Knowledge-Based Systems 282 (2023) 111112

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Relative Synergy Coefficient: A novel way to detect variable interaction in large dataset☆

Yanrui Li [a], Kaiyou Fu [b], Yuchen Zhao [a], Chunjie Yang [a,*]

[a] College of Control Science and Engineering, State Key Laboratory of Industrial Control Technology, Zhejiang University, Zheda Road 38, Hangzhou, China
[b] School of Medicine, Zhejiang University, Hangzhou, China

ARTICLE INFO

Dataset link: https://datatopics.worldbank.org/world-development-indicators/, https://portal.gdc.cancer.gov/

Keywords:
Feature interaction
Knowledge discovery
World development indicators
Gene network

ABSTRACT

Feature interaction, also referred to as feature synergy, denotes the phenomenon wherein interactive features collectively convey more information than their individual contributions, thereby holding paramount significance in the realms of feature engineering and data mining. Many prevailing techniques designed to detect these interactions primarily rely on model-based methods to compute absolute synergy. However, such approaches often prove ill-suited for extensive datasets and overlook variables with comparatively minor primary effects. In response, we introduce a groundbreaking metric known as the Relative Synergy Coefficient (RSC). This novel metric facilitates swift identification and quantification of relative synergy's potency within large datasets. The proposed indicator is a non-parametric metric based on information entropy, and its computation involves the utilization of discretization and normalization techniques. The generality, equitability and robustness of metric is proved on the simulated data. Besides, the indicator is proved to be effective and can cross-validate with domain knowledge on two real world datasets.

1. Introduction

Feature interaction, also called feature synergy, refers to the fact that interactive features could provide more information when combined together than the sum of information provided individually [1] which plays an important role in data-driven modeling. Aristotle's predicate "The whole is greater than the sum of its parts" applies in the presence of interactions. We can always find the synergy relationship between features in reality, for example, water and oxygen are both essential for seed germination. Also, we provide a medical example in Fig. 1, which depicts the drug filtration rate with the pH of the urine. In this example, the properties of drugs and the pH of the urine interactively determine the drug filtration rate. Drugs that act as weak bases are increasingly excreted as the pH of the urine becomes more acidic, and the inverse is true for weak acids. In fact, detecting interactive relationships between variables is helpful for model construction, model explanation and domain knowledge extraction.

In the realm of model construction, the detection of feature interactions constitutes a vital facet of the feature engineering process. Identifying interactive variables and subsequently incorporating them into the selected features can effectively enhance the precision of the model [2–4].

In model explanation aspect, detecting feature synergy is beneficial to explain the model. The interpretation of the model is especially critical in fields that require high model security and accuracy or require interpretable evidence to support decision making like medical, finance and so on [5,6]. Previous approaches mostly focus on the main effects of variables, finding variables that are significant for the results through correlation indicators or model-based interpretable methods [7]. The feature interaction will be a great supplement to explain the model.

For domain knowledge extraction, the detected synergistic relationship can often be abstracted as 'if-else' rules, which can be subjected to in-depth analysis by experts to uncover their underlying causal factors [8,9].

Because of the importance of the synergy, a few methods have been proposed to calculate it, which will be introduced in the Section 2. In summary, existing methods have two shortcomings. First, all of the methods calculate the absolute value of synergy. However, the inability to distinguish between synergistic and main effects causes these methods to overestimate the variables with large main effects while neglecting the other variables, which is detrimental to both knowledge discovery and feature engineering. Second, most of the methods evaluate the information gain through a model. Testing every pair of variables through the model is time-consuming, and the result has a

☆ This work was supported by National Natural Science Foundation of China (61933015).
* Corresponding author.
E-mail addresses: liyanrui@zju.edu.cn (Y. Li), 11818434@zju.edu.cn (K. Fu), 3180101495@zju.edu.cn (Y. Zhao), cjyang999@zju.edu.cn (C. Yang).

https://doi.org/10.1016/j.knosys.2023.111112
Received 12 June 2022; Received in revised form 11 August 2023; Accepted 20 October 2023
Available online 31 October 2023

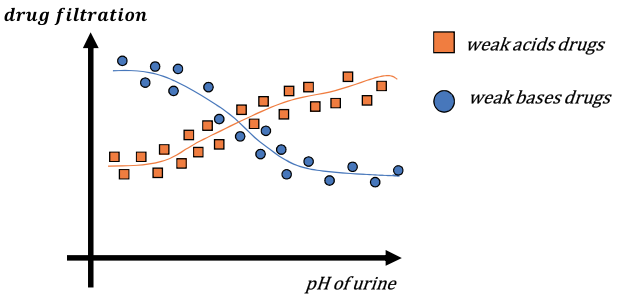


Fig. 1. An intuitive medical example of feature interaction. It describes the relationship between drug filtration and pH of urine.

strong correlation with model fitting, which restricts their application to complex large datasets.

In addressing these challenges, this paper introduces an innovative direct measurement approach for swiftly detecting relative synergy within extensive datasets. The contributions of this study encompass the following:

(1) We design a novel metric to detect relative synergy in large datasets. This indicator measures the ratio of additive information gain brought by one variable in a synergy pair to the amount of information this variable can provide. This design helps to find those synergistic relationships that are neglected because of small main effects.

(2) We present a new method to directly calculate the information gain rather than based on model. The calculation of information gain is difficult due to the difficulty of computing the probability density function of continuous variables. The proposed method calculates it by discretizing the data and then normalizing it. The experiment results on real-world data indicates the proposed method can well fit theoretical values.

(3) We verify the generality, equitability and robustness of this metric on simulated data. Furthermore, on two real-world dataset (World development indicators data and gene data), the indicator proved to be effective and the synergy relationships detected by it can cross-validate with domain knowledge.

The rest of this paper is organized as follows: Section 2 reviews some related works. Section 3 introduces the RSC and concludes its characteristics from simulation. In Sections 4 and 5, several experiments are conducted on real-world datasets. In Section 6, the future work of this method is discussed and finally the conclusion is made in Section 7.

2. Related work

We classify feature interaction detection methods into three distinct categories: direct methods, specific model-based methods, and unspecified model-based methods. In direct methods, feature interactions are directly discerned from the data itself. Specific model-based methods entail the automated construction of a dedicated model from the provided dataset, subsequently extracting interaction insights. As for unspecified model-based methods, both the dataset and a user-built model are prerequisites and interaction strength is computed based on the provided model.

2.1. Direct method

(1) **Two-way ANOVA** [10] involves conducting hypothesis tests for each interaction candidate by checking each hypothesis with F-statistics. However, its premise that the independent variables are categorical limits its application.

(2) **FAST** [11] detects pairwise feature interactions by building a extremely simple model using two cuts on the input space of X_i and X_j. The cuts c_i and c_j on X_i and X_j are parallel to the axes, and the predict value of each part is the mean value of y. The strength of interaction is measured by searching all possible (c_i, c_j) and picking the lowest RSS(residual sum of squares)

$$RSS = \sum_{k=1}^{n}(y_k - T_{ij}(x_k)) \tag{1}$$

where T_{ij} is the predictor. The method is fast neglects variables with small main effect.

2.2. Specific model-based methods

(1) **Groves** [12] detects interactions between features x_i and x_j by comparing unrestricted model $g(x)$ and a restricted model $R_{ij}(x)$. Both $g(x)$ and $R_{ij}(x)$ are trained on a given training dataset by using additive groves of trees, while $R_{ij}(x)$ is prevented from modeling an interaction (x_i, x_j). The strength of the interaction I_{ij} is measured by

$$I_{ij} = [RSME(g) - RSME(R_{ij})]/std(Y) \tag{2}$$

where $RMSE(g)$ denotes the root mean squared error of model g and $std(y)$ denotes the standard deviation of the target values. Although Groves is accurate in practice, building restricted and unrestricted models are computationally expensive and therefore this method is almost infeasible for large high dimensional datasets.

(2) **GUIDE** [13] detects pairwise interactions based on χ^2 tests. GUIDE divides the (X_i, X_j) space into four quadrants by splitting the range of each feature into two halves at the sample median, and an addictive model g on the training data without any feature interaction is fit. Then, GUIDE constructs a 2×4 contingency table using the residual signs as rows and the quadrants as columns. The cell values in the table are the number of "+ "s and "−"s in each quadrant. These counts permit the computation of a p-value to measure the interaction of a pair.

2.3. Unspecified model-based methods

(1) **Model errors-based(MEB) method** [14] detects the feature interactions from model f by comparing the prediction errors of model f on the original data and on permuted data. The strength of the interaction is measured by

$$I_{ij} = E_{ij} - E_i - E_j + E_{\emptyset} \tag{3}$$

where $E, E_i and E_{ij}$ denotes the prediction error of model f on the original data, the data generated from the original by permuting X_i, and by permuting both X_i and X_j, respectively. The RMSE and misclassification error rate are used to evaluate the prediction error for the regressor and classifier, respectively.

(2) **Partial dependence plot** [15] shows the marginal effect a pair of features have on the predicted outcome of a machine learning model. If the contour lines are parallel, the feature interaction exists. However, it require the variable independent to each other, and is often used as a visualization but not a detection method.

(3) **ANOVA decomposition** [16] measures the strength of the feature interaction $X_S(|S| \geq 2)$ via

$$\hat{\sigma^2_u} = \frac{1}{2n}\sum_{t=1}^{n}\left(f(x^{(t)}) - \sum_{k=0}^{|U|}(-1)^{|U|-k}\sum_{v\in\bigcap_k U} f(x_v^{(t)}, x_{\sim v}^{r(t)})\right)^2 \tag{4}$$

where $U = \{\{1,\dots,m\}\backslash\{j\}\}_{\{j\in S\}}$, $\bigcap_k U$ denotes the collection of k-way intersections among the elements of U, and $r(t)$ is an integer that is randomly chosen in $\{1,\dots,n\}$.

(4) **Feature construction method** [17] transfers a given set of input features to generate more powerful one. The construction process is often heuristic, and the generated features are evaluated to be effective if the accuracy of prediction improved a lot. This method can not only find the interaction but also provide explanations through the formulas of constructed features. However, the operator used to generate new features must base on domain knowledge, and the construction process is also time-consuming.

3. Our method

In this section, we describe an exploratory indicator, Relative Synergy Coefficient(RSC) to detect and score the synergistic effect between two variables and analyze its characteristics through data simulation.

3.1. Definition

First, we will briefly introduce the origin of the idea of RSC. To start with, we suppose two variables X_1, X_2 act in concert to affect an outcome Y. The information provided by X_1 and X_2 to predict Y can be represented in the following equation.

$$H(Y \parallel X_1, X_2) = H(Y \parallel X_1) + H(Y \parallel X_2) + H(Y \parallel X_1 \bigcap X_2) \tag{5}$$

where $H(Y \parallel X_1, X_2)$ refer to the information that is provided by X_1 and X_2 to predict Y. $H(Y \parallel X_1)$ and $H(Y \parallel X_1)$ are the information provided by X_1 or X_2 separately and the $H(Y \parallel X_1 \bigcap X_2)$ is the synergy.

To get the synergy value, a simple idea is to calculate it by the following equation.

$$H(Y \parallel X_1 \bigcap X_2) = H(Y \parallel X_1, X_2) - H(Y \parallel X_1) - H(Y \parallel X_2) \tag{6}$$

However, this calculation have two shortcomings. First, it is impossible to get the theoretical value of $H(Y \parallel X_1, X_2)$ and $H(Y \parallel X_1)$, $H(Y \parallel X_2)$ of continuous variables. The existing methods are estimates of the theoretical values or using model accuracy as substitution. Because of that, the values of $H(Y \parallel X_1, X_2)$ and $H(Y \parallel X_1)$, $H(Y \parallel X_2)$ may have different value range, and thus are not comparable. Second, it is actually difficult to divide synergy and main effect. When evaluating synergy, a part of the main effects are often included, which leads to biased results for variables with small main effects. To address these problems, we use the following equation to calculate the relative synergy

$$RSC = \frac{infogain(X_1, X_2)}{H(Y \parallel X_2)} = \frac{H(Y \parallel X_2) + H(Y \parallel X_1 \bigcap X_2)}{H(Y \parallel X_2)} \tag{7}$$

where, $InfoGain(X_1, X_2)$ means the additional information brought by X_2. Compared to the calculation of absolute synergy, this definition uses division rather than subtraction to calculate the relative synergy. Since the information gain includes main and synergistic effects, it is clearer and simpler to define and calculate.

According to the above idea, we proposed the relative synergy coefficient (RSC). The simple introduction is as follow: if the variable X_1 and X_2 are interactive in predicting Y, the information gain of introducing $X_1(X_2)$ to $X_2(X_1)$ should be much greater than the information straightly provided by $X_1(X_2)$ to predict Y themselves. The mathematical expression of the idea is shown as alg 1, where I_1, I_2 are the $H(Y \parallel X_1), H(Y \parallel X_2)$ in Eq. (6), and G_1 and G_2 refer to the information gain relating to Y when add X_1 (X_2) into X_2 (X_1). Finally, RSC is the larger value of two relative information gaining which prevents the algorithm from giving bias to the informative variables, making the outcome focuses on the extra information gaining of synergetic variables.

Algorithm 1: Relative synergy coefficient(X_1, X_2, Y, b)

Data: interactive variables X_1,X_2,predict variable$Y$
Input: binning size n
Output: Relative synergy coefficient

I_1 ←the information provided by X_1 about Y;
I_2 ←the information provided by X_2 about Y;
G_1 ←InfoGain(X_2,X_1,Y,b);
G_2 ←InfoGain(X_1,X_2,Y,b);
return RSC = MAX($\frac{G_1}{I_1}, \frac{G_2}{I_2}$)

3.2. Calculation

The idea is simple. The hard part is how we calculate the main effect of a variable I and the information gain G. Different calculation methods may be used to calculate the information gain and main effect in RSC, but they have to be precise. For example, if the Pearson correlation coefficient is used to calculate the main effect, the information gain might come from the non-linear main effect provided by single variable. Such an approach might not ensure that RSC accurately evaluates the intended synergy effect, contrary to our expectations.

In our method, I is calculated by the maximal information coefficient(MIC). MIC is an indicator for estimating the value of mutual information between two variables which calculates the mutual information of continuous data by first discretizing and then normalizing, so as to solve the difficult problem that the mutual information cannot be calculated due to the uncomputable joint probability density of continuous data. MIC can find all kinds of correlations between two variables rather than linear correlations. A detailed description of the MIC can be found in [7].

As for G, we design the measuring approach as Alg. 2.

Algorithm 2: InfoGain(X_1, X_2, Y, b)

Input: Variable x_1,Variable x_2,Y,binning size n
Output: information gaining adding X_1 to X_2 to predict Y

$X = [x_1, x_2]$;
$[\{X_1, Y_1\}, ..., \{X_b, Y_b\}]$ ← Split the data into b parts according to the value of x_1;
$[\mu_1, ..., \mu_b], [\sigma_1, ..., \sigma_b], [s_1, ..., s_b]$ ← calculate the mean, variance and number of y in each subset;
for *i in b* **do**
  $[\{X_{i1}, Y_{i1}\}, ..., \{X_{ib}, Y_{ib}\}]$ ← split the subset $\{X_i, Y_i\}$ into n parts according to the value of x_2;
  $[\mu_{i1}, ..., \mu_{ib}], [\sigma_{i1}, ..., \sigma_{ib}], [s_{i1}, ..., s_{ib}]$ ← calculate the mean, variance and number of y in each subsubset;
  $g_i = \sum_{j=1}^{b} \frac{s_{ij} \times |\mu_i - \mu_{ij}|}{s_i \times \sqrt{\sigma_i \sigma_{ij}}}$
end
return $\sum_{i=1}^{b} g_i$(pure InfoGain), $\sum_{i=1}^{b} g_i / \sqrt{b}$(adjust InfoGain)

The calculation method is inspired by MIC, which uses discretization to solve the difficult of calculation of continuous variables.

In specific, for computing information gain of adding X_1 to X_2 to predict Y, the method is as follows. First, split the dataset Y into Y_i according to X_2. Then, each part Y_i is further split by X_1 into smaller parts Y_{ij}, the InfoGain will be large if the mean is divergent and the variance is shrink compared with the dataset in Y_i and Y_{ij}. We define the adjust InfoGain and pure InfoGain. The difference is in adjust InfoGain, we divide the pure InfoGain by a coefficient, $\sqrt{b}$, to make the InfoGain with different bin size comparable.

Algorithm 3: Max RSC(X_1,X_2,m))

Input: Variable x_1,Variable x_2,Y,max bin size m
Output: maxRSC

for *i=2 to m* **do**
  $rsc_i \leftarrow RSC(X_1, X_2, Y, i)$
end
return $Max(rsc(2), rsc(3), ..., rsc(m))$

Finally, we define the maxRSC, which is the max value of RSC(with adjust InfoGain) with different bin sizes b. The maxRSC can give an overall value for each variable pairs without worrying whether a specific bin size may have a great adverse effect on specific data or not.

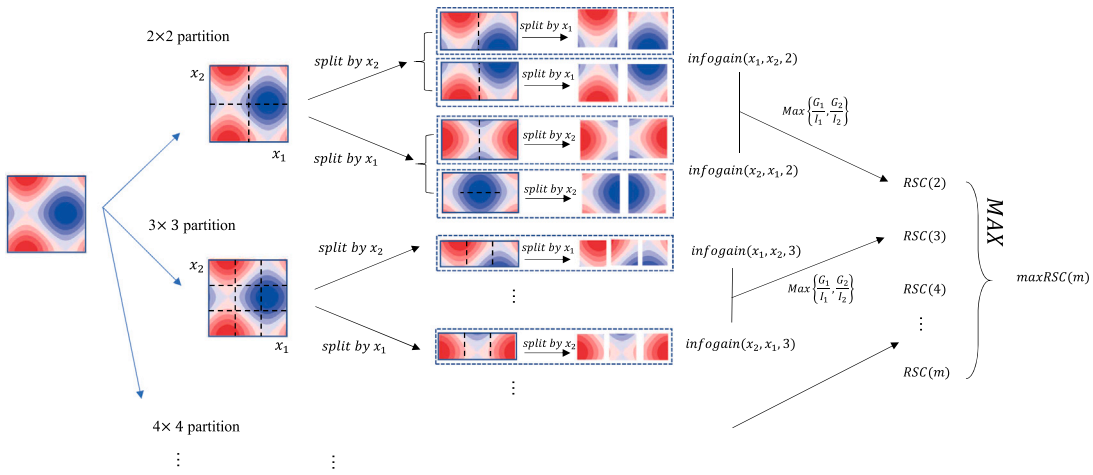


Fig. 2. Schematic diagram of calculating maxRSC. The 2-dimension data is splitted into $m \times m$ squares by splitting the range of each feature evenly. Then, the $InfoGain(x_1, x_2)$ and $InfoGain(x_2, x_1)$ are calculated respectively, leading to the RSC value. Finally, the maxRSC is picked from the RSC with different bin size.

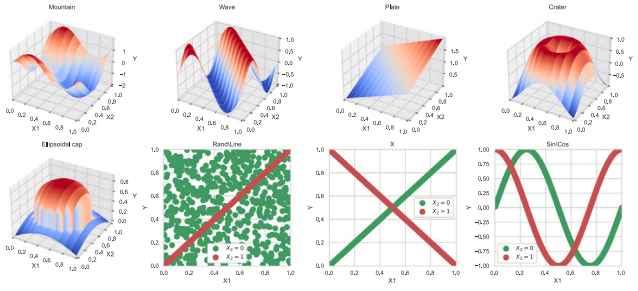


Fig. 3. Visualizations of simulation functions. The two variables of first five functions are both continuous and in the last three functions, the variable x_1 is category variable. The formula of functions are represents in appendix.

An intuitive example of the calculation process of maxRSC is illustrated in Fig. 2.

The complexity of the RSC method and the maxRSC are $O(b^2N)$ and $O(m^3N)$, where b is the bining size, N is the number of data and m is the max binning size set in maxRSC.

3.3. Simulation

The proposed RSC is evaluated on the simulated datasets. Eight different datasets are constructed as shown in Fig. 3.

Through the results shown in Fig. 4, the following paragraphs will discuss the generality, equitability and robustness of the RSC.

By generality, we mean that the RSC can capture a wide range of synergistic types between different variable pairs, not limited to specific types. Also, it can handle both continuous and categorical variables.

By equitability, we mean it in two aspects. First, the statistic should give a similar score for synergy effects of the same strength. However, the definition of the strength of synergy effect is quite ambiguous. From the simulation results, the different types of function might yield significantly RSC scores with even one order of magnitude, but we are not sure it is because the measurement is inequitable or the simulated data have different synergy in nature. However, in the experiment of real-world dataset(Section 4), where the dependent variable y is related to a large number of variables and the noise exists in both x and y, the fact that RSC among millions of pairs did not stratify reflects the equitability of RSC in some extent. Another aspect of equitability is that the RSC should give a higher score for those who have a greater impact on y, in other words, when the noise added to y rises, which means the impacts of x_1, x_2 decline, the RSC should reduce. This is proved by our simulation. When bin size is appropriate, the RSC has a significant drop with the noise rise as Fig. 4(A–E) shows.

By robustness, we mean that the measurement should not be affected by missing values or outliers. Since we use binning in the calculation, outliers will be classified into a certain region and be represented by the statistics of data in that area, so it will not have an evident impact on the results. Concerning missing values, since RSC is derived from statistical calculations involving cross-group samples, it remains unaffected as long as the pattern of missing values does not correlate with the variable y.

Besides, the RSC is also symmetrical due to the max operator in Alg. 1.

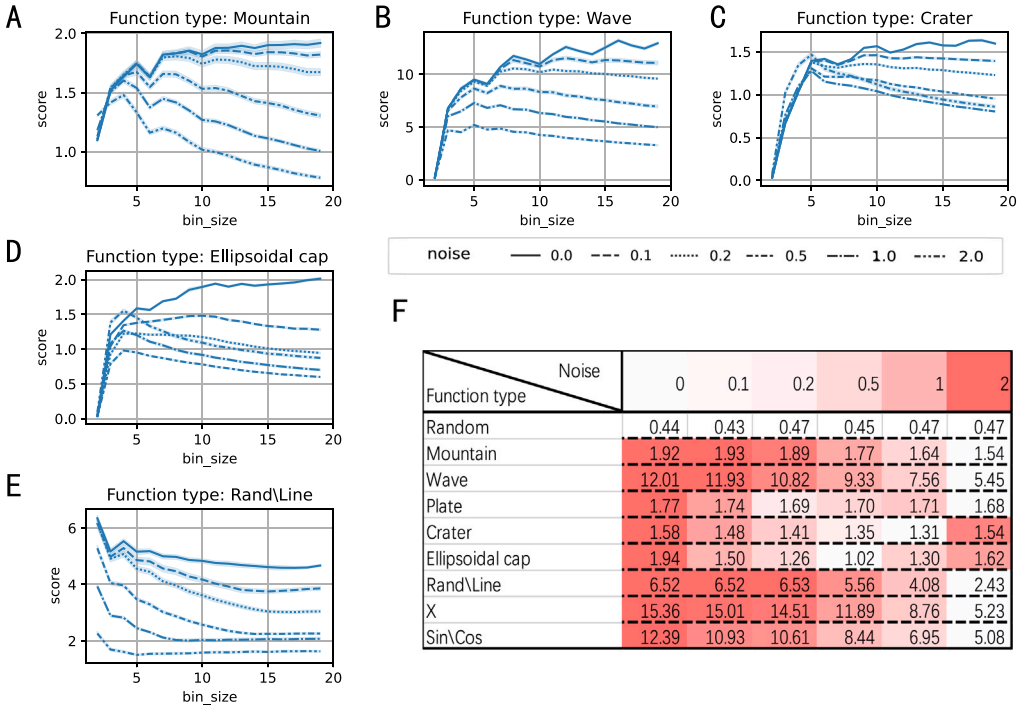
3.4. Comparisons with other methods

Although some methods have been proposed to measure synergy, it is difficult to compare these methods. The most important reason is that unlike regression and classification tasks, where the data is clearly labeled, the strength of synergy is never a label of the data. Therefore, there is no standard dataset for this task. In addition, most methods are based on models, and the choice of different models will greatly affect the outcome.

In this section, we compare the different methods of synthetic data generated by the following function:

$$F_x = \pi^{x_1 x_2}\sqrt{2x_3} - sin^{-1}(x_4) + log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7 + sin(x_3 + x_7) - 0.2 * x_{11} + 0 \times (x_{12} + x_{13} + x_{14}) \tag{8}$$

where variables x_4, x_5, x_8, x_{10} are uniformly distributed in $[0.6, 1]$ and the other variables are uniformly distributed in $[0, 1]$. Variables x_{12}, x_{13}, x_{14} are irrelevant variables so the weights of them in equation are 0. Through this equation, dependent variable y influenced by multiple variables. We define the two variables have synergy if they are in the same term. For example, the term $log(x_3 + x_5)$ means the x_3 and x_5 have synergy effect. This approach renders the task of detecting synergy akin to a classification task, with the Area Under the Curve (AUC) serving as a measure to assess accuracy. For the experiments, a dataset containing 10,000 data points was generated. The average accuracy of the various methods is shown in Fig. 5, along with a comparison of the time required for computation. From the result, the accuracy of the RSC is better than the GUIDE, PDF and MEB methods and slightly inferior to the FAST method. The time cost of RSC is less than other methods. However, the significance of this comparative experiment is somewhat limited, because the synergy effect does not necessarily equivalent to that two variables work in the same term. We display this experiment



F

Function type \ Noise	0	0.1	0.2	0.5	1	2
Random	0.44	0.43	0.47	0.45	0.47	0.47
Mountain	1.92	1.93	1.89	1.77	1.64	1.54
Wave	12.01	11.93	10.82	9.33	7.56	5.45
Plate	1.77	1.74	1.69	1.70	1.71	1.68
Crater	1.58	1.48	1.41	1.35	1.31	1.54
Ellipsoidal cap	1.94	1.50	1.26	1.02	1.30	1.62
Rand\Line	6.52	6.52	6.53	5.56	4.08	2.43
X	15.36	15.01	14.51	11.89	8.76	5.23
Sin\Cos	12.39	10.93	10.61	8.44	6.95	5.08

Fig. 4. The result of simulation. (A to E) the RSC score of functional relationships with different extend of noise added as the bin size increases. The sample data size is 2000 in this experiment. (F) The result of maxRSC with bin size less than 10 on simulated data with different extend of noises.

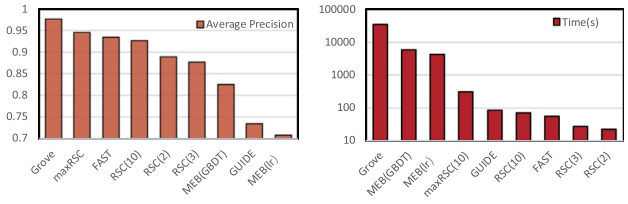


Fig. 5. Precision/Cost on synthetic function.

to establish a point of reference for the sake of comparability with certain other method. The paramount focus of this paper resides in its real-world applications.

4. Application on world development indicators data

4.1. Dataset

We applied the RSC methodology to real-world data, specifically the World Development Indicators (WDI) dataset, in order to identify synergy effects that impact future GDP per capita growth. WDI is the primary World Bank collection of development indicators, compiled from officially recognized international sources. It presents the most current and accurate global development data available and includes national, regional and global estimates. In this experiment, our objective is to investigate which indicators exhibit a synergistic impact on GDP per capita growth in the forthcoming year. Considering the emergence of the COVID-19 pandemic in 2020 and its potential influence on the dataset, we analyze data from the years 2018 and 2019. The dependent variable y in this context pertains to the GDP per capita growth in 2019. Meanwhile, the independent variables x consist of a total of 484 indicators (equivalent to 234,256 pairs) from the year 2018, allowing us to explore all feasible synergy effects.

Table 1 shows some of the most relevant indicators using MIC. Among 484 variables, only 13 variables have MIC greater than 0.3, and most MIC of variables are between 0.2 and 0.3.

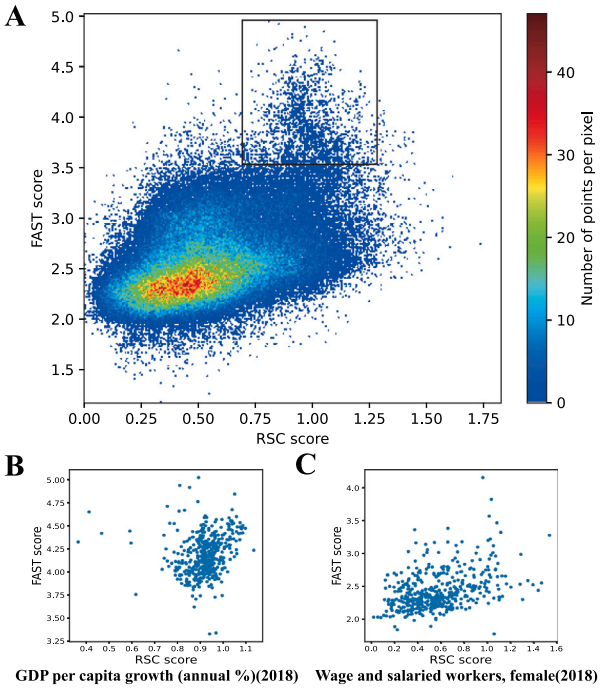


Fig. 6. Comparison of RSC and FAST on the WDI dataset. (A) RSC versus FAST for all pairwise relationships in the WDI dataset. (B) Data points that involve variable *'GDP per capita growth (annual %)(2018)'* from (A). (C) Data points that involve variable *'Wage and salaried workers, female(2018)'* from (A).

4.2. Relative synergy vs. absolute synergy

We compared the relative synergy(calculated by RSC) and absolute synergy(calculated by FAST) on this dataset. Note that we use RSC with pure InfoGain in this experiment and set $b = 2$. Besides, the

Table 1
Some most relevant variables and their MIC.

Indicator	MIC
GDP per capita growth (annual %)	0.5386
GDP growth (annual %)	0.4700
Households and NPISHs Final consumption expenditure per capita growth (annual %)	0.4055
Industry (including construction), value added (annual % growth)	0.3771
Services, value added (annual % growth)	0.3699
Adjusted savings: carbon dioxide damage (% of GNI)	0.3239
Manufacturing, value added (annual % growth)	0.3212
Imports of goods and services (annual % growth)	0.3199
Households and NPISHs Final consumption expenditure (annual % growth)	0.3134
Inflation, GDP deflator (annual %)	0.3112
Communications, computer, etc. (% of service exports, BoP)	0.3100
Gross capital formation (% of GDP)	0.3098
CO2 emissions (kg per 2015 US$ of GDP)	0.3096
Merchandise exports to low-and middle-income economies in Europe & Central Asia (% of total merchandise exports)	0.3029
Container port traffic (TEU: 20 foot equivalent units)	0.3016
Gross fixed capital formation (% of GDP)	0.3006

FAST algorithm is recoded to find the max value of $-RSS$ instead of the minimum value of RSS, so the larger the FAST value, the more interactive the two variables. Fig. 6 represents the comparison results. Each data point on the graph corresponds to a pair of variables, with their positions determined by the FAST and RSC values associated with that specific pair.

From Fig. 6(A), in general, the range of RSC values is between 0 and 1.8, which indicates that the value domain of RSC is relatively stable under real-world datasets and verifies the robustness of this metric. Most of the values are clustered in the range of 0.25–0.75, The distribution of data points indicates that when the RSC value is large, the corresponding FAST value will not be small and vice versa.

Fig. 6(b) and (c) further reflects the difference between RSC and the method that calculate absolute synergy. The Fig. 6(b) represents the values of variable pairs that involve variable "GDP per capita growth (annual %)(2018)", The Fig. 6(c) represents the values of variable pairs that involve variable "Wage and salaried workers, female(2018)". The former variable has MIC of 0.54 which is the most informative variable while the latter one only has MIC of 0.17. For FAST values, the former ones are generally greater than 3.75 and the latter ones are generally less than 3, which indicates that the absolute synergy value is greatly affected by the main effect. For the value of RSC, most of RSC values in Fig. 6(b) are clustered around 0.8–1.1, which is more concentrated, while the distribution of the RSC in Fig. 6(c) is more dispersed. This reflects the difference between relative synergy and absolute synergy. Since the former variable is more informative, the information gain brought by it is basically the main effect, so the RSC is more stable. Because the latter variable is uninformative, the synergy or information redundancy generated by its interaction with other variables is more likely to affect the value of RSC.

4.3. Synergy or redundancy

Remember the calculation of RSC (Eq. (9)), where the F_1 and F_2 represent the measurement functions of theoretical values.

$$RSC = \frac{G_1}{I_1} = \frac{F_1(I_1 + S(X_1, X_2))}{F_2(I_1)} \tag{9}$$

If the theoretical values can be precisely measured, which means $F_1(x) = F_2(x) = x$, then the RSC can be reduced to the following equation

$$RSC = \frac{G_1}{I_1} = \frac{I_1 + S(X_1, X_2)}{I_1} = 1 + \frac{S(X_1, X_2)}{I_1} \tag{10}$$

In this situation, $RSC = 1$ will be a dividing line to determine the presence of synergy. Specifically, RSC values exceeding 1 indicate a synergistic relationship, while values below 1 suggest that the two variables possess redundant information. This observation provides insight into the potential of the proposed calculation method to accurately estimate theoretical values, given the characteristic exhibited by RSC.

From Fig. 6, We find a bump of data points boxed in black, where the FAST values are large and the RSC values are close to 1. In specific, from the distribution of data points where one variable is 'GDP per capita growth (annual %)(2018)', as shown in Fig. 6(B), we may conclude that most of the data points in this area have one variable which is informative to predict y that leads to a high FAST value. As for the RSC value, as most of the information gain results from the main effect, RSC for those pairs of variables that have no significant synergistic effect is close to 1. This just confirms the assumptions of the theoretical analysis and shows that our proposed method might estimate the theoretical values well.

4.4. Explainability and validation

In this section, we illustrate how RSC can be helpful to obtain domain knowledge by digging into examples. Table 2 shows the top 20 variable pairs with the highest synergy score ranked by RSC($b = 2$) on the WDI dataset. Among these pairs, we have chosen three for detailed illustration, each representing distinct types of synergy. These examples shed light on the domain knowledge that can be gleaned from such analyses.

The first example is the augment synergy between indicator "fertilizer consumption" and indicator "strength of legal rights index" as shown in Table 3. From univariate analysis, the higher the fertilizer consumption and the stronger the legal right, the faster GDP grows. As a result of interaction, the trend remains the same, and a subset of data with extremely large or small y will be discovered and useful for knowledge acquaintance. Under this circumstance, the per capita GDP growth is more likely to be very slow when the country has low fertilizer consumption and weak strength of legal rights. Our algorithm can help experts to find this kind of anomaly so that they can explain what might cause it, like the region, religion or something else.

The second interaction effect is cross, which means after cutting two variables, the trend between the different groups is opposite. The example shown in Table 4 is synergy between the variable 'labor participation rate, male' and 'refugee population by country or territory of asylum'. If we only consider single factor variables, we may draw the conclusion that the high labor participation rate will have a negative impact on economic development. This is a reasonable conclusion because the high labor participation rate might mean the lack of educational opportunities for teenagers or the imperfect welfare system and lack of pension system. However, when we include the indicator of refugee population, GDP growth shows different trends based on the growth of employment in high refugee countries and low refugee countries, making the previous conclusion invalid. The result indicates

Table 2
Top scoring 20 of interaction relationships (by RSC with $b = 2$) from the WDI dataset.

RANK	Variable x_1	Variable x_2	MIC1	MIC2	IG1	IG2	RSC
1	Fertilizer consumption (kilograms per hectare of arable land)	Strength of legal rights index (0 = weak to 12 = strong)	0.197	0.203	0.142	0.371	1.827
2	Mammal species, threatened	Unemployment, male (% of male labor force) (modeled ILO estimate)	0.201	0.224	0.348	0.331	1.735
3	Employment in services, male (% of male employment) (modeled ILO estimate)	Prevalence of undernourishment (% of population)	0.251	0.220	0.404	0.212	1.612
4	Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)	Refugee population by country or territory of asylum	0.209	0.194	0.335	0.300	1.598
5	Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)	Population ages 65 and above, female (% of female population)	0.200	0.262	0.314	0.307	1.568
6	Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)	Population ages 65 and above (% of total population)	0.200	0.248	0.314	0.307	1.568
7	Age dependency ratio, old (% of working-age population)	Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)	0.248	0.200	0.307	0.314	1.568
8	Refugee population by country or territory of asylum	Strength of legal rights index (0 = weak to 12 = strong)	0.194	0.203	0.087	0.314	1.547
9	Death rate, crude (per 1000 people)	Life expectancy at birth, total (years)	0.264	0.205	0.333	0.317	1.544
10	Labor force participation rate, male (% of male population ages 15–64) (modeled ILO estimate)	Refugee population by country or territory of asylum	0.217	0.194	0.335	0.298	1.544
11	Gross fixed capital formation (% of GDP)	Wage and salaried workers, female (% of female employment) (modeled ILO estimate)	0.301	0.174	0.243	0.266	1.531
12	Armed forces personnel, total	Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)	0.200	0.209	0.300	0.319	1.525
13	Employment in services, male (% of male employment) (modeled ILO estimate)	Population ages 50–54, female (% of female population)	0.251	0.228	0.377	0.297	1.502
14	Employment in services, male (% of male employment) (modeled ILO estimate)	Population ages 55–59, female (% of female population)	0.251	0.267	0.376	0.301	1.498
15	Final consumption expenditure (constant 2015 US$)	Population density (people per sq. km of land area)	0.198	0.222	0.085	0.330	1.489
16	Mortality rate attributed to unintentional poisoning, male (per 100,000 male population)	Population ages 75–79, female (% of female population)	0.200	0.270	0.298	0.272	1.487
17	Mammal species, threatened	Unemployment, total (% of total labor force) (modeled ILO estimate)	0.201	0.225	0.298	0.308	1.486
18	Air transport, registered carrier departures worldwide	Unemployment, female (% of female labor force) (modeled ILO estimate)	0.203	0.225	0.145	0.333	1.477
19	Strength of legal rights index (0 = weak to 12 = strong)	Time to resolve insolvency (years)	0.203	0.212	0.299	0.053	1.474
20	Air transport, passengers carried	Arable land (% of land area)	0.202	0.251	0.073	0.369	1.472

Table 3
An example of augment synergy selected by RSC.

Fertilizer consumption	Low		High	
Strength of legal rights index	Weak	Strong	Weak	Strong
Country number	48	53	52	49
GDP growth	0.04	2.22	1.12	2.40

Table 4
An example of cross synergy selected by RSC.

Labor participation rate, male	Low		High	
GDP growth	1.53		1.28	
Refugee population by country or territory of asylum	Low	High	Low	High
Country number	56	52	43	51
GDP growth	2.38	0.58	0.36	1.97

Table 5
An example of simpson paradox selected by RSC.

Time to resolve insolvency	Short		Long	
Country number	110		101	
GDP growth	1.48		1.45	
Legal right index	Weak	Strong	Weak	Strong
Country number	45	65	54	47
GDP growth	0.63	2.06	0.74	2.28

that in low refugee countries, low labor participation leads to a high GDP growth whereas in countries with a large amount of refugee, high labor participation leads to high GDP growth. In this type of synergy, one of the variable is often the reason, or related to the reason why another variable have various patterns in predicting y.

Besides, the Simpson paradox, a phenomenon that a trend appears in several groups of data but disappears or reverses when the groups are combined, can also be detected by our method. The Table 5 shows the occurrence of the paradox. Overall, the GDP per capita growth is lower when the time to resolve insolvency is short. However, an intriguing contrast emerges when we delve deeper into the data by partitioning it based on the strength of a country's legal rights. Remarkably, irrespective of whether a country boasts strong or weak legal rights, a prolonged time for insolvency resolution corresponds to higher GDP growth. This observation is contrary to the overarching trend. The underlying reason behind this paradox lies in the composition of countries within the short-insolvency-resolving-time category; a greater proportion of these countries possess robust legal rights.

The above analysis is made based on the binning size $b = 2$, because it is the easiest to understand. With the binning size grow, the result may be complex for understanding, but still we can find something useful.

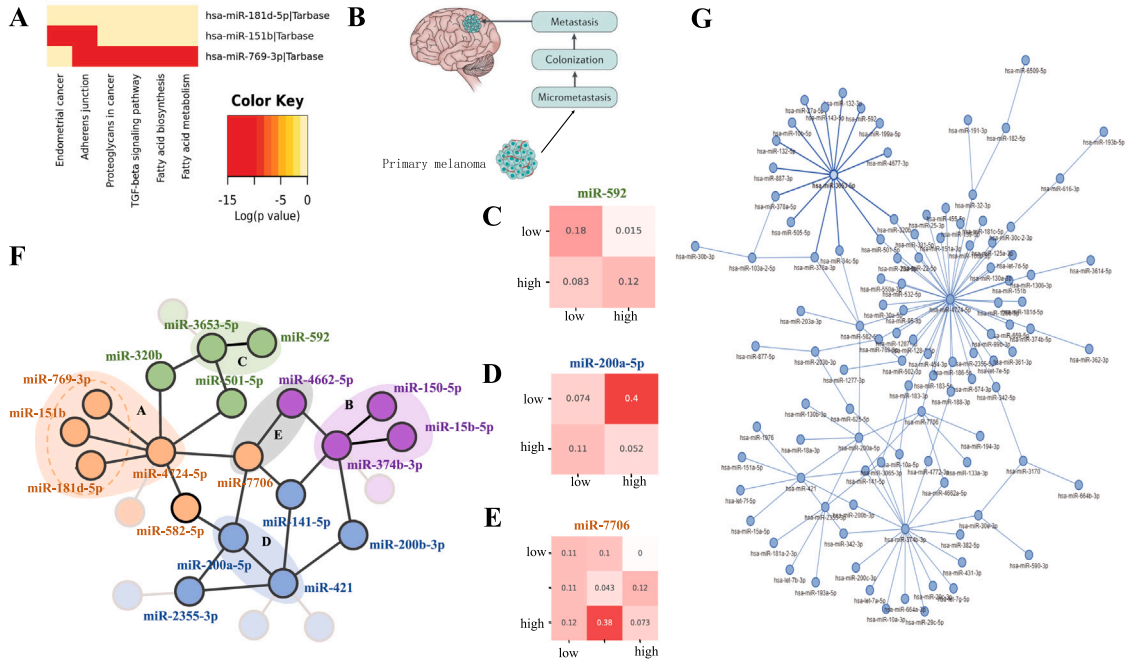


Fig. 7. The gene interaction network and the explainability. (G) Gene interaction network based on top 0.1% interaction relationships. (F) The subset of the network (G), which contains the relationship pairs that can be explained by figure (A–E). (A) The pahthway enrichment analysis of miR-4724-5p related miRNAs. (B) Process of brain metastasis of melanoma. The miR-3653-5p, miR-592 and miR-501-5p synergistically regulate this process. (C–E) The heatmap of LNM rate in the interactive gene pairs.

5. Application on gene data

The occurrence and development of human diseases are almost all related to gene mutation and maladjustment. Gene expression data can provide important insights for the diagnosis and treatment of diseases. However, gene expression data has the characteristics of high dimensionality, limited sample size and imbalanced distribution, making it difficult to identify the interrelationships between genes and their effects on diseases [18]. Our algorithm offers a new way to solve this problem.

The Cancer Genome Atlas (TCGA) is a cancer genomics program that sequences and molecularly characterizes primary cancer samples, including endometrial cancer (EC). Endometrial cancer is a major gynecological malignancy worldwide [19]. Lymph node metastasis (LNM) is a critical prognosis-related risk factor for EC, and the status of lymph nodes is an essential consideration when making clinical decisions [20]. MicroRNAs (miRNAs) are small RNA molecules that play significant roles in the occurrence and development of cancer by guiding target mRNA cleavage or translational inhibition. However, how miRNAs interact with each other in the LNM process is unknown. Thus, we tested our algorithm on the TCGA-UCEC database. The results are shown in Fig. 7.

The Fig. 7(C–E) describe how the LNM rate change with the interaction of two synergic variables. Take miR-200a-5p vs miR-421 for example, the miR-421 was positively related to LNM when the expression of miR-200a-5p is at a low level but was negatively related to LNM while the expression of miR-200a-5p is at a high level, indicating there may be an interaction of mutual restriction. Moreover, the ąřhigh miR-421 ąśand ąřmiR-200a-5pąś expression patterns can detect a fairly high LNM rate, which may help in diagnosis in clinic. However, as the center of the sub-network (Fig. 7(G)), miR-421 has not been reported to have a direct relationship with endometrial cancer in the existing literature, but many studies have proved that it plays an important role in the occurrence and development of liver and lung cancer through the PTEN pathway [21,22]. Thus, exploring the effect of the miR-421/PTEN pathway on endometrial cancer progress might be valuable.

As shown in Fig. 7(B,F), there is considerable interrelationship among miR-150-5p, miR-15b-5p and mir-374b-3p. Also, it is reported that The above three genes together could predict the brain metastases of primary melanoma, its mechanism may be the same as lymph node metastasis [23]. Therefore, it is reasonable to speculate that miR-150-5p, miR-15b-5p and mir-374b-3p may act synergistically in distant metastasis of endometrial cancer and provide insights for future research.

From Fig. 7(A,F), miR-181d-5p, miR-151b and miR-769-3p show similar patterns of interaction with miR-4724-5p, indicating that those miRNAs may engage in common biological processes and share similar function. Indeed, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis reveals that above-mentioned miRNAs were enriched in Endometrial Cancer, Adherens Junction, Proteoglycan in cancer and TGF-ę signaling pathway, proving that those miRNAs actually were involved in the development of carcinoma in common [23].

As a successful example, this experiment shows that RSC has great potential in processing gene data sets which can help researchers analyze the impact of genes on human diseases from the perspective of gene synergy.

6. Future work

This study introduces an innovative approach for detecting relative synergy effects among variables, and there are ample opportunities for further exploration in the future. In this section, we will discuss potential areas for improvement in this method from three distinct perspectives.

(1) Better binning method. In this paper, the two independent variables are split into the same number of buckets that have equal-sized samples. However, the optimal partition may not be accessible by this method. For example, in Fig. 8, our method will split the dataset into 2×2 as the first one or into 3×3 as the fourth one, but the optimal partition might be the third one where the partition number and bucket size are unequal. Therefore, an improved method for determining the optimal partitioning is warranted.

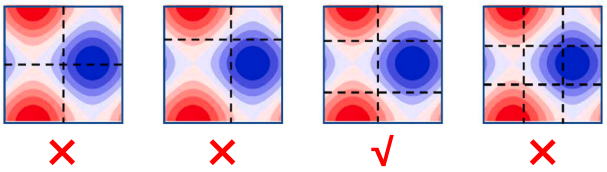


Fig. 8. Different partitions of the dataset.

(2) Automatic pattern recognition of synergy. While our current study manually elucidates various types of synergy effects and analyzes the associated knowledge, a more streamlined approach involves automating the classification of extracted synergy types and transforming them into actionable insights. Using the information contained in the matrix of RSC with different bin size is a promising way to do that. For example, we notice the RSC will change differently with the increase of binning size for different types of synergy. Overall the RSC will fluctuate sharply first and transfer into a smooth decline or increase when binning size is large enough as shown in Fig. 4(A–E). This changing trend of the RSC is determined by the characteristics of different synergy type. Consequently, this phenomenon possesses the potential to serve as a valuable tool for synergy analysis.

(3) Synergy among multiple variables. In this paper, we only present the RSC that detect the relative synergy between two variables. It is also important to detect the synergy involve more than two variables. To do so, the calculation approach which is used in RSC to compute information gain, i.e., discretization and normalization can also be used for calculation of high-dimensional synergy.

7. Conclusion

In this paper, we propose a novel indicator RSC for calculating the relative synergy in the dataset. This metric is based on the information entropy theory and use discretizing and normalization method to compute the information gain of continuous variables. A simulation study proves that the RSC has excellent characteristics of generality, equitability, and robustness. Additionally, we apply it to a dataset of World Development Indicators to identify synergies that help predict GDP per capita growth over the next year. The results demonstrate the difference between RSC and absolute synergy detection methods and the benefits that this method can provide for variables with small main effects. Besides, the variable pairs selected by RSC are proved to be helpful for knowledge discovery by three different synergy type examples. Using the RSC, we also built a gene synergy network that helps researchers understand how genes affect human diseases from a gene synergy perspective. Finally, the future improvement of this method, the better binning method, automatic pattern recognition of synergy and synergy among multiple variables is discussed.

CRediT authorship contribution statement

Yanrui Li: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Kaiyou Fu:** Data curation, Validation, Writing – original draft. **Yuchen Zhao:** Writing – review & editing. **Chunjie Yang:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

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.

Table 6
Formulas of the simulated data.

Function	Formula
Mountain	$y = sin(2\pi x_1) + cos(2\pi x_2)$
Wave	$y = sin(2\pi x_1 + 2\pi x_2)$
Plate	$y = x_1 + x_2$
Crater	$y = sin(2\pi\sqrt{(x_1-0.5)^2+(x_2-0.5)^2})$
Ellipsoidal cap	$y = \begin{cases} \sqrt{0.16 - x_1^{*2} - x_2^{*2}} + 0.5 & \text{if} x_1^{*2} + x_2^{*2} \le 0.16 \\ 0.66 - \sqrt{x_1^{*2} + x_2^{*2}} & \text{else} \end{cases}$
Rand/line	$y = \begin{cases} random & \text{if } x_2 = 0 \\ x_1 & x_2 = 1 \end{cases}$
X	$y = \begin{cases} x_1 & \text{if } x_2 = 0 \\ 1 - x_1 & x_2 = 1 \end{cases}$
Sin/Cos	$y = \begin{cases} sin(2\pi x_1) & \text{if } x_2 = 0 \\ cos(2\pi x_1) & x_2 = 1 \end{cases}$

Data availability

The WDI data can be downloaded in https://datatopics.worldbank.org/world-development-indicators/ and gene data can be downloaded in https://portal.gdc.cancer.gov/.

Acknowledgments

This work is supported by National Natural Science Foundation of China (61933015).

Appendix A. Formulas of simulated data

The formulas of the simulated data shown in the Fig. 3 are as follows: where $x_1^* = x_1 - 0.5$ and $x_2^* = x_2 - 0.5$ (see Table 6).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2023.111112.

References

[1] Alex A. Freitas, Understanding the crucial role of attribute interaction in data mining, Artif. Intell. Rev. 16 (3) (2001) 177–199.
[2] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, Huan Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.
[3] Firuz Kamalov, Orthogonal variance decomposition based feature selection, Expert Syst. Appl. 182 (2021) 115191.
[4] Haitian Wang, Shaw-Hwa Lo, Tian Zheng, Inchi Hu, Interaction-based feature selection and classification for high-dimensional biological data, Bioinformatics 28 (21) (2012) 2834–2842.
[5] João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, Pedro Bizarro, TimeSHAP: Explaining recurrent models through sequence perturbations, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2021, pp. 2565–2573.
[6] Anupam Datta, Matt Fredrikson, Klas Leino, Kaiji Lu, Shayak Sen, Zifan Wang, Machine learning explainability and robustness: Connected at the hip, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, 2021, pp. 4035–4036.
[7] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, Pardis C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524.
[8] John Watkinson, Kuo-ching Liang, Xiadong Wang, Tian Zheng, Dimitris Anastassiou, Inference of regulatory gene interactions from expression data using three-way mutual information, Ann. New York Acad. Sci. 1158 (1) (2009) 302–313.

[9] Orlando Anunciação, Susana Vinga, Arlindo L Oliveira, Using information interaction to discover epistatic effects in complex diseases, PLoS One 8 (10) (2013) e76300.
[10] Ronald Aylmer Fisher, Statistical methods for research workers, in: Breakthroughs in Statistics, Springer, 1992, pp. 66–70.
[11] Yin Lou, Rich Caruana, Johannes Gehrke, Giles Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 623–631.
[12] Daria Sorokina, Rich Caruana, Mirek Riedewald, Daniel Fink, Detecting statistical interactions with additive groves of trees, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 1000–1007.
[13] Wei-Yin Loh, Regression tress with unbiased variable selection and interaction detection, Statist. Sin. (2002) 361–386.
[14] Sejong Oh, Feature interaction in terms of prediction performance, Appl. Sci. 9 (23) (2019) 5191.
[15] Jerome H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. (2001) 1189–1232.
[16] Giles Hooker, Discovering additive structure in black box functions, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 575–580.
[17] Yanrui Li, Chunjie Yang, Domain knowledge based explainable feature construction method and its application in ironmaking process, Eng. Appl. Artif. Intell. 100 (2021) 104197.
[18] Mengyun Wu, Huangdi Yi, Shuangge Ma, Vertical integration methods for gene expression data analysis, Brief. Bioinform. 22 (3) (2021) bbaa169.
[19] Rebecca L. Siegel, Kimberly D. Miller, Ahmedin Jemal, Cancer statistics, 2020, CA: Cancer J. Clin. 70 (1) (2020) 7–30.
[20] Sharyn N Lewin, Thomas J Herzog, Nicanor I Barrena Medel, Israel Deutsch, William M Burke, Xuming Sun, Jason D Wright, Comparative performance of the 2009 international federation of gynecology and obstetrics' staging system for uterine corpus cancer, Obstetr. Gynecol. 116 (5) (2010) 1141–1149.
[21] Yong-Jie Xu, Rui-Shi Wei, Xin-Hua Li, Qiang Li, Jian-Rong Yu, Xiao-Fei Zhuang, MiR-421 promotes lipid metabolism by targeting PTEN via activating PI3K/AKT/mTOR pathway in non-small cell lung cancer, Epigenomics 14 (3) (2022) 121–138.
[22] Yongfeng Hui, Dong Jin, Junzhi Leng, Di Liu, Peng Yuan, Chaofeng Tang, Qi Wang, Hsa_circ_0007059 sponges miR-421 to repress cell growth and stemness in hepatocellular carcinoma by the PTEN-AKT/mTOR pathway, Pathol. Res. Pract. 229 (2022) 153692.
[23] Doug Hanniford, Judy Zhong, Lisa Koetz, Avital Gaziel-Sovran, Daniel J Lackaye, Shulian Shang, Anna Pavlick, Richard Shapiro, Russell Berman, Farbod Darvishian, et al., A miRNA-based signature detected in primary melanoma tissue predicts development of brain metastasis, Clin. Cancer Res. 21 (21) (2015) 4903–4912.