A Parallel Feature Selection Method Based on NMI-XGBoost and Distance Correlation for Typhoon Trajectory Prediction

Baiyou Qiao

qiaobaiyou@mail.neu.edu.cn

Northeastern University

Yuanqing Hao

Northeastern University

Rui Wang

Shenyang Institute of Automation

Peirui Wang

Northeastern University

Donghong Han

Northeastern University

Gang Wu

Northeastern University

Research Article

Keywords: Feature selection, NMI, XGBoost, Distance correlation, Spark

Posted Date: January 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2479525/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at The Journal of Supercomputing on January 23rd, 2024. See the published version at https://doi.org/10.1007/s11227-023-05863-3.

A Parallel Feature Selection Method Based on NMI-XGBoost and Distance Correlation for Typhoon Trajectory Prediction

Baiyou Qiao[1*], Yuanqing Hao[1], Rui Wang[2], Peirui Wang[1], Donghong Han[1] and Gang Wu[1]

[1*]School of Computer Science and Engineering, Northeastern University, Shenyang, 110619, China.
[2]Shenyang Institute of Automation,Chinese Academy of Sciences, Shenyang, 110169, China.

*Corresponding author(s). E-mail(s): qiaobaiyou@mail.neu.edu.cn;
Contributing authors: 2001747@stu.neu.edu.cn; wangrui1@sia.cn; 20711810@stu.neu.edu.cn; handonghong@mail.neu.edu.cn; wugang@mail.neu.edu.cn;

Abstract

Typhoon trajectory related data involves many factors such as atmospheric factors, oceanic factors, and physical factors. It has the characteristics of high dimension, strong spatio-temporal correlation, and non-linear correlation, which increases the difficulty of typhoon trajectory prediction. Using feature selection approaches to select appropriate prediction factors becomes an important means to reduce the dimension of typhoon trajectory related data and improve the performance and accuracy of typhoon trajectory prediction methods. However, the existing feature selection methods based on linear correlation analysis can not well depict the nonlinear correlation between data features, which results in low accuracy of feature selection. The feature selection methods based on nonlinear correlation analysis are computationally expensive, which affects the timeliness of feature selection. To solve the problem, we propose a parallel feature selection method NX-Spark-DC based on the Spark platform for typhoon trajectory related data. The method firstly filters out the redundant

features of typhoon related data by Normalized Mutual Information (NMI) method, subsequently eliminates the useless features by XGBoost machine learning model, and thus reducing the dimension of typhoon related data. On this basis, an improved Spark-based parallel distance correlation algorithm (Spark-DC) is proposed to select the feature combinations with strong correlation. A series of experimental results show that NX-Spark-DC method has high execution efficiency and accuracy, which is significantly better than the existing methods.

Keywords: Feature selection, NMI, XGBoost, Distance correlation, Spark

1 Introduction

A typhoon is a very dangerous and catastrophic tropical weather system, which brings strong winds, heavy rain, and storm surge that can lead to a large number of direct and secondary disasters, seriously affecting the safety of people's lives and properties in coastal areas and the local economic development. Accurate and timely prediction of typhoon trajectory has become one of the important means for typhoon disaster prevention and auxiliary decision-making. The prediction of typhoon trajectory involves many factors such as atmospheric factors, oceanic factors, geographical factors, etc. These factors have complex spatio-temporal correlations and present high-dimensional nonlinear characteristics. Therefore, designing an effective feature selection method to analyze the influencing factors of typhoon trajectory and selecting the combinations of the factors that have a great influence on typhoon moving path becomes the key to making fast and accurate typhoon trajectory prediction, and has been one of the hot spots in typhoon related research fields.

At present, there are three kinds of feature selection methods, which are the filter method, the wrapper method and the embedding method. The filter method [1] mainly uses statistical methods to evaluate the correlation between the tag and features, and then filters out the features with low correlation, which is independent of the subsequent models. The wrapper method [2, 3] is mainly based on the training effect of the subsequent machine learning algorithms for the feature selection, which requires many times of training and has a large computational cost. The embedding method [4–6] integrates feature selection and model training into one process and achieves feature selection while training, but its parameter setting is complicated and the time complexity is high. Typhoon related data is a kind of spatio-temporal series data. For this kind of data, regression analysis and correlation analysis techniques are mainly used to study the importance of each factor for the feature selection. Commonly used regression analysis techniques include stepwise regression, multivariate regression, and so on. Common used correlation analysis methods mainly include statistical correlation coefficient [7–10], canonical correlation analysis[11], mutual information correlation analysis [12], matrix calculation

[13, 14] and distance correlation (DC) [15], etc. Huang et al. [16] analyzed the correlation between the environmental field prediction variables and the prediction errors of typhoon paths in the Northwest Pacific and the South China Sea using the correlation coefficient method. They selected appropriate features to establish a prediction model using the linear regression analysis method. Chen et al. [17] obtained the typical features between factor fields and prediction fields by using canonical correlation analysis approach and established the prediction equation of typhoon moving paths with a stepwise regression method. Obviously, the feature selection methods based on correlation coefficients and canonical correlation analysis have certain limitations. As they cannot well characterize the nonlinear relationships among typhoon trajectory factors, thus affecting the accuracy of feature selection and the precision of typhoon trajectory prediction models. Although the feature selection methods based on Mutual Information (MI) and the methods based on distance correlation [18] can analyze the nonlinear correlations among variables and characterize the high-dimensional nonlinear relationships, their computational efficiency is not high, so they cannot fully satisfy the requirement of the feature selection for large-scale and high-dimensional typhoon trajectory related data.

Aiming at the feature selection problems of the typhoon trajectory related data, we propose a parallel typhoon trajectory feature selection method based on Spark (NX-Spark-DC) by combining various methods such as Normalized Mutual Information (NMI), XGBoost machine learning model [19]and Distance Correlation analysis. Firstly, the two kinds of feature selection methods, the filter and the wrapper methods are used to carry out preliminary feature dimension reduction. On this basis, the feature combinations with strong correlation are selected based on the improved parallel distance correlation analysis method, so as to realize the final feature selection of typhoon trajectory related data and improve the prediction accuracy of typhoon moving trajectory. The main contributions of this paper are as follows:

- For the first time, the two feature selection methods, the NMI based filter method and the XGBoost based wrapper method are combined to achieve the preliminary feature dimension reduction of typhoon trajectory related data, which greatly reduced the calculation cost of subsequent feature selection.
- A parallel distance correlation calculation method based on Spark (Spark-DC) is proposed. This method can well characterize the high-dimensional nonlinear relationship between variables, and has high computational efficiency. Based on this method, high-precision feature selection of typhoon trajectory related data is realized, and the accuracy of typhoon trajectory prediction is improved.
- A series of experiments are carried out on the real data sets, and the results show that the parallel feature selection method proposed in this paper has high execution efficiency and analysis accuracy, which is significantly better than the existing feature selection method based on correlation analysis.

The rest of the paper is arranged as follows. Section 2 introduces the related work. Section 3 presents the proposed parallel feature selection method. In Section 4, we give the performance evaluation of the proposed method. A conclusion and our opinion in future research direction are given in section 5.

2 Related Work

The analysis and prediction of typhoon trajectory involve many environmental factors such as atmosphere factors, ocean factors, geography factors, chemistry factors, etc. Each factor has strong spatial and temporal characteristics. In the early stage, researches mainly used thermodynamic and dynamical knowledge [20], complex topographic and coastline features, conventional weather maps and satellite cloud maps to analyze the factors affecting typhoon movement track, and make feature selection. This feature selection method mainly depends on artificial experience and needs a lot of prior knowledge. Later, with the rapid development of big data and data-driven forecasting methods, the feature selection methods based on correlation analysis and regression analysis have received much attention and have been applied to the analysis and prediction of typhoon movement trajectory. Huang et al. [16] analyzed the correlation between the environmental field variables and the forecast errors of typhoons in the Northwest Pacific and the South China Sea by using the correlation coefficient method, established a forecast model by selecting appropriate features using linear regression analysis method, which achieved good results in 24-hour typhoon trajectory forecast. Chen et al. [17] obtained the typical features composed of various factor fields by using canonical correlation analysis method. Combining with the synoptic experience features, they established the equation of typhoon moving path by using the step-wise regression method, which improved the prediction ability. The above traditional correlation analysis methods are mainly used to analyze the linear correlation but are not suitable for analyzing the nonlinear correlation among data features. Therefore, researchers have proposed a series of nonlinear correlation analysis methods, mainly including mutual information methods, matrix-based correlation analysis methods, and distance-based correlation analysis methods. The mutual information method defines the maximum information coefficient (MIC) to measure the nonlinear correlation between two variables. La et al. [21] define standardized mutual information (NMI) to better describe the nonlinear relationship between variables and proposed a feature selection algorithm based on NMI. Székely et al. [15] used the distance of the eigenfunction to depict the nonlinear relationship between two random variables and proposed the concepts of distance covariance and distance correlation, which have been widely used in the field of correlation analysis and feature selection. Wang et al. [22] proposed a new analysis and prediction method for PM2.5 concentration based on the distance correlation and SVR. This method uses a distance correlation to screen important factors and realizes accurate prediction by the SVR model. Wen et al [23] proposed a weighted distance correlation method

wdCor for assessing the correlation between genetic markers and image data, which solved the problem of correlation analysis and feature selection for high-dimensional data. Zhang et al. [24] proposed a clustering method based on the distance correlation to portray the nonlinear relationship between clusters, which has good measurability and spatial contraction. However, the time complexity of the distance correlation method is high, and the computational efficiency is low in the case of large samples. Based on the biased estimation of Hilbert Schmidt's Independent Criterion (HSIC0), Zhang et al. [25] realized the correlation measurement between clusters and verified the effectiveness of HSIC0 for solving nonlinear correlation problems, but it also has a high computational cost.

3 Methodology

Based on the Spark platform, we propose a parallel feature selection method NX-Spark-DC based on NMI, XGBoost, and DC. The method integrates the filter and the wrapper techniques to solve the feature selection problem of typhoon trajectory related data, expecting to obtain the combinations of features with high correlation, thus improving the accuracy and efficiency of typhoon trajectory prediction approaches. The overall framework of the NX-Spark-DC approach is shown in Figure 1. It mainly consists of data normalization, feature dimension reduction based on NMI and XGBoost, and parallel distance correlation analysis and feature selection based on Spark.

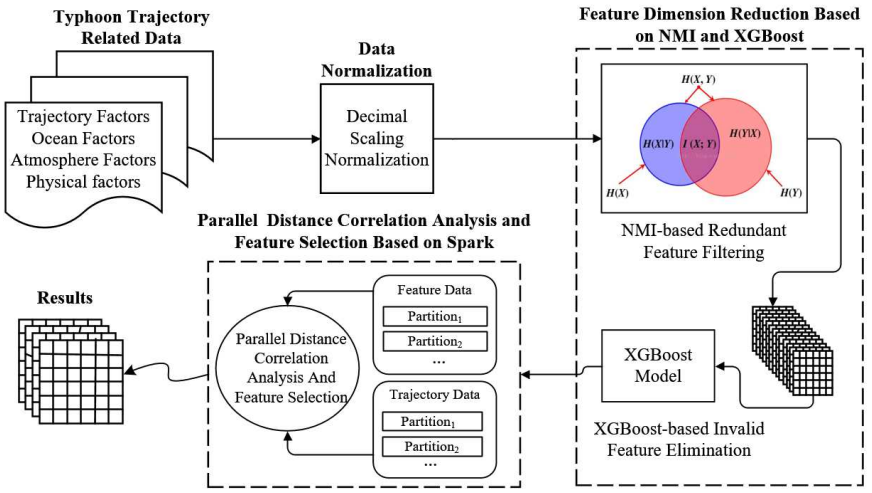

Fig. 1 The overall framework of the feature selection method NX-Spark-DC

At first, the typhoon trajectory related data consisting of many factors are normalized by using decimal scaling normalization. Then the factors having low correlation with typhoon trajectory are filtered out by using NMI

method, the factors that play an inverse role in the typhoon trajectory prediction are eliminated by XGBoost model using the strategy of forward exclusion and recursive deletion, thus realizing the feature dimension reduction. On this basis, an improved parallel distance correlation method is used to analyze the correlation of the remaining factors, and the combinations of factors with the highest correlation are selected as the final features of the prediction model, and each component is described in detail below.

3.1 Data Normalization

To eliminate the influence of different dimensional data features on analysis and prediction, we use the decimal scaling normalization method to preprocess typhoon related data. Suppose that $F = \{(f_{1j}, f_{2j}, \ldots, f_{mj}) \mid j = 1, 2, \ldots, n\}$ is a typhoon dataset containing m features, n is the number of samples and f_{ij} denotes the value of the j-th feature of the i-th sample. When normalizing the feature $f_i (i = 1, 2, \ldots, m)$, it is necessary to determine the number of decimal point shifts k_i of the sample data. The calculation method of k_i is shown in equation (1). After that, the sample value f_{ij} can be normalized into $f_{\mathrm{ij}}^{'}$ by equation (2).

$$k_i = \left\lceil \log_{10}\left(\max_{1 \le j \le n} \mid f_{ij} \mid \right) \right\rceil \tag{1}$$

$$f_{\mathrm{ij}}^{'} = \frac{f_{\mathrm{ij}}}{10^{k_i}} \tag{2}$$

3.2 Feature Dimension Reduction Based on NMI and XGBoost

Typhoon related data are usually high-dimensional data, which can affect the performance and accuracy of analysis and prediction algorithms. So it is necessary to reduce the dimensionality of the data. Considering the linear and nonlinear relationships between data features, a feature dimension reduction method based on NMI and XGBoost is proposed. The process of the feature dimension reduction based on NMI and XGBoost is shown in Figure 2.

From Figure 2, it can be seen that our method includes two stages, NMI-based redundant feature filtering and XGBoost-based invalid feature elimination. The method combines the advantages of filtering and wrapping feature selection algorithms. The core idea of the method is to eliminate features that have less relationship with typhoon trajectory and play an inverse role in typhoon trajectory prediction, so as to reduce feature redundancy and improve the efficiency and accuracy of the feature selection method.

3.2.1 NMI-based Redundant Feature Filtering

Different from the traditional correlation coefficient, NMI can effectively describe the linear and nonlinear relationship between two variables. Therefore,

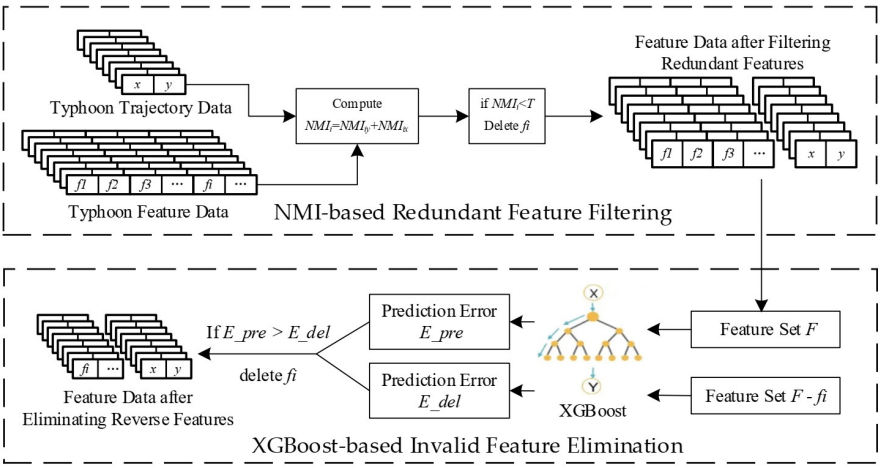

Fig. 2 The process of the feature dimension reduction based on NMI and XGBoost

it can better describe the correlation between features in typhoon trajectory dataset, which is why we choose NMI to filter redundant features. Let $L = \{(x_j, y_j) \mid j = 1, 2, ..., n\}$ be the typhoon sample dataset, where n is the number of samples, x_j and y_j represent the longitude and latitude of the j-th sample respectively.

For each feature f_i $(i = 1, 2, \ldots, m)$, the NMI value NMI_{ix} (between f_i and typhoon trajectory longitude x) and NMI_{iy} (between f_i and typhoon trajectory latitude y) are calculated at first. Then, the overall NMI values NMI_i (between f_i and the typhoon position) is obtained to represent the degree of correlation between the feature f_i and the typhoon position. Information entropy H is an effective measurement tool to describe the information contained in variables. For a data sequence $X = (x_1, x_2, ..., x_i, ..., x_n)$ containing n variables, it is assumed that its probability distribution is $P(X = x_i) = p_i,\ i = 1, 2, ..., n$. Then the calculation of information entropy $H(X)$ is shown in equation (3).

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{3}$$

For $Y = (y_1, y_2, ..., y_i, ..., y_n)$, the calculation of conditional entropy $H(X/Y)$ is shown as equation (4), where p_{ij} is the joint probability distribution of (X, Y), $p_i,\ p_j$ are the marginal distributions of X and Y.

$$H(X/Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p_{\mathrm{ij}} \log \frac{p_{\mathrm{ij}}}{p_j} \tag{4}$$

Mutual information can reflect the degree of dependence between variables. The definition of mutual information value $I(X;\ Y)$ between X and Y is shown in equation (5).

$$I(X;\ Y) = H(X) - H\left(\frac{X}{Y}\right) = H(Y) - H\left(\frac{Y}{X}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{\mathrm{ij}} \log \frac{p_{\mathrm{ij}}}{p_i \cdot p_j} \quad (5)$$

The mutual information value is normalized to [0, 1] to obtain the NMI value $NMI(X;\ Y)$, see equation (6).

$$NMI(X;\ Y) = 2\frac{I(X;\ Y)}{H(X) + H(Y)} \quad (6)$$

Through the above equations, we can get the value of standardized mutual information $\mathrm{NMI}(f_i;\ x)$ and $\mathrm{NMI}(f_i;\ y)$, record as NMI_{ix} and NMI_{iy} respectively. Then NMI between f_i and typhoon moving path will be calculated. We record it as NMI_i and it's shown in equation (7).

$$\mathrm{NMI}_i = \mathrm{NMI}_{ix} + \mathrm{NMI}_{iy} \quad (7)$$

Suppose that the preset threshold of NMI is T. If NMI_i is lower than T, it is regarded as a redundant feature and will be filtered out. By calculating and comparing the NMI value of each feature, all redundant features can be filtered out.

3.2.2 XGBoost-based Invalid Feature Elimination

Although the redundant features have been filtered using NMI in the first stage, there may still be some features that have no contribution or play an inverse role in typhoon trajectory prediction. Therefore, we use XGBoost model to eliminate the reverse features for further reduction of feature dimension.

XGBoost [19] is a machine learning algorithm based on lifting tree, which has high computational efficiency and the characteristics of preventing overfitting. XGBoost sums the results of several weak learners to be the final predicted value and uses gradient lifting decision tree (GBDT) algorithm to train the model. For the XGBoost model composed of k CARTs, the calculation method of its predicted value $\widehat{y}_i$ is shown in equation (8).

$$\widehat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ \ f_k \in \mathcal{F} \quad (8)$$

Among them, $\mathcal{F}$ is the function space of weak learner composed of all CART trees, and $f_k(x_i)$ represents the weight of the leaf node where the i-th sample is classified in the k-th tree.

The objective function of XGBoost algorithm $\mathcal{L}(\phi)$ consists of two parts, its calculation method is shown in equation (9). Among them, $l(y_i, \widehat{y}_i)$ is a loss function, which is used to describe the fitting degree between the predicted value and the actual value. The regular term $\Omega(f_k)$ is the penalty term of XGBoost, which is used to prevent over-fitting. The calculation method of $\Omega(f_k)$ is shown in equation (10).

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \Omega(f_k) \tag{9}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{10}$$

In equation (10), T is the number of leaf nodes in a regression tree, ω is the weight of leaf nodes, and λ, γ are parameters.

After the filtering stage, the feature set F consisting of the remaining typhoon data features is obtained. For each feature in F, it is sorted according to its NMI value from smallest to largest. Then the influence of each feature on the prediction results is evaluated according to the change of prediction error by using XGBoost model. After that, the features that contribute less to the prediction or even play a negative role are eliminated according to the evaluation result. This can effectively reduce the number of features while ensuring the accuracy of the prediction model. The specific steps are as follows:

Step 1. The sliding window method is used to divide the typhoon sample data composed of features in F. The sample data of the previous k moment are used as input, and the typhoon moving position (longitude x_{k+t} and latitude y_{k+t}) at $k+t$ moment are used as labels, and the XGBoost model is used for training and prediction. Euclidean distance is used as a measure of prediction effectiveness, and the corresponding prediction errors E_{pre} are calculated. The loop variable i is set to 1.

Step 2. Remove a feature f_i from F in order, divide the data again according to the method in step 1, train and test XGBoost model, calculate the prediction error E_{del}.

Step 3. If $E_{\text{del}} < E_{\text{pre}}$, delete f_i from F, assign $E_{\text{pre}} = E_{\text{del}}$, set $i = i + 1$, and Go back to step 2 and continue the execution. The number of features in F is reduced one by one until each feature has been screened once.

Algorithm 1 is the description of feature dimension reduction based on NMI and XGBoost. Among them, lines 1-5 calculate the NMI value between each feature and typhoon moving path, and lines 6-9 filter redundant features whose NMI value is lower than the threshold T. In lines 10-16, the features is sorted by their NMI value and the invalid features that play a reverse role in prediction are eliminated by XGBoost. Finally, the feature set after dimension reduction is returned.

Different from traditional feature selection methods, this algorithm filters features based on NMI in the first stage, and eliminates the features having a reverse role in prediction based on XGBoost in the second stage, which ensures

Algorithm 1 Feature Dimension Reduction Based on NMI and XGBoost

Input: F: feature dataset; L: typhoon track dataset; T: threshold of NMI;
Output: *dim_reduct_F*: result feature dataset

BEGIN
FOR each f_i in F **DO**
Calculate $H(f_i), H(x), H(y), H(f_i/x), H(f_i/y), I(f_i; x), I(f_i; y)$ according to equation (3)-equation (5)
Calculate NMI(f_i; x) and NMI (f_i; y) according to equation (6)
$\mathrm{NMI}_{ix} \leftarrow \mathrm{NMI}(f_i; x)$; $\mathrm{NMI}_{iy} \leftarrow \mathrm{NMI}(f_i; y)$
$\mathrm{NMI}\_ll[\,] \leftarrow \mathrm{NMI}_i = \mathrm{NMI}_{ix} + \mathrm{NMI}_{iy}$
END FOR
FOR each NMI_i in $\mathrm{NMI}\_ll[\,]$ **DO**
IF $\mathrm{NMI}_i < T$ **THEN**
delete NMI_i
END FOR
$\mathrm{NMI}\_ll[\,]$*.sort()*
FOR each f_i in F **DO**
$E_{\mathrm{pre}} \leftarrow XGB\_error(F)$
$E_{\mathrm{del}} \leftarrow XGB\_error(F - f_i)$
IF $E_{\mathrm{pre}} > E_{\mathrm{del}}$ **THEN**
Delete f_i
END FOR
dim_reduct_F = F
RETURN *dim_reduct_F*
END

that the feature set after dimension reduction has low redundancy and strong correlation. After the two stages of feature dimension reduction, the number of data features is greatly reduced, and the computational cost of the next stage of correlation analysis and feature selection is reduced.

3.3 Parallel distance correlation analysis and feature selection based on Spark

In the feature selection of typhoon trajectory data, it is necessary to analyze the correlation between the typhoon trajectory and other features, so as to select the features or the combinations of features with high correlation, which requires the analysis of linear and nonlinear correlation between clusters composed of different features. Obviously, the traditional correlation coefficient analysis method and canonical correlation analysis method are not competent. The distance correlation can not only describe the linear and nonlinear relationship between two variables, but also analyze the linear and nonlinear relationship between any feature combination without depending on any model assumptions. For this, DC is used to analyze the correlation between typhoon moving trajectory and the related features and complete the feature selection.

However, the computational complexity of DC is too high to result in poor analysis timeliness. Therefore, the distributed computing platform Spark is used to improve the parallelization of DC algorithm, which greatly improves the computational efficiency and realizes the feature selection. The correlation coefficient calculation method and the parallel feature selection method based on DC under Spark will be described in detail below.

3.3.1 Calculation Method of distance correlation

Distance Correlation coefficient is a measure of random vector correlation. The distance is mainly reflected in the difference between the joint characteristic function and the product of their respective marginal characteristic functions between any two variables [24]. The concepts of distance correlation and distance covariance provide a new method for correlation measurement and independence test, which can realize the correlation measurement and independence test of arbitrary random vectors. The distance correlation between two random vectors X and Y is expressed as $\mathrm{dCor}(X, Y)$. $\mathrm{dCor}(X, Y) = 0$ means that X and Y are independent of each other, the larger the value of $\mathrm{dCor}(X, Y)$, the stronger the correlation between X and Y.

The feature selection of typhoon trajectory related data is to find out the feature combination that can most affect typhoon moving trajectory, which can be realized by calculating the distance correlations between the feature combinations of typhoon sample data and the typhoon moving trajectory (denoted by longitude and latitude features). Here, we redefine the calculation equations of distance covariance, distance variance and distance correlation from the perspective of typhoon data samples, thus realizing the calculation of distance correlation matrix between feature combinations. If the typhoon feature data set $Z = \{f_j \mid (j = 1, 2, \ldots, m)\}$ is composed of m features and n samples, the feature subset is $F = \{s_i \mid (i = 1, 2, \ldots, s),\ s_i \in Z\}$ composed of $s(s<=m)$ features and the typhoon trajectory set is $L = \{x,\ y\}$, in which f_j, s_i, x, y are all non-empty subsets. The feature subset F is a $n * s$ matrix ($F \in R^{n*s}$) and the typhoon trajectory set L is a $n * 2$ matrix ($L \in R^{n*2}$). The number of columns in matrix F and L are different. Each sample can be regarded as a row vector of $s + 2$ dimensions. Next, the matrix calculation of distance correlation will be given.

Suppose that the any two observation samples of the typhoon feature set F are F_k and F_l, $k, l = 1,\ 2,\ \ldots,\ n$. Firstly, the Euclidean distance between the two samples is calculated, the calculation equation is shown in equation (11).

$$a_{kl}^{*} = \| F_k - F_l \| \tag{11}$$

Then, the center matrix is calculated based on the Euclidean distance between samples, and the equation is shown in equation (12).

$$A_{kl}^{*} = a_{kl}^{*} - \overline{a}_{k.}^{*} - \overline{a}_{.l}^{*} + \overline{a}_{..}^{*} \tag{12}$$

Where $\overline{a}^*_{k.}$, $\overline{a}^*_{.l}$ and $\overline{a}^*_{..}$ are the intermediate variables, their calculation are shown in equation (13), equation(14) and equation(15).

$$\overline{a}^*_{k.} = \frac{1}{n}\sum_{l=1}^{n} a^*_{kl} \tag{13}$$

$$\overline{a}^*_{.l} = \frac{1}{n}\sum_{k=1}^{n} a^*_{kl} \tag{14}$$

$$\overline{a}^*_{..} = \frac{1}{n^2}\sum_{k,l=1}^{n} a^*_{kl} \tag{15}$$

For typhoon trajectory set $L = \{x,\ y\}$, we can use the same method as above to calculate the Euclidean distance b^*_{kl} between any two samples k and l, i.e., $b^*_{kl} = \parallel L_k - L_l \parallel$, and then we can obtain the center matrix B^*_{kl}. i.e., $B^*_{kl} = b^*_{kl} - \overline{b}^*_{k.} - \overline{b}^*_{.l} + \overline{b}^*_{..}$. the intermediate variables $\overline{b}^*_{k.}$, $\overline{b}^*_{.l}$ and $\overline{b}^*_{..}$ are calculated in a similar way as $\overline{a}^*_{k.}$, $\overline{a}^*_{.l}$ and $\overline{a}^*_{..}$.

The sample distance covariance $\mathrm{dCov}^2_n(F, L)$ between features F and L is calculated in equation (16).

$$\mathrm{dCov}^2_n(F, L) = \frac{1}{n^2}\sum_{k,l=1}^{n} A^*_{kl}B^*_{kl} \tag{16}$$

Similarly, the sample distance variance $\mathrm{dVar}_n(F)$, $\mathrm{dVar}_n(L)$ of the features F and L are obtained, as shown in equation (17) and equation(18).

$$\mathrm{dVar}^2_n(F) = \mathrm{dCov}^2_n(F, F) = \frac{1}{n^2}\sum_{k,l=1}^{n} {A^*_{kl}}^2 \tag{17}$$

$$\mathrm{dVar}^2_n(L) = \mathrm{dCov}^2_n(L, L) = \frac{1}{n^2}\sum_{k,l=1}^{n} {B^*_{kl}}^2 \tag{18}$$

The calculation of the distance correlation $\mathrm{dCor}_n(F, L)$ between the features F and the typhoon trajectory feature L is shown in equation (19).

$$\mathrm{dCor}_n(F, L) = \begin{cases} \dfrac{\mathrm{dCov}_n(F, L)}{\sqrt{\mathrm{dVar}_n(F)\,\mathrm{dVar}_n(L)}}, & \mathrm{dVar}_n(F)\,\mathrm{dVar}_n(L) > 0 \\ 0 \quad , & \mathrm{dVar}_n(F)\,\mathrm{dVar}_n(L) = 0 \end{cases} \tag{19}$$

With the above calculation equations, the distance correlation between each combination of typhoon features and typhoon trajectory can be calculated. Based on the results, the combinations of data features with the strongest correlation are selected as the features of the typhoon prediction model, and the selection of typhoon trajectory related features can be realized.

3.3.2 Parallel Feature Selection Method Based on DC under Spark

The distance correlation calculation equation given in the previous section can better depict the nonlinear correlation between the two types of feature combinations, so it can more accurately analyze the nonlinear correlation between typhoon data features and typhoon moving position (labeled by longitude and latitude features), thus realizing the selection of typhoon trajectory related data features. However, the complexity of the above calculation equations are very high, which affects the efficiency of feature selection methods. Therefore, we design and implement a parallel distance correlation analysis and feature selection algorithm based on Spark framework, which greatly improves the efficiency of feature selection algorithm. Since all features are involved in the calculation of distance correlation, and its norm matrix and center matrix need to be calculated for each group of features, the calculation efficiency can be improved through parallel processing. The framework of the parallel feature selection method based on DC under Spark is shown in Figure 3.

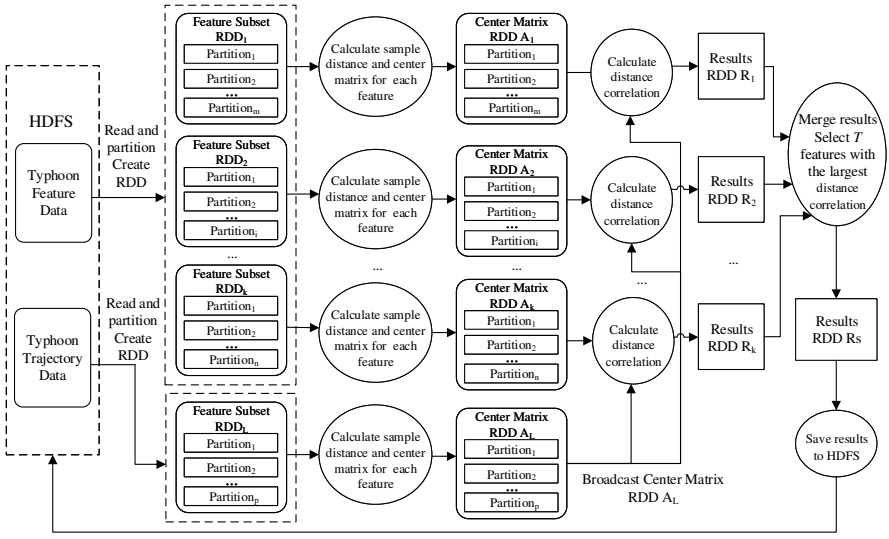

Fig. 3 The framework of the parallel feature selection method based on DC under Spark

It can be seen from Figure 3, the parallel feature selection method proposed in this paper is mainly divided into four stages, as follows:

Stage1 creates a SparkContext, reads the typhoon feature dataset, divides the data set into sub-datasets composed of different data features according to the feature combinations, and forms the corresponding RDDs respectively. Each RDD is composed of corresponding data partitions, each of which is composed of mann data samples; reads typhoon trajectory dataset L composed

of typhoon positions (expressed by longitude and latitude), and forms corresponding RDD_L composed of many partitions, each of which is composed of many samples.

Stage2 Calculates the distance norm matrix on each data partition, merges results on each partition, calculates the corresponding center matrices RDD A_1, RDD A_2, RDD A_k and RDD A_L, Broadcasts the central matrix RDD A_L to other tasks.

Stage3 Aggregates RDD A_L with RDD A_1, RDD A_2 ,..., RDD A_K formed by each combination of features respectively, and calculates distance correlations to form result sets RDD R_1, RDD R_2, ..., RDD R_k.

Stage4 merges the result sets generated in the previous stage and performs feature selection based on the results. The combinations of features whose distance correlation are greater than a given threshold value T are selected as the features of the prediction model, or the k combinations of features with the largest distance correlation are selected as the features of the prediction model. The final results are stored to RDD Rs and outputted to the HDFS file system.

The selection of the threshold T can be calculated by the method in reference [25], as shown in equation (20).

$$t = [\left(\frac{N}{\log N}\right)^{\frac{4}{5}}] \tag{20}$$

Where, N is the number of features and $[\cdot]$ represents rounding.

4 Performance Evaluation

In order to verify the effectiveness of the proposed method NX-Spark-DC, we conducted a series of experiments on real datasets and compared it with the existing methods based on correlation analysis such as Pearson correlation coefficient, NMI and HSIC0. In the following, we will give the specific experimental dataset, experimental environment and result analysis.

4.1 Dataset

The experimental data in this paper are the Best Track typhoon best path dataset provided by the China Meteorological Administration (CMA) and the 6-hourly reanalysis dataset provided by the National Centers for Environmental Prediction (NCEP). The time range covered by the datasets is from July to October of each year from 1990 to 2018. It contains three parts: typhoon movement trajectory data, atmospheric environment data and ocean environment field data generated by the Northwest Pacific and the South China Sea (SCS), with a total of 302 features and 3,428 samples.

4.2 Environment and Metrics

The experimental environment consists of a Spark cluster consisting of five IBM PC rack servers, one of which is a management node and the rest is a compute node. Each server has an E5-2620 CPU (6 cores, 2.0 GHz), 32GB of memory, and 6TB of hard disk. Every server is installed with CentOS 7.0, Python 3.6 and corresponding algorithm modules.

We evaluate the proposed methods on general metrics, including MSE (Mean Square Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), SMAPE (Symmetric Mean Absolute Percentage Error), and R^2 (R-Square).

(1) MSE

MSE is the expected value of the square of the difference between the predicted value and the true value, which is often used as a loss function of regression models. Its value range of $[0, +\infty]$, The calculation is shown in equation (21). The smaller the MSE is, the better the representative effect is. When the value of MSE is 0, the prediction effect is excellent.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2 \tag{21}$$

(2) RMSE

RMSE is often used as a measure of the model prediction results, its calculation method is shown in equation (22). The smaller the RMSE is, the better the representative effect is. When RMSE is 0, the prediction effect is the best.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2} \tag{22}$$

(3) MAE

MAE is calculated as shown in equation (23), and its value range is $[0, +\infty]$. The smaller the MAE is, the better the representative effect is. When MAE is 0, it means that the prediction effect is excellent.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y}_i - y_i| \tag{23}$$

(4) MAPE

MAPE is calculated as shown in equation (24), and its value range is $[0, +\infty]$. When MAPE is 0%, it means that the prediction result is completely correct. When MAPE is greater than 100%, it means that the model is poor.

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y}_i - y_i}{y_i}\right| \tag{24}$$

(5) SMAPE

SMAPE is calculated as shown in equation (25) and its value range is $[0, +\infty]$. When SMAPE is 0%, it means that the prediction result is completely correct. When SMAPE is greater than 100%, it means that the model is poor.

$$\mathrm{SMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|\widehat{y}_i - y_i|}{\frac{(|\widehat{y}_i| + |y_i|)}{2}} \tag{25}$$

(6) R^2

The calculation method of R^2 is shown in equation (26), and the value range is $[0, +\infty]$. It reflects how much the change of the independent variable can explain the change of the dependent variable. The smaller R^2 is, the lower the interpretation degree is, that is, the worse the prediction effect is.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\widehat{y}_i - y_i)^2}{\sum_{i=1}^{n} (\overline{y}_i - y_i)^2} \tag{26}$$

4.3 Experimental results

The proposed method NX-Spark-DC is experimented with the above real datasets, compared with the existing correlation analysis based methods such as Pearson, NMI, and HSIC0, and verified by XGBoost and SVR respectively. The parameter setting of the used model and the analysis of experimental results show in detail below.

4.3.1 Model Parameter Settings

According to the existing experience and knowledge, we first set several groups of parameters. After that, we design a series of experiments to compare the model's performance under different parameters. Finally, we select a group of parameters with the best dimension reduction effect and fitting ability. the specific parameter settings are as follows:

(1) XGBoost model parameters in the feature dimension reduction stage: The training process is stopped after 30 iterations, the maximum depth is 5, the Gamma value is set to the default value, the random sampling ratio is set to 0.8, the regularization parameters are set to 1 and 0.8 respectively, and the learning rate is 0.3.

(2) XGBoost model parameters in the experimental verification stage: the learning rate is XGBoost mode is adjusted to 0.1, other parameters are consistent with those in the factor dimension reduction stage.

(3) SVR model parameters in the experimental verification stage: the penalty function is set to 1, the designated kernel function is Gaussian kernel (RBF), the Gamma is set to "auto," and other parameters use default values.

4.3.2 Result Analysis

(1) Efficiency

To verify the computational efficiency of NX-Spark-DC algorithms, we compare it with the original distance correlation method DC. Figure 4 shows that the execution time of the two algorithms varies with the increase of data volume when the number of Executors is 36. As can be seen from the figure, the execution time of both algorithms increases with the amount of data, which is as expected. The execution time of DC algorithm grows faster, while the execution time of NX-Spark-DC algorithm has a relatively flat growth trend, which is much lower than that of DC algorithm. Obviously, the parallel algorithm NX-Spark-DC in this paper outperforms the traditional DC serial algorithm. This also indicates that the parallel algorithm in this paper is effective and can significantly improve the computational efficiency of the distance correlation.

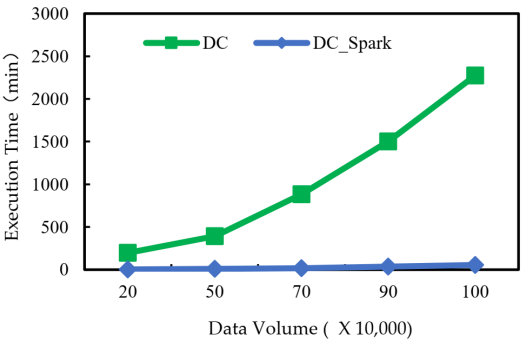

Fig. 4 Comparison of the execution time of the two algorithms with the data volume

Figure 5 shows the change of execution time with the number of executors when the dataset size is 700,000. As you can see from Figure 5, the execution time of the algorithm NX-Spark-DC continues to decrease as the number of executors increases, which is in line with expectations. However, the decrease in execution time slowly slows down with the increase of the number of executors, which is due to the increase in the number of executors, resulting in an increase in the cost of communication between executors, thereby slowing down the decline in execution time, so the number of executors must be appropriate.

(2) Accuracy

In order to further verify the accuracy of the NX-Spark-DC method, we use the NX-Spark-DC method to select the features of the typhoon trajectory prediction on the real data set, and compare it with the representative correlation analysis based methods Pearson, NMI, DC and HSIC0. Each method select 35 features from 300 typhoon features to fit and predict the typhoon moving trajectory.

Table 1 shows the scores of several methods on each index when using XGBoost as a forecasting model. On the basis of the original DC algorithm, the NX-Spark-DC method adds two stages of dimension reduction, the filtering and the screening, and realizes parallelization processing. It can be seen

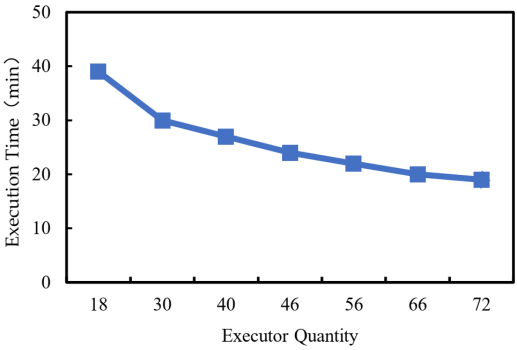

Fig. 5 Change of execution time with the number of executors

Table 1 Comparison of five methods under XGBoost

Method	MSE	RMSE	MAE	MAPE	SMAPE	R^2
NMI	2.79E-04	0.0167	0.0092	4.6319	4.5917	0.8846
Pearson	8.09E-06	0.0028	0.0017	0.9704	0.9694	0.9875
DC	1.15E-05	0.0034	0.0025	1.8102	1.8057	0.7751
HSIC0	7.99E-06	0.0028	0.0018	1.0664	1.0657	0.9801
NX-Spark-DC	**4.23E-06**	**0.0020**	**0.0012**	**0.7555**	**0.7551**	**0.9902**

from the results in Table 1 that the NX-Spark-DC method is obviously better than the traditional DC method in all indicators. This also shows that it is necessary to increase the dimension reduction processing stage, which can significantly improve the accuracy of feature selection. In general, the NX-Spark-DC method is superior to other methods. This is because the method makes full use of the two-stage dimension reduction processing of filtering and screening, so as to eliminate the features of weak correlation and playing a reverse effect on prediction, and adopts the feature selection method based on distance correlation, so the features with high correlation are selected, and the accuracy of prediction is improved.

To avoid losing generality, we take 35 features screened by the above five analysis methods as inputs and use the SVR machine learning model to predict typhoon moving trajectory. The scores of the five methods are shown in Table 2.

Table 2 Comparison of five methods under SVR

Method	MSE	RMSE	MAE	MAPE	SMAPE	R^2
NMI	0.0011	0.0327	0.0218	13.0511	11.6779	0.0582
Pearson	0.0010	0.0317	0.0213	12.9174	11.4888	0.0854
DC	0.0010	0.0321	0.0221	13.3889	11.8563	0.0730
HSIC0	0.0010	0.0324	0.0218	13.2135	11.7643	0.0701
NX-Spark-DC	**0.0010**	**0.0310**	**0.0209**	**12.7535**	**11.3593**	**0.1014**

It can be observed from Table 2 that the performance of the five algorithms is close when SVR is used as a prediction model. Nevertheless, NX-Spark-DC is still the best for the same reasons as above. It shows that the feature selection method proposed in this paper has good accuracy and is superior to the other four correlation analysis based methods. It can be seen from Table 1 and Table 2 that the prediction accuracy of SVR is far less than that of XGBoost, which shows that the XGBoost model is more suitable for prediction under large data volumes.

(3) The influence of the factor dimension reduction stage

To further verify the influence of factor dimension reduction on several feature selection methods based on the correlation analysis, we compared the feature selection effects of four correlation analysis methods before and after adding feature dimension reduction stage. The prediction model adopts XGBoost and the specific results are shown in Table 3. In this experiment, NMI, Pearson, DC, and HSIC0 methods use data sets without dimension reduction processing and NX_NMI, NX_Pearson, NX_HSIC0 and NX-Spark-DC use the data set after dimension reduced processing.

Table 3 Comparison of feature dimension reduction under XGBoost

Method	MSE	RMSE	MAE	MAPE	SMAPE	R^2
NMI	3.01E-04	0.0173	0.0097	4.9954	4.9491	0.8619
NX_NMI	2.79E-04	0.0167	0.0092	4.6313	4.5917	0.8846
Pearson	8.09E-06	0.0028	0.0017	0.9704	0.9694	0.9875
NX_Pearson	7.27E-06	0.0027	0.0015	0.8221	0.8216	0.9892
HSIC0	7.99E-06	0.0028	0.0018	1.0664	1.0657	0.9801
NX_HSIC0	6.74E-06	0.0026	0.0015	0.9162	0.9143	0.9873
DC	1.15E-05	0.0034	0.0025	1.8102	1.8057	0.7751
NX-Spark-DC	**4.23E-06**	**0.0020**	**0.0012**	**0.7555**	**0.7551**	**0.9902**

The prediction accuracy under XGBoost has been improved after reducing dimension. NMI, Pearson, and HSIC0 correlation analysis methods have improved the prediction accuracy after using feature dimension reduction processing. However, the the improvement is not particularly obvious, which may be related to the data set. Compared with the DC method, the accuracy of the NX-Spark-DC is improved obviously, which shows that the dimension reduction processing is suitable for the DC method. Through the above comparison, features screened by the NX-Spark-DC have an excellent prediction effect and is the best among several methods.

When SVR is used as the prediction model, the scores of the several methods are shown in Table 4. The scores of the four methods have increased after adding dimension reduction processing, which shows that the standard dimension reduction method can improve the feature selection accuracy of various methods to some extent. Nevertheless, the improvement range is not significant. Overall, the performance of the NX-Spark-DC method is still the best. At the same time, it can be seen from the comparison between Table 3 and

Table 4 Comparison of feature dimension reduction under SVR

Method	MSE	RMSE	MAE	MAPE	SMAPE	R^2
NMI	0.0011	0.0327	0.0218	13.0511	11.6779	0.0582
NX_NMI	0.0011	0.0325	0.0219	13.0553	11.7292	0.0632
Pearson	0.0010	0.0317	0.0213	12.9174	11.4888	0.0854
NX_Pearson	0.0010	0.0314	0.0212	12.8603	11.4539	0.0915
HSIC0	0.0010	0.0324	0.0218	13.2135	11.7643	0.0701
NX_HSIC0	0.0010	0.0310	0.0209	12.7535	11.3593	0.1014
DC	0.0010	0.0321	0.0221	13.3889	11.8563	0.0730
NX-Spark-DC	**0.0010**	**0.0302**	**0.0202**	**12.3516**	**11.0501**	**0.1202**

Table 4 that the prediction accuracy of XGBoost is obviously better than that of SVR.

5 Conclusion

In this paper, we introduced the distance correlation analysis method into the typhoon trajectory correlation analysis and feature selection for the first time, and proposed a parallel feature selection method NX-Spark-DC based on NMI-XGBoost and distance correlation for typhoon trajectory prediction. The NX-Spark-DC method first uses NMI to filter out the useless features of the original data set, and then uses the XGBoost model to eliminate the counterproductive features, thus realizing the dimension reduction of the original data. On this basis, an improved parallel feature selection method based on distance correlation under Spark is proposed to realize the feature selection of typhoon trajectory related data. Experimental results on real data sets show that the parallel feature selection method proposed in this paper has high computational efficiency and is obviously superior to existing methods in accuracy. As the matter of fact, the method proposed in this paper is still preliminary. Although the NX-Spark-DC method supports multi-feature combination correlation analysis and feature selection, due to the limitations of the experimental environment, these experiments is mainly based on the correlation analysis of a single feature and typhoon moving trajectory to achieve feature selection. The follow-up work will increase the experimental research on multi-feature combination and optimize our method to further improve the efficiency and accuracy.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (No. 2019YFB1405302 and 2016YFC1401900) and the National Natural Science Foundation of China (No. 61872072 and 61073063).

Data availability. The data sets generated and analysed during the current study is available at: https://github.com/qiaoby/Typhoon-related-data.

Declarations

Conflict of interest: The authors have no conflicts of interest to declare that are relevant to the content of this article.
Ethical approval: This article does not contain any studies with human participants performed by any of the authors.
Author Contributions: Conceptualization: B.Q. and R.W.; methodology: B.Q., Y.H., and R.W.; validation: Y.H., R.W. and P.W.; formal analysis: B.Q., Y.H., R.W, G. W.; data curation: Y.H. and P.W.; Writing - original draft preparation: B.Q., Y.H and R.W.; writing - review and editing: B.Q. and Y.H., G.W, D.H; visualization: Y.H., R.W., and D.H.; All authors read and approved the manuscript.

References

[1] Yab, L.Y., Wahid, N., Hamid, R.A.: A modified whale optimization algorithm as filter-based feature selection for high dimensional datasets. In: Ghazali, R., Mohd Nawi, N., Deris, M.M., Abawajy, J.H., Arbaiy, N. (eds.) Recent Advances in Soft Computing and Data Mining, vol. LNNS457, pp. 90–100. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-00828-3_9

[2] Sankar, K., Uma Maheswari, P.: A dynamic wrapper-based feature selection for improved precision in content-based image retrieval. Concurrency and Computation: Practice and Experience **34**(8), 5368 (2019). https://doi.org/10.1002/cpe.5368

[3] Liu, W., Wang, J.: Recursive elimination–election algorithms for wrapper feature selection. Applied Soft Computing **113**, 107956 (2021). https://doi.org/10.1016/j.asoc.2021.107956

[4] Fu, Y., Liu, X., Sarkar, S., Wu, T.: Gaussian mixture model with feature selection: An embedded approach. Computers & Industrial Engineering **152**, 107000 (2021). https://doi.org/10.1016/j.cie.2020.107000

[5] Jiménez-Cordero, A., Morales, J.M., Pineda, S.: A novel embedded min-max approach for feature selection in nonlinear support vector machine classification. European Journal of Operational Research **293**(1), 24–35 (2021). https://doi.org/10.1016/j.ejor.2020.12.009

[6] Thejas, G.S., Garg, R., Iyengar, S.S., Sunitha, N.R., Badrinath, P., Chennupati, S.: Metric and accuracy ranked feature inclusion: Hybrids of filter and wrapper feature selection approaches. IEEE Access **9**, 128687–128701 (2021). https://doi.org/10.1109/ACCESS.2021.3112169

[7] Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classifications, pp. 2–34. Springer, New York, NY (1979). https://doi.org/10.

1007/978-1-4612-9995-0_1

[8] Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classifications. II: Further Discussion and References, pp. 35–75. Springer, New York, NY (1979). https://doi.org/10.1007/978-1-4612-9995-0_2

[9] Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications, iv: Simplification of asymptotic variances. Journal of the American Statistical Association **67**, 415–421 (1972). https://doi.org/10.1007/978-1-4612-9995-0_4

[10] C., S.: The proof and measurement of association between two things. Int J Epidemiol. **39**(5), 1137–1150 (2010). https://doi.org/10.1093/ije/dyq191

[11] Yang, X., Liu, W., Liu, W., Tao, D.: A survey on canonical correlation analysis. IEEE Transactions on Knowledge and Data Engineering **33**(6), 2349–2368 (2021). https://doi.org/10.1109/TKDE.2019.2958342

[12] Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science **334**, 1518–1524 (2011)

[13] Bickel, P.J., Levina, E.: Regularized estimation of large covariance matrices. The Annals of Statistics **36**(1), 199–227 (2008). https://doi.org/10.1214/009053607000000758

[14] Cai, T.T., Zhou, H.H.: Optimal rates of convergence for sparse covariance matrix estimation. The Annals of Statistics **40**(5), 2389–2420 (2012). https://doi.org/10.1214/12-AOS998

[15] Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. The Annals of Statistics **35**(6), 2769–2794 (2007). https://doi.org/10.1214/009053607000000505

[16] HUANG Yiwu, G.S., Shuanzhu, G.: Correlation and regression analysis of typhoon forecast errors and ambient variables by t639. Meteorological Monthly **42**(12), 1506–1512 (2016). https://doi.org/10.7519/j.issn.1000-0526.2016.12.008

[17] Xiaoyuan, C., Shanxian, Y., Hanhui, L.: The application of canonical correlation analysis in the prediction of typhoon tracks. Journal of Tropical Meteorology **1987**(4), 328–332 (1987)

[18] Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with hilbert-schmidt norms. In: Jain, S., Simon, H.U., Tomita, E. (eds.) Algorithmic Learning Theory, pp. 63–77. Springer,

Berlin, Heidelberg (2005). https://doi.org/10.1007/11564089_7

[19] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785

[20] Xiaoya, H.: An artificial intelligence prediction model based on principal component analysis for typhoon tracks. Chinese Journal of Atmospheric Sciences **37**(5), 1154–1164 (2013)

[21] Vinh, L.T., Lee, S., Park, Y.-T., d'Auriol, B.J.: A novel feature selection method based on normalized mutual information. Applied Intelligence **37**(1), 100–120 (2012)

[22] Yu, M., Cai, X., Song, Y., Wang, X.: A fast forecasting method for pm2.5 concentrations based on footprint modeling and emission optimization. Atmospheric Environment **219**, 117013 (2019). https://doi.org/10.1016/j.atmosenv.2019.117013

[23] Wen, C., Yang, Y., Xiao, Q., Huang, M., Pan, W.: Genome-wide association studies of brain imaging data via weighted distance correlation. Bioinformatics **36**(19), 4942–4950 (2020). https://doi.org/10.1093/bioinformatics/btaa612

[24] Zhang, L., Kong, L., Chen, H.: HIERARCHICAL CLUSTERING VIA DISTANCE CORRELATION. Mathematica Numerica Sinica **41**(3), 320–334 (2019). https://doi.org/10.12286/jssx.2019.3.320

[25] Zhang, X., Liu, L., Xinyao, G.: Measurement of Nonlinear Correlation Coefficient among Classes Based on HSIC0. Computer Engineering and Applications **55**(3), 46–49 (2019)