the year, then we can observe the provincial energy efficiency category distribution and change trend.

A. Data Standardization

The most typical method of data standardization is the standardization of 0 and 1(normalized).Through the linear transformation of the original data, the results into interval [0, 1]. When data are positive value, we can use another kind of simplified form to express:

$$x^{*}=\frac{x_i}{\sum_{i=1}^{n}x_i},i=1,2,\cdots,n \tag{1}$$

Considering the use of the characteristics of the data set values are positive, so use simplified conversion function to normalize every components. Eigenvalues of the transformed distribution in [0, 1], eliminates the feature space caused by different domain feature value selection error.

B. Feature Selection

Feature selection is done by searching for a data set of all possible feature set, according to certain rules to select a set of effective features to reduce the dimension of feature space. At the same time, removing the feature space of some redundant information to avoid the impact of these information for classification prediction, then to improve the prediction accuracy and computational efficiency of classification algorithm. Information gain is the most common method of feature selection, the concept is closely related to the information entropy[15,16,17]. According to the research, scholars in the use of information gain to extract features, generally gain value threshold is set to 0.0025[18]. Using data mining software WEKA calculate the characteristics of information gain value, and select the six characteristics of the information gain value is greater than 0.0025 to sort it, eigenvalue reserved to two decimal places. The results are shown in table I:

TABLE I. DIFFERENT CHARACTERISTICS ON THE CLASSIFICATION OF INFORMATION GAIN

sort	Information gain value	feature of the no	feature of the name
1	0.68	F6	energy consumption per unit GDP
2	0.25	F8	sulfur dioxide emission coefficient
3	0.18	F7	capital stock
4	0.15	F4	GDP
5	0.14	F1	primary energy production
6	0.08	F3	energy consumption elasticity coefficient

Can be seen from table I, the six characteristic F6、F8、F7、F4、F1、F3 have strong correlation with the category attribute, as the key influence factors of energy efficiency.F6(energy consumption per unit GDP) the biggest impact, the second is F8(sulfur dioxide emission coefficient),F7(capital stock)、F4(GDP)、F1(primary energy production)、F3（energy consumption elasticity coefficient）These four characteristics have the similar influence degree of the energy efficiency. And the data set F5(the energy industry investment)、F2（the total energy consumption）the two features are filtered. Therefore, we can think the two features are almost no effect on energy efficiency.

C. Classification Analysis

The classification of the two types of problems is the most basic problem, it is more simple and easy handling than other classification problems. In this paper, the analysis of energy efficiency also can be summed up in two types of problems, the example of the data set is divided into two categories, high energy efficiency and low energy efficiency. So here the classification number is set to 2, the column tag of 0 and 1, 0, for high energy efficiency, 1 represents the low energy efficiency.

According to the results of feature selection in section 3.2, F6, F8, F7, F4, F1, F3 is the key influence factors of energy efficiency, but F5, F2 these two factors almost had no effect on energy efficiency. So remove the data set F5 and F2 these two attributes in classification, to eliminate the attribute values on the result of classification. There are a lot of the classification of the data mining method, The decision tree algorithm is the most widely used[19]. The traditional decision tree algorithm is simple and practical, Classification algorithm based on rule learning and learning strategy based on element learning strategies has the advantages of easy to optimize, high efficiency and high accuracy[20,21]. Due to the three kinds of classification methods in many areas are verified its good classification effect, therefore, in the study tries to use the above three methods to classify data sets.

1) Decision Tree

The decision tree, also known as decision tree, it based on the instance of inductive learning algorithm, then from a set without order and rules of tuple, reasoned decision tree representation of classification rules[22,23]. C4.5 algorithm is widely used in the decision tree algorithm[24], it is j. Ross Quinlan put forward on ID3 algorithm. It is based on information gain rate methods to select testing attributes, information gain rate is equal to the information gain ratio of segmentation information[22,25]. C4.5 is implemented in the WEKA J48 decision tree, in this paper, using J48 classifier for energy efficiency factors affecting data set for training, get the decision tree as shown in figure 1.

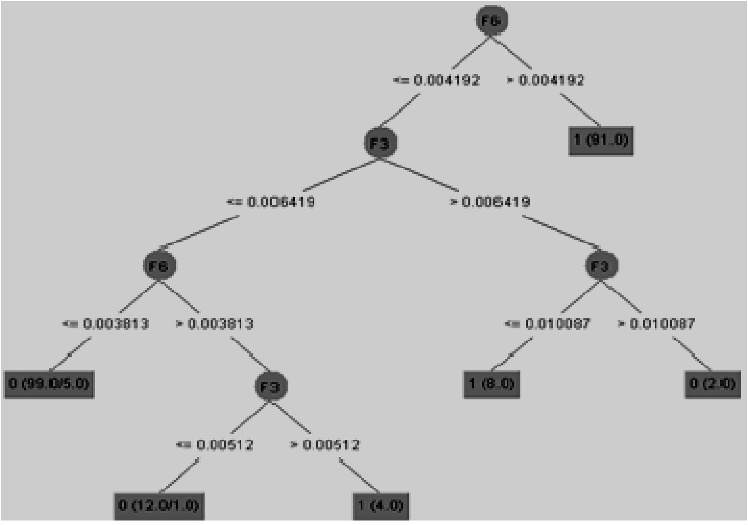

Fig.1 Energy Efficiency Of Classification Decision Tree Branches

2) Based On The Rules of Classification

Based on the rules of classification is to use a set of if…then rules to classify. Algorithm thought: Start with the training set to generate rules set, each rule is to use conjunction conditions, with "⇒ " separate rules before and after: As a rule ri: condition j⇒ yj. If the rule of r and record before x attributes matching, says r overlay x. When r cover a given record, said r is triggered. Set up rules set, then classify. Classification of each stay record and each of a set of rules, if a rule is triggered, the record is to be classified. This article uses JRip classifier established rules, through the RIPPER algorithm to implement. The sorting scheme based on class, belong to the same class rules in rule sets along with them, and then these rules according to which they belong to the class information together[26]. Between the rules of the same kind of relative order is not important, because they belong to the same class. RIPPER algorithm to extract rules directly from the data, when extracting rules, class of all training records is seen as a positive example, other types of training records is seen as a counterexample. The rules of the growth strategy from general to specific, to select the best conjunction item added to the front of the rules of the evaluation standard is FOIL information gain, until the rules cover counterexample, stop adding conjunction. While pruning begins finally add in conjunction, given regulation ABCD⇒ y. Check whether D should be removed first, and then is the CD, BCD, etc.

3) Element Learning

Element learning is conducted based on the results of learning to learn or more and get the final result[27]. While constructing a high accuracy of classification model is difficult, but it is easy to create a discriminant accuracy slightly better than chance would model (weak classifier). Boosting is aimed at the weak classifier learning in advance, it is gradually promoted to "strong classifier" a machine learning method, which can improve effect of classifier. In this paper, Adboost algorithm used by Freud and Schapire improve and widely used in practice[28]. The basic idea: Based on existing sample data sets to build a foundation of "weak classifier", repeated calls the "weak classifier", In each round of wrongful convictions sample gives greater weight, make it more focus on those difficult given samples. After several rounds of circulation, finally adopts the method of weighting each round of "weak classifier" synthetic "strong classifier", so as to get a high precision of classification discriminant model.

4) Classification Results Analysis

In order to guarantee the accuracy of the generated classification model, the model is selected and evaluated by ten-fold cross validation. The data set is divided into 10 subsets, the cross validation is repeated 10 times, and a subset is selected as the test set while the others selected as training set, and the average cross recognition rate of 10 times is the final result.

Precision, recall and F-measure are the common indicators to measure the classification results, using these three indicators to evaluate the performance of the three classifiers used in the experiment. In the calculation of precision and recall, the four indicators used in ROC curve analysis: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). These calculation formulas are as follows:

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

Among them, F-measure is the harmonic mean of precision and recall rate, can be used as a key index to measure the performance of the classifier, and the formula is:

$$F - measure = \frac{2\,precision \times recall}{precision + recall} \quad (4)$$

Table IV is the classification result of the energy efficiency influence factor data set which using J48, JRip, LogitBoost classifiers.

TABLE IV. CLASSIFICATION RESULTS USING THREE CLASSIFIERS

Classifiers	TP Rate	FP Rate	Precision	Recall	F-Measure
J48	0.921	0.078	0.922	0.921	0.921
JRip	0.917	0.083	0.918	0.917	0.917
LogitBoost	0.940	0.059	0.943	0.940	0.940

It can be found from table IV that LogitBoost is better than J48 and JRip classifier. It means that LogitBoost is the best classifier based on the key influence factors on energy efficiency according to the preamble. The F-measure values of the three classifiers are all higher than 0.9. The classification results obtained by the three classification methods are good. The three classification models can be used to predict the energy efficiency of the data in other provinces and cities, which are not included in the data set.

5) Prediction Based Classification Model

Through the analysis of the foregoing, the accuracy of the three classification models is more than 90%, and the good classification results can be achieved. So the three classification models, JRip, LogitBoost and J48 can be used to forecast.

Collects Jilin, Heilongjiang, Guizhou, Yunnan, Gansu and Qinghai province's energy efficiency of the impact factor data in 2013,, feature space is consisted of six factors, a primary energy production, energy consumption elasticity coefficient, GDP, unit GDP energy consumption, capital stock and sulfur dioxide emissions coefficient. After standard the feature values, the above three kinds of classification models are used to predict the results. The results are as follows:

TABLE V. FORECAST RESULTS USED THREE CLASSIFIERS RESPECTIVELY.

Provinces	Classifiers	Predicted	Error prediction	Average probability
Jilin	J48	0	0.949	0.914
	JRip	0	0.922	
	LogitBoost	0	0.870	
Heilongjiang	J48	0	0.949	0.858
	JRip	0	0.922	
	LogitBoost	0	0.704	
Guizhou	J48	1	1	0.993
	JRip	1	0.990	
	LogitBoost	1	0.990	
Yunnan	J48	0	0.949	0.870
	JRip	0	0.922	
	LogitBoost	0	0.739	
Gansu	J48	0	0.917	0.796
	JRip	0	0.923	
	LogitBoost	0	0.547	
Qinghai	J48	1	1	0.997
	JRip	1	0.990	
	LogitBoost	1	0.999	

From the table, it can be seen that the results are consistent with the results obtained by using three different classifiers. Jilin, Heilongjiang, Yunnan and Gansu have been predicted to be 0 kinds of high energy efficiency, while the forecast results of Qinghai and Guizhou are 1 kinds, which are low energy efficiency. The prediction results of the three classification models are consistent with the prediction results of the energy efficiency of each province, but there is a different degree of confidence. For example, using the J48 classifier to predict energy efficiency in Jilin in 2013, the results show that 94.9% to be the 0 categories, while using JRip and LogitBoost classifier to predict the possibility of high energy efficiency in Jilin in 2013 was 92.2% and 87%. Three kinds of classification models were chosen to reflect the average forecast accuracy, so as to choose the forecast results. Here, only the average forecast accuracy of the results more than 85% is considered. Table V shows that the average forecast accuracy of Jilin, Heilongjiang, Guizhou, Yunnan and Qinghai are more than 85%, therefore, the forecast results are adopted, Jilin, Heilongjiang and Yunnan are high energy efficiency provinces in 2013, Guizhou and Qinghai are low energy efficiency provinces. The average forecast accuracy of Gansu is only 79.6%, which is less reliable than the first five provinces, so the forecast result can't be adopted.

D. Cluster Analysis

According to the idea in this work, clustering is made based on the classification algorithm. In WEKA, the classification algorithms are the Simple K-means and EM, respectively. The K-means algorithm, a frequently applied algorithm in cluster analysis, receives the input data k and classifies the n data objects into k clusters, which makes the objects exhibit high similarity in the same cluster but low similarity among different clusters. The EM algorithm, on the other hand, is frequently used in the cluster analysis of machine learning and computer vision. In order to improve the accuracy, the above mentioned algorithms are wrapped by the Meta Density Based Clusterer in WEKA. Mata Density Based Clusterer, fitting a discrete distribution or a symmetrical normal distribution for every cluster, realizes the gradual clustering from the global to the local, thus is powerful in the local search and converges fast. A global population without sub-population is initialized first, and iterative search is conducted in the global population. During this process, the individuals are clustered, and a sub-population is formed when the number of the individuals reached a specified minimum value. Then iterative search is conducted in each sub-population and the individuals are re-clustered, which enhances the ability of the algorithm to jump out of the local optimum.

1) K-means Cluster

Simple K-means, viz. the k means cluster algorithm, first assigns the number of the clusters k and randomly chooses k samples as the center of the initial clusters. Then the distances between each sample and its corresponding cluster center are calculated and the samples are classified. After the classification of all the samples, the centers of clusters are re-calculated. The above process is repeated until the centers of the clusters do not change anymore, the k clusters obtained is the final cluster results [29]. The results based on K-means cluster are listed by year in Table II below:

TABLE II. THE K-MEANS CLUSTER RESULTS OF THE ENERGY EFFICIENCY OF EACH PROVINCE IN CHINA

Year	**Cluster0**	**Cluster1**
2005	Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Henan, Hubei, Hunan, Guangdong, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Shandong
2006	Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Zhejiang, Anhui, Fujian, Jiangxi, Henan, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Jiangsu, Shandong, Guangdong
2007	Beijing, Tianjin, Shanxi, Inner Mongolia, Liaoning, Shanghai, Zhejiang, Anhui, Fujian, Jiangxi, Henan, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Hebei, Jiangsu, Shandong, Guangdong
2008	Beijing, Tianjin, Liaoning, Shanghai, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Hebei, Inner Mongolia, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Guangdong
2009	Beijing, Tianjin, Shanxi, Liaoning, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Hebei, Inner Mongolia, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Guangdong
2010	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi, Hainan, Chongqing, Shaanxi, Ningxia, Xinjiang	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Guangdong, Sichuan
2011	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Hunan, Guangxi, Hainan, Chongqing, Shaanxi, Ningxia, Xinjiang	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan,

		Hubei, Guangdong, Sichuan
2012	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Guangxi, Hainan, Chongqing, Ningxia, Xinjiang	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Hubei, Hunan, Guangdong, Sichuan, Shaanxi
2013	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Guangxi, Hainan, Chongqing, Ningxia	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Hubei, Hunan, Guangdong, Sichuan, Shaanxi, Xinjiang

As illustrated from Table II, the 216 samples are classified into 2 categories by the K-means algorithm. Cluster0 has 140 samples and accounts for 65% of the total, while cluster1 has 76 samples and account for 35%.Comparing the data among different years shows that the energy consumption per GDP (F6) in cluster0 is smaller than that in cluster1, which means lower energy consumption per GDP. And the energy consumption elastic coefficient (F3) in cluster0 is smaller than that in cluster1. Samples in cluster0 have lower energy consumption growth rate and sulfur dioxide emission coefficient than those in cluster1 with the same national economic growth rate, which means that samples in cluster0 have smaller sulfur dioxide emission per energy burning, thus have a lower pollution factor on the environment. Thus, samples in cluster0 are classified as high energy efficient ones while those in cluster1 are classified as low ones.as shown in Table II, the number of high energy-efficiency provinces decreases gradually from 2005 to 2013. Many high energy-efficiency provinces turned into low energy-efficiency ones. For instance, Liaoning was high energy-efficiency province till 2009, but became low energy-efficiency ones from 2010. Similar provinces include Hebei, Shanxi, Inner Mongolia, Shanghai, Zhejiang, etc. Those keeping high energy-efficiency include Beijing, Fujian, Jiangxi and Guangxi, etc., whereas Shandong, Jiangsu and Guangdong were in low energy-efficiency state for a long time.

2) EM cluster

The EM is an algorithm which searches the maximum likelihood estimation or maximum posterior estimation probabilistic model [30]. In EM algorithm, the likelihood probabilities of the parameters and training samples increase gradually through the improvement of the parameters of the model, which end at a maximum. The EM algorithm can be treated as a successive approaching algorithm, in which the parameters of the model are not known. The parameters can be randomly chosen or initialized roughly beforehand. The most probable state of these parameters is determined and the probability of each training sample is calculated. Then the parameters are corrected by the samples at current state and re-estimated, and the models are re-established with the new parameters. In this way, the algorithm ends at a certain convergence criterion after iteration, which ensures the parameters of the model approach the real ones. The cluster results by the EM algorithm are obtained as below:

TABLE III.THE EM CLUSTER RESULTS OF THE ENERGY EFFICIENCY OF EACH PROVINCE IN CHINA

Year	Cluster0	Cluster1
2005	Beijing, Tianjin, Hebei, Inner Mongolia, Liaoning, Shanghai, Zhejiang, Anhui, Fujian, Jiangxi, Henan, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Shanxi, Jiangsu, Shandong, Guangdong
2006	Beijing, Tianjin, Liaoning, Shanghai, Zhejiang, Anhui, Fujian, Jiangxi, Henan, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Hebei, Shanxi, Inner Mongolia, Jiangsu, Shandong, Guangdong
2007	Beijing, Tianjin, Liaoning, Shanghai, Zhejiang, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Shaanxi, Ningxia, Xinjiang	Hebei, Shanxi, Inner Mongolia, Jiangsu, Shandong, Henan, Guangdong
2008	Beijing, Tianjin, Liaoning, Shanghai, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi, Hainan, Chongqing, Sichuan, Ningxia, Xinjiang	Hebei, Shanxi, Inner Mongolia, Jiangsu, Zhejiang, Shandong, Henan, Guangdong, Shaanxi
2009	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi, Hainan, Chongqing, Ningxia	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Guangdong, Sichuan, Shaanxi, Xinjiang
2010	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Hubei, Hunan, Guangxi,	Hebei, Shanxi, Inner Mongolia, Liaoning,

	Hainan, Chongqing, Ningxia	Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Guangdong, Sichuan, Shaanxi, Xinjiang
2011	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Guangxi, Hainan, Chongqing, Ningxia	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Hubei, Hunan, Guangdong, Sichuan, Shaanxi, Xinjiang
2012	Beijing, Tianjin, Anhui, Fujian, Jiangxi, Guangxi, Hainan, Chongqing, Ningxia	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Henan, Hubei, Hunan, Guangdong, Sichuan, Shaanxi, Xinjiang
2013	Beijing, Tianjin, Anhui, Jiangxi, Guangxi, Hainan, Chongqing, Ningxia	Hebei, Shanxi, Inner Mongolia, Liaoning, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong, Henan, Hubei, Hunan, Guangdong, Sichuan, Shaanxi, Xinjiang

The EM cluster classified the 216 samples into 2 categories, where cluster0 has 118 samples and accounts for a percent of 55%, whereas cluster1 has 98 samples and accounts for 45%.

By comparing the clustered data of the same year, the energy consumption per GDP (F6) and the sulfur dioxide emission coefficient of the samples in cluster0 are generally lower than those of the samples in cluster1, which means high efficient energy utilization with economic growth, paralleled by low pollution level of the environment. Thus, samples in cluster0 are classified as high energy efficient ones while those in cluster1 are classified as low ones.

As illustrated by Table III, results of EM cluster algorithm resemble those of the K-means. The number of provinces with low energy efficiency increases with time, and most provinces shifted from high efficient ones to low efficient ones, like Liaoning, Shanghai, Zhejiang, Hubei, Hunan, Sichuan and Shanxi etc. Those keeping high energy-efficiency include Beijing, Fujian, Hainan, Jiangxi etc., whereas Shanxi, Shandong, and Guangdong were in low energy-efficiency state for a long time.

3) Analysis Of The Cluster Results

Based on the results of the two algorithms and the realistic situation of each province, the conclusions can be drawn:

(1) In the years from 2005 to 2013, Beijing has a high energy efficiency, and Tianjin, Hainan, Fujian, Jiangxi were also in a high energy efficiency state; Yet, Hebei, Henan, Liaoning, Shanxi and Inner Mongolia are in low energy efficiency state for a long time. The underlying reason for the horizontal differences can be attributed to the differences in the economic structure. Those regions pillared by technology-intensive industry have higher energy efficiency, whereas those pillared by traditional manufacturing and processing industry have lower energy efficiency.

(2)By analyzing the category changes in each region for different years, it can be shown that the energy efficiency of China in recent years and the number of high energy-efficiency provinces are decreasing these years. Although the energy consumption per GDP is decreasing as demonstrated by the national data, the energy consumption elastic coefficient is in a fluctuation state, the cost of pollution control and energy loss are growing year by year. The underlines reasons is that the energy structure of China with coal as main energy is unreasonable for a long term, and that the economic growth relies on the consumption of the resources rather than on the improvement of technology and management innovation. The above reasons are the obstacles for economic sustainable development and improvement of the energy efficiency. Only by the optimization of the energy structure, the transformation of the economic growth mode and keeping high economic growth with low energy consumption elastic coefficient can the energy efficiency be greatly improved.

IV. CONCLUSION

In this paper, data mining algorithm is used to analyze and evaluate energy efficiency of each province in China. 6 key factors that affect the level of energy efficiency is determined by feature selection. And models with higher classification accuracy is established by three categories of classifiers: J48, JRip and LogitBoost. Then the three categories of classifiers are used to predict energy efficiency level of six provinces including Jilin and Heilongjiang in 2013. Next, in order to eliminate the influence of labels and human experiences, this paper distinguishes high efficiency and low efficiency provinces with K-means and EM clustering method. With the results of clustering. Energy efficiency variation of the whole country is summarized. The following conclusions can be reached:

(1) With feature selection methods, it is available to

find out the determinants for China's energy efficiency. In this paper, a variety of multiple factors mentioned in papers are collected, and six determinants are selected from eight by information gain method. From the analysis, information gain is not only able to identify the factors that affect energy efficiency, but also able to give the level of impact and their importance.

(2) By establishing classification model for each province in China, and evaluating the results of three classification algorithms, we can figure out that LogitBoost based on multiple classifiers fusion can further strengthen the weak classifiers, the energy efficiency classification results is better than decision tree and rule-based classification method. However, in actual forecasting conditions, the method based on decision tree is better than others related to the forecasting of the six provinces.

(3) Through two kinds of clustering methods, it can be noticed that there is a horizontal differentiation in all regions of China. Also, energy efficiency in the same area has longitudinal changes with time. The unbalanced development in different areas of energy efficiency is mainly because of differences in economic structure and energy technologies between regions. While the level of energy efficiency in the same province changes with time is mainly because of the further consumption of traditional energy sources, as well as the adjustment of economic structure and development of energy technologies.

(4) By analyzing energy efficiency category changes of different provinces with clustering algorithm, it can be noticed that in recent years, development of overall China's energy efficiency is on a downward trend. Since 2005, China's high energy efficiency provinces is gradually decreasing. The basic reason is China has long been in an irrational energy structure, coal is the main energy source, and economic development mainly depends on the consumption of resources, rather than relying on technological progress or management innovation.

To sum up, the develop direction for China's energy efficiency is focusing on the key factors influencing energy efficiency, and optimizing the energy structure and transforming the economic growth mode in a more scientific and targeted way. Creation of technical inventions (especially energy technology field) should be encouraged and supported, to promote technological innovation in all aspects of energy usage, in order to achieve the goal of maintaining high economic growth with smaller energy consumption elasticity. In addition, on the basis of comprehensive considering energy supply and demand situation and energy utilization technology. It is also necessary to optimize energy consumption structure considering both traditional energy and new energy. On one hand, we need to constantly strengthen the new energy development to increase the proportion that renewable energy makes up in energy consumption. On the other hand, we should be committed to the development and deployment of clean energy technologies, and improve efficiency of coal, oil, natural gas and other traditional energy sources.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.71501040 and 2014 Nanjing “321 Introduction Plan” Candidate Project.

REFERENCES

[1] Rosen, M.A., “Assessing global resource utilization efficiency in the industrial sector. Science of the Total Environment”, 461: p. 804-807. 2013.

[2] Lee, Y.C. and A.Y. Zomaya,“ Energy efficient utilization of resources in cloud computing systems”. Journal of Supercomputing,. 60(2): p. 268-280. 2012

[3] Subirats, J. and J. Guitart, “Assessing and forecasting energy efficiency on Cloud computing platforms. ”Future Generation Computer Systems-the International Journal of Grid Computing and Escience,. **45**: p. 70-94. 2015

[4] Guan, W. and S. Xu, “Study on spatial pattern and spatial effect of energy eco-efficiency in China.” Acta Geographica Sinica,. **70**(6): p. 980-992. 2015

[5] Shen, N., J. Zhou, and W. Zou, “Energy Efficiency Measures and Convergence in China, Taking into Account the Effects of Environmental and Random Factors. ”Polish Journal of Environmental Studies, **24**(1): p. 257-267. 2015.

[6] Jiankun, H.E. and Z. Xiliang, “Analysis Declining Tendency in China′s Energy Consumption Intensity during the 11^th Five - Year - Plan Period. ”China Soft Science, (4): p. 33-38. 2006

[7] Hu, J.-L. and S.-C. Wang, “Total-factor energy efficiency of regions in China.” Energy Policy, **34**(17): p. 3206-3217. 2006

[8] Wang, Q., et al., “Energy efficiency and production technology heterogeneity in China: A meta-frontier DEA approach.” Economic Modelling,. 35: p. 283-289. 2013

[9] Yang, M., F. Yang, and X. Chen, “On Influencing Factors Affecting China's Energy Efficiency: An Empirical Test Based on the VEC Model.” Resources Science,. **33**(1): p. 163-168. 2011

[10] Shen, Y., L. Lei, and X. Zhang, “Evaluation of energy transfer and utilization efficiency of azo dye removal by different pulsed electrical discharge modes.” Chinese Science Bulletin, 53(12): p. 1824-1834. 2008.

[11] Liu, L., “How large is Chinas Regional Disparity of Energy Efficiency In Industrial Sector?-From the Perspective of Substitution Effect Analysis”. Mathematics in Practice and Theory, 42(12): p. 48-54. 2012

[12] Zou, G., et al., “Measurement and evaluation of Chinese regional energy efficiency based on provincial panel data.” Mathematical and Computer Modelling, 58(5-6): pp. 1000-1009. 2013.

[13] Yang, L. and K.-L. Wang, “Regional differences of environmental efficiency of China's energy utilization and environmental regulation cost based on provincial panel data and DEA method.”

Mathematical and Computer Modelling. 58(5-6): pp. 1074-1083. , 2013

[14] Li, S., “Features of Energy Consumption and Policies for Improving Energy Efficiency in Chinese Industrialization. ”China Soft Science, pp. 23-35. 2010

[15] Lee, C.K. and G.G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization." Information Processing & Management,. 42(1): pp. 155-165. 2006

[16] Shang, C., et al., "Feature selection via maximizing global information gain for text classification." Knowledge-Based Systems. 54: pp. 298-309 , 2013

[17] Daley, D.J. and D. Vere-Jones, "Scoring probability forecasts for point processes: The entropy score and information gain". Journal of Applied Probability, vol. 41 A, pp. 297-312. 2004.

[18] Yang, D.-H. and G. Yu, "A method of feature selection and sentiment similarity for Chinese micro-blogs." Journal of Information Science, vol. 39, No. 4. pp. 429-441. 2013

[19] Dietterich, T.G., "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization". Machine Learning, vol. 40. No. 2, pp. 139-157. 2000

[20] Chen, G., et al., "A new approach to classification based on association rule mining." Decision Support Systems,. vol. 42. No. 2 ,pp. 674-689. 2006

[21] Kim, B., H. Park, and Y. Baek, "Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning." Computers & Education, vol. 52. No. 4, pp. 800-810. 2009.

[22] Quinlan, J.R., "Decision trees and decision-making. Systems, Man and Cybernetics," IEEE Transactions on, vol. 20. No. 2 pp. 339-346. 1990.

[23] Vens, C., et al., "Decision trees for hierarchical multi-label classification". Machine Learning, vol. 73. No. 2 pp. 185-214. 2008.

[24] Polat, K. and S. Guenes, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems." Expert Systems with Applications, vol. 36 No.(2) pp. 1587-1592. 2009.

[25] Quinlan, J.R., "Induction of decision trees. Machine learning", vol. 1 No. 1 pp. 81-106. 1986.

[26] Hühn, J. and E. Hüllermeier, "FURIA: an algorithm for unordered fuzzy rule induction." Data Mining and Knowledge Discovery, vol. 19 No. 3 pp. 293-319. 2009.

[27] Prodromidis, A., P. Chan, and S. Stolfo, "Meta-learning in distributed data mining systems: Issues and approaches."Advances in distributed and parallel knowledge discovery, vol. 3 pp. 81-114. 2000.

[28] Freund, Y. and R.E. Schapire. "A desicion-theoretic generalization of on-line learning and an application to boosting. in Computational learning theory." Springer. 1995.

[29] Kanungo, T., et al., "An efficient k-means clustering algorithm: Analysis and implementation. Ieee Transactions on Pattern Analysis and Machine Intelligence," vol. 24 No. 7 pp. 881-892. 2002.

[30] Do, C.B. and S. Batzoglou, "What is the expectation maximization algorithm?" Nature biotechnology, vol. 26 No. 8: p. 897-900. 2008.