Gai and Zhang
EURASIP Journal on Advances in Signal Processing (2023) 2023:21
https://doi.org/10.1186/s13634-023-00973-9

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Prediction model of agricultural water quality based on optimized logistic regression algorithm

RongLi Gai* and Hao Zhang

*Correspondence:
gairongli@dlu.edu.cn

Department of Information Engineering, Dalian University, Dalian, China

Abstract

Aiming at the problem that the linear regression of the traditional linear water quality prediction model is not robust enough and the prediction accuracy cannot be guaranteed in the presence of interference, this paper proposes an agricultural water quality prediction model based on the Momentum algorithm to optimize the logistic regression algorithm (LRM algorithm). The model uses the Momentum algorithm to optimize the logistic regression algorithm to quickly adjust the misclassified samples. When the object encounters a local optimum in the process of falling, the introduction of momentum makes it easy for the next update to jump out of the local optimum with the help of the last large gradient. In this paper, the performance of the proposed model is evaluated on 4 real data sets. The experimental results show that the LRM algorithm proposed in this paper improves the prediction accuracy of the existing algorithm by an average of 1.11 percentage points.Compared with KNN and other traditional prediction algorithms, LRM not only speeds up the convergence rate of the algorithm, but also reduces the steady-state error and improves the prediction accuracy of water quality, suitable for data mining of complex water quality data, The experiment verifies the feasibility of this method in predicting the actual agricultural water quality and even in predicting and warning the residents drinking water.

Keywords: Momentum algorithm, Logistic regression, Water quality prediction

1 Introduction

In recent years, with the development of agricultural intelligence and modernization in China, the prediction of agricultural water quality directly affects the growth of agricultural products, and it is extremely urgent to propose an efficient agricultural water quality prediction method [1].

Around the 1980s, the water quality prediction theory and related models began to be introduced into China. Water quality prediction is an inevitable requirement of my country's development. With the development of a new generation of information technology, more and more scholars use intelligent algorithms to build water quality prediction models. For example, Li Na [2] proposed a combined grey system (GM) The water quality prediction model based on the principle of metabolism and the principle of water

quality can predict the development trend of river water quality by training data such as potassium permanganate index and NH3-N. Although the model is simple and easy to use, the learning ability of the model is limited and cannot capture nonlinear information. limits the predictive power of the model. Ju et al. [3] established a least squares support vector machine model for predicting ammonia nitrogen content in water, and applied weighted least squares support vector machine (WLS-SVM) to ammonia nitrogen content analysis, although it has a better prediction learning speed, However, its prediction robustness is significantly reduced, and the prediction data is more biased. Liu et al. [4] combined the genetic algorithm and BP neural network algorithm to predict the water quality of the Potomac River in the United States in real time, and made real-time prediction of its water quality parameters turbidity (TURB) and conductivity (SC), and analyzed the performance of the prediction results. In order to verify the accuracy and reliability of the river water quality prediction model based on Genetic algorithm optimization of BP neural network(IGA-BPNN), the interval coverage of the IGA-BPNN model for the prediction results of water quality parameters TURB and SC under normal stationary conditions is 99.81% and 100%, respectively. The prediction results have certain reliability, but water quality prediction involves a multi-factor nonlinear relationship, and the neural network needs a large amount of training sample data [5]. The disadvantage of this method is that in order to achieve a considerable prediction accuracy, complex feature engineering is required. , which improves the difficulty of the model's landing application. Traditional intelligent algorithms have more or less limitations in terms of accuracy, convergence speed and applicability [6]. For this reason, this paper proposes a water quality prediction model (LRM algorithm) based on the Momentum algorithm to optimize the logistic regression algorithm.

Logistic regression is a mathematical concept proposed in 1993, originally to solve the problem of binary classification [7]. On the basis of linear regression, logistic regression constructs a classification model. Since the curve grows faster near the center and slower at both ends [8], this feature will make it possible to quickly adjust the misclassified samples when optimizing the logistic regression algorithm using the Momentum algorithm. On the basis of linear regression, logistic regression adds a layer of nonlinear mapping to the mapping from features to results, that is, the features are summed linearly first, and then the sigmoid function is used to make predictions [9].

In this paper, the data collected by the sensor is used for the experiment. First, the hyperparameter values of the Momentum algorithm are analyzed, and the superiority of the Momentum algorithm is verified. Then, the predicted data of the LRM model is compared with the actual data, and compared with multiple models. A comparative analysis was carried out to verify the reliability and accuracy of the LRM model.

2 Material and methods

2.1 Data set collection

The experiment collects the monitoring data of water quality indicators in Dalian, Liaoning Province from January 2021 to December 2021, installs the smart terminal on the sensor, and collects various data such as temperature, residual chlorine value, and pH value required in the project. In the process of water quality data collection, due to the complexity and change of water quality data, machine failure and data packet loss are

prone to occur in the process of collecting drinking water quality, resulting in continuous or intermittent data loss. Additionally, there may be problems with the server that collects the data, resulting in the loss of data files. Residents have high requirements for drinking water quality, and the quality of drinking water changes with time, resulting in large deviations in water quality collection equipment. Since the water quality collection equipment collects streaming data, the data is continuous and uninterrupted, so data skew is prone to occur during the water quality collection process, resulting in high data outliers. From the data collected from residents' drinking water, it can be seen that there are mainly two types of missing values and outliers in drinking water quality data. Missing values refer to data that have not been collected over a period of time or that have zero values in the data; outliers refer to data with skewed data that deviates significantly from other data. Due to the large amount of drinking water quality data and many parameters, abnormal data may be accompanied by not only one abnormal feature, but also multiple abnormal features. The data set collection is shown in (Fig. 1).

2.2 Dataset preprocessing

The experiment starts by cleaning the collected data. If the deletion rate of the variable is high (greater than 80%), the coverage rate is low, and the importance is low, the missing values in the data can be deleted directly. For data outliers, we use the 3σ criterion to identify data outliers. If the sample is normally distributed or approximately normally distributed, more than 99% of the data are considered to be within 3 standard deviations

id	variableName	dataPointId	err	slaveName	time	value
98853	Temperature	2415028	0	Tem	2021-12-17 19:13:54	12.8
98854	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:13:54	0
98855	Residual chlorine	2415030	0	Reschl	2021-12-17 19:13:55	4
98856	PH	2415027	0	PH	2021-12-17 19:14:57	7.37
98857	Temperature	2415028	0	Tem	2021-12-17 19:14:58	12.8
98858	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:14:58	6
98859	Residual chlorine	241503	0	Reschl	2021-12-17 19:14:59	2
98860	PH	2415027	0	PH	2021-12-17 19:16:05	7.39
98861	Temperature	2415028	0	Tem	2021-12-17 19:16:07	12.8
98862	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:16:09	5
98863	Residual chlorine	2415030	0	Reschl	2021-12-17 19:16:10	0
98864	PH	2415027	0	PH	2021-12-17 19:17:13	7.37
98865	Temperature	2415028	0	Tem	2021-12-17 19:17:13	12.8
98866	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:17:13	5
98867	Residual chlorine	2415030	0	Reschl	2021-12-17 19:17:13	1
98868	PH	2415027	0	PH	2021-12-17 19:18:16	7.37
98869	Temperature	2415028	0	Tem	2021-12-17 19:18:16	12.8
98870	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:18:16	3
98871	Residual chlorine	2415030	0	Reschl	2021-12-17 19:18:17	4
98872	PH	2415027	0	PH	2021-12-17 19:19:24	7.37
98873	Temperature	2415028	0	Tem	2021-12-17 19:19:24	12.8
98874	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:19:24	3
98875	Residual chlorine	2415030	0	Reschl	2021-12-17 19:19:25	1
98876	PH	2415027	0	PH	2021-12-17 19:20:25	7.37
98877	Temperature	2415028	0	Tem	2021-12-17 19:20:26	12.8
98878	Dissolved oxygen	2415029	0	Disoxy	2021-12-17 19:20:26	3

Fig. 1 Data collection graph

of the upper and lower mean. Specifically, the probability that this value is distributed in $(\mu - 3\sigma, \mu + 3\sigma)$ is 99.73%, and the data whose maximum or minimum value exceeds this range are outliers, and outliers are excluded. In the training of the tree model, the tree model has high robustness to extreme values, and there is no information loss, which will not affect the training effect of the model.

In this paper, the cleaned data is pushed to the cloud server, and we label the obtained water quality data set to determine the ownership Category, for example, given the average value of a certain parameter, if the parameter in the water quality data is greater than the corresponding average value, the parameter belongs to a positive sample, otherwise it belongs to a negative sample; use the Momentum algorithm to optimize the logistic regression algorithm to establish a discriminant The model is used for water quality data analysis; the labeled water quality data is input into the discriminant model, and the discriminant model outputs the predicted water quality category; the predicted water quality category is compared with the actual water quality category, if the prediction accuracy is greater than or equal to Threshold, the water quality prediction framework of LRM is obtained; if the prediction accuracy is less than the threshold, go back to step 2 and use the Momentum algorithm to continue to find the minimum value of the maximum likelihood function; input the water quality data to be analyzed into the water quality prediction framework to obtain predictions result.

2.3 Research on logistic algorithm

The core function of the logistic regression algorithm is the Sigmoid function g(z)=$\frac{1}{1+e^{-z}}$, the positive or negative of Z in a determines whether the value of $g(z)$ is greater than 0.5 or less than 0.5 [10] . That is, when Z is greater than 0, $g(z)$ is greater than 0.5, and when Z is less than 0, $g(z)$ is less than 0.5. When the expression corresponding to Z is the classification boundary, it happens that the two sides of the classification boundary correspond to different positive and negative Z, which means that the two points of the classification boundary correspond to $g(z)$ 0.5 and $g(z) < 0.5$ respectively. Therefore, according to the size relationship with 0.5, the classification can be realized [11] .

Although the logistic regression algorithm is robust to noise in the data, it is not particularly affected by slight multicollinearity, and the computational cost is not high, and it is easy to understand and implement, but the logistic regression algorithm is prone to occur. In the case of underfitting, the classification accuracy is not high, and when the data features of the dataset are missing or the data feature space is large, the prediction effect is not good. Therefore, in the next section, it is proposed to use the Momentum algorithm to optimize the logistic optimization. The regression algorithm can make up for the problems existing in the logistic regression algorithm itself.

3 LRM model research

Compared with the traditional deep learning algorithm, the Momentum algorithm introduces the concept of Momentum in physics in principle, and has been very popular in the field of deep learning optimization in the past two years. The training of the traditional gradient descent method is jagged, which greatly lengthens the training time of the model. At the same time, due to the oscillation phenomenon, the learning

rate can only be set small, so as not to deviate from the minimum value because the pace is too large. The traditional gradient descent method is shown in (Fig. 2):

First, let the gradient vector of the t-th iteration be $w^t = \left[w_1^t, w_2^t, w_3^t\right]^T$, This can be done simply by summing the historical values of each component,and use w_{ave}^t as the new gradient update direction. The calculation formula is as formula (1):

$$\begin{bmatrix} w_1^1 + w_1^2 + \cdots + w_1^t \\ w_2^1 + w_2^2 + \cdots + w_2^t \\ w_3^1 + w_3^2 + \cdots + w_3^t \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{t-1} w_1^i \\ \sum_{i=0}^{t-1} w_2^i \\ \sum_{i=0}^{t-1} w_3^i \end{bmatrix} = w_{ave}^t \tag{1}$$

The method of formula (1) is relatively simple, but the calculation weight of each round of gradient is the same, so that the early gradient almost loses the prediction of the overall direction of the gradient, the early gradient jitter is more serious, and the recent gradient jitter is weaker, when the weight is the same, Gradient overall direction prediction may not be accurate; in the late stage of gradient descent, the parameter search space is basically a convex set, the magnitude and direction of the gradient are basically fixed, and the components with basically fixed size and direction are continuously accumulated, and the gradient will become very large, resulting in the least local convergence.

The Momentum algorithm performs a weighted average of the gradient vectors over a period of time, before each update, add part of the previous gradient, so that the entire gradient direction is not too random, because the oscillation directions cancel each other in the accumulation process, suppose that the nth gradient g is - 1 and the gradient g' calculated by the nth+1 is 1, then after the accumulation, the gradient will become 0 when the two update the weight again, instead of updating the weight in the direction where the gradient g is - 1, and then updating the weight in the direction where the gradient g is 1., and then updating the weights in the direction where the gradient g is 1.and calculates the approximate direction of the gradient during the gradient update process, which eliminates uncertain factors such as the swing phenomenon to a certain extent, and makes the gradient update move towards a clearer direction.

The summation average in formula (1) is optimized as an exponential moving weighted average, and the decay rate $\beta \in (0, 1)$ is added, and v'_w is used as the update direction of the gradient. This method makes the earlier gradient weights nearly 0 and the most recent gradient weights close to 1. The calculation formula is as formula (2):

$$\begin{bmatrix} \beta^{t-1} w_1^1 + \beta^{t-2} w_1^2 + \cdots + \beta^0 w_1^t \\ \beta^{t-1} w_2^1 + \beta^{t-2} w_2^2 + \cdots + \beta^0 w_2^t \\ \beta^{t-1} w_3^1 + \beta^{t-2} w_3^2 + \cdots + \beta^0 w_3^t \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{t-1} \beta^i w_1^i \\ \sum_{i=0}^{t-1} \beta^i w_2^i \\ \sum_{i=0}^{t-1} \beta^i w_3^i \end{bmatrix} = v'_w \tag{2}$$

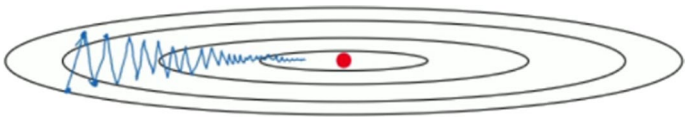

Fig. 2 The traditional gradient descent method training diagram

Through the exponential moving weighted average of formula (2), the vertical components can basically be offset, and the direction is basically reversed due to the existence of one-up and one-down pairing vectors in the jagged shape. From a long-term perspective, the general direction of gradient optimization always points to the minimum value, so the horizontal update direction is basically stable. The trajectory of the new gradient update is shown in (Fig. 3).

It can be seen from Fig. 3 that the Momentum method makes a weighted average of the gradient of the past time step, and the weight decreases exponentially according to the time step. Momentum method makes the update of independent variables of adjacent time steps more consistent in direction.

3.1 Optimizing the logistic regression algorithm

Based on the Momentum algorithm to optimize the logistic regression algorithm, the agricultural water quality prediction method uses the Momentum algorithm to optimize the logistic regression algorithm. Since the curve grows faster near the center and slower at both ends [12], this feature will make it possible to quickly adjust the misclassified samples when optimizing the logistic regression algorithm using the Momentum algorithm. Essentially, the Momentum method is like pushing a ball from a high slope. The ball accumulates Momentum during the downward rolling process, and it will become faster and faster on the way. Finally, it will reach a peak value. Correspondingly, in our algorithm, the Momentum term will increase in the same direction of the gradient direction, and the direction of the gradient direction change will gradually decrease, resulting in faster convergence speed and less vibration.When an object encounters a local optimum in the process of falling, the introduction of Momentum can make the object rush out of the local optimum on the basis of the original Momentum. Furthermore, ordinary gradient descent is completely determined by the gradient, which can lead to severe oscillations and slowdowns in the search for the optimal solution [13, 14]. However, under the Momentum condition, the motion direction of the object is determined by the Momentum and gradient together, which can weaken the oscillation of the object and move to the optimal solution faster.

When using the Momentum algorithm, at the beginning of the iteration, the gradient direction is relatively consistent, and the Momentum algorithm will accelerate and reach the best point faster. At the later stage of iteration, the gradient direction will be determined inconsistently and fluctuate near the convergence value. The Momentum algorithm will slow down and increase stability. we can find the minimum point of the maximum likelihood function, which can allow our discriminant model to couple our input and output data to the greatest extent, and then obtain the LRM model. Figure 4 shows the architecture of the water quality prediction model:

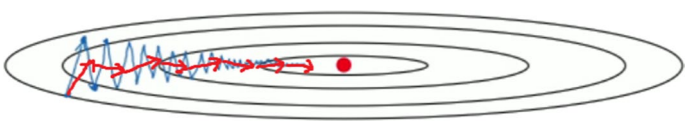

Fig. 3 Momentum algorithm training diagram

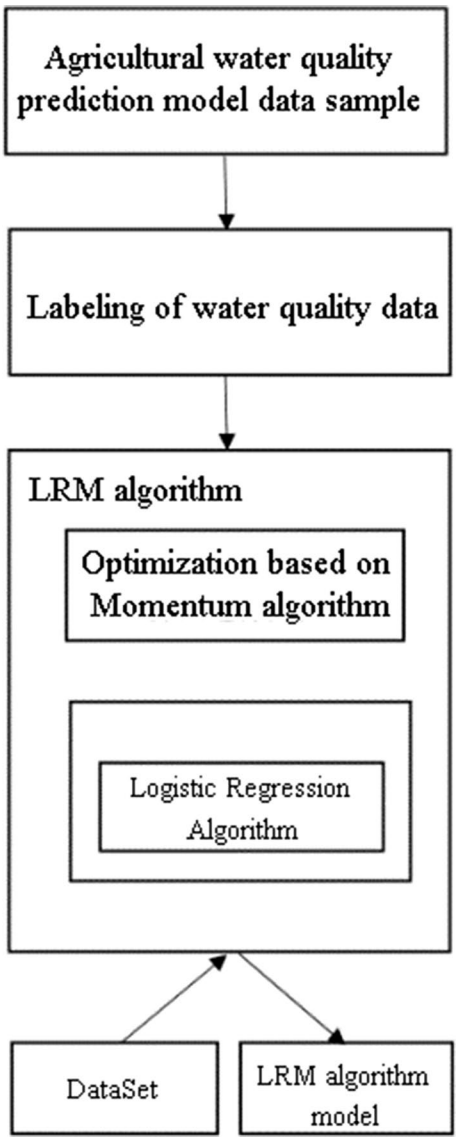

Fig. 4 Architecture diagram of water quality prediction model

We give the following stochastic gradient descent algorithm, the calculation formula is as formula (3):

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{3}$$

$$h_\theta(x) = X^T\theta = \sum_{k=0}^{n}\theta_k X_k \tag{4}$$

$$\theta := \theta - \partial \cdot \frac{\partial J(\theta)}{\partial \theta} \tag{5}$$

where θ is the gradient, $J(\theta)$ is the function of about θ , X^T represents the transpose of the matrix X , m is the number of data points in the water quality data, that is, the number of samples, $i \in (1, 2, \ldots m)$, $k \in (0, 1, 2, \ldots .n)$, y is the real y coordinate value (class label) in the water quality data, $h_\theta(x)$ is the function of x.

The Momentum algorithm is to introduce an exponentially weighted moving average in the ordinary gradient descent method, that is, to define a Momentum, which is the exponentially weighted moving average of the gradient, and then use this value to replace the original gradient direction to update.The momentum calculation formula defined is shown in Eq. (6):

$$v_t = \beta_{v_{t-1}} + (1 - \beta)\nabla L(w) \tag{6}$$

where v_t represents the current Momentum, β is the hyperparameter, $\nabla L(w)$ is the current gradient of the objective function, and this Momentum is used to bring it into the gradient descent formula, such as formula (7):

$$w = w - \alpha_{v_t} \tag{7}$$

where v_t is the Momentum and α is the step size.

3.2 LRM model training

In the ordinary stochastic gradient descent method, since the exact derivative of the loss function cannot be calculated, the noisy data will make the descent process not move in the best direction [15]. The forward direction is closer to the actual gradient. Based on the Momentum algorithm to optimize the logistic regression algorithm for the prediction of agricultural water quality, the Momentum algorithm is proposed to optimize the logistic regression. The Momentum algorithm solves the minimum point of the maximum likelihood function, and gives the overall discriminant model. The calculation formula is as formula (9).

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j} \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2 \tag{8}$$

$$= 2 \cdot \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \tag{9}$$

$$= \frac{2}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot \frac{\alpha}{\alpha \theta_j} \left(\sum_{k=0}^{n} \theta_k x_k^{(i)} - y^{(i)} \right) \tag{10}$$

$$= \frac{2}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \cdot x_j^{(i)} \tag{11}$$

where ∂ is called the learning rate or step size in the Momentum algorithm, and only needs to be updated ∂ each time until the minimum value of the maximum likelihood function is obtained. So according to the Momentum algorithm, every time we just

need to update ∂ , until we reach the maximum. Excellent, in summary, we get the LRM model:

Step1: Obtain water quality data, and label the water quality data to determine the category to which it belongs. For example, given the average value of a certain parameter, if the parameter in the water quality data is greater than the corresponding average value, the parameter belongs to a positive sample, otherwise is a negative sample;

Step2: Use the Momentum algorithm to optimize the logistic regression algorithm, and establish a discriminant model for water quality data analysis;

Step3: Input the labeled water quality data into the discriminant model, and the discriminant model outputs the predicted water quality category;

Step4: The predicted water quality category is compared with the actual water quality category. If the prediction accuracy is greater than or equal to the threshold, the LRM water quality prediction framework is obtained; if the prediction accuracy is less than the threshold, return to step S2 and use the Momentum algorithm to continue to find the maximum likelihood the minimum value of the natural function;

Step 5: Input the water quality data to be analyzed into the water quality prediction framework to obtain a prediction result. The momentum term constructed can not only accelerate the convergence at the initial stage of the algorithm, but also take a negative value at the stable stage, which further reduces the steady-state error.Further, the water quality prediction framework of LRM is as follows, and the calculation formula is as formula (13):

$$P(Y = 1 \mid x) = \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)} \tag{12}$$

$$P(Y = 0 \mid x) = \frac{1}{1 + \exp(w \cdot x + b)} \tag{13}$$

where $x \in R^n$ is the water quality data to be analyzed, $Y \in \{0, 1\}$ is the prediction result data, $w \in R^n$ and $b \in R$ are the parameters, w is the weight vector, b is the bias, and $w \cdot x$ is the inner product of w and x .

Its LRM model process is shown in (Fig. 5):

For the given water quality data x to be analyzed, $P(Y = 1 \mid x)$ and $P(Y = 0 \mid x)$ are obtained, and logistic regression compares the magnitude of the two conditional probability values, and classifies the water quality data to be analyzed into the category with the larger probability value. The final result is determined by the size of $P(Y = 1 \mid x)$ and $P(Y = 1 \mid x)$.

The LRM model can be used for data mining of agricultural water quality data. Compared with KNN and other traditional prediction algorithms, LRM not only speeds up the convergence rate of the algorithm, but also reduces the steady-state error and improves the prediction accuracy of water quality.

4 Experiment and analysis

In order to ensure the accuracy of the experiment, we select the data collected by 10,000 sets of sensors to train the LRM model. First, we analyze the four key variable data of the processing process, and the results are shown in (Fig. 6). In order to ensure

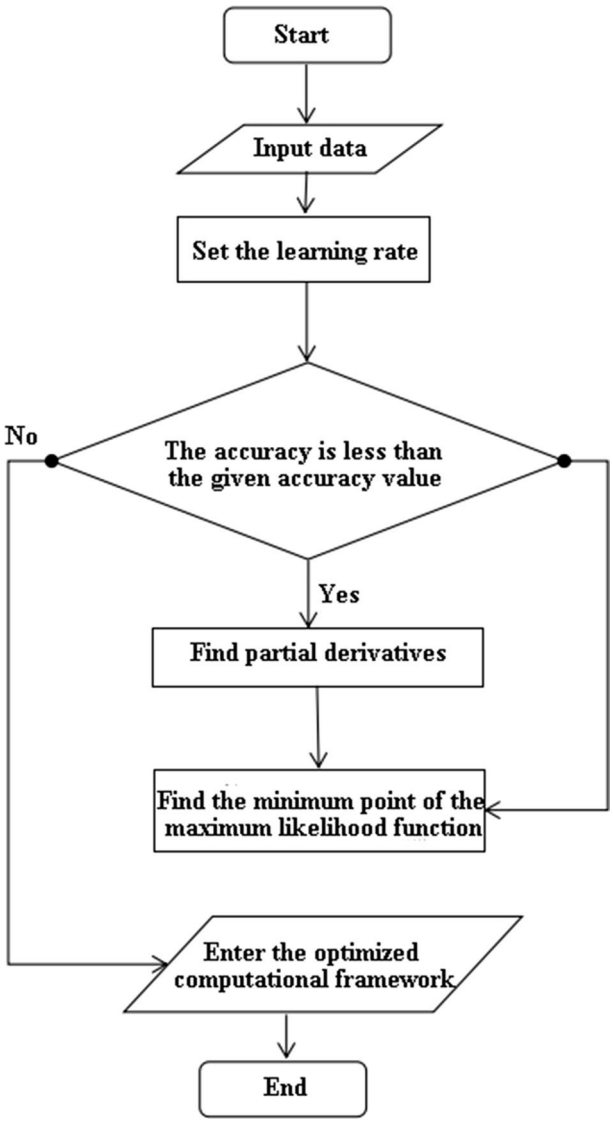

Fig. 5 LRM calculation flow chart

the accuracy of the experiment, we select the data collected by 10,000 sets of sensors to train the LRM model. First, we analyze the four key variable data of the processing process, and the results are shown in (Fig. 6).

It can be seen from Fig. 6 that the trend of dissolved oxygen and residual chlorine is unstable, the pH variable has abnormal values, and the dissolved oxygen variable has missing values. Therefore, we need to clean the data before training to ensure the accuracy of our LRM model training.

4.1 Hyperparameter value experiment

The data points in the graph below were estimated when different values of were used, and the results are shown in (Fig. 7).

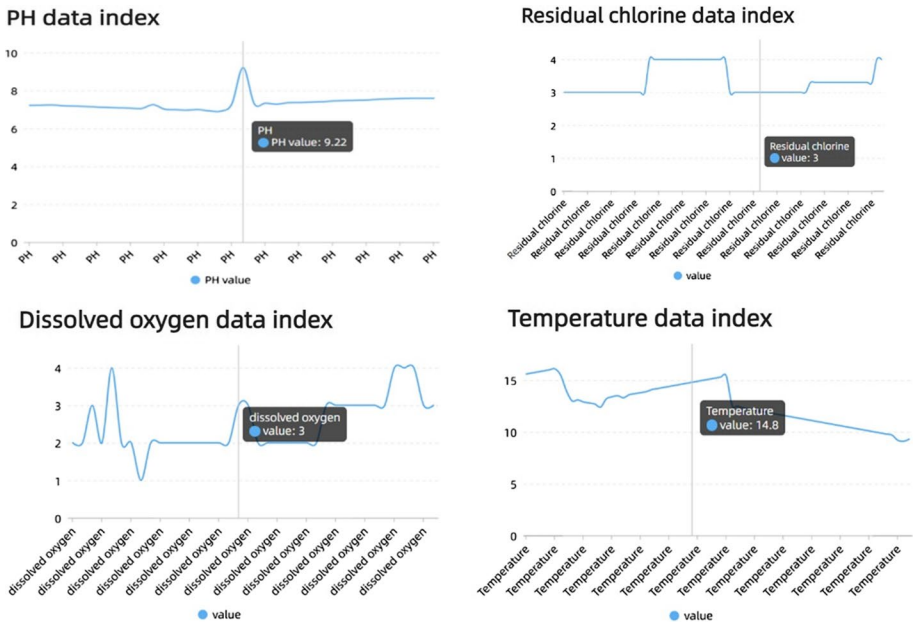

Fig. 6 Variable data curve

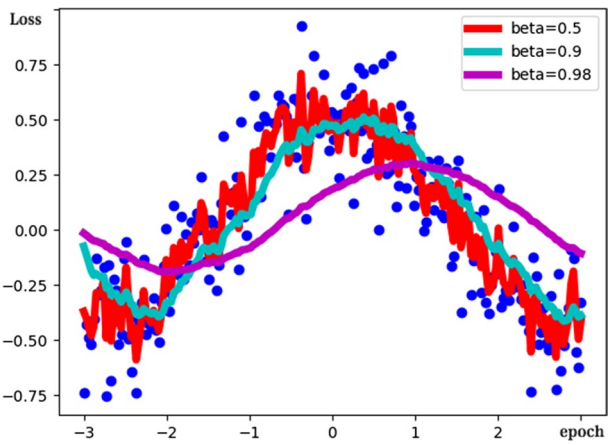

Fig. 7 Hyperparameter value experiment

As shown in Fig. 7, when the β value is too large, it cannot reflect the trend of the overall data well. If it is too small, overfitting occurs, and the fluctuation between adjacent points is too large. When $\beta = 0.9$, it can reflect both The overall trend of the original data, the fluctuation is not so high too big. Therefore, in the Momentum algorithm experiment in this paper, the β value is 0.9. We put β value is 0.9 and the learning rate is 0.01, then the peak value of the update amplitude is 10 times the original gradient times the learning rate.As shown in Fig. 7, when thevalue is too large, it cannot reflect the trend of the overall data well. If it is too small, overfitting occurs, and the fluctuation between adjacent points is too large. When = 0.9, it can reflect both The overall trend of the original data, the fluctuation is not so high too big. Therefore, in the Momentum algorithm experiment in this paper, the value is 0.9. We put value

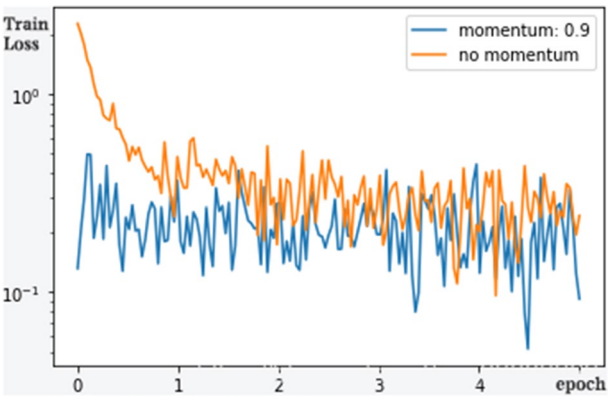

Fig. 8 Momentum contrast experiment

Table 1 Comparison of LRM model and contrastive model

Model	PH prediction		Temperature predictions		Dissolved oxygen prediction		Residual chlorine prediction	
	Recall /%	Correct rate /%	recall /%	Correct rate /%	Recall /%	Correct rate /%	Recall /%	Correct rate /%
LRM	99.3	95.6	92.3	94.6	87.3	90.2	89.6	92.5
SVR	73.5	80.2	85.5	89.2	73.5	80.2	82.6	84.9
ARIMA	76.8	83.6	74.9	86.5	76.8	83.6	78.5	82.1
LSTM	86.5	88.6	88.5	90.6	86.5	88.6	88.5	92.6
KNN	77.6	82.5	82.6	86.5	77.6	82.5	90.6	92.4

is 0.9 and the learning rate is 0.01, then the peak value of the update amplitude is 10 times the original gradient times the learning rate.

4.2 Model comparison

Compare the training effect of logistic regression algorithm model before and after the non Momentum algorithm and the Momentum algorithm.

It can be seen from Fig. 8 that we set the super parameter value as 0.9 and the learning rate as 0.01. After adding Momentum, the loss decreases to a lower degree, which can be understood as an inertial effect. Therefore, each time the algorithm is updated with Momentum, the amplitude of updating will be more than that of the algorithm without Momentum.

4.3 Model comparison

In order to evaluate the performance of each model in the prediction task, it is proved that the LRM model is better than other models in data prediction, the experimental data in the model is compensated, and the data prediction results are analyzed using the recall rate and the correct rate. We select the real data collected by the sensor, randomly select 500 consecutive groups of data, and input the data into the LRM model and the comparison model respectively. Table 1 lists the LRM model and the comparison model

on the four indicators of pH, temperature, dissolved oxygen, and residual chlorine. The recall rate and the correct rate are shown in (Table 1).

It can be seen from Table 1 that the two water quality indicators, temperature and pH, are better, so they are easier to predict, while dissolved oxygen and residual chlorine are more difficult to predict. We compared the traditional algorithm with the agricultural water quality prediction method based on the Momentum algorithm optimized logistic regression algorithm, and found that the LRM model improved the prediction accuracy of all water quality indicators by an average of 1.11%.It is because Momentum algorithm weights the gradient vector within a period of time that it can escape when it falls into the local optimal. In terms of water quality prediction, the LRM model is better than the previous ones The model has been greatly improved, and can better meet the requirements of agricultural water quality prediction.

5 Conclusion and outlook

Aiming at the problem that the linear regression of the traditional linear water quality prediction model is not robust enough and cannot guarantee the prediction accuracy under the condition of interference, this paper proposes a prediction model of agricultural water quality based on the Momentum algorithm optimization logic regression algorithm (LRM algorithm). The model uses momentum algorithm to optimize the logic regression algorithm. The misdivided samples can be adjusted quickly. When the object meets the local optimal in the process of falling, the introduction of momentum makes it easy for the next update to jump out of the local optimal with the help of the last big gradient. In this paper, the performance of the proposed model was evaluated on four real data sets. The experimental results show that the prediction accuracy of the proposed LRM algorithm is improved by 1.11 percentage points on average compared with the existing algorithm. Compared with the traditional water quality prediction algorithm, the proposed LRM algorithm not only speeds up the convergence rate of the algorithm, but also reduces the steady-state error and improves the prediction accuracy, which is suitable for the current complex water quality data mining. The experiment verifies the feasibility of this method in predicting the actual agricultural water quality and even in predicting and warning the residents drinking water.

This paper predicts agricultural water quality data based on Momentum algorithm optimized logic regression algorithm. The traditional water quality monitoring method is not robust enough, and the correctness of data prediction results will be affected, which has certain deficiencies in practical application. The experimental part of this section compares the prediction results of LRM model with SVR, LSTM and other models, and finds that the accuracy and recall rate of LRM model are greatly improved compared with previous models.

Although the prediction model proposed in this paper has obtained good experimental results, it still has some limitations. First of all, only four parameter variables are selected in the model, which is difficult to cope with the increasingly complex changes in agricultural water quality. Moreover, this model has high requirements on data quality. Before model training, data sets should be cleaned with high quality, so it will take a long time to train the model. In future work, based on this algorithm, we will further study how to improve the fault tolerance rate of the model and the diversity of training parameters, and design a better

and stronger fault tolerance agricultural water quality prediction method based on Momentum algorithm optimization logic regression algorithm, so as to further optimize the accuracy of water quality prediction data and the generalization ability of the model. Secondly, due to the confidentiality of the data related to residential drinking water, there are few data that can be studied at present. If the amount of data in the future is sufficient, further data cleaning and prediction research can be carried out. Meanwhile, the correlation between each subsystem can be explored, and the correlation can be added into the diagnostic model as an indicator. Finally, with the increasingly clear explanatory ability of deep learning, the following experiments can try to use neural networks and other methods to establish water quality prediction models. At present, the cleaning and prediction of complex data sets mostly take single system as the research object. In the future, deep learning knowledge can be considered to realize multidimensional system prediction of complex data sets.

Acknowledgements
We thank the School of Information Engineering of Dalian University for its support of the data acquisition hardware.

Funding
The research was funded by Dalian University in China. Fund recipient: Dr. Gai Rong-Li.

Declarations

Competing interests
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 October 2022 Accepted: 5 January 2023
Published online: 06 February 2023

References

1. L.Y. ZhangPeng Li, Validation of logistic model in the study of river valley urban flood. J. Northwest Normal Univ. Natl. Sci. Edn. **53**(1), 128–134 (2017)
2. Li Na, X.G. Wang Lachun, Grey prediction of water quality trends in the huaihe river basin under the jurisdiction of shandong province. Environmental Science and Technology **35**(2), 201–205 (2012)
3. J. Ju, L. Wang et al., Analysis of ammonia nitrogen content in water based on weighted least squares support vector machine (wlssvm) algorithm. J. Softw. Eng. Appl. **9**(02), 45 (2016)
4. Liu Jie, J.D. Zhu Rongjie, Real-time water quality prediction model based on genetic-neural network. South-to-North Water Divers. Water Conserv. Technol. (Chin. Engl.) **18**(6), 93–100 (2020)
5. C. Zhou, M. Liu, J. Wang et al., Water quality prediction model based on cnn-lstm. Hydropower Energy Sci **39**(03), 20–23 (2021)
6. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
7. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
8. Z. Yang, X. Huiting, Z. Zhijie, Real-time river water quality prediction model based on spatial correlation and neural network model. Beijing Da Xue Xue Bao **58**(2), 337–344 (2022)
9. Yang Bo, Y.Z. Liu Yu, Long-term prediction of container throughput by introducing logistic growth model. Journal of Chongqing Jiaotong University (Natural Science Edition) **39**(11), 45–50 (2020)
10. X. Song, X. Liu, F. Liu, C. Wang, Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int. J. Med. Informatics **151**, 104484 (2021)
11. S. Kabir, S. Patidar, G. Pender, Investigating capabilities of machine learning techniques in forecasting stream flow. in: *Proceedings of the Institution of Civil Engineers-Water Management*, vol. 173, pp. 69–86 (2020). Thomas Telford Ltd
12. Tang Yishun, L.Z. Xu Qing, Water quality prediction based on optimized nonlinear autoregressive neural network model. J. Donghua Univ. (Natl. Sci. Edn.) **48**(3), 93–100 (2022)
13. J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, et al. Large scale distributed deep networks. Adv. Neural Inf. Process. Syst. **25** (2012)
14. D. Povey, X. Zhang, S. Khudanpur, Parallel training of dnns with natural gradient and parameter averaging. arXiv preprint arXiv:1410.7455 (2014)
15. Y. Wang, J. Zhou, K. Chen, Y. Wang, L. Liu, Water quality prediction method based on lstm neural network. in: *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 1–5 (2017). IEEE

Publisher's Note