ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

A survey on data mining techniques in agriculture

D. Sabareeswaran[*1], A. Edwin Robert[2]

[*1]*Research Scholar, Dept. of Computer Science, Karpagam University, Coimbatore, India.*
[2]*Assistant Professor, Dept. of Computer Application, Karpagam University, Coimbatore, India.*
[*1]sabareeswarenphd@gmail.com, [2]edrtedwin@gmail.com

Abstract

Objective: To study about different data mining methods utilized for detecting plant diseases, soil moisture and crop growth monitoring.
Methods: Different data mining techniques are used in agriculture for detecting crop diseases, soil moisture and crop growth monitoring such as Support Vector Machine (SVM), Artificial Neural Network (ANN) and Regression model.
Findings: The inclusion of modern technologies can enhance the crop production and resolve major issues in traditional farming. The crop production is mainly depends on the availability of arable land and influenced by yields, macro-economic uncertainty and consumption patterns. The actual yield is mostly depends on crop's genetic potential, amount of sunlight, water and nutrients absorbed by crop, presence of weeds and pests. In addition, the crop production is enhanced by combining crop models with data mining approaches.
Applications/Improvements: Finally, different data mining techniques used in agriculture are compared in order to prove their effectiveness. Hence, the agricultural monitoring system can be enhanced by using data mining techniques.

***Keywords*:** Crop production, Data mining techniques, Crop diseases, Soil Moisture, Crop growth monitoring.

1. Introduction

Agriculture [1], the main backbone of India, is the development of plants for food, bio-fuel, medicinal plants and other products used for sustaining and enhancing individual life. The history of agriculture engages thousands of years back, and its growth has been motivated and defined by several atmospheres, cultures and technologies. Industrial agriculture founded on large-scale monoculture farming has become the most dominant agricultural methodology.

In past century, agriculture has been categorized by improved productivity, the substitution of synthetic fertilizers and pesticides for water pollution and farm subsidies. In modern years, the external environmental effects of traditional agriculture have been removed and resulting in the organic and sustainable agriculture movements. Nowadays, plant diseases and land degradation are the most important problems in agriculture. Therefore, the improvement of agricultural is monitored by data mining techniques achieved through improved information and communication processes.

Data mining in agriculture applications involves the conceptualization, design, development, estimation and application of modern ways for utilizing the information and communication technologies (ICT) in rural domain including with the major objective on agriculture productivity. Different modelling processes [2]and simulation methods have been implemented for dynamic systems in agriculture. The major challenge in agriculture is that no specific measures have been taken out with the large sets of agricultural data. Modern research executes data mining in agriculture.

Data mining is the method of discovering previously unknown and potentially interesting patterns in large datasets. The mined information is typically represented as a model of semantic structure of database, wherein the model may be used on new data for prediction or classification of agricultural data. The foremost issues in agriculture and modern techniques are associated to the overall crop production. The main contribution of this paper is to analyse the different data mining techniques in agriculture for monitoring plant diseases, soil moisture and etc.

The rest of the paper is organized as follows: Section 2 describes the different data mining techniques used for detecting plant diseases. Section 3 describes the different methods used for predicting soil moisture in agricultural fields. Section 4 describes the crop growth monitoring systems in agriculture. Section 5presents the comparison of the techniques in literature. Section 6 concludes the research.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

2. Study of plant/Crop disease detection methods

The semi-automatic segmentation algorithm of plant leaf disease symptoms [3] was developed using digital image processing. The algorithm was based on the grayscale histograms and provided for distinguishing signs and symptoms of plant disease from asymptomatic tissues in leaves. The histograms of H from HSV color space and a* color channels from L*a*b color space were manipulated by this algorithm. An in-depth analysis of the issues such as lesion delimitation, illumination, leaf venation interference and leaf roughness of disease symptoms segregation was discussed. The problem with this algorithm was that some sources of error could not be removed hence the accuracy of the algorithm was reduced.

A machine learning regression techniques [4] were investigated for detecting leaf rust disease by using hyper-spectral measurement. The partial least square regression (PLSR), v-support vector regression (v-SVR) and Gaussian process regression (GPR) techniques were presented for detecting wheat leaf rust disease. The effects of disease symptoms on prediction performances of these techniques were manipulated and compared with spectral vegetation index (SVI).Here, the spectra of infected and non-infected leaves were measured according to the different disease symptoms by using non-imaging spectroradiometer. However, training samples were expensive and the accuracy was reduced due to the scattering of the data.

The computer vision based method[5] was investigated for automatic detection of crop diseases. The features were extracted by combining marker controlled watershed segmentation and super-pixel based segmentation. The extracted features were selected based on textural, Gabor, gradient and biologically inspired features. Then the features were classified based on support vector machines (SVM) and also compared with ANN based classification. This approach was used for reducing the complexity of image processing functions and improving the segmentation quality. However, the robustness of this approach was low and it collects less data from different symptoms.

A robot [6] was developed for crop disease detection using image processing. The ground based agricultural robot called eAGROBOT was presented for monitoring cotton and groundnut fields. The pictures of the plant were captured by the eAGROBOT. The captured image was pre-processed and transformed by artificial intelligence (AI) based embedded algorithms. Then the transformed images were clustered for detecting the cluster image of interest. Then the selective features were classified based on the neural network algorithm to identify the types of symptoms and differentiate early and late leaf diseases. However, the cost to end user and complexity of maintenance were high.

The support vector machine [7] was proposed for plant disease detection. The original image was captured and pre-processed. Then the pre-processed image was segmented as black and background pixels of image and also hue and saturation segments were separated. The features were extracted from segmented images and classified by enhanced support vector machine (SVM) to detect the diseases. The features were matched to images by using SIFT. However, the accuracy of this technique was low.

3. Study of soil moisture prediction methods

A prediction algorithm [8] was proposed for soil moisture based on improved BP. The prediction algorithm based on BP neural network and particle swarm optimization (PSO) was introduced for predicting the time series of soil moisture information acquired from wireless sensor networks. The time series parameters of BP were determined and the weight and threshold of BP were improved by using particle swarm optimization (PSO) algorithm. The soil moisture time series was predicted by BP method. However, the mean square error for BP method is high.

The multi-spectral and FTIR techniques [9] were investigated for estimation of soil moisture. The information of soil samples were collected by using randomized sampling technique. Each sample was placed in sealed polyethylene bags and transferred for soil moisture and soil texture analysis. The soil texture was analyzed by using hydrometer method. The soil moisture was measured by using gravimetric method. The effect of the reflectance on different moisture conditions was analyzed by FTIR spectroscopy method. The triangulation method was also applied for appropriate estimation of soil moisture. However, computing power was high and FTIR have single beam whereas other dispersive methods have double beam.

The soil moisture estimation [10] was investigated by using ensemble kalman filter. The method was developed with assimilating temperature data from distributed temperature sensing (DTS) for estimating soil moisture at high resolution. The correlation between temperature and soil moisture was ensured that soil moisture can be estimated

by soil temperature data. The soil moisture and temperature data were merged by using ensemble kalman filter for estimating soil moisture. However, the performance of this method was still poorer than synthetic approaches.

The dense temporal series of C and L-band SAR data [11] was developed for retrieving soil moisture. The interaction between SAR signal and crops were characterized by scattering mechanisms. The relationship between backscatter and soil moisture content temporal modifications as a function of various SAR bands and polarizations were described. The change in soil moisture content during plants growth period was monitored by using change detection technique which is applied to the multiple temporal C and L-band SAR data. However, the incidence angle was critical issue since it should not exceed moderate values.

Bayesian change detection method [12] was investigated for retrieving the soil moisture under various roughness conditions. The changes in backscattering signals were identified and analyzed them with soil moisture variations through considering the changes in radar signals owing to roughness changeability. Then the change detection method based on Bayesian technique was developed over a long time period for retrieving the soil moisture content under agricultural fields through L and P-band SAR images. However, the backscattering changes analysis was high complexity.

4. Study of crop growth monitoring systems

The clustering method [13] was proposed for plant production system monitoring. The potential of multiple endmember spectral mixture analysis (MESMA) was developed for simultaneously extracting the sub-pixel cover fraction and uncontaminated spectral signature of the crop element from the mixed hyper-spectral signal. The lookup tables (LUT) were constructed by using radiative transfer models for both crop and soil factor. The clustering approach was introduced after segmentation process for improving effectiveness of the utilization of LUT in MESMA model. The most favourable clusters were selected based on the Bayesian selection method. However, the non-linear mixture effects presented in the orchard systems and influence of shadow in mixture were not considered.

A dynamic unmixing model [14] was developed for plant production monitoring system. The observed spectra such as vegetation and soil were considered to be linear combinations of spectra from available spectral libraries by using linear mixing model (LMM). The unmixed problems were tackled by multiple endmember spectral mixture analysis (MEMSA) and sparse unmixing via variable splitting and augmented Lagrangian (SUnSAL). For effective library reduction, modification of Hyper-spectral unmixing via multiple signal classification and collaborative sparse regression (MUSIC-CSR) was developed. This technique was used for pruning the dictionary and understanding high-quality of the vegetation spectra on the ground by means of pruned dictionary as input to available unmixing techniques. However, this method was difficult for acquiring reliable estimation of ground fractional abundances.

The parameter based model [15] was investigated for crop yield prediction. Here, the crop yield was determined by attributes. The yield of wheat was predicted by using fuzzy logic (FL), adaptive Neuro fuzzy inference system (ANFIS)and multiple linear regression (MLR) techniques. The prediction was achieved by considering biomass, extractable soil water, radiation and rain as different input parameters. The database was pre-processed by means of eliminating outliers, redundant, inconsistent and missing values. The yield of wheat was more accurately predicted by ANFIS method however the mean square error value of the method was slightly high.

The ant colony algorithm with centre data aggregation [16] was developed for agricultural monitoring system. The plant growth was monitored by using environmental parameters collected by using Zigbee based weather stations. The entire sensors were incorporated into weather stations and simply single monitoring node was employed for data aggregation. An energy efficient centre data aggregation method was presented in which ant colony optimization algorithm was applied to the production of level gradient field. The remote web-based human machine interface was also developed for monitoring the plant production. However, the average delay of this technique was high.

The rice growth monitoring [17] was investigated by using X-band Co-polar SAR. The plant growth monitoring was achieved by implementing BiologischeBundesanstalt, Bundessortenamt und CHemischeIndustrie (BBCH) scale assignment. The cultivated paddy rice fields were analyzed through TerraSAR-X co-polar SAR information. The similar groups were structured by using K-means clustering algorithm which is collected of backscattering intensities and polarimetric phase differences. Then, the similar groups were classified by clustering approach based on the temporal separability of descriptive parameters. The growth trend based update was used for avoiding misclassification between two distinct growth stages. However, this method was sensitive to noise and outlier features values.

ISSN (online) : 2277-5390
ISSN (Print): 2277-5382

Table 1. Comparison of different methods in plant disease detection system

Ref. No.	Method	Approach used	Merits	Demerits	Number of images considered	Performance Metrics
[3]	Semi-automatic method	RGB to HSV and L*a*b* transformation, Histogram analysis, Segmentation based on H and a channel	High flexibility and robust	Accuracy is reduced due to some sources of error	938 images	Correlation coefficient=0.90 Lin's concordance correlation coefficient=0.86 Standard error of estimate=6.62
[4]	PLSR, v-SVR, GPR	Partial least square regression, Support vector regression, Gaussian process regression	Less RMSE value, High coefficient of determination	training samples are expensive and the accuracy is reduced due to the scattering of the data	175 images	Coefficient of determination: PLSR=87 v-SVR=95 GPR=97 RMSE: PLSR=0.13 v-SVR=0.12 GPR=0.05
[5]	Computer Vision-based method	Segmentation based on marker controlled watershed transformation and super-pixels, Support vector machine	High accuracy, less execution time	Robustness is low	180 images	Accuracy: Septoria disease=70% Yellow rust disease=95% Severity Assessment: Septoria disease=92% Yellow rust disease=67%
[6]	eAGROBOT	Image pre-processing, Neural network classification	High classification accuracy	High maintenance complexity and cost	50 cotton images, 100 groundnut images	Accuracy: Cotton=92% Groundnut=89%
[7]	Enhanced SVM method	Descriptors, Image masking, Feature matching	High area detection and less execution time	Accuracy is low	300 images	Time=0.047s Area detection=82.9% Accuracy=69%

5. Comparison of various techniques

This section provides an overview of advantages and disadvantages in different techniques used in agricultural monitoring systems whose functional scenarios are discussed in brief in previous section. From the following table, a better technique can be determined which provides considerable enhancement in the crop production. The comparison between different plant disease detection methods are shown in Table 1.The comparison between different soil moisture predictions method are shown in Table 2.The comparison between crop growth monitoring systems are shown in Table3.

Table 2. Comparison of different methods in soil moisture prediction system

Ref. No.	Method	Approach used	Merits	Demerits	Number of samples considered/data base	Performance Metrics
[8]	Improved Back-Propagation method	Particle swarm optimization, wavelet transform, BP neural network	High prediction accuracy, Fast convergence speed	Mean square error is high	17520 data	Number of iterations=14 Prediction accuracy=98%
[9]	Multi-spectral and FTIR method	Gravimetric method, Normalized difference vegetation index, Land surface temperature, Temperature vegetation dryness index	Less resource consuming, Less RMS value	High computing power	120 soil samples	Normalized Difference Vegetation Index=0.54 Land Surface Temperature=40.57° C
[10]	Ensemble Kalman Filter (EnKF) method	Hydrus-1D model, EnKF method, Inversion method	Less RMSE value	Performance is still poorer than synthetic approaches	SMAP-database	RMSE=0.039
[11]	Multi-temporal C and L-band SAR data monitoring system	Backscatter at C band, Backscatter at L band, Radiative transfer approach	High feasibility, High accuracy	Incidence angle is critical issue since exceeds its moderate values	250 wheat fields and 140 rape fields	RMSE: Wheat=1.6 Rape=1.7 Correlation: Wheat=0.75 Rape=0.69
[12]	Bayesian change detection method	Bayesian approach, Integral equation model	Backscattering coefficient is reduced	Analysis of backscattering changes is high complexity	32 fields, 10 soybean and 21 corn fields	L band: Correlation coefficient = 0.85 P band: Correlation coefficient =0.80

Table 3. Comparison of different methods in crop growth monitoring system

Ref. No.	Method	Approach used	Merits	Demerits	Number of plant considered/Dat abase	Performance metrics
[13]	MESMA	Segmentation, Clustering approach, Linear spectral mixture analysis,	High computational efficiency	Influence of shadow in mixture is not considered	30 trees	Coefficient of determination=0.62 RMSE=0.02
[14]	MUSIC-CSR	Library pruning, Dynamic unmixing method, Multiple endmember spectral mixture, Sparse unmixing	High accuracy, Less running time	Less reliability	10 citrus trees	Time=0.71s Projection error=0.001
[15]	ANFIS	Fuzzy logic, Multiple linear regression, Adaptive Neuro fuzzy inference system	High accuracy, RMSE value is less	Four parameters such as biomass, esw, rain and radiation are only considered	50 wheat dataset	RMSE=3.328
[16]	Improved algorithm based monitoring system	Ant colony optimization, Centre data aggregation algorithm,	Less energy consumption,	High average delay	360 temperature and humidity data	Average delay=1.69% Energy consumption=74.1%
[17]	Feature clustering method	Bun-dessortenamt und CHemischeIndustrie scale, K-means clustering, Classification, Growth-trend-based update	Better classification accuracy	Sensitive to noise and outlier features values	Suville and Ipsala database	Suvile dataset: Accuracy =92.69% Ipsala dataset: Accuracy =91.27%

6. Conclusion

In this research, the data mining techniques in agriculture such as plant disease detection techniques, soil moisture prediction methods and crop growth monitoring techniques are studied and compared for analysing the performance. From this study, it is clear that the crop production should be enhanced by plant disease prediction, growth monitoring techniques. The problem in plant disease prediction and growth monitoring system for enhancing crop production are developed by optimization techniques in data mining which will be our future work.

7. References

1. R. V. Chandrashekar. Development or end of Agriculture? Implications on Agriculture. *Indian Journal of Innovations and Developments*. 2012; 1(8), 653-658.
2. D. Ramesh, B. V. Vardhan. Data mining techniques and applications to agricultural yield data. *International Journal of Advanced Research in Computer and Communication Engineering*. 2013; 2(9), 3477-3480.
3. J. G. A.Barbedo. A novel algorithm for semi-automatic segmentation of plant leaf disease symptoms using digital image processing. *Tropical Plant Pathology*. 2016; 41(4), 210-224.
4. D. Ashourloo, H. Aghighi, A. A. Matkan, M. R. Mobasheri, A. M.Rad. An Investigation Into Machine Learning Regression Techniques for the Leaf Rust Disease Detection Using Hyperspectral Measurement. 2016; 9(9), 4344-4351.
5. L.Han, M. S. Haleem, M.Taylor. A novel computer vision-based approach to automatic detection and severity assessment of crop diseases. In: *Science and Information Conference (SAI).* 2015 Jul; 638-644.
6. S. K. Pilli, B. Nallathambi, S.J. George, Diwanji, V. eAGROBOT-A robot for early crop disease detection using image processing. In: *Electronics and Communication Systems (ICECS), International Conference on* IEEE. 2014; 1-6.
7. R.Kaur, S. S.Kang. An enhancement in classifier support vector machine to improve plant disease detection. In: *MOOCs, Innovation and Technology in Education (MITE), IEEE 3rd International Conference on* IEEE. 2015; 135-140.
8. Y. Xiaoxia, Z.Chengming. A soil moisture prediction algorithm base on improved BP. In: *Agro-Geoinformatics (Agro-Geoinformatics), 2016 Fifth International Conference on* IEEE. 2016 Jul; 1-6.
9. S. M. Z.Younis, J.Iqbal. Estimation of soil moisture using multispectral and FTIR techniques. *The Egyptian Journal of Remote Sensing and Space Science*. 2015; 18(2), 151-161.
10. J. Dong, S.C. Steele-Dunne, T. E. Ochsner, N.van de Giesen. Determining soil moisture by assimilating soil temperature measurements using the Ensemble Kalman Filter. *Advances in Water Resources*. 2015, *86*, 340-353.
11. A. Balenzano, F. Mattia, G. Satalino, M. W. Davidson. Dense temporal series of C-and L-band SAR data for soil moisture retrieval over agricultural crops. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2011; 4(2), 439-450.
12. C.Notarnicola. A bayesian change detection approach for retrieval of soil moisture variations under different roughness conditions. *IEEE Geoscience and Remote Sensing Letters*. 2014; 11(2), 414-418.
13. L. Tits, B. Somers, P.Coppin. The potential and limitations of a clustering approach for the improved efficiency of multiple endmember spectral mixture analysis in plant production system monitoring. *IEEE Transactions on Geoscience and Remote Sensing*. 2012; 50(6), 2273-2286.
14. M. D. Iordache, L. Tits, J. M. Bioucas-Dias, A. Plaza, B. Somers. A dynamic unmixing framework for plant production system monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2014; *7*(6), 2016-2034.
15. A. Shastry, H. A. Sanjay, M. Hegde. A parameter based ANFIS model for crop yield prediction. In: *Advance Computing Conference (IACC), IEEE International* on IEEE. 2015, 253-257.
16. W. T. Sung, H. Y. Chung, K. Y.Chang. Agricultural monitoring system based on ant colony algorithm with centre data aggregation. *IET Communications*. 2014; 8(7), 1132-1140.
17. O. Yuzugullu, E.Erten, I.Hajnsek. Rice growth monitoring by means of X-band co-polar SAR: Feature clustering and BBCH scale. *IEEE Geoscience and Remote Sensing Letters*. 2015; 12(6), 1218-1222.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:
D. Sabareeswaren, A. Edwin Robert. A survey on data mining techniques in agriculture. *Indian Journal of Innovations and Developments.* 2016; 5 (8), August.