Rice Crop Yield Prediction in India using Support Vector Machines

Niketa Gandhi
University Dept. of Computer Science
University of Mumbai
Mumbai, Maharashtra, India
niketa@gmail.com

Leisa J. Armstrong
University Dept. of Computer Science, University of Mumbai and School of Science, Edith Cowan University
Perth, Western Australia
l.armstrong@ecu.edu.au

Owaiz Petkar
University Dept. of Computer Science
University of Mumbai
Mumbai, Maharashtra, India
owaizpetkar@gmail.com

Amiya Kumar Tripathy
Department of Computer Engineering
Don Bosco Institute of Technology
Mumbai, Maharashtra, India
tripathy.a@gmail.com

Abstract— **Food production in India is largely dependent on cereal crops including rice, wheat and various pulses. The sustainability and productivity of rice growing areas is dependent on suitable climatic conditions. Variability in seasonal climate conditions can have detrimental effect, with incidents of drought reducing production. Developing better techniques to predict crop productivity in different climatic conditions can assist farmer and other stakeholders in better decision making in terms of agronomy and crop choice.**

Machine learning techniques can be used to improve prediction of crop yield under different climatic scenarios. This paper presents the review on use of such machine learning technique for Indian rice cropping areas. This paper discusses the experimental results obtained by applying SMO classifier using the WEKA tool on the dataset of 27 districts of Maharashtra state, India. The dataset considered for the rice crop yield prediction was sourced from publicly available Indian Government records. The parameters considered for the study were precipitation, minimum temperature, average temperature, maximum temperature and reference crop evapotranspiration, area, production and yield for the Kharif season (June to November) for the years 1998 to 2002. For the present study the mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) were calculated. The experimental results showed that the performance of other techniques on the same dataset was much better compared to SMO.

Keywords- artificial intelligence; crop analysis; crop yield; machine learning technique; prediction.

I. INTRODUCTION

With an increasing world population and changing climate, has come the necessity to secure the world food resources. Farmers are faced with having to make difficult decisions as to how to remain productive and sustainable with changing climates and market economic pressure. The provision of accurate and timely information such as meteorological, soil, use of fertilizers, use of pesticides can assist farmers to make the best decision for their cropping situations This could benefit them to attain greater crop productivity if the conditions are suitable or help them to decrease the loss due to unsuitable conditions for the crop yield. A number of studies have investigated how Information and Communication Technologies (ICT) can be applied to improve crop yield prediction and have successfully implemented in various climatic scenarios [1,2,3,4,5,6,7,8]. This paper examines the application of machine learning for the prediction of rice crop yield. The current study used a dataset from 27 districts as representative of Maharashtra, state in India. Various climatic factors which are known to affect the rice crop yield, such as precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration, were considered with the rice yield production for the Kharif season for the years 1998 to 2002. The Sequential Minimal Optimization (SMO) classifier using the WEKA tool has been applied on the current dataset. The analysis of results were undertaken and conclusions made as to its effectiveness for improving rice crop yield prediction

II. RELATED WORK

Support Vector Machines (SVMs) a supervised machine learning technique. There are a number of examples of where it has been used in the agricultural domain. Tripathi *et al.,* (2006) reported on how SVM was applied for reduction of precipitation for climate change scenarios [9]. To minimize the generalization error bound and to achieve generalized performance, SVM was used to forecast the demand and supply of pulp wood [10]. SVM was also applied to provide insights into crop response patterns related to climate conditions by providing the features contribution analysis for agricultural yield prediction [11]. For classification of agricultural datasets the use of discretization based Support Vector Machine was used [12].

Huang *et al.,* (2010) reported the use of SVM to model urban land use conversion. This study reported a relationship between

978-1-5090-2033-1/16/$31.00 ©2016 IEEE

various factors and rural-urban land use [13]. SVM has also been applied for the estimation of crop biophysical parameters with the use of aerial hyper spectral observations [14].

III. RESEARCH METHODS

This section discusses the methods used for this research and includes details of the study area, datasets and methodology.

A. Study Area

Maharashtra has a long coastline stretching nearly 720 kilometers along the Arabian Sea and occupies the western and central part of the country [15]. Figure 1 below shows the study area selected for this research. The state has a geographical area of 3,07,713 sq. km. It is bounded by North latitude 15°40' and 22°00' and East Longitudes 72°30' and 80°30'. The state has 35 districts which are divided into six revenue divisions viz. Konkan, Pune, Nashik, Aurangabad, Amravati and Nagpur for administrative purposes. For the present research, 27 districts were selected as representatives of the state depending on the data availability. The states selected were Ahmednagar, Amravati, Aurangabad, Beed/Bid, Bhandara, Buldhana, Chandrapur, Dhule, Gadchiroli, Gondia, Hingoli, Jalana, Jalgaon, Kolhapur, Latur, Nagpur, Nanded, Nasik, Osmanabad, Parbhani, Pune, Sangli, Satara, Solapur, Wardha, Washim and Yavatmal. Principal crops grown in the state are rice, jowar, bajra, wheat, tur, mung, urad, gram and other pulses [15].

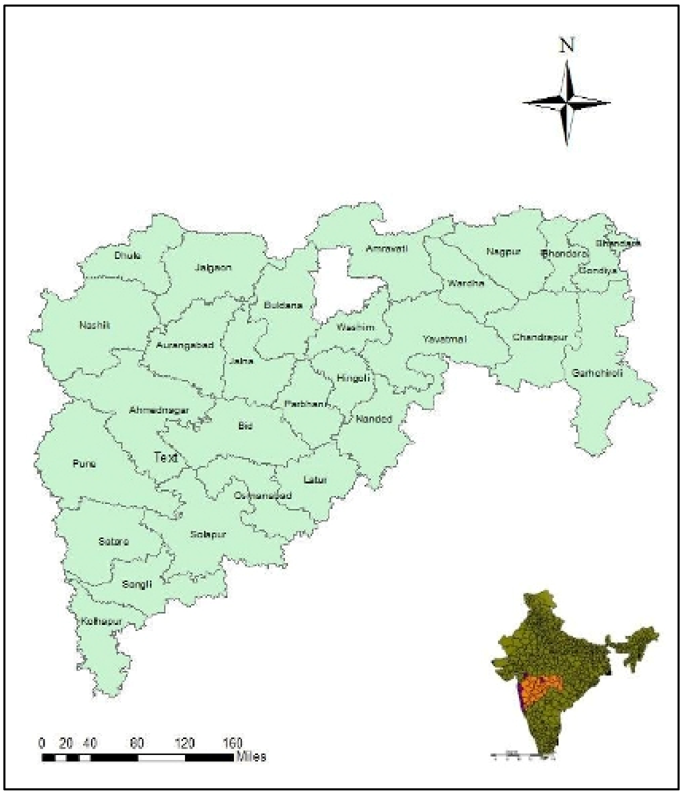

Figure 1 Study Area – Districts of Maharashtra State, India

B. Dataset Used

All the datasets used in the research were sourced from the openly accessible records of the Indian Government. This was sourced for the years 1998 to 2002 for the Kharif season of rice production. From the vast initial dataset, only a limited number of important factors which have the highest impact on agricultural yield were selected for the present research. Figure 2 below shows the parameters selected for the present study.

- Precipitation (mm): The total precipitation for Kharif season (June to November) for each year of every district was calculated from the monthly mean precipitation of that year for a particular district.

- Minimum, Average, Maximum Temperature (degree Celsius): Crop production will definitely have an impact due to variation in the temperature. Hence minimum, average and maximum temperature for each year of every district was considered for the present research. The average temperatures for the Kharif season (June to November) were calculated from the monthly mean temperature for minimum, average and maximum temperatures of that year for every district.

- Reference Crop Evapotranspiration (mm): The reference crop evapotranspiration was calculated on the basis of monthly mean of that year for the Kharif season for every district.

- Area (Hectares): The rice cultivated area in Kharif season (June to November) for every year in each selected district of Maharashtra state was considered for the present research.

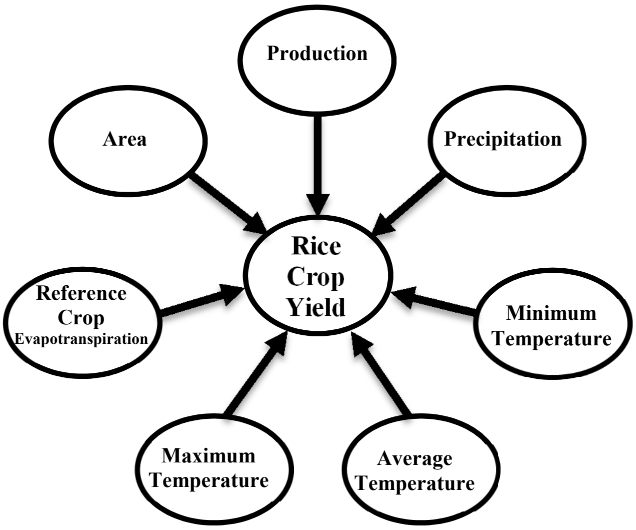

Figure 2 Study of climatic parameters on Maharashtra State, India

- Production (Tonnes): The rice production for the above cultivated area for Kharif season (June to November) for every year in each selected district of Maharashtra state was considered for the present research.

- Yield (Tonnes/Hectare): Depending on the rice production and the area cultivated for rice in Kharif season, for every year of each of the selected district from the Maharashtra state, the calculated yield was considered for the present research.

C. Methodology Used

The following steps were followed to prepare the data for processing after incorporating all the datasets of this study in to Microsoft Office Excel.

Step 1: Acquiring each parameter (precipitation, minimum, average, maximum temperature and reference crop evapotranspiration) monthly mean records of each district from 1998 to 2002 from the Indian Government records.

Step 2: Calculating the total precipitation, average temperature for the minimum, average and maximum temperature and average reference crop evapotranspiration for each year for each district during the Kharif season (June to November) of Maharashtra state.

Step 3: Acquiring each districts area, production and rice crop yield details of the year 1998 to 2002 from the publicly available Indian Government records.

Step 4: The raw data set was then collated in single sheet which consisted of the following columns in Microsoft Excel: sr. no, name of the district, year, precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration, area, production and yield.

Step 5: For some of the districts particular year's climatic parameters or production data was not available hence those year's data was not used for the current research. Record number was added for each record

Step 6: For preparing the data set for applying data mining techniques, unrequired columns were omitted. They were sr. no, name of the district and year.

Step 7: The data set was then sorted on the basis of area. Area less than 100 hectares were not considered for the present research. So those records were omitted.

Step 8: The dataset was then sorted on the basis of yield to classify the records in to low, moderate and high. The low yield was from 0.15 to 0.60 tonnes/hectare, moderate from 0.61 to 1.10 tonnes/hectare and high from 1.11 to 3.16 tonnes/hectare. Class low had 45 records with the range 0.15 to 0.60 tonnes/hectare, class moderate had 46 records with the range 0.61 to 1.10 tonnes/hectare and class high had 44 records with the range 1.11 to 3.16 tonnes/hectare.

Step 9: The yield has been calculated on the basis of area and production hence these two columns were omitted.

Step 10: This data set was then saved in .csv format for further applying data mining techniques. This file had following columns: precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration, crop yield and class.

The WEKA tool is a freely available and open source data mining tool available under the GNU General Public License [16]. The data set prepared for the present exploration was saved in .arff file format and processed through WEKA to build the algorithm on the current data set.

Support Vector Machine

The SVM approach [17] is used to create functions from a set of labeled training data. These functions can be a classification function or it can be general regression function. For the current study SMO algorithm was used to study the performance of this approach on the dataset used for the present study. The results were generated in WEKA by using the SMO algorithm. The results were further evaluated using various performance measures as discussed in the next section.

IV. PERFORMANCE EVALUATION

Each instance is classified into two classes in a bunary classification model. The two classes are true and false class. This gives rise to four possible classifications for each instance namely:
True Positive (TP): The number of correct predictions that an instance is positive.
False Positive (FP): The number of incorrect predictions that an instance is positive.
False Negative (FN): The number of incorrect predictions that an instance is negative.
True Negative (TN): The number of correct predictions that an instance is negative.
This situation can be depicted as a confusion matrix also called contingency table as shown in table 1 below.

Table 1 A confusion matrix

		Observed	
		True	**False**
Predicted	**True**	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

The observed classifications for a phenomenon are compared with the predicted classifications of a model in a confusion matrix. In table 1 the classification that are shown along the major diagonal of the table are the correct classifications refereed as true positives and true negatives. The model errors are signified by the other fields. Only the true positive and true negative fields would be filled out for a perfect model and the other fields would be set to zero. From the confusion matrix, a number of model performance metrics can be derived.

The most common metric is accuracy which is defined as the overall success rate of the classifier and is computed as

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN)$$

Other performance metrics include Sensitivity/Recall and Specificity/Precision. Sensitivity is defined as percentage of correctly classified instances. Specificity is defined as percentage of incorrectly classified instances. These can be computed as

$$\text{Sensitivity/Recall} = TP / (TP + FN)$$

$$\text{Specificity/Precision} = TP / (TP + FP)$$

F1 Score is a measure of test's accuracy. To compute the score it considers precision and recall. F1 score is the harmonic mean of precision and recall. F1 score can be computed as

$$F1 = (2TP)/(2TP + FP + FN)$$

The weighted average of the precision and recall is referred as F1 score. It reaches its best value at 1 and worst at 0.

Mathews Correlation Coefficient (MCC) is computed as a measure of the quality of classification. It is considered as a balanced measure that can be used even if the classes are of different sizes by considering true and false positives and negatives. The MCC returns a value between −1 and +1 which is a correlation coefficient between the observed and predicted binary classifications. A perfect prediction is represented by coefficient of +1, random prediction by 0 and total disagreement between prediction and observation by -1. It can be computed as shown below.

$$MCC=((TP*TN)-(FP*FN))/\surd(TP+FP)(TP+FN)(TN+FP)(TN+FN)$$

For reference and evaluation, the relative mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) are also computed.

V. EXPERIMENTAL RESULTS

This section discusses the results obtained after applying the SMO technique on rice crop yield dataset of Maharashtra state, India. WEKA was used to construct the algorithm. The different parameters set for SMO algorithm were as follows: biildLogisticModels = false; c = 1.0; checksTurnedOff = false; debug = false; epsilon = 1.0E – 12; filterType = Normalize training data; kernel = PolyKernel; numFolds = 1; randomSeed = 1; toleranceParameter = 0.001. The algorithm achieved the accuracy of 78.76%, sensitivity of 68.17% and specificity of 83.97%.

The F1 score was computed to measure the test's accuracy and achieved a score of 0.69. Mathews Correlation Coefficient was used to measure the quality of classification which resulted in 0.54. The error results of the classifier are mean absolute error of 0.23, root mean squared error of 0.39, relative absolute error of 67.38% and root relative squared error of 82.51%.

VI. DISCUSSION AND CONCLUSIONS

In recent years, great efforts have been undertaken on the challenging task of predicting rice crop yield. Developing accurate models for crop yield estimation using Information and Communication Technologies may help farmers and other stakeholders improve decision making in relation to national food import/exports and food security.

Rice is one of the most important food crops of India. It is cultivated all over the country and contributes more than 40% of total food grain production [18]. Given the importance of rice to world's food security, any improvements in the forecasting of rice crop yield under different climatic and cropping scenarios will be beneficial.

This research has demonstrated the prediction of rice crop yield by applying one of the machine learning technique, support vector machine (SVM). The experimental results showed that the other classifiers such as Naïve Bayes, BayesNet and Multilayer Perceptron performed better by achieving the highest accuracy, sensitivity and specificity compared to SMO classifier with lowest accuracy, sensitivity and specificity that has been reported earlier for the same data set [19,20]. In terms of test's accuracy and quality also BayesNet and Multilayer Perceptron showed the highest accuracy and best quality and SMO showed the lowest accuracy and worst quality. It can be concluded that other classifiers used on the current study dataset and reported earlier should be recommended for further development of a rice prediction model [19, 20].

REFERENCES

[1] R. Medar, V. Rajpurohit, "A survey on data mining techniques for crop yield prediction", International Journal of Advance Research in Computer Science and Management Studies, vol. 2, no. 9, pp. 59-64, 2014.

[2] S. Bejo, S. Mustaffha and W. Ismail, "Application of artificial neural network in predicting crop yield: A review", Journal of Food Science and Engineering, vol. 4, pp.1-9, 2014.

[3] S. Dahikar and S. Rode, "Agricultural crop yield prediction using artificial neural network approach", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 2, no. 1, pp. 683-686, 2014.

[4] W. Guo and H. Xue, "An incorporative statistic and neural approach for crop yield modelling and forecasting", Neural Computing and Applications, vol. 21, pp. 109-117, 2012.

[5] W. Guo and H. Xue, "Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models", Mathematical Problems in Engineering, pp.1-7, 2014.

[6] D. Ramesh and B. Vardhan, "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473, 2015.

[7] K. Tanaka and T. Kiura, "Crop yield prediction systems for rainfed areas and mountainous areas in Thailand", Proceedings of the 9th Conference of the Asian Federation for Information Technology in Agriculture "ICT's for future Economic and Sustainable Agricultural Systems", 2014.

[8] G. Yengoh and J. Ardo, "Crop yield gaps in Cameroon", AMBIO, Springer, vol. 43, pp. 175-190, 2014.

[9] S. Tripathi, V.V. Srinivas and R.S. Nanjundiah, "Downscaling of precipitation for climate change scenarios: a support vector machine approach", Journal of Hydrology, vol. 330, no. 3, pp.621-640, 2006

[10] V. Anandhi and R.M. Chezian, “Support Vector Regression in forecasting”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, no. 10, October 2013.

[11] S. Brdar, D. Culibrk, B.Marinkovic, J.Crnobarac and V. Cmojevic, “Support Vector Machines with Features Contribution Analysis for Agricultural Yield Prediction”, 2011.

[12] A. Bharadwaj, S. Dahiya, and R. Jain, “Discretization based Support Vector Machine (D-SVM) for Classification of Agricultural Datasets”, International Journal of Computer Applications, vol. 40, no. 1, pp.8-12, 2012.

[13] B. Huang, C. Xie, and R. Tay, “Support vector machines for urban growth modeling” Geoinformatica, vol. 14, no. 1, pp.83-99, 2010.

[14] Y. Karimi, S.O. Prasher, A. Madani, A. and S. Kim, “Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations”, Canadian Biosystems Engineering, vol. 50, no. 7, pp.13-20. 2008.

[15] Report on Economic Survey of Maharashtra 2012-2013, Directorate of Economics and Statistics, Planning Department, Government of Maharashtra, Mumbai (2013).

[16] Weka 3:Data Mining Software in Java, Machine Learning Group at the University of Waikato, Official Web: http://www.cs.waikato.ac.nz/ml/weka/index.html, accessed on 26th March 2016.

[17] C. Saunders, M.O. Stitson, J. Weston, L. Bottou, and A. Smola, Support vector machine-reference manual, 1998.

[18] Gain Report on Global Agricultural Information Network, India Grain and Feed Annual, USDA Foreign Agricultural Service (2014).

[19] N.Gandhi, L.J. Armstrong and O. Petkar, “Predicting Rice Crop Yield using Bayesian Networks”, communicated, 2016.

[20] N.Gandhi, L.J. Armstrong and O. Petkar, “Rice Crop Yield Prediction in India using Artificial Neural Network”, International Conference on 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India scheduled on 15th and 16th July 2016.