Malaya Journal of Matematik, Vol. 8, No. 4, 1771-1776, 2020

https://doi.org/10.26637/MJM0804/0074

Enhanced crop yield prediction using Monte Carlo method and binary cuckoo search

Chellammal Surianarayanan[1]*, Kodimalar Palanivel[2] and K. Mani[3]

Abstract

The yield of crops is influenced by various factors such as weather conditions, soil characteristics, irrigation facility, solar radiation, fertilizer application, tillage, etc. Accurate prediction of crop yield is an important issue in agriculture as un-presented changes in yield will significantly influence food supply and market prices. Data pre-processing and selection of relevant features is an essential step while perform prediction using machine learning algorithms. In this work, Monte Carlo simulation for random selection of data and binary cuckoo search for relevant feature selection are used with an objective of enhancing the accuracy of prediction using multiple linear regression technique. Experimental results are discussed.

Keywords

Binary cuckoo search, Monte Carlo method, multiple linear regression, prediction of crop yield.

AMS Subject Classification

11K45, 65C05.

[1,2] *Department of Computer Science Bharathidasan University Constituent Arts & Science College, Tiruchirappalli-621303, Tamil Nadu, India. Affiliated to Bharathidasan University Tiruchirappalli, Tamil Nadu, India.*

[3] *Department of Computer Science, Nehru Memorial College, Puthanampatti-621007, Tiruchirappalli, Tamil Nadu, India.*

***Corresponding author**: [1]drschellammal@gmail.com; [2]pkodimalar.np@gmail.com; [3]nitishmanik@gmail.com

Article History: Received **13** August **2020**; Accepted **29** September **2020**

Contents

1. Introduction

Agriculture is one of the country's principal economic sectors. Use of machine learning techniques in agriculture helps to gain effective regulation of irrigation, fertilizer, diseases, elimination of insect pests in crop growing and crop yield prediction. Modern agriculture, by itself, produces huge amounts of data from sensors such as soil-related, crop-related, intercultural management, crop patterns, and data related to harvesting. In addition, there are many official databases maintained and governed by weather departments and agricultural departments where data related to weather patterns, soil, water and crop yield can be analyzed and correlated with each other. The yield of a crop is affected by various climatic parameters, soil parameters, water parameters and other environmental conditions[1]. Manually extracting knowledge from the archived data is tedious and machine learning plays a crucial role in yield prediction[2-4]. In addition, the data collected from different data sources needs to be preprocessed and relevant features need to be identified for effective prediction of yield. In this research work, Monte Carlo method is used to select the data at random which enables to arrive as single valued function between various features and yield of the crop. Using this method, data collected from 2007 to 2016 are preprocessed and given as input for identifying relevant features using binary cuckoo search algorithm. After pre-processing and feature selection, the prediction is done using multiple linear regression and experimentation results are discussed.The paper is organized as follows. Section II highlights the literature related to the theme of the work. Section III describes the proposed method for prediction of crop yield. Section IV presents the experimentation results and discussion. Section V concludes the work.

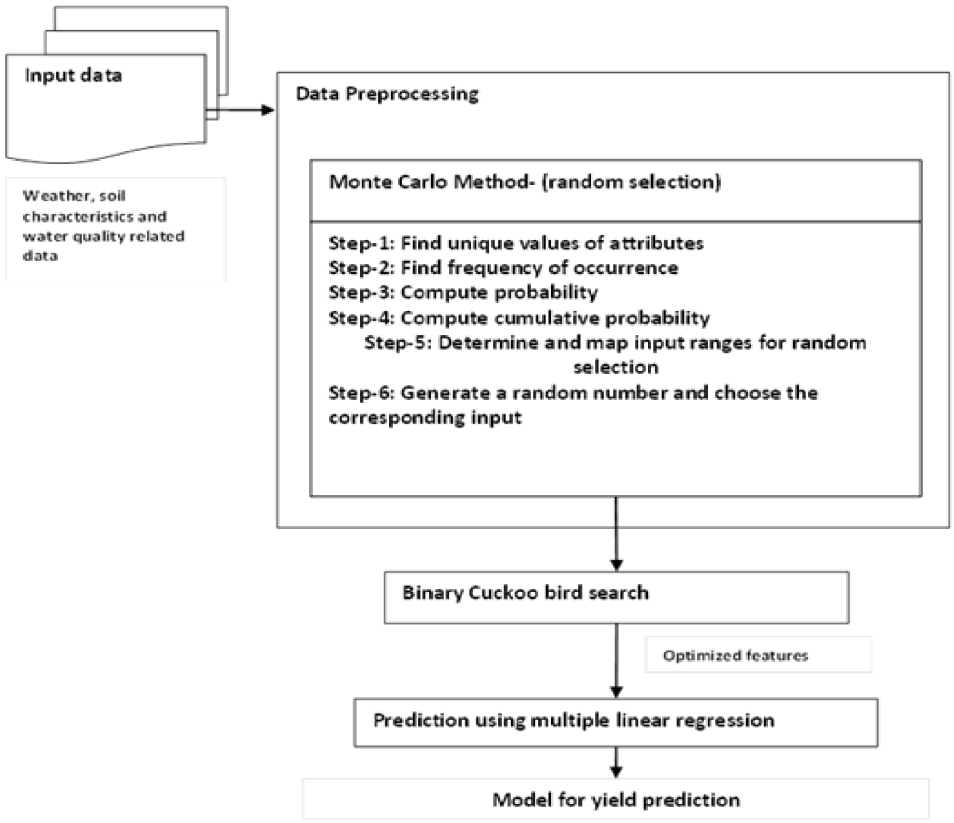

Figure 1. Proposed approach enhanced yield prediction

2. Related Work

There are various research works that discuss about the importance of preprocessing. The research work[5] discusses about different activities, namely, data cleaning, relevance analysis, data transformation, data reduction and generalization for the preparation data for prediction. In [6], the authors discussed that the amount of work involved in preprocessing accounts for 80% of the total workload and there are several preprocessing methods which are to be used according to the data available and problem in hand. In another research work [7], the authors used data set collected from agricultural web sites which is like to have standard format and thus requires only the removal of unwanted data. Selection of relevant features can be viewed as optimization problems. Cuckoo Search (CS) algorithm[8] is being used for selecting optimized features due to its large generalization capabilities. It imitates the reproduction of the cuckoo and combining the cuckoo nest's behaviour with L*é*vy's preference for flight.

Different categories and application of cuckoo search algorithm are discussed in [9]. CS algorithm is found to outperform Differential Evolution algorithm with respect to convergence speed and efficiency of computation[10-11]. An extended binary version of the Cuckoo Search, namely BCS, is being used for mapping continuous solutions produced by CS to binary ones while performing feature selection. The usefulness of BCS in identifying optimized features which yields to enhanced classification accuracy is discussed in [12-13]. Regression techniques are extensively used in prediction of crop yield [14], [15], [16], [17], [18].

Table 1. Data set description

Variable	Type
Station Code	String
Station Name	String
District	String
Latitude	String
Longitude	String
Year	Number
Month	Number
Day	Number
Hour	Number
$a1$-$a11$ (weather parameters)	Absolute pressure, min. temperature, max temperature, temp dry bulb, temp wet bulb, relative humidity, Instant wind speed, Ave. wind speed, wind direction, evaporation, rainfall
$a12$-$a28$ (Soil and water related parameters)	$TDS, No2+No3, Ca, Mg, Na, K, Cl, So4, Co3, HCo3, F, Ph, EC, HAR, SAR, RAC, Na\%$

Table 2. Sample Record

THANJAVUR		ThanjavurThanjavur			10°46′22”		79°08′09”		2007	1	1
08:30	1015.80	22.00	29.50	24.20	21.00	74.00	4.00	NE	0.80	10.00	
0.00	391	15	44	16	81	7	124	34	1	114	0.5
7.8	700	175	2.67	0.00	48.95						
THANJAVUR		ThanjavurThanjavur			10°46′22”		79°08′09”		2008	1	1
08:30	1012.40	21.00	30.00	23.00	21.60	74.00	2.00	2.52	NE	1.00	
0.00	641	10	40	58	117	11	17	67	0	323	0.1
8.0	1220	340 2.76	0.00	41.83	9.20						

3. Proposed Approach

An approach is proposed to enhance the prediction accuracy of crop yield by efficiently preprocessing the input data using Monte Carlo(MC) method for random selection of input and by selecting optimized features using binary cuckoo search algorithm. The proposed approach is shown in Figure. 1. The site-specific agriculture crop data set collected from Thanjavur, Tamil Nadu in India for the years from 2007 to 2016. It contains ten files with totally 7245 records. The description of the data set and sample record are given in Table 1 and Table 2 respectively.

3.1 Monte Carlo based Data Preprocessing

Monte Carlo(MC) method consists of solving a problem by constructing a random process with required parameters to that problem. It gives an approximate solution for the problem in hand quickly. Monto Carlo method is typically being applied when deterministic solution may break down. Monte Carlo method is used to simulate real time processes and phenomena and the simulation helps to choose one of the many different possible outcomes with randomness.

In the case of yield prediction, various parameters such as temperature, pressure, humidity, wind speed, wind direction, evaporation, water quality parameters, soil quality parameters, etc., are being typically collected from different regions and maintained by weather departments and agricultural departments are routine activity.

One of the important aspects regarding the data collection is that different parameters are being collected at different acquisition rate. For example, temperature, pressure, wind speed and wind direction are being routinely collected whereas parameters such as water quality parameters and soil quality parameters are collected at different rate say for example, once in 3 months. So, for an attribute such as temperature, there may be around 365 readings per year (with number of readings taken per days is 1) whereas for an attribute such as soil quality parameters there may be 4 values with number of reading taken per year is 4. In addition, for each crop, yield values are recorded once per year. In its raw form, the archived data has the form of multiple valued function characteristics. So, to arrive at a data study for prediction, it is required to preprocess the data in such way that will provide a one-to-one mapping between attributes which are relevant for yield prediction and yield(i.e. single valued function). How the input data is preprocessed using MC method is also illustrated in Figure.1

Algorithm-1
Monte Carlo-based Data Reduction & Random selection

Input: $D^t_{i,j}$
Output: $RD_{i,j}$
1. For $t = 1$ to F
2. Read File t
3. For $j = 1$ to M
4. Find Unique Value (UV)
5. Find Frequency of UV
6. Compute Probability
7. Compute cumulative Probability
8. Determine Input ranges corresponding cumulative probability
9. Select the Random Number (RN) with respect to input ranges
10. Get the Record based on RN
11. End For
12. End For

Listing. 1 Monte Carlo method for data preprocessing and random selection of inputs

Let consider the multiple data set $D^t_{i,j}$ where t = 1,2,...,F (number of files), i = 1,2,...M (number of rows) and j = 1,2,...N (number of columns). The proposed method uses two different techniques for preprocessing the data, (i) Monte Carlo based random selection and (ii) Average based method to reduce the data set $D^t_{i,j}$ into $RD_{i,j}$. Monte Carlo based method is a probability-based data reduction and selection method. It facilitates the random selection of inputs according to the algorithm given in Listing. 1

The above method is illustrated with five attributes, namely, $A1$, $A2$, $A3$, $A4$ and $A5$. Consider the sample data set with 5 columns and 15 records, given in Table 3. Further, the unique values of attributes, probability and cumulative probability are column $(A1)$ is shown in Table 4.

From the computed values of cumulative probability, the unique values of attributes are grouped into distinct range as given in Table 5. The grouped cumulative probability and unique values are shown in Table 6. After arriving the groups and cumulative probability, a random number with the range

Table 3. Sample Dataset

a1	a2	a3	a4	a5
1012.80	30.00	43.00	2.00	2.00
1013.50	23.00	85.00	1.00	1.00
1010.90	21.00	43.00	2.00	1.60
1012.60	24.00	81.00	2.00	1.40
1012.70	21.80	55.00	3.00	2.00
1013.30	23.00	83.00	2.00	1.60
1011.90	21.00	64.00	3.00	1.40
1013.30	23.00	76.00	2.00	0.00
1011.80	21.20	64.00	3.00	1.20
1016.50	22.40	80.00	2.00	1.00
1012.20	21.20	63.00	3.00	1.00
1015.50	23.60	90.00	2.00	0.00
1012.20	21.20	63.00	0.00	0.00
1016.10	24.80	87.00	3.00	1.00
1012.20	21.20	63.00	0.00	1.00

Table 4. Unique Values, probability and cumulative probability for Column-1 $(a1)$

Unique Value	Fre -que -ncy	Probability	Cumulative Probability
1012.6	1	0.06666666666666667	0.06666666666666667
1013.5	1	0.06666666666666667	0.13333333333333333
1010.9	1	0.06666666666666667	0.2
1011.8	1	0.06666666666666667	0.26666666666666666
1012.7	1	0.06666666666666667	0.3333333333333333
1011.9	1	0.06666666666666667	0.39999999999999997
1012.8	1	0.06666666666666667	0.4666666666666666
1015.5	1	0.06666666666666667	0.5333333333333333
1016.5	1	0.06666666666666667	0.6
1012.2	3	0.2	0.8
1013.3	2	0.13333333333333333	0.9333333333333333
1016.1	1	0.06666666666666667	1.0

Table 5. Range Values for cumulative probability

#	Range From	Range To
1	0	≤ 0.1
2	≥ 0.1	≤ 0.2
3	$\geq$0.2	≤ 0.3
4	$\geq$0.3	≤ 0.4
5	$\geq$0.4	≤ 0.5
6	$\geq$0.5	≤ 0.6
7	$\geq$0.6	≤ 0.7
8	$\geq$0.7	≤ 0.8
9	$\geq$0.8	≤ 0.9
10	$\geq$0.9	1

of groups is generated and the data against the generated random number is chosen as data corresponding to that year.

3.2 Average based preprocessing

The Average defined as the mean value, which is equal to the ratio of the sum of the number of a given set of values to the total number of values present in the set. In this preprocessing the conventional method of determining average is used.

Table 6. Average of unique values with corresponding average cumulative probability

Group	Unique Value	*UV* Average	Cumulative probability	*CF* Average
1	1012.6	1012.6	0.06666666666666667	0.06666666666666667
2	1013.5 1010.9	1012.2	0.13333333333333333 0.2	0.16666666666666669
3	1011.8	1011.8	0.26666666666666666	0.26666666666666666
4	1012.7 1011.9	1012.3	0.3333333333333333 0.39999999999999997	0.36666666666666664
5	1012.8	1012.8	0.4666666666666666	0.4666666666666666
6	1015.5 1016.5	1016.0	0.5333333333333333 0.6	0.5666666666666667
7	0	0	0	0
8	1012.2	1012.2	0.8	0.8
9	0	0	0	0
10	1013.3 1016.1	1014.7	0.9333333333333333 1.0	0.9666666666666667

3.3 Optimized feature selection

After performing MC method based preprocessing, the data is given to binary cuckoo search algorithm for optimized feature selection. Binary cuckoo search algorithm employs a boolean n-dimensional lattice in which the feature set solutions are represented using binary values, 0 and 1. It converges faster than conventional CS algorithm. The optimized feature set identified is given in Table 7.

Table 7. Selected features

Selected Attributes List						
a 1	a2	a3	a4	a5	a6	a7
a8	a9	a10	a12	a17	a19	a21
a23	a24	a25	a28			

Crop yield has been predicted using multiple regression algorithm and the results obtained using the proposed Monte Carlo based method of preprocessing and binary cuckoo search based optimized feature selection are compared with accuracy obtained using the conventional average based preprocessing. Predicted yield using the above methods are given in Table 9.

Table 8. Prediction result obtained using the proposed approach and average based preprocessing

Actual Data	Prediction using Monte Carlo based preprocessing and optimized feature set	Prediction using average based preprocessing and optimized feature set
541.0	495.09349330372953	538.0046831511645
462.0	351.1366986534291	517.193660309646
645.0	644.9999999695762	581.4499945712823
540.0	559.4885687599892	278.85307870939596
551.0	510.51757443163945	538.7549283271719
782.0	894.6067369385564	1014.6842257735452
705.0	694.4392422780593	600.6873664873578
1100.0	1220.3148535584405	1128.8953948625945
1350.0	1198.4880068261928	1001.9971903398555
12.6	119.51482528038753	488.0794774679962

Table 9. Correlation coefficient comparison

Dataset	All Attributes	Selected Attributes
Proposed approach	0.16	0.94
Average based method	0.110	0.124

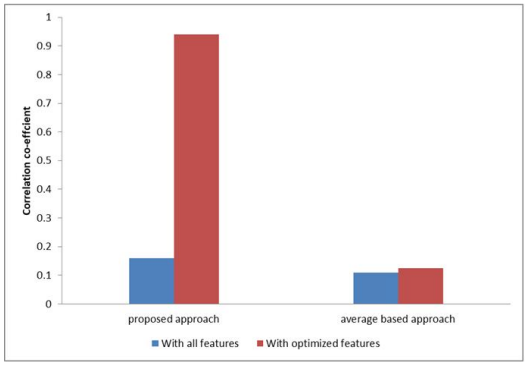

Figure 2. Comparison of correlation coefficient

The predicted values obtained using the above methods are evaluated using correlation coefficient. The correlation coefficient is compared as shown in Figure.2. It is found that the Monte Carlo based preprocessing with optimized feature selection outperforms the conventional average based preprocessing with optimized feature selection.

4. Conclusion

In this research paper an approach is proposed to enhance the prediction accuracy of crop yield prediction using multiple linear regression technique. In this approach Monte Carlo based method is used to pre-process the input data to transform multi valued function mapping between various attributes and crop yield into single valued function mapping. In addition, binary cuckoo search method is used to identify an optimized feature set for crop yield prediction. The proposed method has been tested with a site specific data collected from 2007 to 2016. From experimentation, the proposed approach is found to outperform the conventional average based preprocessing.

References

[1] Annachlingaryan, Brettwhelan Salahsukkarieh, Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and Electronics in Agriculture, 151(2018), 61–69.

[2] Yogesh Gandge, Sandhya, A study on various data mining techniques for crop yield prediction, *Electrical Electronics Communication Computer and Optimization Techniques (ICEECCOT) International Conference*, (2017), 420–423.

[3] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa, Machine Learning Methodologies for Paddy Yield Estimation in India: a Case Study, *Geoscience and Remote Sensing Symposium IGARSS, IEEE International*, (2019), 7254–7257.

[4] Potnuru Sai Nishant, Pinapa Sai Venkat, Bollu Lakshmi Avinash, B. Jabber, Crop Yield Prediction based on Indian Agriculture using Machine Learning, *Emerging Technology (INCET), International Conference*, (2020), 1–4.

[5] Jyotshna Solanki, Yusuf Mulge, Different Techniques Used in Data Mining in Agriculture, *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(5), (2015), 1223–1227.

[6] Chunling Li, Ben Niu, Design of smart agriculture based on big data and Internet of things, *International Journal of Distributed Sensor Networks*, 16(5), (2020), https://doi.org/10.1177/1550147720917065.

[7] Anusha A.Shettar, Shanmukhappa A. Angadi,Efficient data mining algorithms for agriculture data, *International Journal of Recent Trends in Engineering and Research*, 2(9), (2016), 142–149.

[8] X.S. Yang, S. Deb, Cuckoo search via Levy flights, *World Congress on Nature & Biologically Inspired Computing*, (2009), 210–214.

[9] Venkata Vijaya Geeta, Pentapalli, P. Ravi Kiran Varma, Cuckoo Search Optimization and its Applications: A Review, *International Journal of Advanced Research in Computer and Communication Engineering*, 5(11), 2016.

[10] M.I. Solihin, M.F. Zanil, Performance comparison of Cuckoo search and differential evolution algorithm for constrained optimization, *International Engineering Research and Innovation Symposium (IRIS)*, 160(1), (2016), 1–7.

[11] M.A. Adnan, M.A. Razzaque, A comparative study of particle swarm optimization and Cuckoo search techniques through problem-specific distance function, *2013 International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia*, 2013.

[12] D. Rodrigues, BCS: A Binary Cuckoo Search algorithm for feature selection, *2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing*, (2013), 465–468.

[13] S. Salesi and G. Cosma, A novel extended binary cuckoo search algorithm for feature selection, *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA), London,* (2017), 6–12,

[14] D.S. Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta, Machine Learning based Crop Prediction System Using Multi-Linear Regression, *International Journal of Emerging Technology and Computer Science,* 3(2), (2018), 31–37.

[15] Aditya Shastry, H.A. Sanjay, E. Bhanusree, Prediction of crop yield using regression techniques, *International Journal of Soft Computing*, 12(2), (2017), 96–102.

[16] Betty J. Sitienei, Shem G. Juma, and Everline Opere, On the Use of Regression Models to Predict Tea Crop Yield Responses to Climate Change: A

Case of Nandi East, *Sub-County of Nandi County, Kenya, Climate,* 5(54), (2017). doi:10.3390/cli5030054 www.mdpi.com/journal/climate

[17] V. Sellam and E. Poovammal, Prediction of Crop Yield using Regression Analysis, *Indian Journal of Science and Technology,* 9(38), (2016).

[18] M. Lavanya, R. Parameswari, A Multiple Linear Regressions Model for Crop Prediction with Adam Optimizer and Neural Network Mlraonn, *International Journal of Advanced Computer Science and Applications*, 11(4), (2020).

★★★★★★★★★
ISSN(P):2319 – 3786
Malaya Journal of Matematik
ISSN(O):2321 – 5666
★★★★★★★★★