Enhancing Seed Selection and Providing Guidance for Cultivation using Random Forest Technique

Ayushi Gupta, Nikhil Narayan, Kanmani Sivagar

Abstract*: Seed Selection is a very challenging job because for a selection of a seed multifarious parameters are to be taken under consideration. Also seed analysis require a prediction of which seed is suitable which needs a great accuracy as there are numerous things to be taken into account like soil type, ph of soil, nutrient content of soil, elevation of land, weather of the area, etc. Several algorithms have been devised from time to time but each of the methods differs in their own way. The algorithms, which are discussed, are K-Means Algorithm, K-Nearest Neighbor Algorithm, Naïve Bayes Classifier, Decision Tree, Regression Model, etc. Data mining techniques can overcome this challenging job.

Keywords: **classification, decision, estimation, features, knowledge, management, monitoring, parameters**

I. INTRODUCTION

Agriculture is a portentous part in the Indian economy. Agriculture is an unscientific process but in order to get the best out of the soil, one needs a good scientific approach to it. Indian government has loads of data related to agriculture but in raw form, which is not understandable or fruitful to farmers. In the era of digitalization data mining techniques can get the desired result.

Seed analysis is a very much a generalized term and here it can take various form, like, crop yield estimation, crop monitoring, crop support system, crop pest and disease analysis, seed selection analysis and crop classification. As mentioned there is an enormous scope in analysis of crop. Each devised for analyzing crops in a different perspective and thus perceiving things in a broad way. Our project deals with selection of seed.

The input to our project will be weather condition parameters like temperature, ph, air humidity and rainfall. Several methods are used to get the desired output. Each method has its own pros and cons. In our project random forest technique will be used. Random forest can predict crop appropriate for a given area with accuracy above 90%. Not only selection of crop but also a guide of how to cultivate the particular crop is available. Even information regarding management of crop which includes a good knowledge of the correct amount, quality and time during which pesticides need to be added.

Revised Manuscript Received on April 15, 2020.

***** Correspondence Author

Ayushi Gupta*, CSE, SRM Institute of Science and Technology, Chennai, India. Email: ayushi.gupta097@gmail.com

Nikhil Narayan, CSE, SRM Institute of Science and Technology, Chennai, India. Email: nikhilnarayan527@gmail.com

Kanmani Sivagar, CSE, SRM Institute of Science and Technology, Chennai, India. Email: kanmanis@srmist.edu.in

Following difficulties are faced when one does a crop analysis: The parameters are enormous and finding correlation and then drafting an algorithm which covers every things calls for many challenges and to collect data set itself is a very difficult job.

To solve these difficulties, techniques are there such as: Moving average and cluster correlation techniques, K-Means Algorithm, K-Nearest Neighbor Algorithm, Naïve Bayes Classifier, Decision Tree, Regression Model, etc. There are so many techniques as we can see above. Accuracy of each differs in the way they are applied. Data Mining techniques can overcome greatest of challenging analysis.

There is a comparison table included beneath the literature survey, which gives a brief view about the algorithms presented in this survey paper.

II. LITERATURE SURVEY

A study which does crop produce estimation for crops in a given area by using models is presented in [1]. Firstly NDVI data is analyzed and its correlation is conducted with the crop yield. The paper establishes a regression model for prediction of crop yields. The aim is to tell whether the yield is sufficient to fulfil the needs of the future. If crop yield is low then, that region will be at risk for food. Three major crops are taken into account like paddy rice, winter wheat and summer corn. Correlation coefficient analysis is done between crop yields and NDVI data. Also crop identification and distribution of crops in a given area is calculated apart from the crop yield estimation. The assumption that is taken into account is that if there is a strong correlation between the data taken then it indicates that there is a good crop yield. It works well for small regions with mixed type of cultivation and also regions where it is not known regarding the crop type. Further the regression model can be improvised by taking into account the physical factor like geographical information and type of soil. Also it can further includes the past crop yield in that particular region.

The influence of illumination and viewing angular effects on crop types both for wide swath and narrow swath and surface reflectance of typical surface is done in [2]. Three types of angular effects are taken into consideration namely mean local time drift effect, view angle effect and day of year effect.

These effects were analyzed based on satellite BRDF (bi-directional reflectance distribution function) and field measurements. A look up map is being made for the crops using the MODIS (Moderate Resolution Imaging Spectroradiometer) and CDL (cropland data layer) to correct these angular effects. The result provides a data source for time series analysis and crop condition monitoring. It correctly views the angle effects but depends on the sensor swath width, acquisition time, image and applications. It should monitor the growth by taking other parameters also like nutrient content of the soil.

Using the Crop Environmental Resource Synthesis (CERES) – Wheat Model, a data assimilation strategy was developed using a particle filter for better performance of crop models. The model in [3] is an excellent agro-ecological dynamic model because it takes into account the effects of management, soil, carbon, genetics, weather, water and nitrogen. Two experiments were performed. The first one focused on checking the feasibility of the data assimilation strategy. The second experiment was conducted on regional scale, which remotely sensed LAI with fine spatial resolution. In this case it analyzed the regional winter wheat yields. It can provide real time information on regional crop growing states. The uncertain elements need to be analyzed for further research like weather, multi sourced remotely sensed observations, soil, and regional field management information.

Crop type classification is a challenging job when it comes to identifying images sets with low spatial resolution as shown in [4]. To solve such a thing, a method is developed which combine high-resolution images with low spatial resolution in high time frequency to get a superb classification of the crop types. If one has very few high-resolution images for location then also up to 20% improvement in classification can be one using this approach. It covers whole land data using high and low resolution image. It uses data from multiple sensors simultaneously. It can help in taking correct decision by making crop classification map.

Crop growth monitoring system is presented using the existing polarimetric decomposition in [5]. The proposed method uses uninhabited aerial vehicle synthetic aperture radar (UAVSAR). It uses scattering mechanisms and vegetation orientation. This is taken into consideration because it directly affects the crop growth status. Performance of the system is evaluated using the SMAPVEX12 data. For wheat, corn and canola, significant scattering was present in early growth stage whereas it was dominant in later stage. For pasture and soybean, dominant scattering was present in the ground matrix. Correlation is performed between the dataset and crop height and biomass. It is the first paper, which elaborates and demonstrates the potential of polarimetric decomposition.

Specifically based analysis on cotton crops is performed as inferred by the title of [6]. Using SAR data, investigation of vegetation canopies is done using backscattering techniques. Model is built around radiative transfer theory. It is used for growth profile studies, yield estimation, crop monitoring, range lands management, crop discrimination, soil moisture estimation, etc. LAI (Leaf Area Index) model, water cloud model and attenuation model are being used for this purpose. To analyze the crop growing stage, a linear zone is chosen from the relationship between crop height and LAI data semi empirically. Since the detailed ground data isn't present therefore it proves to be a con.

Knowledge graph is a method, which is chosen for analyzing crop pests and diseases executed in [7]. Knowledge representation uses RDF and ontology and expressing learning. Knowledge extraction is done based on entity recognition, relation extraction, attribute extraction, common resource, knowledge fusion and reasoning. Intelligent semantic search is performed. Knowledge graphs weren't perfect and require more in depth analysis. The future scope can be to construct large-scale domain knowledge map to achieve more accurate result. Automation and expansion of these maps for large-scale networks can prove to be an improvised and a better version of what system they have proposed.

Environmental factors are directly monitored by intelligent agriculture IOT equipment. The data collection is performed which then undergoes an analysis using 3D cluster in [8]. It has features like data normalization, crop clustering in an appropriate group and advice on crop, which is suitable for a particular region. The result, which came, says that it is indeed flexible. The algorithm uses cluster correlation and moving average. It can't help in automatic cultivation to farmers. If it incorporates artificial intelligence then this system can prove to be a better and improvised version of it.

Pest management, which in itself is a difficult job, moreover not much light has been thrown on this subject till now. In [9] it is used for real time detection and then identifies the invertebrate prevailing in crops. But spectral information limits the capability as camouflage pests make it difficult to detect. Now using local variance and detection logic one can overcome such an obstacle. Local variance of normal algorithm can differentiate broad leaves from large pests in noisy environment also. A multispectral 3D vision system is developed using the multispectral images of blue, red, ultraviolet, near infrared and green, which can create denser point cloud of pest and plants. It can be integrated with an automatic seedling stage, which is the future scope of this paper.

Subset selection technique is used in crop analysis in [10]. The main emphasis of this paper is on crop classification. PolSAR (full polarimetric synthetic aperture radar) data is taken into picture. A RF (random forest) based technique is used for Multitemporal PolSAR classification. The feature selection technique includes three steps namely, using RF based partial probability plot, and effects of feature on a class was identified. A feature subset was chosen which provide separation of ODR on same data for different crops. The last step was to remove highly correlated features. It has very high scope to be featured in the snow and land phenology. The classification accuracy is very high marking it up to 99% and thus is a very novel strategy.

III. COMPARISON TABLE

The comparison table below compares 10 different paper based on their main emphasis, algorithm, advantages and disadvantages. Algorithms used in the paper come with their own flaws. Many aspects in crop analysis have been covered such as crop monitoring, growth status, etc. Quite many data mining techniques have been used but further improvising can be made in this field by adopting random forest technique.

IV. IMPLEMENTATION

Random forest is the approach being used here because it is an ensemble approach. Random forest can be used both for classification and regression problems. Random forest operates by constructing multiple decision trees and based on the output class by each, it calculates mode or mean of the classes in case of classification. It has two key concepts which makes it random. Firstly, random subsets of features considered when splitting nodes and random sampling of training data points when building trees. Random forest's accuracy score can be further improved by tuning the hyper parameters. Tuning the hyper parameters leads to getting optimal values. Random forest helps in overcoming the overfitting problem of decision trees. Therefore, random forest gives an accuracy score of about 92% which proves to be much better than the other algorithm's accuracy score. Random forest is best suitable for categorical values.

V. FUTURE SCOPE

The future scope includes taking into account every parameter that relates to selection of crop. The data set can be taken by real time implementation using IOT. Also each project, which is developed on the basis of the algorithms mention in the survey paper, should be further improvised to run without an Internet connection. There can be a voice recognition system, which uses intensive machine learning and artificial intelligence technique to detect and understand what the farmer's needs.

VI. CONCLUSION

The above literature survey which discusses about all the main data mining techniques. Well, it is evident that data mining techniques can help in crop analysis to a great extent. A random forest prediction data mining technique also exist which can help in prediction of crop suitable for a given region. Random forest can help in classification methods.

VII. RESULTS AND DISCUSSIONS

The result obtained from implementing Random Forest Technique was indeed better than the already existing algorithms. Moreover, an ensemble learning algorithm proves to be better in this scenario. Since here we have been implementing a classification algorithm that is, we need to classify which seed is suitable so in such a scenario random forest proves to be better. Also random forest avoids the risk of overfitting a problem and hence generalizes much better for a new testing data or example. It runs better in large databases and that's y produces a high accuracy score.

Table- I: Comparison Table

YEAR	REFERENCE NO.	MAIN EMPHASIS	ALGORITHM	ADVANTAGES	DISADVANTAGES
2014	[1]	Crop Identification and Yield Estimation	Regression Model	It works well in small area where the type of crop is not known precisely.	Physical factors are not incorporated in the regression model.
2014	[2]	Crop Monitoring	BRDF Model	It correctly views angle effects.	Depends on the sensor swath width, the image acquisition time, and applications.
2014	[3]	Estimating winter wheat yields	Particle filter based data assimilation strategy	In real time, remote sensing can provide actual information on regional crop growing states.	It doesn't use strategies of simultaneous state-parameter estimation
2014	[4]	Crop Type Classification	Projected Gradient Descent	It can create crop type classification maps.	Computational performance and need for ideal crop curves.
2016	[5]	Crop Growth Status	Matrix Decomposition	Crop growth monitoring via the vegetation orientation and scattering mechanisms.	Their pattern are dependent on phenological development stage
2004	[6]	Analysis of cotton crops.	Parameter Estimation, Iterative Methods.	Model depends on a very important parameter, LAI.	Model depends on a very important parameter, LAI.

2019	[7]	Analysis of Crop Pest and Diseases.	Knowledge Graph	It provides a new way for knowledge management and is more flexible method.	Knowledge maps were not perfect and in-depth
2019	[8]	Analysis of Crop Selection.	Moving Average and Cluster Correlation	It cleans out data with more drastic variation.	Analysis result can't help farmer reach automatic cultivation.
2017	[9]	Invertebrate Detection on Crops.	LVN (Local Variance of Normal) Algorithm	It can detect even the camouflage pests.	Accuracy is dependent on the complexity of scenarios.
2018	[10]	Crop Classification.	Subset selection technique.	It has potential application in the field of snow phenology.	It is used to identify the specific range of parameter where probability is high.

REFERENCES

1. G. Jing Huang, Huimin Wang, Qiang Dai, Dawei Han, "Analysis of NDVI Data for Crop Identification and Yield Estimation", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (Volume: 7, Issue: 11, Nov. 2014).
2. Feng Gao, Tao He, Jeffrey G. Masek, Yanmin Shuai, Crystal B. Schaaf, Zhousen Wang, "Angular Effects and correction for medium resolution sensors to support crop monitoring", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (Volume: 7, Issue: 11, Nov. 2014).
3. Zhiwei Jlang, Zhongxin Chen, Jin Chen, Jia Liu, Jlanquiang Ren, Zongnan Li, Liang Sun, He Li, "Application of Crop Model Data Assimilation with a Particle Filter for Estimating Regional Winter Wheat Yields", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (Volume: 7, Issue: 11, Nov. 2014).
4. Mark W.Liu, Mutlu Ozdogan, Xiaojin Zhu, "Crop type Classification by Simultaneous use of Satellite Images of Different Resolutions", IEEE Transactions on Geoscience and Remote Sensing (Volume: 52, Issue: 6, June 2014).
5. Hongquan Wang, Ramata Magagi, Kalifa Goita, "Polarimetric Decomposition for Monitoring Crop Growth Status." IEEE Geoscience and Remote Sensing Letters (Volume: 13, Issue: 6, June 2016).
6. S. Maity, C. Patnaik, M. Chakraborty, S. Panigrahy, " Analysis of Temporal Backscattering of Cotton Crops using a Semiempirical Model", IEEE Transactions on Geoscience and Remote Sensing (Volume: 42, Issue: 3, March 2004).
7. Liu Xiaoxue, Bai Xuesong, Wang Longhe, Ren Bingyuan, Lu Shuhan, Li Lin, "Review and Trend Analysis of Knowledge Graphs for Crop Pest and Diseases", IEEE Access (Volume: 7).
8. Fan-Hsun Tseng, Hsin-Hung Cho, Hsin-Te Wu, "Applying Big Data for Intelligent Agriculture-Based Crop Selection Analysis", IEEE Access (Volume: 7).
9. Huajian Liu, Sang-Heon Lee and Javaan Singh Chahl, "A multispectral 3D Vision System for Invertebrate Detection on Crops", IEEE Sensors Journal (Volume: 17, Issue: 22, Nov.15 2017).
10. Siddharth Hariharan, Dipankar Mandal, Siddhesh Tirodkar and Vineet Kumar, "A Novel Phenology Based Feature Subset Selection Technique Using Random Forest for Multitemporal PolSAR Crop Classification", IEEE Journal Selected Topics in Applied Earth Observations and Remote Sensing (Volume: 11, Issue: 11, Nov. 2018).
11. N. Hemageetha, "A survey on application of data mining techniques to analyze the soil for agriculture purpose", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom).
12. Gao Yi-yang, Ren Nan-ping, "Data Mining and analysis of our agriculture based on the decision tree", 2009 ISECS International Colloquium on Computing, Communication, Control, and Management.
13. Yogesh Gandge, Sandhya, "A study on various data mining techniques for crop yield production", 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT).
14. Rakesh Shirsath, Neha Khadke, Divya More, Pooja Patil, Harshali Patil, "Agriculture decision support system using data mining", 2017 International Conference on Intelligent Computing and Control (I2C2).
15. Alexandre Bouvet, Thuy Le Toan, Nguyen Lam-Dao, "Monitoring of the Rice Cropping System in the Mekong Delta using ENVISAT/ASAR Dual Polarization Data", IEEE Transactions on Geoscience and Remote Sensing (Volume: 47, Issue: 2, Feb. 2009)

AUTHORS PROFILE

Ayushi Gupta is a Computer Science undergraduate of B.Tech (Bachelor of Technology) from SRM Institute of Science and Technology, Chennai. Ayushi is a final year student. Ayushi has presented a conference survey paper in ICIOT (International Conference on Internet of Things) 2020 before. Ayushi's area of work includes Data Science, Data Mining and Machine Learning. Ayushi is a member of IET (Institution of Engineering and Technology) since 2016.

Nikhil Narayan is a Computer Science undergraduate of B.Tech (Bachelor of Technology) from SRM Institute of Science and Technology, Chennai. Nikhil is a final year student. Nikhil has presented a conference survey paper in ICIOT (International Conference on Internet of Things) 2020 before. Nikhil's area of work includes Machine Learning. Nikhil is a member of IET (Institution of Engineering and Technology) since 2016.

Kanmani Sivagar, is an Assistant Professor in department of CSE from SRM Institute of Science and Technology, Chennai. Kanmani has completed her B.Tech in Information Technology from SSN College of Engineering, Anna University, 2010 and M.E in Computer Science and Engineering from the same. Kanmani has selected publications in International Journal of Pure and Applied Mathematics, 2018, International Journal of Innovative Science, Engineering & Technology, 2017, International Journal of Computer Science Trends and Technology, 2017, International Journal of recent trends in engineering and research, 2017, International Journal of Advanced Research in Basic Engineering Sciences and Technology, 2016, International Journal of Computer Trends and Technology (IJCTT), 2014. Kanmani received "Best Organizer Award" for organizing a technical seminar – Perl scripts and Programming in Infosys Technologies Limited, Gold Medalist in Post Graduate Degree, 2010, Won scholarship for complete 2 years during 2008 – 2010 to pursue M.E – CSE in SSN College of Engineering. Kanmani is a member of IET 2012 and ISCA, 2017.