International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

AGRICULTURAL CROP CLASSIFICATION MODELS IN DATA MINING TECHNIQUES

J. Saranya[1], Dr. M. Mohamed Sathik[2], Dr. S. Shajun Nisha[3]

[1]M.phil Research Scholar, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India
[2]Principal, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India
[3]Assistant Professor & Head, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India

---

Abstract - *Data mining is relatively a new approach in the field of agriculture. Extraction of knowledge in agricultural data is a difficult task. These difficulties are due to climate, geographical and biological factors. In order to handle these problems, remote sensing technology offers traditional methods to categorize agricultural crops. This paper focuses on the analysis of the agricultural landsat data and finding optimal parameters to maximize the accuracy level of classified crops using data mining techniques like Naive Bayes. The performance metrics used for analysis are kappa statistics, Correctly classified instances, Incorrectly classified instances, Mean absolute Error, Root Mean Squared Error, Average precision and Average Recall.*

***Key Words*: Data mining, Naive Byes classifier, Landsat, Accuracy**

1. INTRODUCTION

Agriculture is the most important area of application, particularly in developing countries. The use of technology in agriculture can change the decision-making situation and farmers can produce a better way. Data mining plays a crucial role in decision making on several issues related to the field of agriculture. Data mining is a useful technique for finding the relevant pattern of the huge data set. The agriculture field contains many data such as soil data, harvest data, metrological data, geographical and Landsat data. Landsat data are used to analyze and handle with algorithms in data mining such as Naive Byes. This method is used in the Landsat data and produce extraordinary significant benefits and predictions that can be used for commercial and scientific purposes.

1.1 LANDSAT

The Landsat satellite series has provided a continuous earth observation data record since the early 1970 s. Landsat1-3 carried the multi spectral scanner system instrument. In addition to MSS, Landsat 4 and 5 carried out the thematic mapper instrument in the age of the Landsat 30m pixel resolution. [6]Landsat 5 was launched on March 1, 1984 and functioned over 28 years until 2012. Landsat 7 was launched on April 15, 1999 and with Enhanced Thematic Mapper Plus the favorable 30 m spatial resolution with better accuracy in radiometric and geometric calibration. Both Landsat 5 and 7 included a thermal infrared band at a spatial resolution of 120m and 60 m respectively. Landsat 8 was launched on February 2013. The Landsat 8 includes the operational land imager and the thermal infrared sensors.

The Landsat data are corrected radio metrically and geometrically by the USGS Earth resources Observation and science center. Terrain correction is applied using a digital elevation model and ground control points. USGS distributes shortwave Landsat data digital number and surface reflectance. They can be downloaded through the USGS (United States Geological Survey) global visualization viewer to make the entire archive of Landsat data available at no charge[5]. A fundamental Challenge in pursuit of either of these goals is the considerable spatial and temporal heterogeneity of agriculture landscapes[1].

1.2 LITERATURE REVIEW

Yang, Z., Mueller [2] have proposed a choice tree-administered arrangement technique is utilized to create the openly accessible state level trim cover groupings and give edit real estate's gauges dependent on the CDL and NASS June Agricultural Survey ground truth to the NASS Agricultural Statistics Board. This paper gives a diagram of the NASS CDL program. It depicts different information, handling strategies, order and approval, precision evaluation, CDL item particulars, scattering settings and the product real estate estimation procedure.

Duro, D.C., Franklin et. al [3] have proposed Pixel-based and protest based picture examination approaches for ordering wide land cover classes over farming scenes are looked at utilizing three administered machine learning calculations: Decision tree (DT), Random Forest (RF), and vector machine (SVM). By and large grouping correctnesses between pixel based and question based arrangements were not measurably critical ($p>0.05$) when a similar machine learning calculations were connected. Utilizing object-based picture examination, there was a factually noteworthy distinction in grouping precision between maps delivered utilizing the DT calculation contrasted with maps created utilizing either RF ($p=0.0116$) or SVM calculations ($p=0.0067$). Utilizing pixel-based picture examination, there

was no measurably critical contrast (p>0.05) between results delivered utilizing distinctive arrangement calculations.

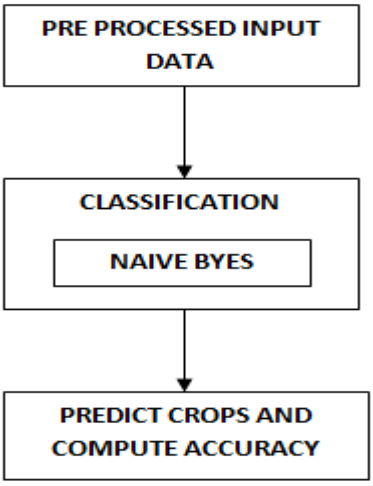
Kathija and Dr.S.Shajun Nisha. [8] have proposed the smallest subset of features from Wisconsin Diagnosis Breast Cancer (WDBC) dataset by applying confusion matrix accuracy and 10-fold cross validation method to ensure high accurate classification of breast cancer that can ensure highly accurate ensemble classification of breast cancer in benign or malignant.

A.Ameer Rashed Khan, Dr.S.Shajun Nisha and Dr.M.Mohamed Sathik.[9]have proposed the performance of different clustering algorithm such as Expectation Maximization (EM), Farthest Fast and K-means by correctly clustered instances and time taken to build the model for mushroom dataset using data mining tool WEKA.

1.3 Motivation Justification

Bayes' rule is a general technique in classification, but costly in terms of requiring large training sets. Naïve Bayes classifiers are based on the assumption that likelihoods of each feature are independent of those of other features. By making independence assumptions, much less training data is required. Often the results are very good.

1.4 Outline of the proposed work



1.5 Organization of the paper

This paper is organized as follows: In Section II describes the classification methods. Section III Display the Experimental results, and in section IV conclusion is placed.

2. METHODOLODY

In this paper Naïve Byes classification methodology are used to find the accuracy of crop Classification.

2.1 Dataset

Dataset consists of different attributes that have complex relationships between dynamic variables. There are two types of datasets training and testing. Training data set consists of 449 instances, Testing data set consists of 94 instances.

2.2 Classification

2.2.1 Naïve Byes

Naive Bayes Approach is Bayesian approach for classification is a statistical and linear classifier which predicts class label for data instance on the basis of distribution of attribute values. This is a parametric classification where the size of the classifier remains fixed. Distribution can be normal (Gaussian), kernel, multivariate or multi nominal. While normal distribution for weather data, Bayesian classifiers use Bayes theorem to find posterior probabilities of input data instance of all classes. Class label having maximum conditional probability is assigned to data instance. Naive Bayes attributes have no effect on each other and they have independent distribution of values.

Let the training data X, the posterior probability of a hypothesis h, P (h|x) follows the Bayes theorem as:

$$P(h|x)=\frac{p(h|x)p(h)}{p(x)}$$

Where P(h|x) is the posterior probability of hypothesis h over training data X, P(x|h) is the probability of likelihood that the X is conditionally independent of other variables and also called prior probability. Maximum the posterior probability strengthens the selected arguments for prediction.

3. EXPERIMENTAL RESULTS

3.1 Performance Metrics

3.1.1 True Positive Rate

A true positive is an outcome where the model correctly predicts the positive tuples. It measures the proportion of actual positives that are correctly identified.

$$\textbf{True Positive Rate} = \frac{TP}{TP+FN}$$

3.1.2 False Positive Rate

The false positive rate is ratio between the numbers of negative events wrongly classified as positive.

$$\textbf{False Positive Rate} = \frac{TP}{FP+TN}$$

3.1.3 Precision

It is the proportion of instances that are truly of a class divided by the total instances classified as that class.

Precision = $\frac{TP}{TP+FP}$

3.1.4 Recall

It is the proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)

Recall = $\frac{TP}{TP+FN}$

3.1.5 F-Measure

A combined measure for precision and recall is calculated as

F-measure= $2\left(\frac{Precision*Recall}{Precision+Recall}\right)$

3.1.6 Matthews Correlation Coefficient

MCC has a range of values from -1 to 1 where -1 indicates completely wrong binary classifier while 1 indicates a completely a correct binary classifier. Using the MCC allows one to gauge how well their classification model/function is performing.

MCC = $\frac{TP*TN-FP*FN}{[(TP+FP)*(FN+TN)*(FP+TN)*(TP+FN)]^{\left(\frac{1}{2}\right)}}$

3.1.7 Precision Recall Curve

The PRC is called as Precision recall characteristics curve. It is a comparison of two operating characteristics (PPV and sensitivity) as the criteria. A PRC curve is a graphical plot which explain the performance of binary classifiers as its discrimination threshold is varied.

3.1.8 Receiver Operating Characteristics

ROC is comparison of two operating characteristics TPR and FPR. A receiver operating characteristic curve is a graphical action which analyses the performance of a classified as its partiality threshold is varied. It is an completion of plotting the true positive rate vs. false positive rate at varied threshold settings.

3.1.9 Mean Absolute Error

Mean Absolute Error is the common difference between the Original Values and the Predicted Values. It gives us the capacity how far the predictions were from the actual output. They don't gives us any idea of the direction of the error. we are under predicting the data or over predicting the data. Mathematically, it is written as

Mean Absolute Error = $\frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y_j}|$

3.1.10 Mean Squared Error

Mean Squared Error(MSE) is similar to Mean Absolute Error, the only difference being that MSE takes the fair and square difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

Mean Squared Error = $\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y_j})^2$

3.2 Performance Evaluation

Table 1: Detailed Accuracy On Training Dataset

	CORN	BTCORN	SOYBEAN	WEIGHTED AVERAGE
TP RATE	1	0.404	0.419	0.47
FP RATE	0.538	0.005	0.066	0.08
PRECISION	0.175	0.99	0.756	0.83
RECALL	1	0.404	0.419	0.47
F-MEASURE	0.298	0.574	0.539	0.534
MCC	0.284	0.468	0.429	0.436
ROC	0.856	0.738	0.731	0.748
PRC	0.571	0.831	0.659	0.748

Table 2: Detailed Accuracy On Testing Dataset

	CORN	BTCORN	SOYBEAN	WEIGHTED AVERAGE
TP RATE	1	0.52	0.267	0.521
FP RATE	0.486	0.159	0	0.166
PRECISION	0.414	0.542	1	0.728
RECALL	1	0.52	0.267	0.521
F-MEASURE	0.585	0.531	0.421	0.492
MCC	0.461	0.365	0.399	0.406
ROC	0.945	0.764	0.905	0.878
PRC	0.795	0.705	0.909	0.826

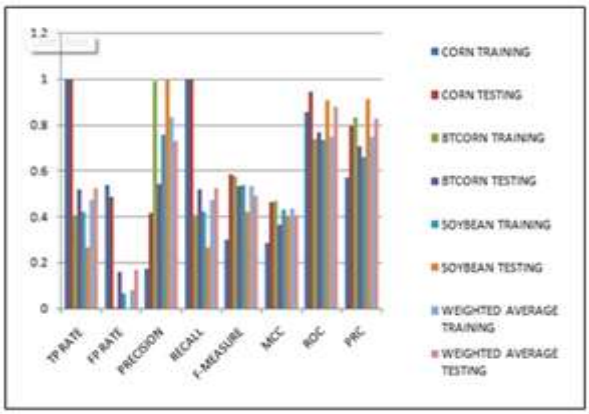


Chart 1: Performance Evaluation Comparison for Crop Dataset

Table 3: Dataset Classification Based On Its Properties

PERFORMANCE METRICS	NAIVE BYES	
	TRAINING	TESTIING
CORRECTLY CLASSIFIED INSTANCES	211	49
INCORRECTLY CLASSIFIED INSTANCES	238	45
KAPPA STATISTICS	0.2916	0.329
MEAN ABSOLUTE ERROR	0.3827	0.302
ROOT MEAN SQUARED ERROR	0.5596	0.5092
RELATIVE ABSOLUTE ERROR	97.69%	71.25%
ROOT RELATIVE SQUARED ERROR	93.73%	98.69%

3. CONCLUSIONS

Various data mining techniques are implemented on the input data to assess the best performance yielding method. The present work used data mining techniques Naive byes to obtain the optimal parameters to achieve higher accuracy of crop classification. In future enhancements, we have a comparative analysis of different classification algorithm with the accuracy levels and choose the best algorithm in classification models.

REFERENCES

1) Bolton, D.K., Friedl and M.A., "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics., " Elsevier, 2013.

2) Boryan, C., Yang, Z., Mueller, R., Craig and M., "Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program.," Elsevier, 2011.

3) Duro, D.C., Franklin, S.E., Dube and M.G., "A comparison of pixel-based and object based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery.," Elsevier, 2012.

4) Hansen, M.C., Loveland and T.R., "A review of large area monitoring of land cover change using Landsat data.," Elsevier, 2012.

5) Lobell, and D.B., "The use of satellite data for crop yield gap analysis.," Elsevier, 2013.

6) Yaping Cai, Kaiyu Guan and Jian Peng., "A high Performance and in season classification system of field level crop types using time series landsat data and a machine learning approach.," Elsevier, 2014.

7) Kathija and Dr. S.Shajun Nisha., "Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques., "International Journal of Innovative Research in Computer and Communication Engineering, 2016.

8) A.Ameer Rashed Khan, Dr.S.Shajun Nisha and Dr.M.Mohamed Sathik., "Clustering Techniques For Mushroom Dataset.," International Research Journal of Engineering and Technology (IRJET), 2018.

BIOGRAPHIES

J.Saranya, M.Phil Research scholar, currently pursiung at sadakathullah appa college, I had completed my PG at Apollo arts and Science college, Madras University in IT and completed my B.Sc., Computer Science at Govindammal Aditanar College for Women, Manonmaniam Sundaranar University, Tirunelveli. I had a certification of NPTEL courses. My research area is in Data Mining.

Dr.M.Mohamed Sathik, Principal Sadakathullah Appa college,Tirunelveli.He has completed Ph.D(Computer science & engineering) Ph.D(Computer science), M.Phil. (Computer Science),M.Tech(Computer Science and Information Technology) in Manonmaniam Sundaranar University, Tirunelveli. He has so far guided more than 40 research scholars. He has published more than 100 papers in International Journals and also two books. He is a member of curriculum development committee of various universities and autonomous colleges of Tamil Nadu. He is a syndicate member Manonmaniam Sundaranar University, Tirunelveli. His specializations are VRML, Image Processing and Sensor Networks.

Dr.S.Shajun Nisha, Assistant Professor and Head of the PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli. She has completed M.Phil. (Computer Science) M.Tech (Computer and Information Technology) in Manonmaniam Sundaranar University, Tirunelveli and She had completed Ph.D (Computer Science) in Bharathiar university, Coimbatore. She has involved in various academic activities. She has attended so many national and international seminars, conferences and presented numerous research papers. She is a member of ISTE and IEANG and her specialization is Medical Image Processing and Neural Networks