AN INTENSIVE STUDY ON PRECISION AGRICULTURE: CROP YIELD PREDICTION

UN ESTUDIO INTENSIVO SOBRE LA AGRICULTURA DE PRECISIÓN: PREDICCIÓN DEL RENDIMIENTO DE CULTIVOS

UM ESTUDO INTENSIVO SOBRE A AGRICULTURA DE PRECISÃO: PREDIÇÃO DA PRODUÇÃO DE CULTURAS

P. Suvithavani [1], Dr. S. Rathi [2]

[1] Research Scholar(Part Time), Assistant Professor, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering & Technology, Coimbatore - 641062. E-mail: suvitha.gct@gmail.com; psvani.gct@gmail.com;

[2] Assistant Professor (sr. grade), Department of Computer Science and Engineering, Government College of Technology, Coimbatore-13. E-mail: rathi@gct.ac.in

Abstract

Farming is largely a challenging and cumbersome profession. The unstable and volatile commodities market squeezes the life out of a farmer who already is experiencing unprecedented scarcity of water on top of ever-rising operational costs. Farmers are constantly subjected to restrictive regulations on irrigation, pesticide use and fertilizer application, which leads them to explore and find new ways to boost the agricultural yield. Fortunately, a huge amount of data is available on modern farms ranging from yield monitors to infrared imaging, but the sad state of affairs that agricultural profession is ages behind other industries in utilizing data to make professional decisions. Soon using data to optimize decision making will no longer be a novelty, but an essential practice to stay afloat in business. This paper discusses different decision making algorithms such as Support Vector Machine (SVM), Bayes Model, Neural Network (NN), Random Forest (RF), and methods. The challenge identifying predictive abilities using promising methods with a small dataset and the characteristics of different machine learning algorithms have been discussed. It has many issues such as learning performance, computation time and scalability. These issues are also discussed detail in this review work.

Keywords: Farming, agricultural , machine learning algorithms, crop yield prediction, dataset, and classification.

Resumen

La agricultura es en gran medida una profesión desafiante y engorrosa. El inestable y volátil mercado de productos básicos exprime la vida de un agricultor que ya está experimentando una escasez de agua sin precedentes además de los crecientes costos operativos. Los agricultores están constantemente sujetos a regulaciones restrictivas sobre el riego, el uso de pesticidas y la aplicación de fertilizantes, lo que los lleva a explorar y encontrar nuevas formas de aumentar el rendimiento agrícola. Afortunadamente, hay una gran cantidad de datos disponibles en granjas modernas que van desde monitores de rendimiento hasta imágenes infrarrojas, pero el triste estado de los asuntos que la profesión agrícola está envejeciendo detrás de otras industrias en la utilización de datos para tomar decisiones profesionales. Pronto usar datos para optimizar la toma de decisiones ya no será una novedad, sino una práctica esencial para mantenerse a flote en los negocios. Este documento discute diferentes algoritmos de toma de decisiones como Support Vector Machine (SVM), Bayes Model, Neural Network (NN), Random Forest (RF) y métodos. Se discutió el desafío de identificar capacidades predictivas usando métodos prometedores con un pequeño conjunto de datos y las características de diferentes algoritmos de aprendizaje automático. Tiene muchos problemas, como el rendimiento de aprendizaje, el tiempo de cálculo y la escalabilidad. Estos problemas también se discuten en detalle en este trabajo de revisión.

Palabras clave: agricultura, agricultura, algoritmos de aprendizaje automático, predicción del rendimiento de los cultivos, conjunto de datos y clasificación.

Abstrato

A agricultura é em grande parte uma profissão desafiadora e incômoda. O instável e volátil mercado de commodities espreme a vida de um agricultor que já está experimentando uma escassez de água sem precedentes, além dos crescentes custos operacionais. Os agricultores estão constantemente sujeitos a regulamentações restritivas sobre irrigação, uso de pesticidas e aplicação de fertilizantes, o que os leva a explorar e encontrar novas maneiras de aumentar o rendimento agrícola. Felizmente, uma enorme quantidade de dados está disponível em fazendas modernas que vão desde monitores de rendimento até imagens de infravermelho, mas o triste estado de coisas que a profissão agrícola está atrás de outras indústrias ao utilizar dados para tomar decisões profissionais. Em breve, o uso de dados para otimizar a tomada de decisões não será mais uma novidade, mas uma prática essencial para se manter à tona nos negócios. Este artigo discute diferentes algoritmos de tomada de decisões, como a Máquina de Vetor de Suporte (SVM), o Modelo de Bayes, a Rede Neural (NN), a Floresta Aleatória (RF) e os métodos. O desafio de identificar habilidades preditivas usando métodos promissores com um pequeno conjunto de dados e as características de diferentes algoritmos de aprendizado de máquina tem sido discutido. Tem muitos problemas, como desempenho de aprendizado, tempo de computação e escalabilidade. Essas questões também são discutidas em detalhe neste trabalho de revisão.

Palavras-chave: Algoritmos de agricultura, agrícolas, aprendizado de máquina, predição de produção agrícola, conjunto de dados e classificação.

1. Introduction

The world population is constantly in an upward momentum in the background of unpredictable climatic changes. Farmers are kept on their toes as they face the conundrum of making tedious decisions on staying fruitful and sustainable in the ever varying climatic and economic conditions. Using computers could help farmer ward of these challenges; they can use computers in assessing the fertility of the soil nutrient, what fertilizers and pesticides best suit their land, etc. This could help them in attaining excellent crop productivity in case the conditions are appropriate and decrease their loss due if the conditions do not favor them (Niketa Gandhi et al, 2016).

Crop growth (Mo et al, 2005; Folberth et al, 2012) quality and yield greatly rely on weather and environmental factors such as seasonal temperature and precipitation changes, day-to-day temperature ranges and, water cycles existing between soil and atmosphere. As in every vocation, less productivity in agriculture translates to increased price of the produce in the market. In this backdrop, machine learning algorithms (Snoek et al 2012; Nasrabadi 2017; Pedregosa et al 2011) will help the farmers to double the crop production. Having better crop predictions can aid the farmers in improving their nitrogen management to satisfy the needs of the new crop and mill managers could make better plans about the mill's labor needs and also the maintenance scheduling activities, and likewise, the marketers can carry out the management of the forward sale and storage of the crop with more confidence. Therefore, accurate yield forecasts can enhance the sustainability of the industry by providing better environmental and economic results. The predictor variables include variables that are dependent on indices used for simulated biomass, previous yields, local climate

1. Introducción

La población mundial está constantemente en un momento ascendente en el contexto de cambios climáticos impredecibles. Los agricultores se mantienen alerta mientras enfrentan el enigma de tomar decisiones tediosas para mantenerse fructíferos y sostenibles en las condiciones climáticas y económicas siempre cambiantes. Usar computadoras podría ayudar al agricultor a enfrentar estos desafíos; pueden usar computadoras para evaluar la fertilidad del nutriente del suelo, qué fertilizantes y pesticidas se adaptan mejor a su tierra, etc. Esto podría ayudarlos a lograr una excelente productividad de cultivos en caso de que las condiciones sean apropiadas y disminuir sus pérdidas si las condiciones no favorecen ellos (Niketa Gandhi et al, 2016).

El crecimiento de los cultivos (Mo et al, 2005; Folberth et al, 2012) dependen en gran medida del clima y los factores ambientales, como la temperatura estacional y los cambios de precipitación, los rangos de temperatura diarios y los ciclos del agua existentes entre el suelo y la atmósfera. Como en toda vocación, una menor productividad en la agricultura se traduce en un mayor precio del producto en el mercado. En este contexto, los algoritmos de aprendizaje automático (Snoek et al 2012; Nasrabadi 2017; Pedregosa et al 2011) ayudarán a los agricultores a duplicar la producción de cultivos. Tener mejores predicciones de cultivos puede ayudar a los agricultores a mejorar su gestión de nitrógeno para satisfacer las necesidades de la nueva cosecha y los gerentes podrían hacer mejores planes sobre las necesidades laborales de la planta y las actividades de programación de mantenimiento, y asimismo, los comercializadores pueden llevar a cabo la gestión de la venta y almacenamien information consisting of rainfall, radiation, and maximum and minimum temperature. Big data technologies maximize the industry directions over the important industry decisions, which impact the sustainable agricultural systems and thus bring in a partial solution to issues relating food shortages kept as scope of future work.

The structure of this work is provided as below: Section 2 explains the Architecture of Smart Agriculture Management System (SAMS). Section 3 explains the importance of Predictive Analytics in Precision Agriculture. Section 4 explains the Survey on Crop yield Prediction. Section 5 explains the inference and conclusion of this study.

to a futuro del cultivo con más confianza. Por lo tanto, los pronósticos de rendimiento precisos pueden mejorar la sostenibilidad de la industria proporcionando mejores resultados ambientales y económicos. Las variables de predicción incluyen variables que dependen de los índices utilizados para biomasa simulada, rendimientos anteriores, información climática local que consiste en lluvia, radiación y temperatura máxima y mínima. Las tecnologías de Big Data maximizan las instrucciones de la industria sobre las decisiones importantes de la industria, que impactan en los sistemas agrícolas sostenibles y, por lo tanto, aportan una solución parcial a los problemas relacionados con la escasez de alimentos que se mantienen como alcance del trabajo futuro.

La estructura de este trabajo se proporciona a continuación: la Sección 2 explica la Arquitectura del Sistema de Gestión de Agricultura Inteligente (SAMS). La Sección 3 explica la importancia de Predictive Analytics en Precision Agriculture. La Sección 4 explica la Encuesta sobre la Predicción del rendimiento de los cultivos. La Sección 5 explica la inferencia y conclusión de este estudio.

2. Architecture of Smart Agriculture Management System (SAMS)

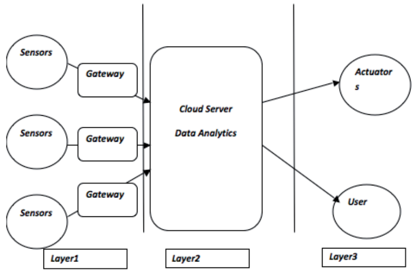

Figure 1: Smart Agriculture Management System

Figure 1 shows the architecture of Smart Agriculture management Systems, which has three layers. The first layer comprises sensors and gateways where the physical environment changes are monitored and sent to a cloud server to perform analytics through the gateway. The second layer comprises a cloud server in which Analytics Algorithm can be run and smart decisions be made. Then the decision would be sent to the third layer, where the user or actuators will respond to the decision. In this survey we are focusing on algorithms involved in Data Analytics of Layer 2.

In the concept of Internet of Things (IoT), the server should be intelligent enough to make decisions appropriately. The sensor monitors the soil moisture, leaf wetness, temperature, and humidity level in the environment and sends the data to the server through the gateway in which it performs the analytics and then the server sends the recommendations to the farmer's mobile phone on which the actions of the farmers can be based. (Raheela, 2016).

3. Predictive Analytics in Precision Agriculture

Because of increasing demand of decision support systems, Precision Agriculture can be used as an effective tool. The services that can be obtained using Precision Agriculture are information services, traceability systems, precision irrigation, monitoring, controlling and management of the field (Shailaja Patil,2016). With the help of Predictive Analytics, we can get realtime data on climate, soil and air quality, planting, crop maturity, equipment and labour costs and availability. These data will help us in making ideal decisions regarding a sustainable agricultural development plan. The goal is to get better understanding of the land, weather, climate and planting, which can help in prediction of events with greater accuracy which in turn can easily translate to sustainable agriculture (Khushboo Babaria,2015). We have 4 different steps in this process; (1) fresh data gathering and cleaning; (2) renovating the cleaned data into a desired format that can be used by the machine learning method; (3) generating a predictive model (training) using the renovated data; (4) reporting predictions to the user based on the previously created predictive model. The learning model applies the following areas in agriculture domain:

1. Prediction of Crop Yield
2. Plant Diseases Detection and Classification
3. Management of Fertilizers and Pesticides
4. Ranking and Categorizing of Agriculture Products
5. Supervision of Soil Fertility

4. Literature Survey

This survey will give a brief narration about classification algorithm in machine learning, which can help in building a model to predict crop yield. These predictions will help the farmer to produce good yield and maintain the sustainability of soil. A few classification algorithms are discussed in this survey in the perspectives of crop prediction; they are Support Vector Machine (SVM), Naïve Bayes (NBs), Neural Network (NN) and Random Forest (RF).

Support Vector Machine:

Support Vector Machines (SVMs) are employed for detecting and exploiting thecomplicated patterns existing in data through the clustering, classification and ranking of the data. This algorithm is applied to Downscaling in hydro-climatology application. The data relating to the climate variables (predictors) identified at every region are classified with the help of cluster analysis to create two groups, which represent wet and dry seasons. For every region, SVM- based Downscaling Model (DM) is designed for season(s) with considerable rainfall making use of primary components obtained from the predictors in the form of input and the contemporary precipitation seen at the area in the form of an output. The SVM would be an ideal choice for the downscaling task due to its capability of providing a better generalization performance in acquiring non-linear regression associations between predictors and predictions in spite of the fact that it does not include the knowledge about the problem domain. These techniques are applicable for considerably smaller data sets (N _ 2000, based on the computer memory). For big data sets, using standard SVM and Least Squares Support Vector Machines (LS-SVMs) is an issue. The quality of algorithm is measured with Normalized Mean Square Error (NMSE). The NMSE value for rainfall prediction on Bihar Plateau is 0.56(training) and 0.46(testing), Kerala is 0.24(training) and 0.46(testing), Coastal Karnataka is 0.42(training) and 0.96(testing), and Orissa is 0.45 (training) and 0.27 (testing) (Tripathi et al.,2006).

The SVM predicts crop response patterns related to climate conditions. This algorithm is applied to agricultural yield prediction application. This model estimates the contribution made by every feature on the entire range of its input values. Over a range beginning from the minimal to maximal value for a certain feature, 'm' equally distributed points are drawn and the contribution of the feature in these points are calculated. Modification lies in the random sampling that is used here for the complete instance, with an exception made for ith feature. Linear regression is generally utilized because of its simple nature, but it depicts poor performance if used for complex issues. Non-linear methods can be more suitable for these problems. The comprehensibility of models can be achieved with better sampling method, which would deal with the issue of correlations and interactions happening among attributes. The quality of the prediction algorithm is with Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values. The values are 449.41 551.67 for maize dataset, 281.23 342.69 for soybeans and 4477.70 6315.52 for sugar beet. (S. Brdar et al.,2011)

The SVM algorithm is applied to Rice Crop Yield Prediction application. The SVM Model was applied to labeled and processed dataset collected from the publicly available records of Government of India. The selected parameters for this model were hail, temperature (min, max, avg), evapotranspira tion, production and area. The Self-Organizing Map (SOM) algorithm was used from the WEKA tool to perform the SVM Classification on selected dataset. Every instance was divided into two classes present in a binary classification model. This method minimized generalization errors and helped achieve generalized performance.It needs improvements with small dataset. The accuracy of this algorithm is 78.76% (Niketa Gandhi et al,2016).

Bayes Model:

Natural resource problems are generally modeled with the help of data, which is mostly incomplete, non-synchronous and gathered at various spatial and temporal scales with diverse degrees of uncertainty. Changes owing to climate, soil, pests and management decisions add up more to the structural and functional complications of the ecosystems managed. Bayesian networks are ideally suitable for these conditions by facilitating the diagnostic-reasoning on conditional dependencies for assessing the model, structural in addition to parameter uncertainty. The author proposed the Naive-Bayes (NB), Tree-structured Naive-Bayes (TAN) and General Bayes (GB) learning on the Annual yield of barley straw data. The GB technique, depending on every correlation as specified from expert information, also replicates the central peak, but overestimates peak yields higher than Tree-structured Naive-Bayes (TAN). It lacks in general model framework to take more fertilizer kinds (P - phosphorus, Ca - calcium), into consideration and to distinguish various kinds of yield losses by isolating separate nodes for pests/weeds/pathogens and severe weather conditions. The NB classifier performs the prediction of a multi-modal distribution with highest yields that are predicted between 2300 and 2700 kg/ha, TAN with largest yield 2800 kg/ha and the GB predicted at higher yields approximating to 3600 kg/ ha (Nathaniel K. Newlands, 2010).

A spatial model of maize yields in the US Corn Belt from 1970-2012 was made using a Bayesian prior, which induces spatial smoothness among the regression coefficients to mitigate the effects of noisy data across regions and improve yield forecasting to estimate an individual set of regression coefficients for every state and use a Bayesian prior over these coefficients that is spatially smooth. This approach is straightforward and has been shown to be an effective way to model crop yield. Because daily climate observations are high-dimensional, it is unwise to include all the observations directly into the regression model. It reduces the dimensionality of Growing Degree Days (GDDs) by including average GDD and squared average GDD in the regression models. Simple least squares model does not do a good job of capturing the complexity of crop yield. Fitting a separate least squares model for each state improves the fit of the model, but makes generalization to a new year problematic because of overfitting. The spatial model with the Multivariate Conditional Auto Regressive (MCAR) prior does almost as well as the multiple least squares approach in terms of variance explained and does drastically better in terms of predictive performance. This algorithm will not work with very different climate time series; it is likely the case that we are not properly modeling the relationship between yield and climate. The simple LS with R2:0.44 RMSE:42.96, LS by State with R2:0.50 RMSE:65.36, MCAR with R2: 0.50 RMSE:26.11. (Charles L. Hornbaker II, 2013).

This research work studies the usage of Bayesian Networks for predicting the rice crop yield for the state of Maharashtra, India. The parameters selected for the study include precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration, area, production and yield for the Kharif season (June to November) for the years, 1998 to 2002. The BayesNet and Naïve Bayes algorithm from WEKA tool were used to implement the selected parameters. Of these two algorithms, BayesNet is the better one with good accuracy, specificity, and sensitivity. Also, BayesNet is better than SMO, Support Vector Machine (SVM). Bayesian Network (BN) is the best option to build the model for crop prediction. These methods can be utilized for assessing the model structures and also the uncertainty pertaining to the parameters. The systems with rising complexities also consider this method to be greatly desirable. BayesNet with Accuracy of 97.53%, MAE of 0.0425, RMSE of 0.1449. Naïve Bayes (NBs) with accuracy with 84.69%, MAE of 0.1456 and RMSE of 0.2999 (Niketa Gandhi et al, 2016).

The real NBs have a critical drawback that is the production of repetitive predictors. The regularization method was utilized forgetting a computationally effective classifier dependent on NBs. The construction suggested, used L1-penalty has the capability of eliminating repetitive predictors, in which a modified version of the Least-Angle Regression (LARS) algorithm is designed for solving this issue, rendering this technique suitable for an extensive range of data. This method supports both numerical and categorical predictors. The new data are generated from the existing by including the redundant and irrelevant values. The issues of irrelevant predictors and redundant predictors were handled in this modi fied algorithm. Selective Naïve Bayes (SNB) is observed to be much better compared to Naïve Bayes (NB) model, as the L1-penalty tackles with redundancy. NB accuracy is 0.37(±0.01) with number of predictors is 100.0(±0.0) and SNB accuracy is 0.31(±0.02) with number of predictors is 60.5(±6.2) (Kefaya Qaddoum et.al, 2014).

A spatio-temporal yield model is estimated with the help of Bayesian method referred to as Markov Chain Monte Carlo. By standardizing the simulated variables over the Normalized Difference Vegetation Index (NDVI), the impact of drought related variables on wheat yield is explored and two variables are obtained.

They are ratio of the original evapotranspiration for the difference computed between the NDVI values noticed at the beginning and at the peak, and also the ratio of the absolute value of water deficiency for the difference computed between the NDVI values noticed at the beginning and at the peak. It shows the important relationship between wheat yield and these indicators at the initial phenological period in addition to the flowering and ripening phenological periods added together. The size of the data set is not sufficient for quantifying the underlying the real correlation between the yield and its predictors. The association between two newly created variables is checked and the yield for province based data gathered for the same period of time.

The farming provinces with cultivating farms contribute the greatest correlations among all the provinces indicating that the usage of these variables must be restricted to these particular geographical units (KasirgaYildirak et.al,2015).

Bayesian networks can exhibit superior prediction accuracy even in the case of much small sample sizes. The estimation of the conditional probabilities of the model can be done from data employing an Expectation-Maximization (EM) algorithm. It needs just the structure of the model to be known in prior, and iteratively computes the maximum likelihood estimates for the parameters, with the data and the model structure given. Environmental data frequently include the values that are missing, as the problems in sampling may indicate that some distinct event or point in time gets missed. Dissimilar to several estimation techniques, EM algorithms can deal with conditions related to missing observations; it could be that the data is missing in random or a missing observation depends on the states of the remaining variables. The distributions corresponding to the partially complete data can be approximated making use of Dirichlet distributions. One significant characteristic of Bayesian techniques is the usage of information obtained beforehand. Priors are a reflection of knowledge about the subject before the conduction of the research, and could be both hugely informational and comprehensive in the event, where there is prolific information about the subject in prior or on the other hand if not much information is present. These priors are thereafter updated with the data to get a synthesis about the old knowledge and fresh data. Then this synthesis can be utilized in the form of a prior in a fresh new research. This technique renders the scientific learning process to be open, and also keeps the assumptions created by the scientists to be transparent and overt to discussion. As BNs are analytically solved, they can yield quicker responses to the queries, when the model gets compiled. The compiled form of a BN comprises of a conditional probability distribution for all the combinations of variable values and can therefore render any distribution instantaneously, in contrary to simulation models, where the results have to be simulated that can be a very lengthy process. In environmental research and also several other fields, data and parameters frequently have continuous values. However, Bayesian networks can manage continuous variables only in a limited fashion. (L. Uusitalo,2007)

A Bayesian network model is introduced for the forecasting of monthly rainfall at 21 stations located in Assam, India. Bayesian Network (BN) is typically a probabilistic graphical model, which exhibits conditional probabilities between multiple variables/nodes. Rainfall at a station is considered to be a variable for this model and the dependencies between rainfalls at individual stations is indicated by BN. Rainfall dependencies existing between various stations are computed employing K2 algorithm that, in turn, gets BN on the basis of a greedy search algorithm. Five local and global atmospheric parameters, including Temperature, Relative Humidity,

Wind Speed, Cloud Cover and Southern Oscillation Index (SOI) are utilized in the form of proofs for this model. Conditional probabilities between stations and atmospheric parameters are computed with Maximum Likelihood Parameter Estimation (MLE). Monthly data of 20 years from the period of 1981 to 2000 for every parameter is utilized for this research work, obtained from various sources. Bayesian model executes on discretized data, and therefore for this work, three discretized values have been taken into account for every variable on the basis of their distribution. Thirteen diverse combinations consisting of five atmospheric parameters are analyzed, providing a comparison on the efficiency of various parameters involved in rainfall prediction. It deals with continuous variables in just a limited fashion. The BN model is effective when the correlations between the variables show non-linearity and complexity. The model is tested across 21 different stations. The highest accuracy station is Kokrajhar 95.8333 and the lowest accuracy station is Kamrup 86.1111. (Ashutosh Sharma et al, 2016).

Neural Network:
Artificial Neural Network(ANN) Models were evolved with the help of historical yield data obtained at several locations all throughout Fujian for Rice yield prediction. Field-aggressive rainfall data and the weather variables (everyday sunshine hours, every day solar radiation, daily temperature sum and daily wind speed) were used for all the locations. Adjusting the ANN parameters such as learning rate and number of hidden nodes had an effect over the accuracy with respect to the predictions of rice yield. Optimal learning rates were noticed between 0.71 and 0.90. Smaller data sets require less number of hidden nodes and lower learning rates in the case of model optimization. ANN models provided a constant generation of yield predictions with more accuracy in comparison with regression models. ANN rice grain yield models devised for Fujian resulted in R2 and RMSE of 0.67 and 891 vs. 0.52 and 1977 respectively for linear regression. Although it consu

mes more time to be developed, when compared to a variety of other linear regression models, ANN models exhibited remarkable performance in predicting the rice yields accurately under the general Fujian climatic conditions. They require little memory and are generally fast. The complicated initial parameterization processes of ANN networks and overfitting problem necessitate more focus(B. J I et al, 2007).

This model took a total of about 15 districts of Bangladesh into consideration. With the purpose of grouping the districts into individual clusters, the assumption, which had to be used was that the districts having the similar values of necessary attributes must belong to the same cluster. Based on this assumption, the chosen attributes were categorized for the clustering of the districts as below; cluster Type-1 is dependent on the attributes of rainfall, minimum temperature, maximum temperature, humidity and sunshine. These include the environmental or climatic attributes that are taken into consideration. Cluster Type-2 relies on the attributes belonging to soil pH and soil salinity. Cluster Type-3 is dependent on the area irrigated. Clustering done on the basis of the area attributes for every district was taken since the individual clusters can be obtained depending on the different ranges of areas, which were irrigated for every district. Cluster Type-4 is on the basis of the different crop yields consisting of amon, aus, boro, potato and wheat. K-means clustering was used in the selected districts according to the categorized types mentioned previously. The classification/regression models given below were utilized forgetting the results of crop yield prediction: Linear Regression, k-Nearest Neighbor (k-NN), Neural Net (NN). Owing to the smaller training set, the prediction was not with the accuracy as it was anticipated and at times, anomalies were faced. The RMSE value for Amon is 24270.2(Linear),22578.36(K-NN), 24791.96(Neural),Aus is 4754.944(Linear), 13873.16(K-NN), 2788.586(Neural),Boro is 9653.76(Linear) 20122.28 (k-NN), 21128.44(Neural) , Potato is 9279.495(Linear), 23264.65(k-NN), 7553.811(Neural) and Wheat is 2776.443(Linear) 3414.207(K-NN) 2996.593(Neural). RMSE comparison clearly shows that different models provide better results for different crops (A.T.M ShakilAhamed et al,2015).

The models were developed using the earlier yield data extracted at various places all through Maryland. Field-specific rainfall data and the USDA-Natural Resources Conservation Service (NRCS) Soil Rating for Plant Growth (SRPG) values were used for all the locations. SRPG and weekly rainfall means were necessary for obtaining an efficient prediction on corn and soybean yield. Adjustment of the ANN parameters such as learning rate and the number of hidden nodes impacted the prediction accuracy of crop yield. Optimal learning rates dropped between 0.77 and 0.90. Compact data sets required less number of hidden nodes and reduced learning rates in the optimization of the model. ANN models consistently rendered yield predictions with more accuracy in comparison with regression models. ANN corn yield models for Maryland resulted in r2 and RMSEs of 0.77 and 1036 versus 0.42 and 1356 for linear regression, respectively. ANN soybean yield models for Maryland resulted in r2 and RMSEs of 0.81 and 214 versus 0.46 and 312 for linear regression, respectively. Although more amount time is consumed for their design when compared to the linear regression models, ANN models was revealed to be a much better technique for making accurate prediction of corn and soybean yields under the common Maryland climatic scenario (MonishaKaul et al,2004).

The artificial neural networks were developed and used for predicting the rice yield by making use of six meteorological factors; rainfall, evapotranspiration, temperature, humidity, water distribution and wind speed. The evapotranspiration (ET0) was obtained by making use of Pennman-Montieth equation. The monthly dataset of Phimai district from 2002 to 2007 was utilized for training the model and employed for predicting the rice yield from 2008 to 2012. The result proves that the Back Probagation (BP) neural computing technique could be applied with success in ANN modeling. The precision of prediction approximated the original data. The empirical rice yield predicting model ANN generated uniformly greater R2 (0.99) and lesser RMSE value (9.94) compared to linear regression based yield models. ANN can be utilized for predicting the result of fresh independent input data and has excellent capability in predictive modeling, i.e., all the characters that describe the unknown scenario can be provided to the ANN that is trained, and then prediction of agricultural system may be feasible. (SaisuneeJabjone et al,2013).

The improvement made to overcome the presence of local minima in addition to global minima, the error surface greatly curved along a weight dimension, and the direction of the negative gradient vector that might direct away from the minimum of the error surface are performed using heuristics techniques (Quick Propagation, Conjugate Gradient Descent(CGD) and numerical optimization techniques (Levenberg-Marquardt(LM)). The largest number of nodes present in hidden layer will produce the highest absolute error in back-propagation error. The Conjugate Gradient Descent (CGD) is based on heuristic approaches performing the linear searches of minimum error value; it is suitable for this kind of problem due to its quadratic convergence property. Three parameters are considered in prediction - they are pest, disease and weed. The CGD algorithm has only 2 hidden nodes. This is suitable for rice prediction due to its quadratic convergence property. It avoids local minima phenomena and exhibits slower convergence. The performance of CGD algorithm is outstanding compared to back-propagation algorithm(S. Putch et.al, 2004).

An ANN model was used for approximating a nonlinear function associating corn yield to soil, weather, and management factors. The network training utilized data from the Morrow Plots. The assumed input factors that have an influence over corn yield and for which data available from the Morrow Plots includes: Soil (pH, P, K, 0rganic matter), Weather(Growing season GDD, May rainfall, June rainfall, early July rainfall, late July rainfall, August rainfall, previous year rainfall) and Management(Genetic yield potential of the hybrid, N fertilizer applied, planting density, rotation factor). The ANN consists of 15 input nodes, 20 hidden-layer nodes, and one output node. The RMS errors typically decreased with rising numbers of hidden-layer elements, but the training time shoot up. The dynamic learning rate was huge for initial epochs and was made lesser for later subsequent epochs. It is not efficient at the prediction of extremes and lack in computing the yields depending on historical weather patterns. The RMS error obtained for 60 verification patterns was approximately 20% (J. Liu et.al, 2001).

Random Forest:

A data mining technique like Random Forests (RFs) can be exploited when creating a prediction model in case the search space associated with predictor variables is huge. Research activities exploring the accuracy corresponding to RFs to describe the annual changes observed in sugarcane productivity and the suitability of predictor variables generated from crop models combined with the climate and seasonal climate prediction indices identified is not more. Simulated biomass received from the Agricultural Production Systems Simulator (APSIM) sugarcane crop model, seasonal climate prediction indices and rainfall recorded, maximum and minimum temperature, and radiation were taken to be the inputs for a RFs classifier and a Random Forest regression model for describing the annual variation observed in regional sugarcane yields at the location of Tully in northeastern Australia. Prediction models were created on September the first in the year prior to the harvest, and then on January the first and March the first during the year of harvest that generally runs between June and November. The results indicate that in 86.36 % of years, there are chances to determine as early as September in the year prior to harvest whether the production would be greater than the median. This accuracy boosted to 95.45 % by the month of January during the year of harvest. The R-squared of the Random Forest regression model gradually rose from 66.76 to 79.21 % from September in the year before harvest through to March in the same year of harvest. The model was not capable of considering the amount of damage incurred to sugarcane owing to wet weather harvesting.(Yvette Everingham et.al,2016).

Random Forests (RF) were used for the estimation of mango fruit yields in accordance with water supply under diverse irrigation regimens. In order to deal with the variability pertaining to the mango fruit yields seen in the field, a group of RF models was designed in order to estimate the minimum, mean and maximum values for each one the mango fruit yields, which are “total yield” and “number of marketable mango fruits”. In the case of RF modelling, both 10-day rainfall and irrigation data combined was utilized in the form of the model input for evaluating the impacts of water sources over the mango fruit yields. The RF models provided an accurate estimation of the maximum and mean values of mango fruit yields, and exhibited medium accuracy for the minimum of mango fruit yields. The variable importance measure, which is calculated in the RF computation, proved that the timing of water supply has an effect over the mango fruit yields while rainfall and irrigation have diverse impacts on the mango fruit yields. This case study over the estimation carried out on mango fruit yields shows the suitability of RF in the field of agricultural engineering with a specialized attention on water management. The RF models yielded an estimation that was least accurate about minimum mango fruit yields. Addition of more environmental factors like temperature and radiation may enhance the performance of the model and exhibit more impacts on mango yield. The model performance and the information obtained from the RF models permits for exact modelling and the design of enhanced management practices in targeted agricultural systems (Shinji Fukuda et.al,2013).

Biomass is a significant indicator of growth in crops. In order to estimate biomass quickly and nondestructively, an enhanced technique, which mixes vegetation indices dependent on HJ-CCD and Random Forest (RF) regression technique is used. The accuracy of estimation and reliability of the RF model were validated for every stage (i.e., jointing, booting, and a thesis). Moreover, the comparison of the RF model results were then carried out with Support Vector Regression (SVR) and ANN models. The estimation accuracy of RF performed better than that of SVR and ANN at every stage. RF provided similar kind of reliability with SVR at diverse growth stages in both of the training and testing datasets, and exhibits better reliability compared ANN at every stage. At the same time, the RF model for every stage offers a slightly better generalization capability in comparison with the ANN model that behaves in a relatively unpredictable manner when utilized with individual input data, which differs from what was provided during the training phase. In comparison with the RF and SVR results for every stage, ANN exhibits performance that is much poorer in testing rather than in training. This is because of the fact that ANN is frequently used on very huge chunks of sampling data, but SVR and RF are applicable for small chunks of sampling data. The estimation accuracy and reliability of the RF model were verified for every stage (i.e., jointing, booting, and a thesis). In the case of RF models, the R2 values for the estimated-against the measured biomass regression for the three stages arrived to 0.533, 0.721 and 0.79, correspondingly, and the respective RMSE values were 477, 1126.2 and 1808.2 kg ha−1. The estimation accuracy of RF performed better than that of SVR and ANN at every stage (Li'aiWanga et.al,2016).

A combination created randomly of features is chosen at all the nodes to carry out the splitting. The bagging technique asserts to maximize the accuracy of the Random Forest algorithm by reducing the generalization error for the ensemble trees because of the usage of random features. This generalization error estimates are conducted with the Out of Bag (OOB) technique. This procedure of bootstrapping helps improve the model performance as it also reduces the variance of the model with no increase in the bias. It signifies that the predictions of a single tree are hugely reactive to noise inside the training set while the average of several trees is not correlated. The Random Forest algorithm is suited for wide range of datasets. This model provides a better accessibility to determine required soil N P-K content by utilizing one time soil testing data like available soil N-P-K content, soil type, crop type and yield target. The RMSE obtained during the prediction of soil N -P-K required is 6.118521(N), 5.195799 (P) and 4.710358(K)[Ambarish G. Mohapatra et al,2017].

RF does not make any assumptions as to explaining the phenomenon or impose a subdivision of the problem space. It clusters automatically and it feeds either the entire problem, or any number of discretely categorized or continuous explanatory variables. The RF Model provides an efficient balance in calculation of times and in accuracy of predictions. It has ability to explore an open range of covariants, as made available by the user [Diogo Vitorino,2013].

5. Inference and Conclusion

Precision Agriculture (PA) is a method to manage a farm with the help of mining algorithms to guarantee that the crops and soil obtain precise what they require for having an optimum health and productivity. The objective of PA is to guarantee profitability, sustainability and environmental safety. The study was done by analyzing the few important algorithms like Support Vector Machine, Bayesian Networks, Neural Networks and Random Forest. The Performance of the predicting algorithm is purely dependent on the quality and quantity of data which is used for analysis. A smaller data set is not capable of revealing every underlying correlation existing among the variables. Smaller data sets must be aided by outside information regarding the interrelations among variables in addition to the distributional characteristics of every variable. Bayesian networks can yield a good prediction accuracy even in the case of considerably smaller sample sizes. As BNs are solved in an analytical manner, they can render quicker responses to the queries during the compilation of the model. The compiled form of a BN comprises of a conditional probability distribution for all the combinations of variable values, and can therefore render any distribution instantaneously in contrary to the simulation models, where the results have to be simulated that can consume lot of time. The rate of learning, number of hidden nodes, and the training tolerance had an effect on the design of the ANN model and the accuracy pertaining to predictions of ANN crop yield. As the amount of data, which is being modelled decreased because of lesser spatial levels, limited number of hidden nodes were necessary. With the decrease in the number of hidden nodes, the optimum learning rate reduced. An artificial neural network (ANN) was used for modelling the correlation between yield and the factors that influence yield. The impacting factors consist of soil factors, weather factors, and management factors. Random factors were eliminated. The ANN parameters that were tested comprised oftype of network, network topology, learning rate, initial weights, kind of transfer function, and number of training epochs. A part of the data set, chosen by stratified sampling, was excluded from training and utilized for verifying the accuracy of the yield predictions. Random Forest, where the variable significance is estimated, the yield can also be assessed. The RF model for every stage offers a slightly better generalization capacity compared to the ANN model that behaves in comparatively unpredictable manner when employed with individualistic input data, which diverge from what was provided during the training stage. In comparison with the RF and SVR results for every stage, ANN exhibits a performance much poorer in testing rather than in training. This is because of the fact that ANN is frequently used on huge chunks of sampling data, but SVR and RF are applicable for smaller chunks of sampling data. One cause for this is probably that the learning capability is of too much potential. RF is not sensitive to collinearity. Based on the size of the dataset, sensitiveness of predictor variable, we need to decide the algorithm to build the model to predict the crop yield. It has many issues such as learning performance, computation time and scalability. These issues have been also solved in the future work in crop yield prediction system. The scalability and computation time complexity issue is solved via the use of parallel processing methods. The prediction rate issue is solved via the use of classification methods.

References

1. RaheelaShahzadi, Muhammad Tausif, JavedFerzund, Muhammad Asif Suryani, Internet of Things based Expert System for Smart Agriculture, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9, 2016.

2. Neil Gershenfeld, RaffiKrikorian, and Danny Cohen. The Internet of Things. Scientific American, 291(4):76–81, 2004.

3. Kerry Taylor et al. Farming the web of things. Intelligent Systems, 28(6):12–19, 2013.

4. Dennis Pfisterer et al. SPITFIRE: toward a semantic web of things. IEEE Communications Magazine, 49(11):40–48, 2011.

5. Freddy Lecu ´ e et al. Smart traffic analytics in the semantic web with ´ STAR-CITY: Scenarios, system and lessons learned in Dublin City. Web Semantics: Science, Services and Agents on the World Wide Web, 27:26–33, 2014.

6. John Havlin et al. Soil fertility and fertilizers: An introduction to nutrient management, volume 515. Pearson Prentice Hall, 2005.

7. Muhammad Intizar Ali, FengGao, and Alessandra Mileo. CityBench: A Configurable Benchmark to Evaluate RSP Engines Using Smart City Datasets. In Proc. of ISWC, pages 374–389. Springer, 2015.

8. ShailajaPatil, Anjali R. Kokate, Dhiraj D. Kadam:Precision Agriculture: A Survey , Volume 5, International Journal of Science and Research,August 2016

9. KhushbooBabariaetal: Survey on Predictive analysis for formulating real time data in precision agriculture. Volume No.: III, Special Issue on IEEE Sponsored International Conference on Innovations in information,2015

10. 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

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. Niketa Gandhi , Leisa J. Armstrong , OwaizPetkar , Amiya Kumar Tripathy “Rice Crop Yield Prediction in India using Support Vector Machines”,2016.

13. Nathaniel K. Newlands , Lawrence Townley-Smith “Predicting Energy Crop Yield Using Bayesian Networks”,2010

14. Charles L. Hornbaker II , J. Benjamin Cook “Predicting Yield in the Corn Belt”, 2013

15. NiketaGandhi ,Leisa J. Armstrong , OwaizPetkar “PredictingRice Crop Yield Using Bayesian Networks”,2016.

16. KefayaQaddoum, “Modified Naïve Bayes Based Prediction Modeling for Crop Yield Prediction”, Modified Naïve Bayes Based Prediction Modeling for Crop Yield Prediction, 2014.

17. KasirgaYildirak, ZeynepKalaylıoglu, · Ali Mermer “Baye

sian estimation of crop yield function: drought based wheat prediction model for tigem farms" , Springer Science+Business Media New York, 2015.

18. L. Uusitalo, "Advantages and challenges of Bayesian networks in environmental modelling", Ecological modelling, 2007.

19. A. Sharma and M. Goyal, "Bayesian network model for monthly rainfall forecast", 2015 IEEE International Conference on Research in computational Intelligence and Communication Networks 2015.

20. B. J I, Y. SUN, S.YANG AND J. WAN, "Artificial neural networks for rice yield prediction in mountainous regions", Journal of Agricultural Science 2007.

21. A.T.M ShakilAhamedNavid, TanzeemMahmood, Nazmul Hossain, Mohammad TanzirKabir, Kallal Das, Faridur Rahman, Rashedur M Rahman, "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh" IEEE,2015.

22. MonishaKaul, Robert L. Hill, Charles Walthall, "Artificial neural networks for corn and soybean yield prediction", Elsevier, 2004.

23. SaisuneeJabjone, Chatchai-Jiamrum, "Artificial Neural Networks for Predicting the Rice Yield in Phimai District of Thailand ", International Journal of Electrical Energy, 2013

24. .S. Putch, M. Rizon, M. Juhari, J. Nor Khairah, S. SitiKamarudin, B. Aryati, R. Nursalasawati, "Backpropagation algorithm for rice yield prediction", Proceedings of 9th of the Ninth International Symposium on Artificial Life and Robotics, Beppu, Japan, Oita, 2004.

25. J. Liu, C. Goering and L. Tian, "A neural network for setting target corn yields", Transaction of the ASAE, 2001.

26. Yvette Everingham,JustinSexton,DanielleSkocaj, Geoff Inman-Bamber, "Accurate prediction of sugarcane yield using a Random Forest algorithm" , INRA and Springer-Verlag France 2016.

27. Shinji Fukuda, Wolfram Spreer , Eriko Yasunaga, Kozue Yuge , VichaSardsud, Joachim Müller, "Random Forests modelling for the estimation of mango (Mangiferaindica L. cv. Chok Anan) fruit yields under different irrigation regimes", Agricultural Water Management 116, 2013.

28. Li'aiWanga , XudongZhoub , XinkaiZhua , Zhaodi Donga , WenshanGuoa, "Estimation of biomass in wheat using Random Forest regression algorithm and remote sensing data", The Crop Journal,2016.

29. Ambarish G. Mohapatr, Dr. Bright Keswani , "Soil N-P-K Prediction using Location And Crop Specific Random Forest Classification Technique in Precision Agriculture ", International Journal of Advanced Research in Computer Science.

30. D. Vitorino, S.T.Coelho, P.Santos, S.Sheets, B.Jurkovac, C.Amado, "A Random Forest algorithm applied to condition based wasterwater deterioration modeling and forecasting", 16th conference on water distribution system analysis (WDSA), Procedia Engineering, 2014.

31. Mo, X., Liu, S., Lin, Z., Xu, Y.,

Xiang, Y. and McVicar, T.R., 2005. Prediction of crop yield, water consumption and water use efficiency with a SVAT-crop growth model using remotely sensed data on the North China Plain. Ecological Modelling, 183(2-3), pp.301-322.

32. Folberth, C., Gaiser, T., Abbaspour, K.C., Schulin, R. and Yang, H., 2012. Regionalization of a large-scale crop growth model for sub-Saharan Africa: Model setup, evaluation, and estimation of maize yields. Agriculture, ecosystems & environment, 151, pp.21-33.

33. Snoek, J., Larochelle, H. and Adams, R.P., 2012. Practical bayesian optimization of machine learning algorithms. In Advances in neural information processing systems (pp. 2951-2959).

34. Nasrabadi, N.M., 2007. Pattern recognition and machine learning. Journal of electronic imaging, 16(4), p.049901.

35. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825-2830.