Application of a Multi-factor Model in the Grain Yield Prediction

Lili Chen[1, 2]

[1]*Laboratory of Intelligent Information Processing, Suzhou University, Suzhou 234000, Anhui, China*
[2]*The Key Laboratory of Intelligent Computing & Signal Processing of MOE, Anhui University, Hefei 230039, Anhui, China*

Hongjun Guo

Laboratory of Intelligent Information Processing, Suzhou University, Suzhou 234000, Anhui, China

Abstract
A multi-factor model for the grain yield prediction based on the previous data and relevant impacting factors is reported. In this model, the weight coefficient of each individual factor that affected the grain yield historically is analyzed by the variation coefficient method and the conversion degree function in the attribute theory. The effects of various factors on the predicted grain yield are then used as standard vectors and subjected to a similarity-based search in the matrix of historical values. The predicted total grain yield is then determined by multiplying the sown area by the unit grain yield which is obtained by the highest similarity between the historical data and those predicted. Compared to the results obtained by the BP nerve network method, this method is simpler, more flexible, less time-consuming, and more accurate.

Key words: PREDICTION OF GRAIN YIELD, CONVERSION DEGREE FUNCTION, COEFFICIENT OF VARIATION, SIMILARITY OF VECTORS

1. Introduction

Grain is one of the most important strategic materials that is essential for national economy and people's livelihood [1]. An accurate prediction of grain yield not only ensures the country's food security, but also provides a crucial reference for the government to make rational food policy. However, grain yield is affected by many factors such as the government policy, resources, natural environment, and so on. The prediction of grain yield is therefore a complex problem covering multiple disciplines. Many scholars and researchers have done a lot of research. Especially, the prediction of Chinese grain production has received extensive attention from our national scientific researchers.

In the world, the commonly used methods for grain yield prediction include meteorological forecasting, remote sensing, statistical dynamic growth simulation, etc. At present, in China, there are some representative methods: neural network models [2], time series analysis model, gray Markov model [3], support vector machine model [4], etc. Different prediction models have different specificity and applicability according to different theories and methods involved, and most of them are focused on the fitting of historical data. However, the grain yield and the influencing factors are uncertain and nonlinear, so the model based on the influence factors of grain yield can objectively reflect the intrinsic relationship.

In this paper, the factors affecting grain yield were analyzed firstly, and the variation coefficient method was used to determine the weight of each factor. The attribute theory method was used to establish the qualitative mapping model [5] of grain yield, and the conversion degree function was used to determine the degree of similarity between the prediction model and the known pattern. This combined method provided a novel way for both qualitative and quantitative prediction of grain yield.

2. The influencing factors of grain production and the collection of original data

2.1. Analysis of the factors influencing grain yield

There are multiple factors affecting grain yield, such as the natural factors including soil, meteorological conditions, pests and diseases. There are also economic and social factors, such as the level of agricultural technology, seed quality, governmental policies [6], etc. According to statistical analysis, the main factors affecting grain output are: sown area (1000 hectares), total power of agricultural machinery (10000 kW), effective irrigated area (1000 hectares), the amount of chemical fertilizer (10000 ton) and disaster area (1000 hectares)[7].

2.2. The original data of grain output and every influence factors

The statistical data of the grain yield in China from 1980 to 2011 are selected as samples for the analysis herein. Total grain output and the raw data [8] of the influencing factors in each year are shown in table 1.

Table 1. The original data

Year	Grain Yield	sown area	total power of agricultural machinery	effective irrigated area	amount of chemical fertilizer	disaster area
1980	32055.5	117234	14745.7	44888.1	1269.4	29777
1981	32502.0	114958	15679.8	44573.8	1406.9	18743
1982	35450.0	113462	16614.2	44176.9	1513.4	15985
1983	38727.5	114047	18022.1	44644.1	1659.8	16209
1984	40730.5	112884	19497.2	44453.0	1739.8	15607
1985	37910.8	108845	20912.5	44035.9	1775.8	22705
1986	39151.2	110933	22950.0	44225.8	1930.6	23656
1987	40473.3	111268	24836.0	44403.0	1999.3	20393
1988	39408.0	110123	26575.0	44375.9	2141.5	23945
1989	40754.9	112205	28067.0	44917.2	2357.1	24449
1990	44624.3	113466	28707.7	47403.1	2590.3	17819
1991	43529.3	112314	29388.6	47822.1	2805.1	27814
1992	44265.8	110560	30308.4	48590.1	2930.2	25859
1993	45648.8	110509	31816.6	48727.9	3151.9	23133
1994	44510.1	109544	33802.5	48759.1	3317.9	31383
1995	46661.8	110060	36118.1	49281.2	3593.7	22267
1996	50453.5	112548	38546.9	50381.4	3827.9	21233
1997	49417.1	112912	42015.6	51238.5	3980.7	30309
1998	51229.5	113787	45207.7	52295.6	4083.7	25181
1999	50838.6	113161	48996.1	53158.4	4124.3	26731
2000	46217.5	108463	52573.6	53820.3	4146.4	34374
2001	45263.7	106080	55172.1	54249.4	4253.8	31793
2002	45705.8	103891	57929.9	54354.9	4339.4	27319
2003	43069.5	99410	60386.5	54014.2	4411.6	32516
2004	46946.9	101606	64027.9	54478.4	4636.6	16297
2005	48402.2	104278	68397.8	55029.3	4766.2	19966
2006	49804.2	104958	72522.1	55750.5	4927.7	24632
2007	50160.3	105638	76589.6	56518.3	5107.8	25064
2008	52870.9	106793	82190.4	58471.7	5239.0	22283
2009	53082.1	108986	87496.1	59261.4	5404.4	21234
2010	54647.1	109876	92780.5	60347.7	5561.7	18538
2011	57120.8	110573	97734.7	61681.6	5704.2	12441

There is a direct relationship between the size of sown area and the corresponding value of factors above, and the factors can also be divided into two categories: benefits and costs. Therefore, we not only use the unit area of total power of agricultural machinery, effective irrigation area, and amount of chemical fertilizer, but also convert the value of disaster area into a benefit type data which can reflect on unit area. Therefore, after this conversion, the main factors that affect grain yield can be divided into agricultural machinery total power (X1), effective irrigation area (X2), chemical fertilizer (X3) and non-disaster area (X4), which are all based on unit areas.

3. The multi-factor combination model for grain yield prediction

3.1. Data preprocessing

When using the original data to construct the prediction model of the grain yield, in order to avoid the different prediction results because of the different dimension, we need to carry on the dimensionless processing to make the law of the model independent of the dimension.

There are many methods to achieve dimensionless data, such as the mean method, standard method, extreme value method and standard deviation method [9], etc. Because the mean method can eliminate the influence of different dimensions and magnitudes while retaining the information of differences in the degree of each index value and ensuring the comparability of data, it was employed for the dimensionless processing of raw data in our study.

It is assumed that there are historical data for n years, and the data of each year contains m indexes or influence factors which are referred to $X_1, X_2, \cdots, X_m$. We use $x_{ij}(i=1,2,\cdots,n; j=1,2,\cdots,m)$ to represent the value in the *i*-th year of the first J index. After the dimensionless processing, it will be changed into. $y_{ij}(i=1,2,\cdots,n; j=1,2,\cdots,m)$

In the mean method, the benefit index was processed by equation (1), in which $\bar{x}_j$ is the mean of index X_j. After the treatment, the mean value of each index is 1, and the variance is the squared coefficient of variation of each index.

$$y_{ij} = \frac{x_{ij}}{\bar{x}_j} \tag{1}$$

According to the above method, when we forecast the grain yield for 2011, various influence factors between 1980 and 2011 need to be preprocessed. The data per unit area after preprocessing is shown in table 2.

Table 2. Data after dimensionless processing

Year	total power of agricultural machinery *X1*	effective irrigated area *X2*	amount of chemical fertilizer *X3*	non disaster area *X4*
1980	0.2798972707	0.7719324882	0.3153605482	1.0000007956
1981	0.3095298634	0.7971802377	0.3634970287	0.9999998784
1982	0.3366813044	0.8110537387	0.4013921686	0.9999995871
1983	0.3614748615	0.8112441978	0.4357165572	0.9999996369
1984	0.3991608169	0.8245017051	0.4661767189	0.9999995251
1985	0.4604997755	0.8785068905	0.5117915590	0.9999998881
1986	0.4865210218	0.8493944999	0.5356570442	1.0000000669
1987	0.5233371002	0.8476703887	0.5513830824	0.9999998189
1988	0.5716861111	0.8648611103	0.6029454819	1.0000000527
1989	0.5815834116	0.8432250862	0.6392483596	1.0000001887
1990	0.5817110925	0.8702229182	0.6869649987	0.9999997298
1991	0.6077871932	0.8960167259	0.7592706168	1.0000004603
1992	0.6468556657	0.9395220897	0.8184972761	1.0000002296
1993	0.6796712602	0.9430563854	0.8812378998	1.0000000042
1994	0.7348726212	0.9603593151	0.9440655209	1.0000006455
1995	0.7778687335	0.9615625049	1.0129752565	0.9999999113
1996	0.7938790285	0.9400477850	1.0318131404	0.9999999501
1997	0.8597472708	0.9498859630	1.0660934697	1.0000006799
1998	0.9108933967	0.9546300876	1.0769227954	1.0000003155
1999	0.9981789229	0.9811459381	1.0996962127	1.0000004095
2000	1.1658560419	1.0812799549	1.2034387886	1.0000008612
2001	1.2790659236	1.1394182726	1.2907022861	1.0000005574
2002	1.4001912581	1.1902497822	1.3727449856	1.0000000646
2003	1.5941168313	1.2918231509	1.5242352988	1.0000003899
2004	1.6179719388	1.2472137722	1.5334759047	0.9999988956
2005	1.6409569330	1.1960900156	1.4965902473	0.9999994054
2006	1.7174326480	1.1961149992	1.5273170551	0.9999998673
2007	1.7904817429	1.1970271821	1.5628222212	0.9999999387
2008	1.8800783863	1.2117566447	1.5684795455	0.9999997489
2009	1.9217093573	1.1791953095	1.5535387594	0.9999997712
2010	2.0048942184	1.1814362824	1.5729608458	0.9999995933
2011	2.0854079983	1.1923745774	1.5929883264	0.9999991327

3.2. The weight of each factor

The grain yield per unit area should be evaluated different when taking into accounts of each index individually, because each index can impact the grain yield to a different extent and the degrees of importance of each index are not identical. Therefore, each index should be given a unique weight coefficient.

The variation coefficient method is an objective weight method which is commonly used in the actual work for determining the weight of index. In order to illustrate the degree of difference between the values of the indicators, we use equation (2) to calculate the coefficient of variation of each index.

$$V_i = \frac{\sigma_i}{\bar{x}_i} \quad (i = 1, 2, \cdots, n) \tag{2}$$

In equation (2), V_i is the coefficient of variation of the first i index, σ_i is its standard deviation, and $\bar{x}_i$ is its mean value.

The weight coefficient of each index between 2001 and 2011 can be calculated by formula (3). They are shown in Table 3.

$$W_i = \frac{V_i}{\sum_{i=1}^{n} V_i} \tag{3}$$

Table 3. Weight coefficient of each index

Year	*X1*	*X2*	*X3*	*X4*
2001	0.4497675373	0.1007674160	0.4494645760	0.0000004708
2002	0.4572393072	0.1108433681	0.4319168835	0.0000004411
2003	0.46263195013	0.1203047115	0.4170629352	0.0000004031
2004	0.4651807649	0.1321856591	0.4026332085	0.0000003675
2005	0.4677329497	0.1363632810	0.3959033560	0.0000004133
2006	0.4714394894	0.1373031988	0.3912568973	0.0000004147
2007	0.4748945790	0.1377521867	0.3873528328	0.0000004015
2008	0.4778958750	0.1378999243	0.3842038119	0.0000003888
2009	0.4811941236	0.1383414830	0.3804640120	0.000003814
2010	0.4846245179	0.1379602362	0.3774148703	0.0000003756
2011	0.4880334699	0.1375139159	0.3744522410	0.0000003730

3.3 Forecasting model of multi factor combination based on conversion degree function

3.3.1 The basic principle of conversion degree function

Any property of an object has a qualitative characteristic and a quantitative characteristic, and they can transform into each other. If there are two different quantitative features called x_1 and x_2, although their corresponding properties after conversion between quantity and quality characteristics belong to the same quality characteristic class $p_i(o)$, that is $p_i(x_1), p_i(x_2) \in p_i(o)$, the extent of conversion $\eta(p_i(x_1))$ and $\eta(p_i(x_2))$ will make a great difference because x_2 is not equal to x_2. Generally speaking, if ξ_i is the midpoint of α_i and β_i, the corresponding quantitative feature $p_i(\xi_i)$ is most stable. It would most unlikely change into other quality characteristics, and it is the best embodiment of the essence of qualitative feature. Therefore, $p_i(\xi_i)$ is the intrinsic character of $p_i(o)$, and ξ_i is called the intrinsic point of $p_i(o)$.At the same time, two quantitative features $p_i(x_1), p_i(x_2) \in p_i(o)$ corresponding to boundary points of α_i and β_i are easy to convert to other features such as $p_j(o)$ or $p_k(o)$ [10].

If we make $k_1, k_2 \in [0,1]$ as the degree of $p_i(x)$ deviating from $p_i(\xi_i)$, and $\eta(x)$ represents the degree of close to $p_i(\xi_i)$, that is, the degree of similarity between x and the quality characteristic class $p_i(o)$ $\eta(x) \in [-1,1]$, then the following equations (4)-(6) are established.

$$k_1(x) = \frac{p_i(\xi_i) - p_i(x)}{p_i(\xi_i) - p_i(\alpha_i)} \quad (x < \xi_i) \tag{4}$$

$$k_2(x) = \frac{p_i(x) - p_i(\xi_i)}{p_i(\beta_i) - p_i(\xi_i)} \quad (x > \xi_i) \tag{5}$$

$$\eta(x) = \begin{cases} -(1 - k_1(x)) & x < \xi_i \\ 1 - k_2(x) & x > \xi_i \end{cases} \tag{6}$$

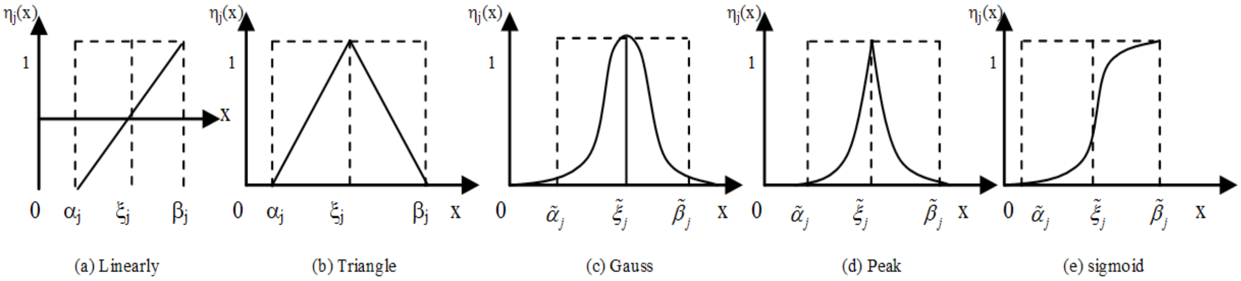

Figure 1. Typical conversion degree functions

Consequently, we can see that the mathematical essence of conversion degree function $\eta_i(x)$ is a comparison between $|x-\xi_i|$ and δ_i, and the degree of similarity of character $p_i(x)$ corresponding to x and intrinsic character $p_i(\xi_i)$ can be reflected according to the result of the comparison [11]. The rule of transformation between quantity and quality characteristics are not identical, so generally there are different types of conversion degree function $\eta_i(x)$. Several kinds of common conversion degree functions have been shown in Fig.1.

In the prediction of grain yield, we chose the peak type of conversion degree function and introduced the weight of influencing factors. So, if there are two vectors X_i and Y which are composed of various factors affecting grain yield, their similarity degree(conversion degree) of can be expressed as equation (7).

$$\eta(Y,X_i)=\exp\{-\frac{\sum_{j=1}^{m}\omega_j\left|y_j-x_{ij}\right|}{\sum_{j=1}^{m}\omega_j\delta_j}\} \tag{7}$$

3.3.2. Prediction technique of multi factor combination

We use the influence factors of each past year as a row vector to construct a matrix

$$\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & x_{ij} & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$$

in which the number of factors is m and the number of years is n. Then, the matrix is used as a qualitative benchmark to construct attribute coordinate, and the influence factors of each year will be a vector or a point in the coordinate system. In this matrix, x_{ij} represents the value of the j influence factor in the i year, the vector of factors for forecast is $(y_1,y_2,\cdots,y_m)$.

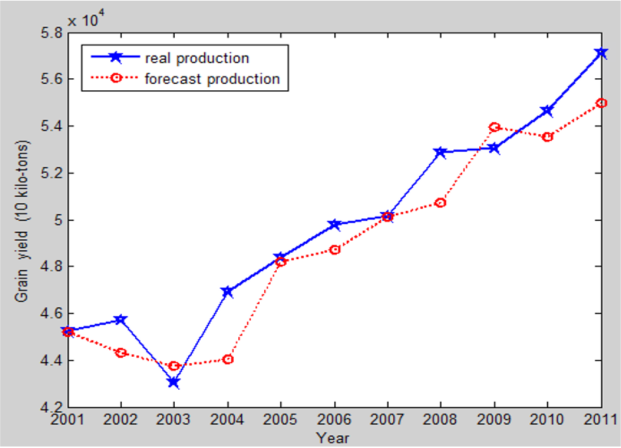

Figure 2. Prediction result of multi factor combination

In the matrix of

$$\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & x_{ij} & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix},$$

we perform similarity search on the vector which is composed of factors effecting the per unit area yield of grain to measure the degree of similarity in the $(y_1,y_2,\cdots,y_m)$ and every row vector $x_i, i=1,2,\cdots,n$ of matrix

$$\begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & x_{ij} & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix},$$

then we can find the year with the greatest similarity, and put the unit area yield in this year as the predicted value.

According to the method above, we have carried on the forecast to the total output of grain from 2001 to 2011. The result has been shown in Figure 1.

4. Comparative analysis of the results of grain yield prediction

4.1. BP neural network model

BP (Back Propagation) network is a multilayer and forward type network, which consists of one input layer, one output layer and several hidden layers. BP network can learn and store a lot of mapping relationship between the input model and output mode. When all the neurons in the input layer receive input information from outside, neuronal activation values will spread from the input layer through the middle layer, then to the output layer. In order to reduce the error between the actual output and the expected output, when each neuron in the output layer has received the response to the input, the error will go through each intermediate layer to make a reverse transmission starting from the output layer and finally back to the input layer. With the revision process of this error back propagation algorithm going on, weights and thresholds will be constantly adjusted, and the correct rate responded to the input mode will be rising, until the error is reduced to an acceptable level or reaches a preset times of learning [12].

Through the analysis, six factors including the sown area (1000 hectares), labor force(1000 people), total power of agricultural machinery (10kilo-kW), the effective irrigated area (1000 hectares), the amount of chemical fertilizer (10 kilo-tons) and disaster area (1000 hectares) are selected as the input variables, and grain yield is selected as the output variable. They are used to construct a three-layer BP neural network, in which the input layer contains six neurons and the output layer contains one neuron.

We selected the grain yield and factors from 1980 to 2000 as the training set of BP neural network. After 6945 steps of training, the network is convergent and the error is 9.99586e-006. The training has achieved a good fitting effect as show in Figure 3.

Using the data of grain yield and all the factors above between 2001 and 2011 as test samples, we have predicted the grain yield and compared it with the actual data of grain yield. It has been shown in Figure 4.

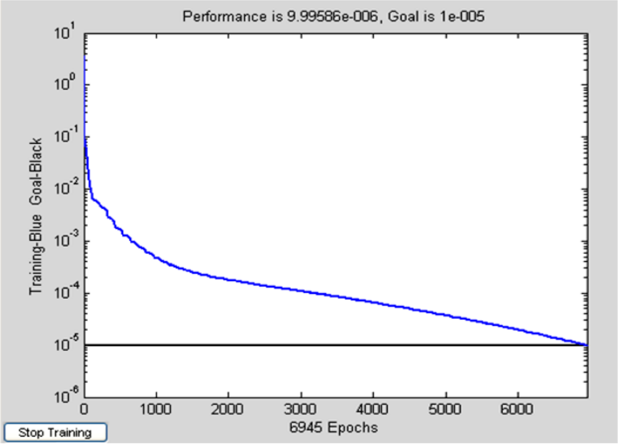

Figure 3. The training results of BP neural network

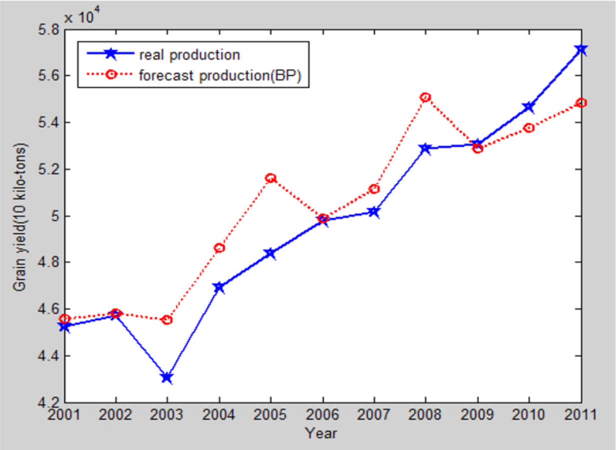

Figure 4. Prediction result of BP neural network

4.2. The comparison between the prediction results of the two models

The prediction results obtained by the above two models was shown in Table 4, as well as the actual grain yield and the relative errors.

Table 4. Error analysis of predicted value and actual value

Year	Real production (10 kilo-tons)	Prediction result of multi factor combination (10 kilo-tons)	Relative error	Prediction result of BP neural network (10 kilo-tons)	Relative error
2001	45263.7	45202.1	0.14%	45599	-0.74%
2002	45705.8	44329.7	3.10%	45814	-0.24%
2003	43069.5	43734.4	-1.52%	45541	-5.74%
2004	46946.9	44020.9	6.65%	48598	-3.52%
2005	48402.2	48181.5	0.46%	51603	-6.61%
2006	49804.2	48717.8	2.23%	49879	-0.15%
2007	50160.3	50126.9	0.07%	51163	-2.00%
2008	52870.9	50708.7	4.26%	55074	-4.17%
2009	53082.1	53956.6	-1.62%	52892	0.36%
2010	54647.1	53515.6	2.11%	53771	1.60%
2011	57120.8	54993.8	3.87%	54861	3.96%

From these experimental results, we find that the average relative error of results predicted by the multi-factor combination model is 2.37%, which can be attributed to the use of influence of various factors on grain yield. It overcomes the limitation of many prediction method [13] focusing on historical data fitting, which only considers the relationship between grain yield and time, while the impact of other factors on grain yield are not taken into account. At the same time, compared with the BP neural network [14] whose the average relative error is 2.64%, this multi-factor combination model not only reflects the inherent relationship of the changes of grain yield more objectively, but also effectively avoids the over-fitting or under-learning phenomenon. Consequently, our method is simpler, more time-efficient, and more accurate than the BP network method.

5. Conclusions

Grain yield forecast is an important measure to prevent food crises and ensure food security. On the basis of in-depth study of existing prediction methods on grain yield, multiple factors were taken into account in order to establish the matrix composed of the influencing factors and the corresponding qualitative mapping model. Conversion degree function induced by the qualitative mapping were used for similarity search in the vector matrix, then the prediction of multi factors combination was achieved with a good prediction effect. However, there are too many influencing factors on grain yield and there is a complex nonlinear relationship between them. So, a more detailed discrimination and in-depth analysis should be made on the variety of factors that affect grain yield, and the relationship between factors should be fully considered. We should learn from various existing forecasting methods and integrate them to improve the prediction accuracy. This will be the direction of our future research.

Acknowledgements

This work was supported by Major Project of Outstanding Yong Talent Foundation at Suzhou University of China(No.2013XQRL01), Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2013YKF17) and Horizontal Project at Suzhou University of China (No.2015hx025).

References

1. Huang Z, Li P, Shang X. (2011) Analysis on Influencing Factors of Grain Production in China. *Journal of AnHui Agricultural Sciences*, 39(21), p.p.13158-13160.
2. Niu Z. X., Li W. P., Zhang W.J. (2012) Prediction of grain yield using AIGA-BP neural network. *Computer Engineering and Applications*, 48(2), p.p.235-237.
3. Liu A., Zhao S., Zhang Y. P. (2007) Yield Forecast Based on Grey-Markov Model. *Computer technology and Development*, 17(6), p.p.191-196.
4. Zai S. M., Jia Y. H., Wen J. (2009) Grain Yield Prediction for Irrigation District Based On LS-SVM. *Agricultural Science & Technology*, 10(6), p.p.1-3,6.
5. Kou J. J., Feng J. L. (2011) Prediction of Grain Yield Based on Attribute Theory. *Computer Knowledge and Technology*, 7(16), p.p.3814-3815.
6. Fan D. J. (2011) An Empirical Analysis on Grain Yield Impact Factors and Measurement of the Contribution Rate. *Journal of Hunan University of Technology*, 25(5), p.p.55-61.
7. Li L. F., Li M. G. (2012) The Analysis on the Influencing Factors of Grain Production in the Light of the Current Situation of China. *China Business and market*, (4), p.p.109-115.
8. Department of Rural Surveys, National Bureau of Statistics (2012) *China Rural Statistical Yearbook 2012*. China Statistics Press: Beijing.
9. Zhang X. M. (2012) Comparative analysis of data nondimensionalization in decision analysis. *Journal of MinJiang University*, 33(5), p.p.21-25.
10. Feng J. L. (2006) Qualitative Mapping Model from Judgment to Recognition and Fuzzy Artificial Neuron. *Pattern Recongnition and Artificial Intelligence*, 19(1), p.p.35-46.
11. Chen L. L., Guo H. J. (2013) Application of Conversion Degree Function and SVM Regression in Precipitation Forecasting. *International Journal of Applied Mathematics and Statistics*, 50(20), p.p.525-532.
12. Guo Q. C., He Z. F., Li L. (2011) Forecast Model for Grain Yield based on BP Neural Network. *Hunan Agricultural Sciences*, (17), p.p.136-138.
13. Wang Y. T., Du Y. L., Jia L. X. (2011) The Applications of Time Series Analysis in Grain Yield Prediction. *Henan Science*, 29(5), p.p.520-523.
14. Yao Z. F., Liu X.T., Yang F., Yan M.H. (2010) Comparison of several methods in grain production prediction, *Agricultural Research in the Arid Areas*, 28(4), p.p.264-268.