Extended Belief Rule-based Model for Environmental Investment Prediction with Indicator Ensemble Selection

Fei-Fei Ye [a, d], Suhui Wang [c], Peter Nicholl [d], Long-Hao Yang [a, d], Ying-Ming Wang [a, b, *]

[a]Decision Sciences Institute, Fuzhou University, Fuzhou, 350116, PR China

[b]Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou University, Fuzhou, 350116, PR China

[c]School of Business Administration, Zhejiang University of Finance & Economics, Hangzhou, 310018, China

[d]School of Computing, Ulster University, Northern Ireland, UK

*Corresponding author. Email: msymwang@hotmail.com

Abstract: Environmental investment prediction is an effective solution to reduce the wasteful investments of environmental management. Since environmental management involves diverse environmental indicators, investment prediction modeling usually causes the curse of dimensionality and uses irrelevant indicators. A common solution to solve these problems is the use of indicator selection methods to select representative indicators. However, different indicator selection methods have their relative strengths and weaknesses, resulting in different selected indicators and information loss of real representative indicators. Hence, in the present work, a new environmental investment prediction model is proposed on the basis of extended belief rule-based (EBRB) model along with the indicator ensemble selection (IES) and is called IES-EBRB model, The EBRB model is a white-box designed decision-making model and has the specialty on using prior knowledge to enhance data analytics for autonomous decision making; and the IES is an extension of ensemble learning to cooperatively integrate different kinds of indicator selection methods for selecting representative indicators. In a case study, the real world environment data from 2005 to 2018 of 31 provinces in China are applied to verify the effectiveness and accuracy of the IES-EBRB model. Results show that the IES-EBRB model not only can obtain desired environmental investments, but also produces satisfactory accuracy compared to some existing investment prediction models.

Keywords: Extended belief rule-based model; indicator ensemble selection; environmental investment prediction; white-box design; knowledge enhanced data analytics

1. Introduction

Due to increasing pollution emission and ecological damage, government and society have recently paid greater attention to environmental protection [13]. The resulting investments have grown annually to ensure that urgent pollution control and ecological remediation have targeted effect. The greatest impact typically depends on the ability of environment managers to formulate effective investment schemes for achievable environmental managements [6][26]. For the sake of facilitating implementation of effective investment schemes, many environmental investment prediction models have been proposed based on environmental indicators and experts' knowledge, while the challenge is still how to build reliable prediction models based on various environmental indicators.

The use of representative environmental indicators to monitor management effectiveness is often identified as one of the successful factors for effectively predicting environmental investments, as they can indicate improvement opportunities on accurate investment planning. In previous studies, many methods have been used for environmental indicator selection, *e.g.*, principal component analysis (PCA)-based method [25] and correlation-based feature selection (CFS)-based method [16]. However, they mainly focused on a single indicator selection method, which has its own advantages as well as weaknesses on indicator selection for different kinds of environmental investment problems. This results in the undesired outcome that irrelevant indicators are selected for environmental investment prediction modeling. Making effective indicator selection is the challenge that must be considered to propose a new environmental investment prediction model.

Based on the selected environmental indicators, a certain decision-making methodology and historical environmental data could be included to construct an investment prediction model, *e.g.*, time series forecasting-based [6][12] and input-output relationship-based models [16][25]. It is worth noting that the specialty of investment prediction models would inherit from the selected decision-making methodology, so the use of a specific decision-making methodology must take into consideration human involvement for environmental managements, which indicates that environment managers can embed experts' knowledge into the modeling of environmental investment prediction and also know exactly well how the model predict outcomes for investments. As a result, environment managers can have greater confidence to formulate an investment scheme according to the predicted environmental investments.

As discussed above on investment prediction modeling, two critical challenges can be summarized as follows: 1) the use of a single method to select indicators is the main way to screen indicator information in previous studies, but the difference of indicator selection methods will increase the probability of the information loss of representative indicators; 2) since environmental management requires human involvement, the modeling of investment prediction must have the feature of white-box design and the prediction process of the model must be explainable for environment managers.

To overcome the two challenges on investment prediction modeling, a new environmental investment prediction model is proposed on the basis of extended belief rule-based (EBRB) model and indicator ensemble selection (IES) and the new model is called the IES-EBRB model, in which the EBRB model is a white-box designed decision-making model proposed by Liu *et al*. [10] and has the ability on using experts' knowledge to enhance data analytics for autonomous decision making; the IES is an extension of ensemble learning [2] to cooperatively integrate different kinds of indicator selection methods for selecting indicators, so representative indicators can be accurately selected for investment prediction modeling. Accordingly, the proposed IES-EBRB model has the following advantages:

(1) The IES-EBRB model is an unlocking structure so that any kind of indicator selection methods can be added and used together to rank the relative importance of different environmental indicators. In other words, environmental indicators can be sorted according to the various perspectives derived from the advantages of different indicator selection methods.

(2) Three ranking combination functions are introduced to integrate the rankings of each indicator, which are obtained from different kinds of indicator selection methods. According to the combined ranking of all indicators, the indicators with top ranking are selected as representative indicators and thus used for the construction of an investment prediction model.

(3) The IES-EBRB model can be regarded as a data-driven and knowledge-driven hybrid model, because its extended belief rule can be generated from historical data and revised according to experts' knowledge. Hence, experienced managers can embed experts' knowledge into the IES-EBRB model to enhance its ability on investment prediction.

(4) The IES-EBRB model imports the benefits from the EBRB model so that it not only has a high interpretability due to the explainable processes of generating extended belief rules and predicting environment investments, but also is easy to achieve a low complexity because the total number of extended belief rules in the IES-EBRB model does not increase exponentially with the increasing number of indicators and/or referential values.

In order to demonstrate the effectiveness of the proposed IES-EBRB model, a real case regarding actual environmental investment data derived from 2005 to 2018 in 31 China provinces is used to illustrate the development procedure of the proposed IES-EBRB model and also provide the comparative analysis of some existing time series forecasting-based and input-output relationship-based investment prediction models.

The remainder of this work is as follows: ***Section 2*** is the literature review and outlines the challenges of previous environmental investment prediction modeling. ***Section 3*** introduces the basic methodology of conventional EBRB model. ***Section 4*** proposes the IES-EBRB model for environmental investment prediction. ***Section 5*** provides a case study to perform model validation. ***Section 6*** concludes this study.

2. Literature Review and Challenges

In this section, the previous studies of the application of the EBRB model and modeling of environmental investment prediction are reviewed, and then the challenges of these studies are summarized to illustrate the necessity of this study.

2.1. Previous applications of the EBRB model

At the beginning of proposing the EBRB model, it was verified its effectiveness on some large-scale data, including the software defect prediction of National aeronautics and space administration [10] and the public health assessment of Northern Ireland [3]. Thereafter, many researchers started paying more attention to the EBRB model. For example, Espinilla *et al*. [7] proposed the adaption of the EBRB model to handle binary sensor data in smart environments. The results showed that the EBRB model can provide a desired accuracy better than the most popular classifiers in terms of robustness. Yang *et al*. [22] introduced data envelopment analysis (DEA) to downsize the rule scale of EBRB model, and the reduced EBRB model demonstrates its performance on the problem of oil pipeline leak detection. Similarly, Yang *et al*. [24] embedded a parameter learning model into the EBRB model to optimize the basic parameters. The case study of bridge risk assessment showed that the accuracy of the EBRB model can be improved using the parameter learning model. The most common application of the EBRB model is the classification problems derived from the well-known UCI database in the field of machine learning [23][28][30], and these classification problems mainly includes the diagnosis of diseases, frequency of blood donation, income level, forest coverage, satellite image classification, etc.

Recently, the EBRB model was applied to the field of environmental management and showed excellent performance on prediction accuracy and explainability, *i.e*., Wang *et al*. [16] introduced the EBRB model with joint learning for the first time to predict environmental investments, The comparative results revealed that the EBRB model has higher accuracy than adaptive neural fuzzy inference system (ANFIS)[26]-based and grey model (GM)[6]-based models. However, the indicators used for predicting three kinds of environmental investments are selected by using the single indicator selection method, which sometime do not have enough ability to select indicators for any kind of investments. Ye *et al*. [27] extended the EBRB model by considering consequence reliability for predicting investments under interval uncertainty, but there exists the similar issue that the indicators are selected according to experts' knowledge, which may be difficult or even impossible to effectively select indicators because of lack of contextual information and/or data. Hence, in order to enhance the application of the EBRB model on environmental investment prediction, an effective method to select indicators for any kind of investments should be proposed to improve the EBRB model.

2.2. Previous modeling of environmental investment prediction

Previous modeling of environmental investment prediction can be summarized into two aspects: time series forecasting -based and input-output relationship-based models. The former one is the use of a model to predict future investments based on previously observed investments. The latter one is the use of a model to predict future investments based on previously observed input and output data of investments.

For the time series forecasting-based models, GM and the auto regressive integrated moving average (ARIMA) is the most common used models in environmental management, *i.e*., Xu *et al*. [20] proposed an optimized hybrid GM model to improve prediction accuracy of electricity energy consumption; Chen *et al*. [6] also proposed the GM-based investment prediction scheme for environmental management for the next ten years of China. Furthermore, Kaytez [12] developed a hybrid method using ARIMA and support vector machine to predict the net electricity consumption of Turkey until 2022. From the above studies, time series forecasting-based models have achieved some successes in the modeling of investment prediction. However, these models inevitably ignored the hidden logic relationship among different factors in environmental managements, *e.g.*, the influence of environmental pollution and economic development on investment prediction.

Therefore, the input-output relationship-based investment prediction models are becoming a trend for environmental investment prediction, because it considers both economic development and pollution emission to build prediction models. The representative studies have: Ye *et al*. [25] utilized the fuzzy rule-based system (FRBS) to develop a new investment prediction model, in which the FRBS is comprised of fuzzy rules generated from input-output data pairs of environmental investments. Similarly, the ANFIS was also used by Ye *et al*. [26] to propose the investment prediction model with consideration of environmental efficiency. From the above studies, although input-output relationship-based models have showed better accuracy than time series forecasting-based models, the used decision-making methodologies are usually difficult to have enough explainability, *i.e.*, the modeling of both FRBS and ANFIS does not make a distinction of different indicators. Hence, in order to enhance the explainability of investment prediction modeling, a white-box design of decision-making methodologies should be used to propose investment prediction models.

2.3. Challenges of constructing new investment prediction model

Previous studies have proposed lots of prediction models for environmental investments. However, the following two challenges still should be considered before constructing a new investment prediction model.

Challenge 1: The information loss of representative indicators in environmental indicator selection.

Environmental investment prediction modeling involves a large number of environmental indicators and the use of a single method to select representative indicators is the existing way to construct an environmental investment prediction model. However, each indicator selection method has its inherent strengths and weaknesses, resulting in an undesired result that the information of representative indicators would be abandoned in the process of indicator selection and also impact the accuracy of prediction models.

Challenge 2: The lack of human involvement and sufficient explainability in prediction models.

The goal of environmental investment prediction is to serve environment managers for making an effective investment scheme. Hence, it is necessary to consider if the decision-making methodology used for investment prediction modeling has the ability for exploiting experts' knowledge to enhance data analytics and provide an explainable process of investment prediction. However, apart from the EBRB model discussed in ***Section 2.1***, existing environmental investment prediction models typically lack human involvement and sufficient explainability.

The above-mentioned two challenges clearly indicate the necessary conditions to propose a new model for predicting environment investments. Hence, based on the previous applications of the EBRB model on environmental investment prediction modeling [16][27], a new environmental investment prediction model is proposed on the basis of the EBRB model and IES in the coming sections.

3. Basic Methodologies of the EBRB Model

The EBRB model [10] is an advanced rule-based system extended from the belief rule-based (BRB) system [21] by embedding belief structures into the IF part of belief rules. The extended belief rules therefore have ability to represent hybrid information of experts' knowledge and historical data under uncertainty.

Suppose that there are *M* antecedent attributes U_i (i=1,…, *M*) with each attribute having J_i reference values $A_{i,j}$ (j=1,…, J_i) and one consequent attribute *D* with *N* consequents B_n (n=1,…, *N*). Hence, the *k*th (k=1,…, *L*) extended belief rule R_k is written as:

$$\begin{aligned} R_k: IF\ U_1\ is\ \{(A_{1,j},\alpha_{1,j}^k); j=1,\ldots,J_1\} \wedge \ldots \wedge U_M\ is\ \{(A_{M,j},\alpha_{M,j}^k); j=1,\ldots,J_M\} \\ THEN\ D\ is\ \{(B_n,\beta_n^k); n=1,\ldots,N\},\ with\ \ \theta_k\ \ and\ \{\delta_1,\ldots,\delta_M\}. \end{aligned} \tag{1}$$

where $\{(A_{i,j},\alpha_{i,j}^k); j=1,\ldots,J_i\}$ and $\{(B_n,\beta_n^k); n=1,\ldots,N\}$ denotes belief structures in antecedent attribute U_i and consequent attribute *D*, respectively; $\alpha_{i,j}^k\,(0\le\alpha_{i,j}^k\le 1)$ and $\beta_n^k\,(0\le\beta_n^k\le 1)$ denote the belief degrees of reference value $A_{i,j}$ and consequent B_n in R_k, and they meets $\sum_{j=1}^{J_i}\alpha_{i,j}^k\le 1$ and $\sum_{n=1}^{N}\beta_n^k\le 1$; $\delta_i\,(0\le\delta_i\le 1)$ and $\theta_k\,(0\le\theta_k\le 1)$ denote the weight of U_i and R_k, respectively.

It is worth noting that, as shown in Eq. (1), an extended belief rule is comprised of linguistic terms (*e.g.*, reference values $A_{i,j}$ and consequents B_n) and numerical terms (*e.g.*, belief degrees $\alpha_{i,j}^k$ and β_n^k and weights δ_i and θ_k). All of these terms not only have its explainable meaning, but also can be determined using experts' knowledge, historical data, or both.

Based on the rules shown in Eq. (1), the evidential reasoning (ER)-based inference method is used to produce outputs for replying any given input data. The detailed steps of ER-based inference method are as follows:

***Step 1*:** To calculate activation weights for each rule. For the given input data $\boldsymbol{x}=\{x_1,\ldots,x_M\}$, each input x_i can be transformed into belief distribution $S(x_i)=\{(A_{i,j},a_{i,j}); j=1,\ldots,J_i\}$ by

$$a_{i,j}=\frac{u(A_{i,j+1})-x_i}{u(A_{i,j+1})-\mu(A_{i,j})}\ \text{and}\ \ a_{i,j+1}=1-a_{i,j}\,,\ \text{if}\ \ u(A_{i,j})\le x_i\le u(A_{i,j+1}) \tag{2}$$

$$a_{i,k}=0\,,\ \text{for}\ \ k=1,\ldots,J_i\ \text{and}\ \ k\ne j,j+1 \tag{3}$$

where $u(A_{i,j})$ denotes the utility value of $A_{i,j}$.

Next, the activation weight of the *k*th (k=1,…, *L*) rule can be calculated by

$$w_k=\frac{\theta_k\prod_{i=1}^{M}(S^k(x_i,U_i))^{\bar{\delta}_i}}{\sum_{l=1}^{L}(\theta_l\prod_{i=1}^{M}(S^l(x_i,U_i))^{\bar{\delta}_i})} \tag{4}$$

where

$$S^k(x_i,U_i)=1-\sqrt{\frac{\sum_{j=1}^{J_i}(a_{i,j}-\alpha_{i,j}^k)^2}{2}} \tag{5}$$

$$\bar{\delta}_i=\frac{\delta_i}{\max_{j=1,\ldots,M}\{\delta_j\}} \tag{6}$$

Step 2: To integrate rules for producing outputs. Based on activation weights, the rules with $w_k > 0$ can be integrated using the analytical ER algorithm [17]:

$$\beta_n=\frac{\prod_{k=1}^{L}(w_k\beta_n^k+1-w_k\sum_{i=1}^{N}\beta_i^k)-\prod_{k=1}^{L}(1-w_k\sum_{i=1}^{N}\beta_i^k)}{\sum_{i=1}^{N}\prod_{k=1}^{L}(w_k\beta_i^k+1-w_k\sum_{j=1}^{N}\beta_j^k)-(N-1)\prod_{k=1}^{L}(1-w_k\sum_{j=1}^{N}\beta_j^k)-\prod_{k=1}^{L}(1-w_k)} \tag{7}$$

Finally, the output of EBRB model for replying the given input data $\boldsymbol{x}$ can be represented as follows:

$$f(\boldsymbol{x})=\sum_{n=1}^{N}u(B_n)\beta_n+(1-\sum_{n=1}^{N}\beta_n)\frac{u(B_1)+u(B_N)}{2} \tag{8}$$

4. An Improved EBRB Model Using IES for Environmental Investment Prediction

In this section, the procedure of IES to select indicators is proposed in ***Section 4.1***, followed by the introduction of the EBRB modeling based on the selected indicators in ***Section 4.2***. Finally, the framework of the new model for environmental investment prediction, called the IES-EBRB model, is provided in ***Section 4.3***.

4.1. IES procedure for environmental investment prediction

In order to overcome ***Challenge 1*** outlined in ***Section 2.3*** for proposing a new investment prediction model, this section provides an ensemble approach for indicator selection, namely IES, whose procedure mainly includes: 1) the use of different feature selection methods to obtain the weights and rankings of the indicators of environmental investments, and 2) the integration of rankings and weights of each indicator using combination functions to select representative environmental indicators. Accordingly, the steps of IES procedure are showed as follows:

Step 1**:** To obtain the individual ranking of environmental indicators by different methods. Considering that any kind of indicator selection method inevitably has its strengths and weaknesses, a smart strategy to select indicators is based on various kinds of methods so that representative indicators can be selected from various perspectives. Based upon this viewpoint, existing indicator selection methods are used together to obtain the individual ranking of the given indicators.

Step 2: To select representative indicators using ranking combination functions. On the basis of the individual rankings of the given indicators, the ranking combination functions, *e.g.*, minimum, average, and geometric average functions, are used to integrate the individual rankings of each indicator to obtain their integrated ranking, so that the indicators with the top integrated ranking can be selected as representative indicators for environmental investment prediction modeling.

In order to provide the more details of the above two steps, the corresponding pseudo-code is shown in Algorithm 1.

Algorithm 1: an algorithm to select the representative indicators for environmental investment prediction

Inputs: $\boldsymbol{IS}=\{IS_1,\ldots, IS_N\}$: a set of N indicator selection methods; $\boldsymbol{U}=\{U_1,\ldots, U_M\}$: a set of M environmental indicators; CF: a certain combination function; T: number of indicators to be selected.

Outputs: A set of T selected indicators $\boldsymbol{\Omega}$.

```
for each $n$ from 1 to $N$ do
    for each $m$ from 1 to $M$ do
        To obtain the importance of indicators $U_m$ using method $IS_n$, denoted as $\varpi_{n,m}$.
    end for
    for each $m$ from 1 to $M$ do
        To obtain the ranking of indicators $U_m$ using importance $\varpi_{n,m}$ in ascending order, denoted as $r_{n,m}$.
    end for
end for
for each $m$ from 1 to $M$ do
    To obtain the integrated ranking of indicator $U_m$ using function $CF$ with $r_{n,m}$, denoted as $r_m$ .
end for
for each $t$ from 1 to $T$ do
    To select the $t$th representative indicator using $\boldsymbol{\Omega}=\boldsymbol{\Omega}\cup\{U_i\},\boldsymbol{U}=\boldsymbol{U}-\{U_i\},i=\arg\min_{U_m\in\boldsymbol{U}}\{r_m\}$ .
end for
```

In the above-mentioned indicators selection algorithm, it is worth noting that the number of representative indicators to be selected, that is *T*, is a crucial threshold because it determines the accuracy of the investment prediction model. In other words, if *T* is too large or small, the model accuracy will be weakened due to the involvement of noise indicators or the information loss of representative indicators. Hence, the value of threshold *T* should be determined carefully according to experts' knowledge or modeling assessment criteria, *e.g.*, Akaike information criterion or generalization error.

4.2. EBRB modeling for environmental investment prediction

In order to overcome ***Challenge 2*** detailed in ***Section 2.3*** for proposing a new investment prediction model, this section provides an EBRB modeling using the input-output data of environmental indicators and environment experts' knowledge, in which the EBRB modeling includes two major parts: 1) determination of basic parameters using learning model or experts' knowledge; 2) generation of belief distributions using basic parameters; 3) calculation of rule weights using belief distributions. Accordingly, the steps of EBRB modeling are showed as follows:

Step 1**:** To determine the basic parameters of the EBRB model. Suppose that an EBRB model has *M* antecedent attributes and one consequent attribute. Hence, based on experts' knowledge, the basic parameters of the EBRB model can be given as: *M* attribute weights δ_i (i=1,…, *M*), J_i reference values $A_{i,j}$ (j=1,…, J_i) for the *i*th antecedent attribute, and *N* consequents for the consequent attribute. Moreover, according to actual demands, the utility values for all reference values and consequents, namely $u(A_{i,j})$ and $u(B_n)$, can be initialized using experts' knowledge or the following learning model.

$$\min \sum_{t=1}^{T} \left| y_t - f(\boldsymbol{x}_t) \right| \tag{9a}$$

$$s.t.\ 0 \le \delta_i \le 1; i = 1,\ldots,M \tag{9b}$$

$$u(A_{i,j}) \le u(A_{i,j+1}); j = 1,\ldots,J_i - 1; i = 1,\ldots,M \tag{9c}$$

$$u(A_{i,1}) = lb_i; u(A_{i,J_i}) = ub_i; i = 1,\ldots,M \tag{9d}$$

$$u(B_n) \le u(B_{n+1}); n = 1,\ldots,N-1 \tag{9e}$$

$$u(B_1) = lb; u(B_N) = ub \tag{9f}$$

where lb and ub are the lower and upper bounds of the consequent attribute; lb_i and ub_i are the lower and upper bounds of the *i*th antecedent attribute; <$\boldsymbol{x}_t$, y_t> (t=1,…, *T*) is the actual input-output data pairs; and $f(\boldsymbol{x}_t)$ is the prediction output of the EBRB model for replying input $\boldsymbol{x}_t$.

Step 2**:** To generate the belief distributions using basic parameters. Based on the basic parameters provided by ***Step 1***, *T* actually input-output data pairs <$x_{k,1}$,…, $x_{k,M}$, y_k> (k=1,…, *T*) should be used to generate *T* sets of input and output belief distributions using the utility-based information transformation technique shown in Eqs. (2) and (3), *i.e.*, the *k*th input-output data pair can generate belief distributions $S(x_{k,i}) = \{(A_{i,j}, a_{i,j}^k); j = 1,\ldots,J_i\}$ (i=1,…,*M*) and $S(y_k) = \{(B_n, a_n^k); n = 1,\ldots,N\}$. All these belief distributions constitute the main component of the *k*th extended belief rule. Additionally, it is worth noting that a set of belief distributions can be also given by experts because the determination of belief distributions is independent of each other, so that experts can involve the modeling of an EBRB model.

Step 3**:** To calculate the rule weight of extended belief rules. Based on the belief distributions provided by ***Step 2***, the rule weight can be calculated for each extended belief rule according to the following two definitions:

Definition 1 **(*****Distance of belief distributions*****)**: Suppose there are two belief distributions ***P*** = (p_s; s=1,…, *S*) and ***Q*** = (q_s; s=1,…, *S*),thus the distance of ***P*** and ***Q*** is as follows:

$$D(\boldsymbol{P}, \boldsymbol{Q}) = \sqrt{\frac{\sum_{s=1}^{S} (p_s - q_s)^2}{2}} \tag{10}$$

Definition 2 **(*****Consistency of extended belief rules*****)**: Suppose there are *L* extended belief rules R_l (l=1,…, *L*), thus the consistency of extended belief rules R_l and R_k (k=1,…, *L*; k≠l) can be calculated as follows:

$$C(R_l, R_k) = \exp\left\{ -\frac{\left(\frac{SA(R_l, R_k)}{SC(R_l, R_k)} - 1 \right)^2}{\left(\frac{1}{SA(R_l, R_k)} \right)^2} \right\} \quad (11)$$

where $SA(R_l, R_k)$ and $SC(R_l, R_k)$ denote the similarity of antecedent (SA) and consequent (SC) attributes between extended belief rules R_l and R_k, respectively, and they can be calculated by:

$$SA(R_l, R_k) = 1 - \max_{i=1,\ldots,M}\{D(S(x_{l,i}), S(x_{k,i}))\} \quad (12)$$

$$SC(R_l, R_k) = 1 - D(S(y_l), S(y_k)) \quad (13)$$

Therefore, based on ***Definitions 1*** and ***2***, the rule weight of the *k*th extended belief rule is calculated as follows:

$$\theta_k = 1 - \frac{\sum_{l=1, l\neq k}^{L}(1 - C(R_l, R_k))}{\sum_{j=1}^{L}\sum_{l=1, l\neq j}^{L}(1 - C(R_l, R_j))} \quad (14)$$

In the above-mentioned EBRB modeling, it can be found from ***Definition 2*** that if two extended belief rules have the same belief distributions in antecedent attributes, namely $SA(R_l, R_k)$=1, then the consistency of two rules tends to be high when there are similar belief distributions in consequent attribute, namely $SC(R_l, R_k)$=1. Otherwise, the consistency ranges from 0 to 1. These fundamental characteristics actually are the rationale of ***Definition 2***. To date, ***Definition 2*** has been used in the EBRB model [10] and FRBS [11] for measuring the consistency of two rules. Additionally, some existing rule consistency-related studies for rule-based systems can be found in [1][4].

4.3. Framework of the IES-EBRB model for environmental investment prediction

In this section, the framework of the new model based on combining the IES procedure with the EBRB modeling and EBRB prediction, and given the name IES-EBRB model, is introduced to illustrate how to address environmental investment prediction and also overcome the challenges shown in ***Section 2.3***. As shown in **Fig. 1**, the IES-EBRB model can be constructed based on environment experts' knowledge, environmental indicators, and historical data, and its final goal is to predict environmental investments based on new environment data.

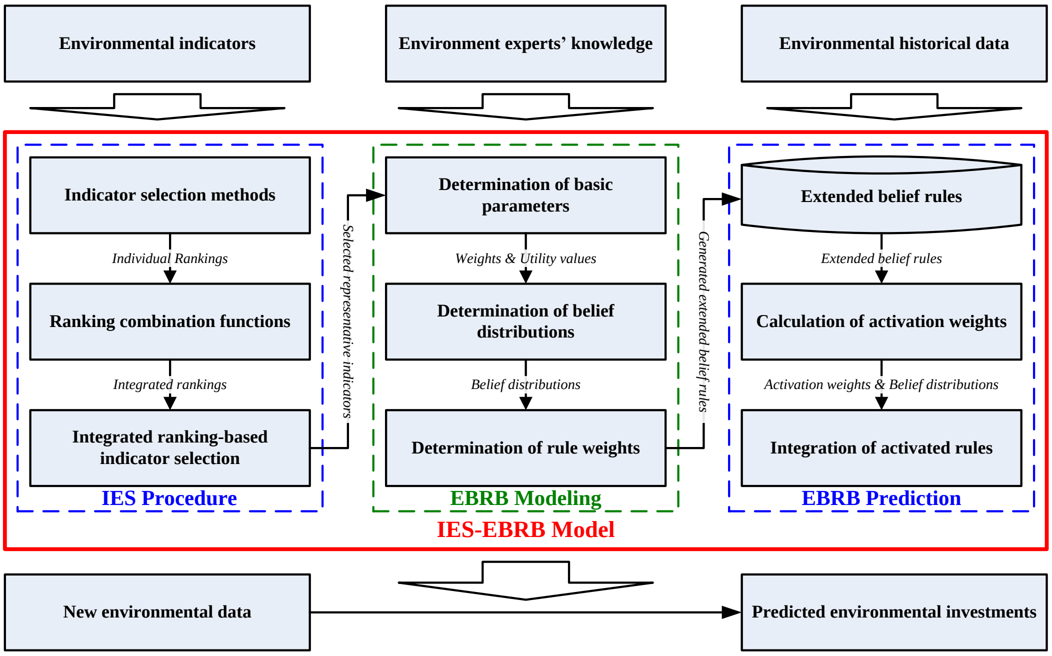
Environmental indicators
Environment experts' knowledge
Environmental historical data
Indicator selection methods
Individual Rankings
Ranking combination functions
Integrated rankings
Integrated ranking-based indicator selection
IES Procedure
Selected representative indicators
Determination of basic parameters
Weights & Utility values
Determination of belief distributions
Belief distributions
Determination of rule weights
EBRB Modeling
Generated extended belief rules
Extended belief rules
Extended belief rules
Calculation of activation weights
Activation weights & Belief distributions
Integration of activated rules
EBRB Prediction
IES-EBRB Model
New environmental data
Predicted environmental investments

Fig. 1. Framework of the IES-EBRB model

From **Fig. 1**, the detailed steps of the IES-EBRB model can be described as follows:

Step 1*:** To select representative environmental indicators using IES procedure. Assume that there are NI environmental indicators $\{U_1,\ldots, U_{NI}\}$ and NO environmental investments $\{D_1,\ldots, D_{NO}\}$. For each environmental investment D_s (s=1,…, NO), M_s representative indicators, denoted as $\{U_1,\ldots, U_{Ms}\}$, can be selected according to the use of indicator selection methods, ranking combination functions, and integrated ranking-based indicators selection shown in ***Section 4.1. Note that the IES procedure provides a strategy for the IES-EBRB model to avoid the information loss of representative indicators.

Step 2: To construct the IES-EBRB models based on representative indicators and EBRB modeling. Assume there are T historical environmental data $\{(x_1^t,\ldots,x_{NI}^t, y_1^t,\ldots,y_{NO}^t); t=1,\ldots,T\}$. For each investment with its representative indicators, *e.g.*, $\{U_1,\ldots, U_{Ms}, D_s\}$, the T input-output data pairs $\{(x_1^t,\ldots,x_{M_s}^t, y_s^t); t=1,\ldots,T\}$ are extracted and thus T extended belief rules can be generated to construct an IES-EBRB model according to the determination of basic parameters, generation of belief distributions, and calculation of rule weights shown in ***Section 4.2***. Note that the EBRB modeling allows the IES-EBRB model to enhance data analytic using experts' knowledge.

Step 3: To predict environmental investments using IES-EBRB models. Based on ***Step 1*** and ***Step 2***, NO IES-EBRB models can be constructed for NO environmental investments. Hence, when a new environmental data, *e.g.*, $\boldsymbol{x} = (x_1,\ldots,x_{NI})$, is given, NO predicted environmental investments $f_s(\boldsymbol{x})$ (s=1,…, NO) can be obtained for replying the new data $\boldsymbol{x}$ according to the calculation of activation weights and integration of activated rules shown in ***Section 3***. Note that the EBRB prediction ensures the ability of the IES-EBRB model having sufficient explainability for environment managers.

In the above-mentioned steps, it can be found that an IES-EBRB model has advantage in terms of high interpretability and low complexity. On the one hand, the IES procedure depends on indicators' weights to select representative indicators, which not only provides a panoramic view to explain the relative importance of different indicators, but also decreases the complexity of EBRB modeling because of the reduced number of indicators. On the other hand, the EBRB modeling helps the IES-EBRB model imports the benefits from the EBRB model, which has been demonstrated to be a white-box designed decision-making model and get rid of the combination explosion problem, which indicates that the number of rules or the complexity of the EBRB model increases exponentially with the increasing number of indicators and/or referential values, so that the IES-EBRB model has the ability to ensure high interpretability and low complexity during the process of EBRB modeling and prediction for environmental investment prediction.

5. Case Study of Environmental Investment Prediction in China

In order to verify the effectiveness of the proposed IES-EBRB model, the actual environmental data of 31 provinces in mainland China were used to perform an empirical case study. The introduction of data source and model setting is in ***Section 5.1***, the development of the IES-EBRB model is in ***Section 5.2***, and the comparative analysis is in ***Section 5.3***.

5.1. Data source and model setting

The ten environmental indicators and three environmental investments widely used in the previous studies [6][16][25][26] are collected to construct the model of environmental investment prediction, and the corresponding environmental historical data related with 31 provinces in the mainland of China from 2005 to 2018 are derived from *China Statistical Yearbook* and *China Environmental Statistical Yearbook*, respectively, in which both of them are the most commonly used and reliable public database for the study of environmental management in China [5][9][19]. The main characteristics of environmental indicators, investments, and data are summarized in **Table 1** and **Table 2**.

Table 1. Introduction of environmental indicators in investment prediction

No.	Environmental indicators	Abbr.	Specific interpretation of indicators	Max	Min	Average	Unit
1	gross domestic product	GDP	Value of gross domestic product	89705	220	15627	10^8yuan
2	Total profit	TP	Total profit of Enterprises above Designated Size	10574	-91.89	1530.9	10^8yuan
3	Garbage clean-up	GCU	Garbage removal and transportation volume	2645	16.3	548	10^4ton
4	Sulfur dioxide	SO_2	Emission of sulfur dioxide	2002000	1000	662937	ton
5	Smoke and dust	SM	Emission of smoke and dust	1797683	1000	364270	ton
6	Carbon dioxide	CO_2	Emission of carbon dioxide	4678	7.07	1096	10^4ton
7	Waste water	WW	Total emission of waste water	938261	2685	201795	10^4ton
8	Chemical oxygen demand	COD	Emission of chemical oxygen demand	198.25	1.38	53.6	10^4ton
9	Lead emission	LE	Lead emission in waste water	42466	0.002	1328	10^3kg
10	Petroleum emissions	PE	Petroleum emissions in waste water	2937	0.03	498	ton

Table 2. Introduction of environmental investments in investment prediction

No.	Environmental investments	Abbr.	Specific interpretation of investments	Max	Min	Average	Unit
1	Energy consumption	EC	Total electricity consumption	5959	9	1397	10^4ton
2	Capital investment	CI	Fixed assets investment	55203	162	10306	10^8yuan
3	Labor investment	LI	Total number of employees	1973	15	469	Person

From **Table 1** and **Table 2**, it is obvious that there are significant regional differences in environmental data for 31 provinces of China in terms of maximum and minimum values. For example, as shown in **Table 1**, the minimum value of TP is only -91.89, while the maximum value of TP is 10574. As shown in **Table 2**, the maximum value of EC is 5959, while the minimum value of EC is 9. Moreover, it is also obvious that the annual CI in China is large, which indicated that the economic development in China has significantly regional difference from the maximum and minimum value of CI, in which the maximum and minimum values of CI are 55203 and 162, respectively.

Additionally, in order to construct and validate the IES-EBRB model for environmental investment prediction using the above environmental data, the environmental data from 2005 to 2017 of each province in China are used as training data for model construction and the environmental data in 2018 are used as testing data for model validation. Furthermore, six kinds of indicator selection methods, namely Pearson correlation-based, ReliefF algorithm-based, random forest classifier-based, simple linear regression-based, correlation coefficient standard deviation-based, and entropy algorithm-based indicator selection methods, three kinds of ranking combination functions, namely minimum-based, geometric average-based, and average-based combination functions, are introduced to select representative environmental indicators. The specific descriptions for each method and function are showed in **Table 3** and **Table 4**, respectively. It should be noted that the selected methods used to select representative indicators are all able to identify the relative importance of different indicators, instead of selecting a subset of indicators, *e.g.*, CFS-based method [16], or extracting the principal components of all indicators, *e.g.*, PCA-based method [25]. Owing to the feature of the selected methods, the results of the indicator selection methods can be accessible to the application of combination functions and they are also able to illustrate the relative importance of different indicators.

Table 3. Introduction of six indicator selection methods

Core of indicator selection	Abbr.	Descriptions
Pearson correlation[8]	PC-IS	To evaluate the worth of indicators by measuring its Pearson's correlation with respect to investments
ReliefF algorithm[8]	RA-IS	To evaluate the worth of indicators by repeatedly sampling a data and considering the value of a given indicator for the nearest data of the same and different investments
Correlation coefficient standard deviation[15]	CCSD-IS	To evaluate the worth of indicators by measuring its correlation coefficient and standard deviation with respect to investments
Entropy algorithm[29]	EA-IS	To evaluate the weight of indicators by measuring its information entropy with respect to investments
Random forest classifier[8]	RFC-IS	The current set of indicators is applied to train a random forest classifier iteratively by removing each indicator, so that the performance can evaluate the worth of indicators
Simple linear regression[8]	SLR-IS	The current set of indicators is applied to do a simple linear regression iteratively by removing each indicator, so that the performance can evaluate the worth of indicators

Table 4 Introduction of three ranking combination functions

Core of combination	Formula	Abbr.	Descriptions
Minimum	$r_m = \min_{n=1,\ldots,N}\{r_{n,m}\}$	MIN-RC	To combine the *m*th indicator's rankings obtained from *N* indicator selection methods based on the minimum function.
Average	$r_m = \sum_{n=1}^{N} r_{n,m}$	AVG-RC	To combine the *m*th indicator's rankings obtained from *N* indicator selection methods based on the average function.
Geometric average	$r_m = \sqrt[N]{\prod_{n=1}^{N} r_{n,m}}$	GAVG-RC	To combine the *m*th indicator's rankings obtained from *N* indicator selection methods based on the geometric average function.

5.2. Development process of the IES-EBRB model

In this section, the main process of developing an IES-EBRB model is provided via the following three sub-processes: 1) indicator selection using the IES procedure, 2) model construction using the EBRB modeling, and 3) model application using the EBRB prediction.

5.2.1. The 1st sub-process: indicator selection using the IES procedure

To illustrate the process of indicator selection, the indicator selection methods in **Table 3** and the ranking combination function in **Table 4** are used to perform the IES procedure shown in ***Step 1*** from ***Section 4.3***. Taking the prediction of CI as example, the importance and ranking of each environmental indicator can be calculated by using the six indicator selection methods and they are shown in **Table 5**. From **Table 5**, GDP is the most important indicator by PC-IS, RFC-IS, and SLR-IS methods, SO_2 is the most important indicator by CCSD-IS and EA-IS methods, and COD is the most important indicator by RA-IS method. The reason why there are different kinds of the most important indicators because different methods have different perspectives on representative indicator selection. Meanwhile, LE and PE have the lower importance and higher ranking compared with other indicators, because the pollution emissions of LE and PE are lower than other pollutants, they mainly affect the quality of water in different regions.

Table 5. Importance and ranking of ten indicators by six indicator selection methods

Indicators	PC-IS		RA-IS		CCSD-IS		EA-IS		RFC-IS		SLR-IS	
	Ranking	Importance	Ranking	Importance	Ranking	Importance	Ranking	Importance	Ranking	Importance	Ranking	Importance
GDP	1	0.892	2	0.029	6	0.091	6	0.076	1	4761.540	1	5524.999
TP	2	0.852	3	0.028	4	0.099	5	0.083	2	4573.554	2	4812.724
GCU	5	0.613	4	0.023	7	0.081	7	0.069	6	551.706	5	2084.033
SO_2	9	0.177	6	0.020	1	0.137	1	0.184	10	-1648.715	9	131.191
SM	7	0.365	7	0.020	8	0.073	8	0.068	8	-1197.295	7	634.845
CO_2	3	0.823	8	0.020	3	0.108	3	0.117	3	3125.520	3	4342.130
WW	4	0.674	5	0.021	5	0.093	4	0.094	4	1906.743	4	2599.113
COD	6	0.549	1	0.030	2	0.115	2	0.134	7	406.387	6	1647.584
LE	8	0.193	10	0.011	10	0.033	10	0.020	5	1393.877	8	142.980
PE	10	0.143	9	0.012	9	0.072	9	0.062	9	-1579.056	10	48.580

Based on the three ranking combination functions in **Table 4**, the individual ranking of ten indicators derived from six methods can be combined for integrated rankings. Taking MIN-RC function for example, the integrated ranking of GDP, CO_2, and SO_2 is 1 because these three indicators are the most selected indicator by one of six indicator selection methods at least. The final results of the three ranking combination functions are shown in **Table 6**. It can be found from **Table 6** that the ascending order of ten indicators has slight difference among MIN-RC, AVG-RC, and GAVG-RC functions, in which GDP, SO_2, and COD are the top three indicators for MIN-RC function, while GDP, TP, and CO_2 for AVG-RC function and GDP, TP, and COD for GAVG-RC function. Additionally, the overall assessment of each indicator is a quantitative weight, which provides an explainable view for environment managers to determine which indicator is the most important indicator and also reflect a relative importance for each indicator.

Table 6. Integrated ranking of ten indicators by three ranking combination functions

Functions	Indicators (Integrated rankings)
MIN-RC	GDP (1.00) = SO_2 (1.00) = COD (1.00) > TP (2.00) > CO_2 (3.00) > GCU (4.00) = WW (4.00) > LE (5.00) > SM (7.00) > PE (9.00)
AVG-RC	GDP (2.83) > TP (3.00) > CO_2 (3.83) > COD (4.00) > WW (4.33) > GCU (5.67) > SO_2 (6.00) > SM (7.50) > LE (8.50) > PE (9.33)
GAVG-RC	GDP (2.04) > TP (2.80) > COD (3.17) > CO_2 (3.53) > SO_2 (4.12) > WW (4.31) > GCU (5.55) > SM (7.48) > LE (8.27) > PE (9.32)

5.2.2. The 2nd sub-process: model construction using the EBRB modeling

To illustrate the process of model construction for the IES-EBRB model when taking the prediction of CI as example, experts' knowledge should be firstly considered into the EBRB modeling, so that the ability of data analytics for the IES-EBRB model can be enhanced with the inclusion of human involvement. Hence, according to the experts' knowledge used in [16], the top four important indicators are selected as representative indicators and all of these selected indicators are assumed to have three reference values, *e.g.*, {*Low*, *Middle*, *High*}, as well as investment CI having three consequents, *e.g.*, {*Low*, *Middle*, *High*}. **Table 7** shows the initial value of basic parameters when the EBRB modeling is performed using the GAVG-RC function for predicting the CI of Hunan province.

Table 7. Initial value of basic parameters for the CI prediction of Hunan

Indicator	Correspondent relationship	Weight	*Low*	*Medium*	*High*
GDP	Antecedent attribute	1.0000	5641.9400	18596.6550	31551.3700
TP	Antecedent attribute	1.0000	154.7700	1101.3200	2047.8700
COD	Antecedent attribute	1.0000	60.2600	95.3900	130.5200
CO_2	Antecedent attribute	1.0000	520.1489	847.1171	1174.0853
CI	Consequent attribute	-	2072.5600	15212.9450	28353.3300

In order to revise the initial value of basic parameters shown in **Table 7** according to historical data, the learning model in Eqs. (9a) to (9f) and the differential evolution (DE) algorithm [24] are used to optimize the value of basic parameters, in which the number of individuals and iterations used in DE algorithm is set as 100 and 600, respectively. **Fig. 2** shows the change of the target value obtained from the learning objective shown in Eq. (9a). It is clear from **Fig. 2** that the target value of EC prediction in Hunan is significantly decreased and gradually tends to converge after 600 iterations, whose value decreases from 11456.85 to 3764.71. The optimized value of basic parameters for EC prediction is shown in **Table 8**.

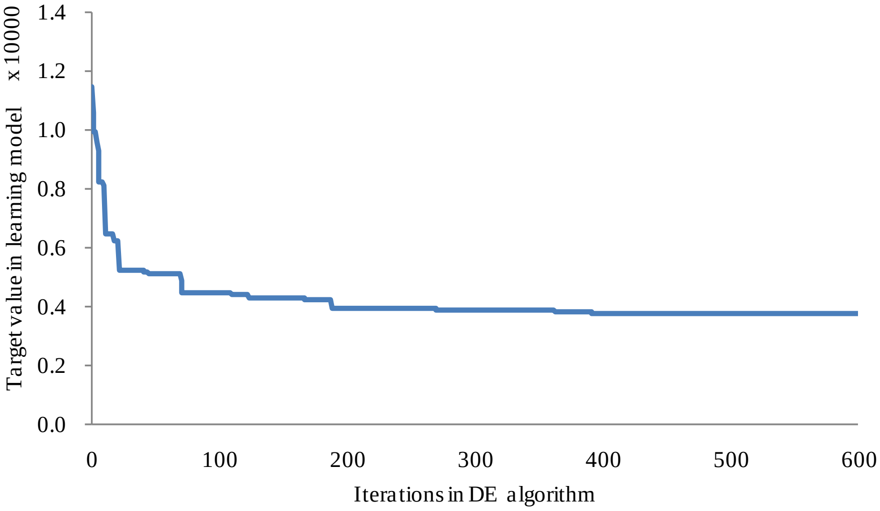

Fig. 2. Change of target values at each iteration

Table 8. Optimized value of basic parameters for the CI prediction of Hunan

Indicator	Correspondent relationship	Weight	*Low*	*Medium*	*High*
GDP	Antecedent attribute	0.9698	5641.9400	13004.8786	31551.3700
TP	Antecedent attribute	0.9816	154.7700	1810.3484	2047.8700
COD	Antecedent attribute	0.9813	60.2600	121.4135	130.5200
CO_2	Antecedent attribute	0.9606	520.1489	1010.0413	1174.0853
CI	Consequent attribute	-	2072.5600	9805.9879	28353.3300

It is clear from **Table 7** and **Table 8** that the value of basic parameters is different after using parameter learning based on historical environmental data, *i.e.*, the weights of four antecedent attributes all are 1.0000 at **Table 7** and 0.9698, 0.9816, 0.9813, and 0.9606, respectively, at **Table 8**. The reason of this difference is because the learning model can adjust the value of basic parameters using historical data to further improve the EBRB modeling. Afterwards, the environmental data from 2005 to 2017 of Hunan province are used to generate the belief distributions and rule weights of extended belief rules according to ***Step 2*** and ***Step 3*** shown in ***Section 4.2***, the corresponding result can be found in **Table 9**. It is worth noting that each rule shown in **Table 9** is explainable for environment manager because of the high explainability of the EBRB modeling. Taking R_1 for an example, when 100% sure that GDP is *High*, 8.1% sure that TP is *Middle* and 91.9% is *High*, 100% sure that COD is *Low*, and 100% sure that CO_2 is *High*, then 100% sure that CI is *High*. Meanwhile, the relative importance of R_1 is lower than other rules because its rule weight is smaller than that of other rules.

Table 9. Extended belief rules for the CI prediction of Hunan

R_k	Rule weight	GDP			TP			COD			CO_2			CI		
		Low	*Middle*	*High*	*Low*	*Middle*	*High*	*Low*	*Middle*	*High*	*Low*	*Middle*	*High*	*Low*	*Middle*	*High*
R_1	0.710	0.000	0.000	1.000	0.000	0.081	0.919	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000
R_2	0.994	0.000	0.143	0.857	0.001	0.999	0.000	0.011	0.989	0.000	0.000	0.230	0.770	0.000	0.178	0.822
R_3	0.994	0.000	0.243	0.757	0.074	0.926	0.000	0.000	0.837	0.163	0.000	0.310	0.690	0.000	0.383	0.617
R_4	0.869	0.000	0.374	0.626	0.000	0.000	1.000	0.000	0.617	0.383	0.000	0.347	0.653	0.000	0.567	0.433
R_5	0.982	0.000	0.507	0.493	0.012	0.988	0.000	0.000	0.459	0.541	0.000	0.714	0.286	0.000	0.746	0.254
R_6	0.862	0.000	0.641	0.359	0.000	0.905	0.095	0.000	0.000	1.000	0.000	0.968	0.032	0.000	0.888	0.112
R_7	0.927	0.000	0.836	0.164	0.217	0.783	0.000	0.680	0.320	0.000	0.184	0.816	0.000	0.018	0.982	0.000
R_8	0.986	0.000	0.997	0.003	0.635	0.365	0.000	0.599	0.401	0.000	0.442	0.558	0.000	0.272	0.728	0.000
R_9	0.984	0.197	0.803	0.000	0.693	0.307	0.000	0.538	0.462	0.000	0.612	0.388	0.000	0.552	0.448	0.000
R_{10}	0.985	0.484	0.516	0.000	0.799	0.201	0.000	0.507	0.493	0.000	0.635	0.365	0.000	0.731	0.269	0.000
R_{11}	0.985	0.722	0.278	0.000	0.929	0.071	0.000	0.476	0.524	0.000	0.830	0.170	0.000	0.857	0.143	0.000
R_{12}	0.985	0.870	0.130	0.000	0.979	0.021	0.000	0.522	0.478	0.000	0.981	0.019	0.000	0.928	0.072	0.000
R_{13}	0.737	1.000	0.000	0.000	1.000	0.000	0.000	0.596	0.404	0.000	1.000	0.000	0.000	1.000	0.000	0.000

5.2.3. The 3rd sub-process: model application using the EBRB prediction

Continuing with the CI prediction of Hunan province, the environmental data in 2018 are used as testing data to predict its investment using EBRB prediction, in which the data are x(GDP)=33902.96, x(TP)=2093.98, x(COD)=57.58, and $x(CO_2)$ =1241.49, respectively. According to ***Step 1*** and ***Step 2*** shown in ***Section 3***, the activation weights w_k (k=1,…, 13) of all extended belief rules in **Table 9** can be calculated and then all these activation weights together with the belief distributions of 13 extended belief rules are used to generate the integrated belief degrees β_n (n=1,…, 3). The corresponding activation weights and belief degrees are shown in **Table 10**, in which the activation weights reflect the relative importance of each rule on predicting Hunan's CI and the belief degrees show the distributed assessment of Hunan's CI prediction when inputs are e x(GDP)=33902.96, x(TP)=2093.98, x(COD)=57.58, and $x(CO_2)$ =1241.49. From **Table 10**, the predicted CI of Hunan province in 2018 is $f(\boldsymbol{x})$ = 0.00011 × 2072.5600 + 0.003068 × 9805.9879 + 0.996822 ×28353.3300 = 28293.5355.

Table 10. Activation weights and belief degrees for the CI prediction of Hunan

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	β_n
Belief degree	*Low*	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.272	0.552	0.731	0.857	0.928	1.000	1.10E-4
	Middle	0.000	0.178	0.383	0.567	0.746	0.888	0.982	0.728	0.448	0.269	0.143	0.072	0.000	3.07E-3
	High	1.000	0.822	0.617	0.433	0.254	0.112	0.000	0.000	0.000	0.000	0.000	0.000	0.000	9.97E-1
	w_k	9.30E-1	5.00E-6	1.84E-3	6.49E-2	1.58E-4	0.00E0	1.10E-3	4.50E-5	9.60E-4	1.06E-3	1.90E-4	5.00E-6	0.00E0	-

Similarly, through the same processes shown in ***Section 5.2.1*** to ***Section 5.2.2*** to construct the IES-EBRB model for three kinds of environmental investments, namely EC, CI, and LI, using three kinds of ranking combination functions, the predicted investments of 31 provinces can be obtained and they are shown in **Fig. 3** to **Fig. 5**. From **Fig. 3**, the predicted ECs of three IES-EBRB models with MIN-RC, AVG-RC, and GAVG-RC functions are very close. From the prediction accuracy of different regions, it can be found that the predicted ECs of Tianjin, Jilin, Heilongjiang, Shanghai, Shandong, Hainan, Xizang, and Qinghai provinces are much lower compared to the actual ECs, while the predicted ECs of Hebei, Shanxi, Neimenggu, Zhejiang, and Guangdong are much higher than other provinces. The main reason is the difference of pollution emission and energy consumption investment in different provinces, which will lead to the difference of prediction accuracy under the same environmental indicators.

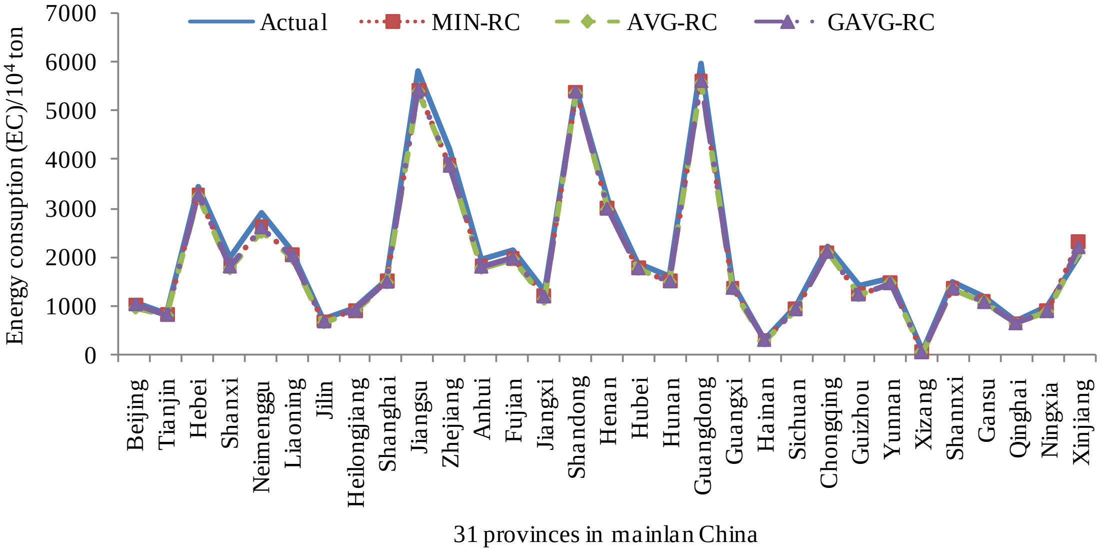

Fig. 3. Predicted EC for 31 provinces in 2018 under three ranking combination functions

The predicted CIs in **Fig. 4** show that the prediction accuracy of the IES-EBRB models is slight difference. The predicted CIs in Hebei are much higher than those in other provinces, while the predicted CIs by MIN-RC in Shannxi are higher than those predicted by AVG-RC and GAVG-RC. The prediction results of regional difference show that Beijing, Heilongjiang, Hainan, Xizang, Qinghai, and Ningxia are much closer to actual CIs and the predicted CIs in Guangdong, Shanxi, and Shandong are much higher than other provinces, indicating that the regional policy, technology development, and economic development of different provinces are also important factors affecting the accuracy of investment prediction.

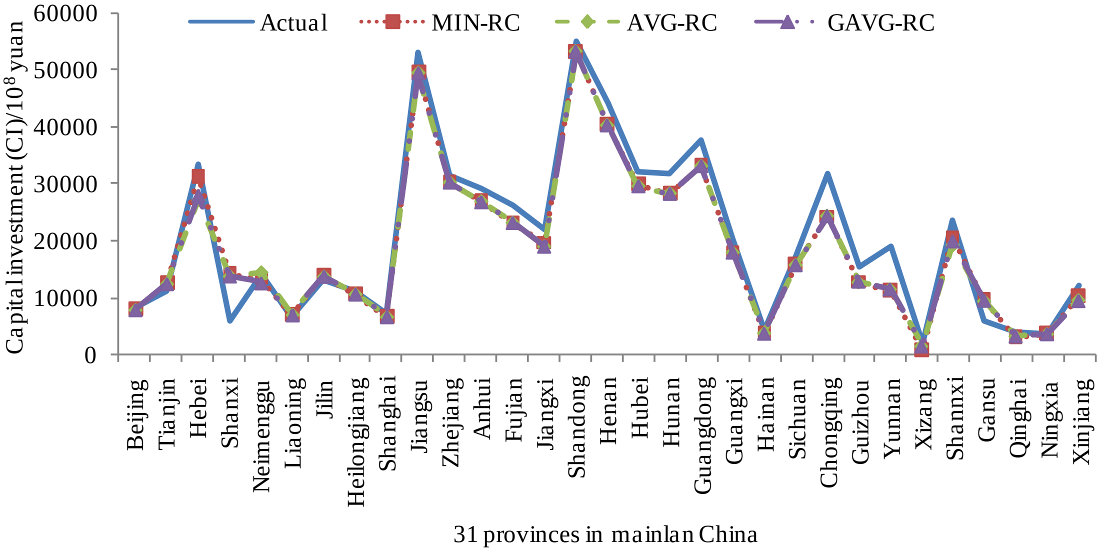

Fig. 4 Predicted CI for 31 provinces in 2018 under three ranking combination functions

Fig. 5 also shows that the predicted LIs of three ranking combination functions are slight difference. Additionally, the predicted LIs of most western regions are much closer to actual LIs compared to the provinces in eastern provinces, and the predicted LIs by AVG-RC and GAVG-RC are lower than MIN-RC. From the view of regional differences, the predicted LIs in Beijing, Hebei, and Liaoning are much higher than other provinces, indicating that the regional factors significantly affect the prediction accuracy in those provinces compared to other provinces under the same indicators.

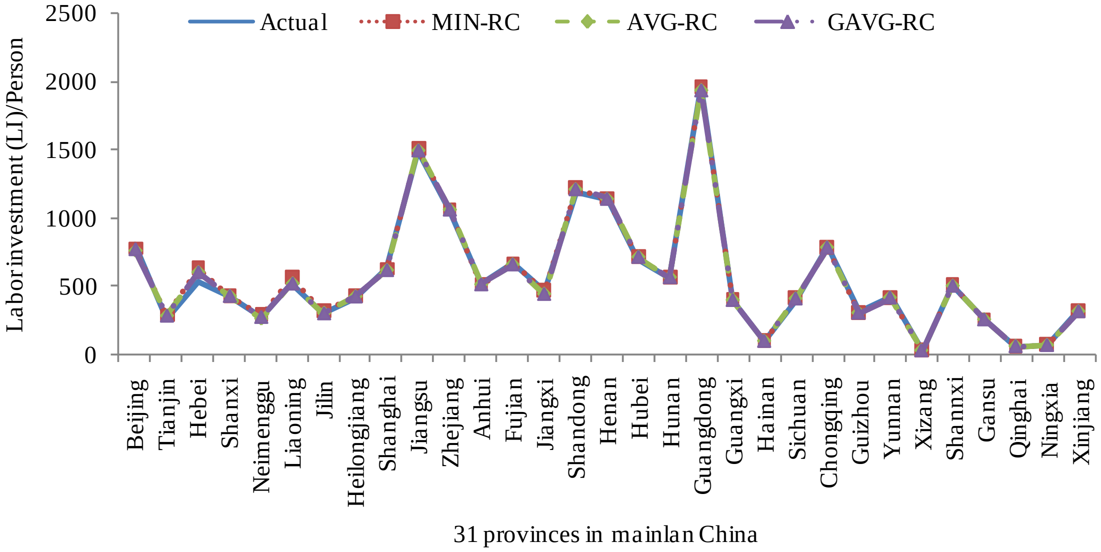

Fig. 5 Predicted LI for 31 provinces in 2018 under three ranking combination functions

5.3. Comparative analysis of existing investment prediction models

In order to demonstrate the performance of the IES-EBRB model, three different comparative analyses are carried out in the terms of three commonly used evaluation criteria, namely mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (R). It is worth noting that the larger R and the smaller MAPE and MAE are considered to be a better performance for an environmental investment prediction model.

For the first comparative analysis, the main target is to compare the performance of the IES-EBRB model under MIN-RC, AVG-RC, and GAVG-RC functions. **Table 11** shows the comparison of prediction accuracies based on MAE, MAPE, and R for the three kinds of IES-EBRB models. From **Table 11**, the IES-EBRB model with MIN-RC produces the best MAE compared with the other two models in CI prediction and the best MAE of EC and LI prediction is obtained from the IES-EBRB model with GAVG-RC, *e.g.*, 2505 (MIN-RC for CI), 125 (GAVG-RC for EC), and 11 (GAVG-RC for LI). Meanwhile, the comparison of MAPEs and Rs shows that the prediction results by AVG-RC and GAVG-RC are better than MIN-RC, *i.e.*, the best MAPE of EC and CI are obtained by AVG-RC and the best R of EC, CI, and LI by AVG-RC.

Table 11. Comparison of investment prediction based on three ranking combination functions

	Investment	MIN-RC	AVG-RC	GAVG-RC
MAE	EC	129	132	**125**
	CI	**2505**	2680	2670
	LI	13	12	**11**
MAPE	EC	7.70%	**6.98%**	7.52%
	CI	18.04%	**17.10%**	17.23%
	LI	2.53%	2.27%	**2.10%**
R	EC	0.9976	**0.9983**	**0.9983**
	CI	0.9828	0.9821	**0.9831**
	LI	0.9986	0.9990	**0.9991**

For the second comparative analysis, the main target is to discuss the influence of the time-series nature of the yearly environmental investments on the modeling of the IES-EBRB model under MIN-RC, AVG-RC, and GAVG-RC functions. Thus, in the proposed modeling process of the IES-EBRB model shown in ***Section 4.2***, the environmental investment in the previous year is regarded as an antecedent attribute, *i.e.*, for the investment prediction shown in ***Section 5.2***, which uses four indicators GDP, TP, COD, and CO_2 as antecedent attributes to predict Hunan's CI at 2018, the new modeling process not only considers GDP, TP, COD, and CO_2 as antecedent attributes, but also takes into account Hunan's CI at 2017 as a new antecedent attribute. In order to compare with the difference of the IES-EBRB models constructed by different modeling, the prediction error ratio is used as an evaluation criterion, whose smaller value is considered to be a more similar accuracy. **Figs. 6** to **8** show the comparison of the prediction error ratio of the IES-EBRB models constructed by different modeling for 31 provinces' EC, CI, and LI prediction under MIN-RC, AVG-RC, and GAVG-RC functions. It is clear from **Figs. 6** to **8** that the prediction error ratio for the IES-EBRB models with or without the consideration of the time-series nature of the yearly environmental investment is almost equal to 0% for three investments under three functions, which means that the time-series nature of the yearly environmental investment has a few influence on the modelling of the IES-EBRB model.

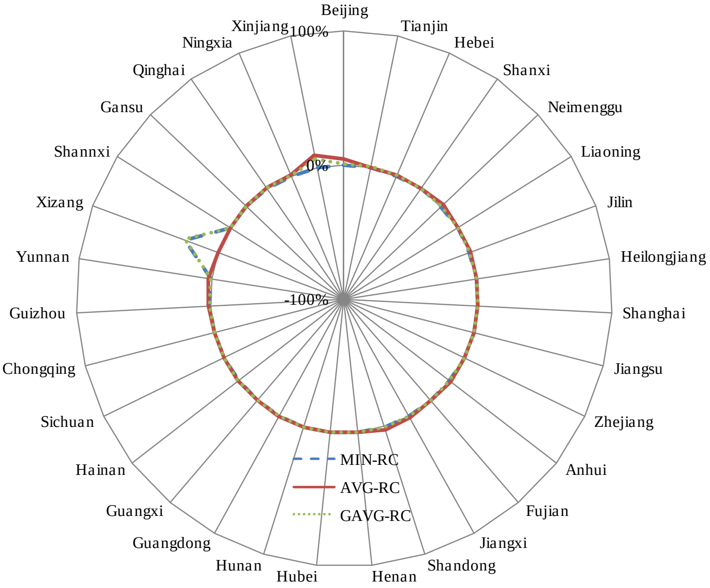

Fig. 6. EC's prediction error ratio for 31 provinces in 2018 under three ranking combination functions

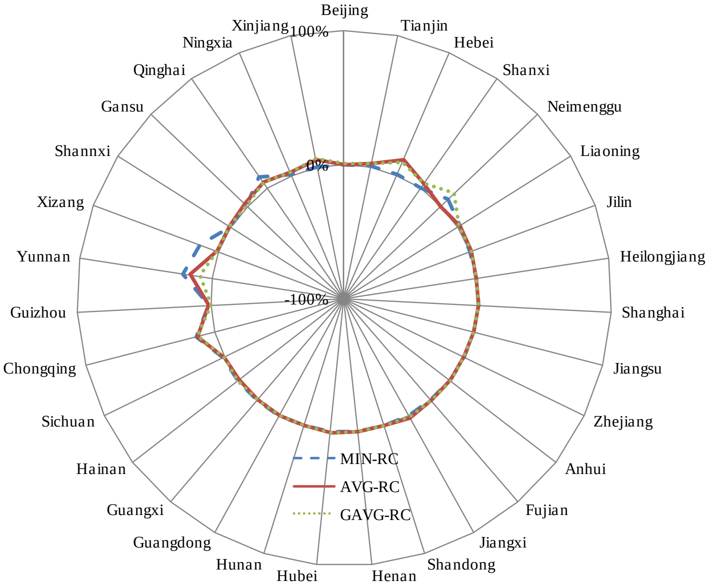

Fig. 7. CI's prediction error ratio for for 31 provinces in 2018 under three ranking combination functions

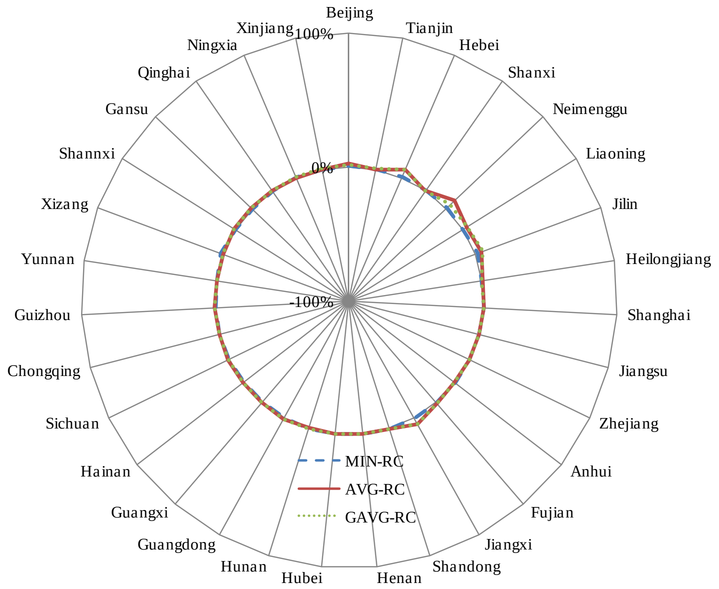

Fig. 8. LI's prediction error ratio for for 31 provinces in 2018 under three ranking combination functions

For the third comparative analysis, the main target is to compare the best IES-EBRB model with existing investment prediction models, including ARIMA-based model [14], GM-based model [6], FRBS-based model [25], ANFIS-based model [26], and EBRB-based model [16], in which the review of these studies can be found in ***Section 2.2***. **Table 12** shows the comparison of different models in investment prediction based on MAE, MAPE, and R. From **Table 12**, the IES-EBRB model produces satisfactory prediction results for three investments compared with existing models, and the MAEs of IES-EBRB model are 129, 2505, and 10, respectively. Comparatively, the EC, CI and LI prediction using the EBRB-based model are much close to those using the IES-EBRB model comparing with other existing models, which reveal that the EBRB model has a better performance than the other decision-making methodologies currently used in environmental investment prediction, and the proposed IES procedure can further improve the performance of the EBRB model. This is because the ARIMA-based and GM-based models belong to the time series forecasting-based model, which inevitably ignore the hidden logic relationship between environmental indicators and investments. Although the FRBS-based and ANFIS-based models take into account the hidden logic relationship to improve the accuracy of predicting environmental investments, the rule representation in these models is based on the IF-THEN rule with singleton fuzzy label, which fail to consider distributed assessment so they are still inferior to the EBRB model.

Table 12 Comparison of different investment prediction models for three investments

	Investment	ARIMA	GM	ANFIS	FRBS	EBRB	IES-EBRB
MAE	EC	1361	666	235	347	174	125
	CI	36182	19613	4119	4225	2742	2505
	LI	304	651	37	195	13	11
MAPE	EC	71.23%	36.32%	16.06%	12.97%	9.17%	6.98%
	CI	202.51%	160.36%	32.54%	30.55%	18.38%	17.10%
	LI	42.06%	67.18%	9.12%	33.90%	2.69%	2.10%
R	EC	0.9644	0.9213	0.9625	0.8950	0.9920	0.9983
	CI	0.9098	0.5314	0.9085	0.9272	0.9733	0.9831
	LI	0.8945	0.8404	0.9968	0.8538	0.9987	0.9991

All in all, the above three comparative analyses of different investment prediction models indicates that the IES-EBRB model is ability to accurately predict environmental investments and has a higher accuracy than existing models. Meanwhile, MIN-RC can improve the prediction accuracy of IES-EBRB model in term of MAE, while AVG-RC and GAVG-RC are the better choice to improve the IES-EBRB model in term of MAPE and R.

6. Conclusions

In this study, a new model called IES-EBRB was proposed for environmental investment prediction. The components of the IES-EBRB model combine the IES procedure and EBRB modeling, where the former process can select representative environmental indicators based on various indicator selection methods, and the latter provides a white-box modeling mechanism for constructing prediction models using experts' knowledge and historical data. The main conclusions of this study can be further summarized as three aspects below:

(1) By focusing on the information loss of representative indicators in environmental indicator selection, different kinds of indicator selection methods are used together to obtain the individual ranking of indicators, followed by three kinds of ranking combination functions to obtain their integrated rankings. This allows better representative indicators to be selected on the basis of the advantages of different indicator selection methods.

(2) By aiming at the human involvement and sufficient explainability in environmental investment prediction modeling, a white-box designed EBRB model is introduced to for investment prediction based on the selected representative indicators, experts' knowledge, and historical data. Owing to advantages of the EBRB model, the proposed investment prediction model has the ability of using experts' knowledge to enhance data analytics for explainable decision making.

(3) On the basis of the proposed IES-EBRB model, the real environmental indicators and data of Chinese 31 provinces were collected to perform an empirical case study under the involvement of experts' knowledge. The results of the case study not only provided a detailed process of developing an IES-EBRB model, but also revealed that the IES-EBRB model had a more powerful ability in predicting investment compared to other models in previous studies.

In future, owing to the fact that the black-box design of investment prediction models has led to resurgence in interest in the explainability of decision-making methodology, the EBRB model can be regarded as an effective tool to be used in more fields of environment managements. Moreover, modeling assessment criteria should be considered to further determine the optimal number of representative indicators for environmental investment prediction.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Nos. 61773123, 71701050, 71801050, and 71872047), the Humanities and Social Science Foundation of the Ministry of Education of China (Nos. 20YJC630188, 19YJC630022, and 20YJC630229), and the Social Science Foundation of Fujian Province, China (No. FJ2019C032).

Reference:

[1] Adilova N. E., 2019. Consistency of Fuzzy If-Then rules for Control System, 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions, 137-142.

[2] Bolón-Canedo V., Alonso-Betanzos A., 2019. Ensembles for feature selection: A review and future trends, *Information Fusion*, 52: 1-12.

[3] Calzada A., Liu J., Wang H., Nugent C., Martínez L., 2015. Application of a Spatial Intelligent Decision System on Self-Rated Health Status Estimation. *Journal of Medical Systems*, 39, 1-18.

[4] Calzada A., Liu J., Wang H., Kashyap A., 2015. A New Dynamic Rule Activation Method for Extended Belief Rule-Based Systems, *IEEE Transactions on Knowledge and Data Engineering*, 27(4): 880-894.

[5] Cao H. J., Fujii H., Managi S., 2015. A productivity analysis considering environmental pollution and diseases in China. *Journal of Economic Structures*, 4(6):11-25.

[6] Chen L., Wang Y. M., Lai F.J., Feng F., 2017. An investment analysis for China's sustainable development based on inverse data envelopment analysis. *Journal of Cleaner Production*, 142, 1638-1649.

[7] Espinilla M., Medina J., Calzada A., Liu J., Martínez L., Nugent C., 2017. Optimizing the configuration of an heterogeneous architecture of sensors for activity recognition, using the extended belief rule-based inference methodology, *Microprocess and Microsystems*, 52: 381-390.

[8] Frank E., Hall M. A., Witten I. H., 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

[9] Lee M., Zhang N., 2012. Technical efficiency, shadow price of carbon dioxide emissions, and substitutability for energy in the Chinese manufacturing industries. *Energy Economics*, 34 (5), 1492–1497.

[10] Liu J., Martínez L., Calzada A., Wang H., 2013. A novel belief rule base representation, generation and its inference methodology, *Knowledge-Based Systems*, 53, 129-141.

[11] Jin Y. C., Seelen W. V., Sendhoff B., 1999. On Generating FC3 Fuzzy Rule System from Data Using Evolution Strategies, *IEEE Transactions on Systems Man Cybernetics-Part B: Cybernetics*, 29(6): 829-845.

[12] Kaytez, F., 2020. A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption, *Energy*, 197: 1-12.

[13] Song M. L., Peng J., Wang J. L., Dong L., 2018. Better resource management: An improved resource and environmental efficiency evaluation approach that considers undesirable outputs. *Resources, Conservation and Recycling*, 128, 197-205.

[14] Valipour M., Banihabib M. E., Behbahani S. M. R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology*, 476, 433-441.

[15] Wang Y. M., Luo Y., 2010. Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making, Mathematical and Computer Modelling, 51(1–2): 1-12.

[16] Wang Y. M., Ye F. F., Yang L. H., 2020. Extended belief rule based system with joint learning for environmental governance cost prediction. *Ecological Indicators*, 111, 1-14.

[17] Wang Y. M., Yang J. B., Xu D. L., 2006. Environmental impact assessment using the evidential reasoning approach. *European Journal of Operational Research*, 174, 1885–1913.

[18] Willett P., 2013. Combination of Similarity Rankings Using Data Fusion, *Journal of Chemical Information and Modeling*, 53(1): 1-10.

[19] Wu J., Li M. J., Zhu Q. Y., Zhou Z. X., Liang L., 2019. Energy and environmental efficiency measurement of China's industrial sectors: A DEA model with non-homogeneous inputs and outputs. *Energy Economics*, 78, 468-480.

[20] Xu N., Dang Y. G., Gong Y. D., 2017. Novel grey prediction model with nonlinear optimized time response method for forecasting of electricity consumption in China. *Energy*, 118, 473-480.

[21] Yang J. B., Liu J., Wang J., Sii H. S., Wang H. W., 2006. Belief rule-base inference methodology using the evidential reasoning approach - RIMER, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 36(2): 266-285.

[22] Yang L. H., Wang Y. M., Lan Y. X., Chen L., Fu Y. G., 2017. A data envelopment analysis (DEA)-based method for rule reduction in extended belief-rule-based systems, *Knowledge-Based Systems*, 123: 174-187.

[23] Yang L. H., Wang Y. M., Fu Y. G., 2018. A consistency analysis-based rule activation method for extended belief-rule-based systems. *Information Sciences*, 445, 50-65.

[24] Yang L. H., Liu J., Wang Y. M., Martínez L., 2019. New activation weight calculation and parameter optimization for extended belief rule-based system based on sensitivity analysis, *Knowledge and Information Systems*, 60(2): 837-878.

[25] Ye F. F., Yang L. H., Wang Y. M., 2019. Fuzzy rule based system with feature extraction for environmental governance cost prediction. J*ournal of Intelligent & Fuzzy Systems*, 37, 2337-2349.

[26] Ye F. F., Yang L. H., Wang Y. M., 2020. A cost forecast method of environmental governance based on the input-output relationship and efficiency. *Control and Decision*, 35(4): 993-1003.

[27] Ye F. F., Yang L. H., Wang Y. M., Chen L., 2020. An environmental pollution management method based on extended belief rule base and data envelopment analysis under interval uncertainty, *Computers & Industrial Engineering*, 144: 1-15.

[28] Zhang A., Gao F., Yang M., Bi W. H., 2020. A new rule reduction and training method for extended belief rule base based on DBSCAN algorithm, *International Journal of Approximate Reasoning*, 119: 20-39.

[29] Zheng K., Wang X., 2018. Feature selection method with joint maximal information entropy between feature and class Pattern Recognition, 77: 20-29.

[30] Zhu H. Z., Xiao M. Q., Yang L. H., Tang X. L., Liang Y. J., Li J. F., 2020. A minimum centre distance rule activation method for extended belief rule-based classification systems, *Applied Soft Computing*, 91: 1-14.

Manuscript File

Declaration of interests

☑The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Credit author statement-R1:

Fei-Fei Ye: Writing–Original Draft, Conceptualization, Methodology, Data Curation, Formal Analysis.

Suhui Wang: Formal analysis, Writing - review & editing, Supervision

Peter Nicholl: Writing - review & editing, Supervision

Long-Hao Yang: Investigation, Writing - review & editing, Writing - review & editing, Supervision.

Ying-Ming Wang: Writing - review & editing, Supervision, Conceptualization, Investigation, Validation.