Research on the Construction and Evaluation of Green Industry System Based on Data Mining

Zhu Long*
Fuyang Normal University, Fuyang, Anhui, China

*Corresponding author: ahfywln@126.com

Abstract. By establishing a regional green innovation system, realizing the integrated development of industries across administrative divisions, and developing emerging industries, continuation industries, and resource reuse industries are effective ways to realize the transformation of economic development mode. In the process of shifting from the industrial economy of the industrial society to the industrial economy of the information society, green industry informatization is an inevitable requirement for realizing industrial restructuring and industrial upgrading, improving the core competitiveness of green enterprises, and a necessary condition for realizing regional national economic informatization. Data mining technology can discover and extract hidden, novel, meaningful and understandable information and patterns from a large amount of data, and realize the transformation from simple data to information to knowledge. According to the general requirements of data mining technology and the characteristics of the green industry field, this research constructed a system framework for green industry data mining, discussed key issues such as data preparation and data selection in data mining of green industry. Meanwhile, using data mining software SAS and Malmquist index as implementation tools, the basic theories and methods of green industry data mining were applied to the innovation evaluation of my country's green industry. The results showed that data mining technology had its unique advantages in dealing with the massive data of the green industry. It can not only quickly realize the diversified statistical analysis of the massive data in the green industry, but also display it to users with rich graphics, which is conducive to the interpretation and understanding of the analysis results.

Keywords: Data mining, Green industry system, System construction and evaluation

1. Introduction
The green industry is an emerging industrial form that has developed relative to the characteristics of traditional industries in history, and its core is guided by sustainable development and green concepts. That is to say, the "green" concept runs through the main links of the industrial production process, and various industries characterized by the production of environmentally friendly products [1]. The development of green industry is mainly to solve outstanding problems of environmental pollution, implement end-point pollution control, and continuously meet the market's increasing demand for green products and food. Therefore, For new enterprises, it is necessary to adopt a radical strategy to eliminate...
enhancing regional innovation capabilities, focusing on green and sustainable development, and making full use of the favorable opportunities for industrial restructuring are important ways to realize asset reorganization and industrial upgrading of green industries [2]. Real-time data is spreading across many areas of the green industry. If we can make good use of these data, we can better promote the development of society. Therefore, it is necessary to actively explore the potential value of massive data and provide more reliable decisions and suggestions for the development of green industries. Making full use of modern information and computer technology and striving to tap the basic information of each class of the green industry is the key to building a green industry system and one of the ways to realize the intelligentization of the green industry [3]. At present, there are many difficulties in the development of our country's green industry, and the emergence of the data mining network system brings hope. Its construction and development provide feasible solutions to solve these problems.

So, this research applied data mining technology to the field of green industry based on a full understanding of the basic theories of data mining. Starting from the characteristics and analysis needs of green industry data, the construction of sustainable development of green industry was taken as an example to illustrate the method of establishing a green industry data warehouse; Based on the general process of data mining, a framework for data mining in the green industry was constructed; Through case analysis, an evaluation was made on the constructed green industrial system. The improvement of green sustainable innovation capabilities based on data mining is the key to accelerating the transformation of economic development, which is of great significance to the sustainable development of green industries.

2. Application of data mining technology in the construction of green industry system

In actual use, the green industry has the characteristics of sharing difficulties and diverse data, which brings various challenges to the use of green industry data. Data mining technology can not only integrate different forms of data sources, but also has its own unique advantages for processing large and complex data. At the same time, data mining is also based on data, without clear assumptions, mining information and discovering knowledge from a large amount of data, providing new ideas for solving complex problems [4]. Based on the general process of data mining, this part analyzes and constructs the system framework of green industry data mining based on the characteristics of green industry data and application requirements, and discusses key issues in data mining for green industry.

2.1 The process of data mining in the construction of the green industry system

Many data mining system developers have proposed some "process reference models" that apply data mining techniques. 5A emphasizes the functions and capabilities of tools that support the data mining process; SEMMA emphasizes the application method combined with data mining system; CRISP-DM emphasizes the methods and steps of implementing data mining projects from a methodological perspective, and is independent of each specific data mining algorithm and data mining system [5]. Combining the related concepts of green industry, 5A and CRISP-DM describe the data mining process from the perspective of support function and methodology respectively. Therefore, this study uses it as a model to study the process of data mining in the construction of a green industrial system, as shown in Figure 1.

Figure 1. The process of data mining in the construction of green industry system
It can be seen from Figure 1 that the process of data mining in the green industry to discover knowledge is not completed at one time, but a cyclic, repeated non-linear process, and each step may be repeated. In the repeated process, we are constantly approaching the nature of the green industry, and constantly optimizing the solutions to problems.

2.2 Analysis of the characteristics of the green industry system
Green industry has four main components: green industry, green products, green consumption and green services. Green industry is the product of the progress of the times, so the development of green industry must make full use of modern science and technology, focusing on promoting the harmonious development of man and nature, and the coordination of population, resources and environment. Only by analyzing the characteristics of the green industry system can we better provide technical support for the construction of the green industry system based on data mining technology. The characteristics analysis of the green industry system is shown in Figure 2 [6].

![Figure 2. Analysis of the characteristics of the green industry system](image)

2.3 Data mining structure of green industry system
The determination of the green industry data mining system structure is the basis for the success of the green industry data mining project. According to the characteristics of green industry data and the requirements of data mining technology, the overall framework for constructing green industry data mining is shown in Figure 3. The framework consists of four parts: data layer, organization layer, mining layer and decision-making layer.

![Figure 3. The overall framework of the green industry based on data mining technology](image)

(1) Data layer
The data layer includes data sets for data analysis and mining of green industries, including both historical data and real-time data statistical yearbook, China Science and Technology Statistics Database, China Energy Statistics Database, etc. It also includes a collection of information oriented to...
different topics established by means of purification, synthesis, classification, and identification based on the data warehouse. Green industry data warehouse can integrate and aggregate heterogeneous data sources, providing a unified and complete data foundation for data mining.

(2) Organizational level

These data cubes cover various topics of concern for decision-making, such as green industry energy consumption, green industry production technology, and green industry environmental protection. In multi-dimensional data cubes oriented to different topics, not only should there be enough macro-grained data, but also local fine-grained data.

(3) Excavation layer

The mining layer is the core of the whole system, which is mainly divided into two parts: data mining and online analysis. Online analysis generally faces middle-level applications for data pre-processing and advanced application query, while data mining faces high-level applications for knowledge discovery.

(4) Decision-making layer

The decision-making layer is a user interface layer for decision-makers. It combines the new knowledge acquired by the mining layer with the accumulated relevant knowledge in the knowledge base to provide decision-makers in various forms, interprets and expresses the mining knowledge, and stores the new results in the knowledge base.

From the overall structure of the entire green industry data mining system, the above four layers are responsible for the tasks of different stages of green industry data mining. That is, from data pre-processing, data mining to knowledge expression, a complete system is formed.

2.4 Data mining tools for green industry

According to the application scope of data mining in the green industry, this research chooses a horizontal data mining system. Typical horizontal data mining tools include IBM Intelligent Miner, SPSS Clementine, SAS Enterprise Miner, etc. For SAS has great advantages in data statistical analysis and data mining, this research uses SAS (Statistics Analysis System) data mining software.

3. Evaluation of green industry system based on data mining

One of the most important characteristics of the green industry is technological innovation. Technological innovation is a kind of technological innovation that meets the needs of sustainable development. Its purpose is not only to improve the environmental benefits of the natural environment, but also to obtain the economic benefits of potential profits, as well as the social benefits of improving the quality of human healthy life. Among them, environmental benefits are the foundation, economic benefits are the means, and social benefits are the ultimate goal. Without economic benefits, there will be no social benefits, and the realization of environmental benefits loses material means. It can be seen that environmental benefits and economic benefits are two important aspects to measure the innovation performance of green industries. Therefore, this research evaluates the green innovation capability of the constructed green industry system based on the aforementioned data mining technology. This article studies green innovation capabilities. In terms of output, environmental pollutant emissions must be considered, and industrial wastewater and solid waste emissions are selected to investigate the greenness of the innovation process. In this paper, energy consumption is included in the input and industrial pollution emissions are included in the output. In the data analysis, considering that the industrial pollution emissions are undesired output, the reciprocal is taken to solve this problem.

3.1 Data selection and mining

Affected by the availability of data, the research period of this article is 2014-2019. The research object is the green innovation capacity of 30 provinces in China (some data in Tibet are not complete, so the analysis is not included). The calculation method adopts the widely used method in recent years. Based on the productivity index. All the data analyzed in this paper comes from the "China Statistical Yearbook", "China Science and Technology Statistical Yearbook", "China Energy Statistical Yearbook"
and the National Research Network database. Among them, the industrial "three wastes", energy consumption per unit of GDP, and regional GDP data come from the "China Statistical Yearbook", the remaining variable data are from the "China Science and Technology Statistical Yearbook".

3.2 Data extraction, transformation and loading

The main task of data extraction, conversion and loading is to inspect and sort the data after extracting data from the data source, and reorganize and process the data according to the design requirements of the data warehouse, and finally load it into the data warehouse. Therefore, under the premise that data is available, and guided by the principles of science, rationality and multi-objectiveness, this article starts from the three aspects of innovation input, innovation output, and innovation environment, and constructs an index that includes 9 indicators. Green industry evaluation index system, see Table 1.

**Table 1. The evaluation index system of green industry innovation ability in this study**

| First level indicator | Secondary indicators | Level three indicators |
|-----------------------|----------------------|-----------------------|
| Innovation ability of green industry | Innovation investment | Expenses for the introduction of foreign technology |
| | | Expenses for purchasing domestic technology pointed out |
| | | Technical transformation expenditure |
| Innovation ability of green industry | Innovation output | Output value of new products |
| | | Total discharge of industrial wastewater |
| | | Number of professional authorizations |
| Innovation environment | Practitioners of research and development institutions |
| | Internal expenditures of research and development institutions |
| | Number of research and development institutions |

Data conversion needs to organize the data in related data sources into complete data files according to the design requirements of the data warehouse, unify the measurement units of similar data, and check the consistency of the data. This research uses the DTS tool in SQL Server2018 to load the data in the data file into the data warehouse for green industry innovation evaluation. After the data is loaded into the data warehouse, a SQL query needs to be used to verify the referential integrity between the latitude and fact tables to ensure that all records are related to appropriate records in other tables. In this study, based on the green industry productivity and the idea of non-parametric methods in data mining, the Malmquist index is combined with the SAS theory in data mining, and the total factor productivity (TEP) is defined using Malmquist input and output distance functions. That is, the frequently used Malmquist index [7]. Define the Malmquist productivity index from period \(t\) to period \(t+1\) under the technical conditions of period \(t+1\):

\[
M(a_{t+1}, b_{t+1}; a_t, b_t) = \frac{Q_{t+1}(a_{t+1}, b_{t+1})}{Q_t(a_t, b_t)} \times \frac{Q_t(a_t, b_t)}{Q_{t+1}(a_{t+1}, b_{t+1})} \times \sqrt{\frac{Q_t(a_t, b_t)}{Q_{t+1}(a_{t+1}, b_{t+1})}}
\]

(1)

If the index is greater than 1, it indicates that the total factor productivity from period \(t\) to period \(t+1\) is increasing, and vice versa. In equation 2, it is recorded as

\[
TEF = \frac{Q_{t+1}(a_{t+1}, b_{t+1})}{Q_t(a_t, b_t)} \times \frac{Q_t(a_t, b_t)}{Q_{t+1}(a_{t+1}, b_{t+1})} \times \sqrt{\frac{Q_t(a_t, b_t)}{Q_{t+1}(a_{t+1}, b_{t+1})}}
\]

(2)

Among them, TEF stands for technological efficiency change, and TECG stands for technological progress. Technical efficiency change PE is further decomposed into pure technical efficiency change and scale efficiency change QE, namely TEF=PE*QE, where
Based on the above process, the Malmquist index is finally decomposed into:

\[ M(TEP) = TEF \times TECG = PE \times QE \times TECG \]  

(4)

The Malmquist index thus obtained has a good form. Based on the actual needs of the research problem, this article decomposes the Malmquist Productivity Change Index into the Technical Efficiency Change Index and the Technical Progress Change Index, and defines TEF as the efficiency improvement of green innovation, and TECG as the technological progress of green innovation. TEP is the Malmquist Productivity Change Index, defined as green innovation capability.

3.3 Data mining and evaluation result analysis

Since this part evaluates the innovation capability of the green industry, it must take into account the input level and output level of innovation and the environmental factors affecting innovation. This indicator is calculated by using energy consumption per unit of GDP and the proportion of new product output in GDP. In the innovation output, in addition to the output value of new products and the number of patent authorizations, the industrial "three wastes" emissions of industrial waste water, waste gas and solid waste are also considered. Since they are undesired outputs, take the reciprocal to solve the problem. Using the input-output panel data selected in this paper, and with the aid of data mining SAS software, the green innovation capability index of each province in China is calculated. In order to observe the technical efficiency of green innovation, the technological progress of green innovation, and the trend of green innovation capability and their interrelationships, the three are drawn in Figure 4.

It can be seen from Figure 4 that from 2014 to 2019, the overall level of innovation in China's green industry has performed well. This is because technological progress has promoted the improvement of green innovation capabilities, that is, the frontier of production has risen significantly. Although the efficiency of green innovation has improved, its strength is still relatively weak. From the perspective of efficiency decomposition, although the scale efficiency has increased, the pure technical efficiency has decreased, which has caused the efficiency improvement of China's green innovation capabilities to be weak. It can be seen that China's innovation capability is the improvement of the production frontier brought about by the introduction, digestion, absorption, and imitation of advanced technologies from developed countries. It has not done enough to improve its own organization and management and optimize the industrial structure. This is also an important growth point for the next round of green innovation.

Figure 4. Trends of China's green industry innovation capabilities from 2014 to 2019

4. Conclusion

The current development of green industries pays more and more attention to the application of information technology. Data mining technology will be used as a favorable tool to concentrate and purify a large amount of green industry data and to scientifically process and analyze the data through
statistical analysis and knowledge discovery. Provide multi-level and multi-functional information services for green industry management decision-making. Therefore, based on a full understanding of the basic theories of data mining, this research applied data mining technology to the field of green industry. Starting from the characteristics and analysis needs of green industry data, the construction of sustainable development of green industry was taken as an example to illustrate the method of establishing a green industry data warehouse; Based on the general process of data mining, a framework for data mining in the green industry was constructed; Through case analysis, the innovation of the constructed green industrial system was evaluated. The research results showed that data mining technology had its unique advantages in processing massive data. It could not only quickly realize diversified statistical analysis of massive data, but also displayed it to users in rich graphics, which was conducive to the interpretation and understanding of analysis results. This is a good simulation evaluation model, which can be applied to evaluate the innovative energy of green industries and provide a reliable basis for the rational development and utilization of green industries.

Acknowledgements
This research was financially supported by Fuyang Humanities and Social Sciences Key Projects: “A study on the development of strategic emerging industries in Fuyang” (No.FYSK2019ZD09), Anhui Philosophy and Social Science Planning Project: “Research on the Construction of strategic emerging industry Cluster Base in Anhui Province under the background of Internet” (No.AHSKQ2016D31) and Anhui Education Department Humanities and Social Sciences Key Research Project: “A study on the cooperative mechanism of social organizations undertaking the transfer of government functions under the modern state governance system” (NO.SK2018A0280).

References
[1] Kim Sangyong, and Shim Jae Heon. “Combining Case-Based Reasoning with Genetic Algorithm Optimization for Preliminary Cost Estimation in Construction Industry.” Canadian Journal of Civil Engineering, vol. 41, no. 1, pp: 65–73, 2014.
[2] Yang, Dongmin, et al. “Research on Green Development of Shaanxi Industry Under The Background of Big Data.” Proceedings of the 2018 2nd International Conference on Big Data and Internet of Things, vol. 3, no. 5, pp: 13–17, 2018.
[3] Runkler, Thomas A., et al. “Fuzzy Clustering for Data Compression, Modeling, and Optimization in Recovered Paper Industry.” European Journal of Control, vol. 7, no. 1, pp: 67–74, 2001.
[4] Ghazal, Ahmad, et al. “BigBench: Towards an Industry Standard Benchmark for Big Data Analytics.” Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp: 1197–1208, 2013.
[5] Castillo, Enrique del, and Marco S. Reis. “Bayesian Predictive Optimization of Multiple and Profile Response Systems in the Process Industry: A Review and Extensions.” Chemometrics and Intelligent Laboratory Systems, vol. 1, no. 2, pp: 104121, 2020.
[6] Sharma, S. P., and Yashi Vishwakarma. “Availability Optimization of Refining System of Sugar Industry by Markov Process and Genetic Algorithm.” 2014 International Conference on Reliability Optimization and Information Technology (ICROIT), pp: 29–33, 2014.
[7] Kauskale, L., et al. “Environmental and Energy Aspects of Construction Industry and Green Buildings.” Latvian Journal of Physics and Technical Sciences, vol. 54, no. 2, pp: 24–33, 2017.