Crop Analytics Using Machine Learning

Kirtiraj Kadam

Department of Technology, Savitribai Phule Pune University (SPPU), Ganeshkhind Rd., Pune, India

---

***Abstract*—**Crop classification and yield estimation has become an important activity across the globe. Identifying different types of crops using satellite imagery can help to formulate better yield output in the field of agriculture. Sentinel-2 data is high resolution satellite image data comprising 13 optical bands (with a resolution varying from 10 to 60 meters) and can be used to identify and classify different types of crops. The research work uses machine learning techniques like Random Forests, Support Vector Machines and Naïve-Bayes classifier to classify different crops using Sentinel-2 data. After obtaining the ground truth data concerning different land cover classes, different classifiers have been trained. The trained classifiers have been used to perform the classification task and calculate the classification accuracies.

Keywords—Crop classification, Machine Learning, Naïve Bayes, SVM, Random Forests, Confusion Matrix, Kappa Coefficient

I. INTRODUCTION

Crop classification and yield estimation have become an important research activity in the field of agriculture as it can provide important information about the crops cultivated and harvested in a particular region. By studying different crop patterns the agricultural researchers can get an idea about the maximum estimate of different crops that are cultivated in a particular region and derive vital information that would be helpful in performing various decision making problems for managing various agricultural resources. Crop classification makes use of remote sensing data.

The image data needed to perform crop classification and yield is typically downloaded from freely available sources such as MODIS and Landsat. However, MODIS data is typically characterized by a mix-pixel problem (pixel that represents two of more entities on the ground) because of its low spatial resolution (250-500 m).

Hence, better results can be achieved using a 30-meter resolution Landsat data, specifically, for the regions characterized by small agricultural fields. Earlier crop identification was performed by using optical remote sensing data but the accuracy for a given region usually depended on the noise and cloud free pixels, which hampered the identification of the crops during monsoon season.

Currently, Europe's Sentinel-2 data is being popularly used for crop identification and classification as its cameras can capture different colors of light (bands of electromagnetic spectrum) in different spatial resolutions, which can portray a different story about the agriculture environment. It has 13 spectral bands, comprising visible, near-infrared and the shortwave infrared at different spatial resolutions ranging from 10 to 60 meters on the ground. Sentinel-2 can distinguish between dead vegetation (barren land) and live vegetation or between rich and poor crop yield as it has infrared footprints.

The research work attempts to use machine learning algorithms to classify different crops using the sentinel-2 data and finding the crop patterns over a region. The system would differentiate between different types of crops and also find other entities such as barren land, water bodies, urban sprawl, etc. A comparative study of different crop classification algorithms has been done and analysed.

II. THEORY AND LITERATURE SURVEY

A. Machine learning algorithms

Decision Trees: A decision tree lets you take decisions to get the desired outcomes by choosing the lowest path or resource cost. The tree consists of internal nodes which represent test on an attribute, the branches represents the outcome, leaf node represents a class label while the root to leaf represent rules which would help to classify. Decision trees are powerful and are used as supervised machine learning algorithms.

Random forest (RF): RF generates a forest of decision trees, enabling predictions to be more powerful and robust. Random forest is an ensemble method for classification, where response of several classifiers is combined to get final prediction result. It has two important parameters, namely, *n_tree* (number of trees to form ensemble) and *m_try* (number of variables/predictors used to split the nodes). The best split for a node increases the accuracy of the classification.

Support Vector Machine (SVM): SVM is a supervised learning model used in machine learning which makes use of statistics for classification of data. A support vector machine works by drawing a best hyperplane (in 2D it would a simple line), which separates two distinct groups of data samples. The best hyperplane is the one that is equidistant from boundary points of two classes (called support vectors) to get maximum separating margin between the classes.

Naive Bayes: Naive Bayes algorithm classifies pattern with attribute array A into class with highest probability $P(C_k \mid A)$. By Bayes theorem this equals $(P(C_k)P(A \mid C_k))/P(A)$. Ignoring the denominator (which is constant) the final classification expression reduces to $\text{argmax}_k\ P(C_k)\prod P(a_i \mid C_k)$, where a_i's are actual attribute values of pattern A to be classified.

B. Literature survey

Orynbaikyzy et al. [2] combined Sentinel-2 and SAR data and used Random forests to improve crop classification accuracy. Liang et al. [3] have used high-resolution, multi-

spectral data to get spectral, textural and structural features for mapping target vegetation using SVM and Random forests. Yi et al. [4] have used multi-temporal Sentinel-2 data and RF algorithm to generate the crop classification map for the Shiyang River Basin with improved accuracy. Jitendra Singh et al. [7] also used Sentinel-1 and SAR data for crop classification in the Indian context. They got an accuracy of 85% using Random forests

Neetu and S. S. Ray [5] have used Sentinel-2A data and various ML algorithms for crop classification. The True Color, False Color and NDVI images were used to get crop features. Their results: CART- 73.4 %, RF-93.3 % and SVM-74.3 %. However, in their approach the vegetable class got mixed with other classes

Zheng et al. [8] and Waldner et al. [9] have used MODIS and Landsat data for vegetation mapping. Sonobe1 et al. [11] have established that band 4 (Red) of Sentinel-2 and VV polarization data of Sentinel-1 are important for crop classification. Korhonen [12] has used Sentinel-2 data to estimate boreal forest canopy cover and leaf area index. R. Saini and S.K. Ghosh [6] classified crops using four Sentinel-2 data bands (NIR, Red, Green, Blue). They got accuracy of 84.22 % by RF and 81.85 % with SVM

III. METHODOLOGY

Crop classification involves three entities, namely; data, model (classifier), and the experts who provide the ground truth. In this research work we have used 10-metre resolution Sentinel-2 data spawning Arag and Bedag villages in Sangali District, Maharashtra (Fig. 1). The ground truth data comprising GPS points and the corresponding crops associated with each GPS point was collected manually. The GPS points were later mapped to pixel positions in the image to create polygons depicting each crop (and non-crop) category.

The classification algorithms used in the study comprised Naïve Bayes, SVM and Random forests. Amongst various bands in the Sentinel-2 data, we have used three bands (NIR, Red, and Green) for classification creating a false color composite as shown in Fig. 2. The NIR band was the most useful band as vegetation reflects the radiation that was necessary for the classification of the crops.

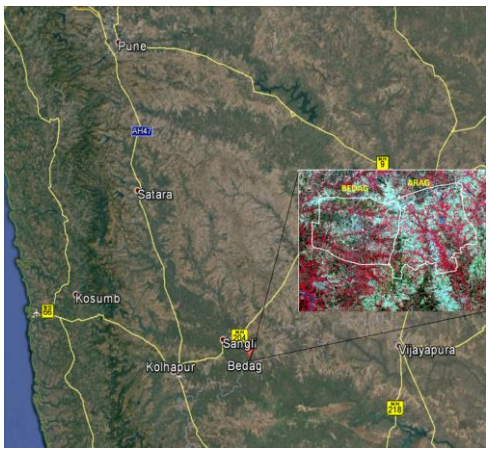

Fig. 1: Study Area

The next step is to train the data using this information, once the models are trained, we will then classify the crops using classifiers such as Naïve Bayes, Support Vector Machine (SVM), Random Forest (RF), etc. and compare the accuracy of every classifier.

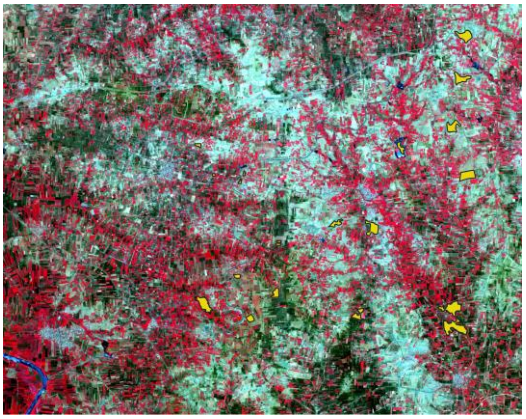

Fig. 2: FCC of Study Area and overlaid polygons

The total number of sample pixels extracted under the polygons (representing different classes) overlaid over the crop image were 17,056. Of these, a total of 10,000 samples were randomly selected for training and testing. The 10,000 samples were split in the ratio 70:30 for training and testing. That is, 70% of the 10,000 samples were used for training while the remaining was used for testing.

IV. RESULTS AND DISCUSSION

A. *Classification using Naive Bayes*

After performing the training and testing process, the overall accuracy for Naive Bayes classifier was 97.32%. The confusion matrix for this classifier is as shown in Table 1. The kappa factor was 0.9602. The statistics by class is as depicted in Table 2.

	Reference				
Prediction	1	2	3	4	5
1	2023	9	0	0	3
2	10	505	3	0	0
3	0	11	1926	36	6
4	0	0	26	69	0
5	2	0	28	3	456

Table 1: Confusion matrix for Naïve Bayes

Barren land	Water body	No crop	Sugarcane	Grapes
0.9951	0.9795	0.9772	0.8168	0.9867

Table 2: Class wise accuracy

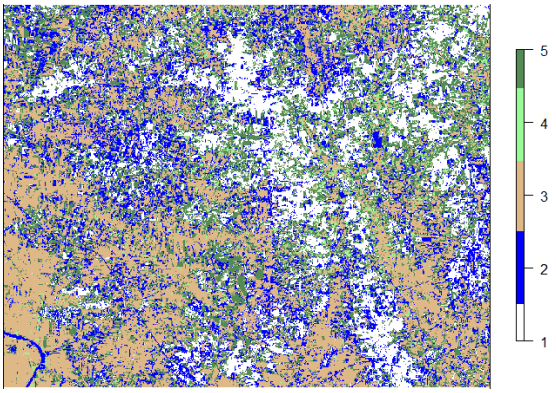

Fig. 3: Classification Results for Naïve Bayes

The pixel classification using the trained Naive Bayes model is as depicted in Fig. 3. The color code and the class indices represent 1-Barren Land, 2-Water Body, 3-No Crop, 4-Sugarcane, and 5-Grapes.

B. *Classification using SVM*

After performing the training and testing process, the overall accuracy for the SVM classifier was 98.26%. The confusion matrix for this classifier is as shown in Table 3. The kappa factor was 0.9741. The statistics by class is as depicted in Table 4.

	Reference				
Prediction	1	2	3	4	5
1	2032	0	0	0	2
2	0	523	0	0	0
3	0	2	1944	33	5
4	0	0	14	72	2
5	3	0	25	3	456

Table 3: Confusion matrix for SVM

Barren land	Water body	No crop	Sugarcane	Grapes
0.9989	0.9981	0.9838	0.8317	0.9869

Table 4: Class wise accuracy

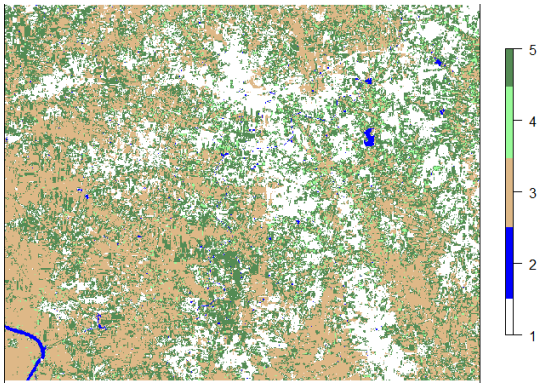

Fig. 4: Classification Results for SVM

The pixel classification using the trained SVM model is as depicted in Fig. 4. The color code and the class indices represent 1-Barren Land, 2-Water Body, 3-No Crop, 4-Sugarcane, and 5-Grapes.

C. *Classification using Random forests*

After performing the training and testing process using the Random forests classifier, the overall accuracy observed was 98.59 %. The confusion matrix for this classifier is as shown in Table 5. The kappa factor was 0.9791. The statistics by class is as depicted in Table 6.

	Reference				
Prediction	1	2	3	4	5
1	2031	9	0	0	2
2	1	521	3	0	0
3	0	4	1947	21	4
4	0	0	17	86	0
5	3	0	19	1	459

Table 5: Confusion matrix for Random forests

Barren land	Water body	No crop	Sugarcane	Grapes
0.9987	0.9961	0.9863	0.8964	0.9910

Table 6: Class wise accuracy

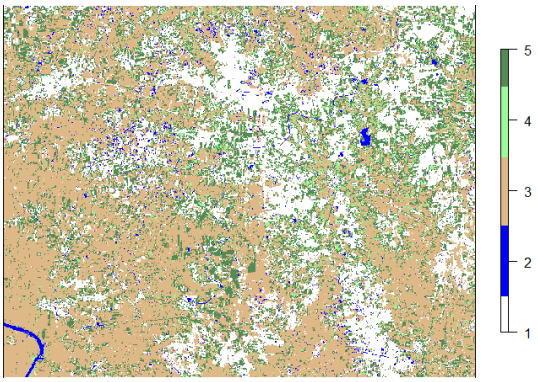

Fig. 5: Classification Results for Random forests

The pixel classification using the trained Random forests model is as depicted in Fig. 5. The color code and the class indices represent 1-Barren Land, 2-Water Body, 3-No Crop, 4-Sugarcane, and 5-Grapes.

D. *Classification using Uniform samples*

Since the number of ground truth samples in the sugarcane class (class 4) is less in comparison with other classes, we consider the complete set of pixels values for training and testing, that is, 17,056 sample pixels. These 17,056 sample pixels were split in the ratio 70:30 for training and testing. That is, 70 % of the 17,056 samples were used for training while the remaining was used for testing.

	Reference				
Prediction	1	2	3	4	5
1	99	0	0	0	1
2	0	100	0	1	0
3	0	0	93	2	3
4	0	0	3	96	0
5	1	0	4	1	96

Table 7: Confusion matrix for Uniform samples and RF

Barren land	Water body	No crop	Sugarcane	Grapes
0.9938	0.9988	0.9587	0.9763	0.9725

Table 8: Class wise accuracy

Finally, 250 sample pixels from training set of each class were randomly selected for uniform training. Similarly, 100

sample pixels from testing set of each class were randomly selected and used for testing. After performing the training and testing process, the overall accuracy for the RF classifier was 96.80%. The confusion matrix for this classifier is as shown in Table 7. The kappa factor was 0.9791. The statistics by class is as depicted in Table 8.

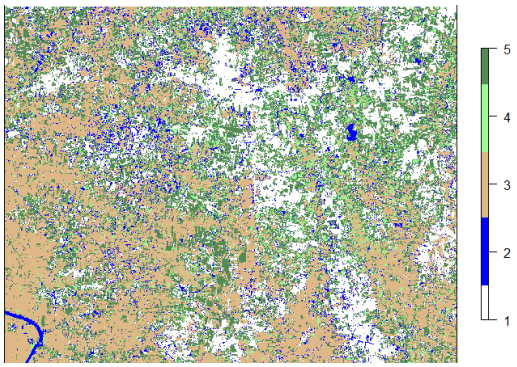

Fig. 6: Classification Results for Uniform samples and RF

The pixel classification using the trained Random forests model and uniform sampling is as depicted in Fig. 6. The color code and the class indices represent 1-Barren Land, 2-Water Body, 3-No Crop, 4-Sugarcane, and 5-Grapes.

All the classifiers also classify other entities such as barren land, water bodies, urban sprawl, etc. The Random forest classifier is an ensemble classifier and hence, as expected, it has a better classification accuracy.

V. CONCLUSION

The overall accuracy for Naïve Bayes classifier was 97.32%, while the SVM classifier gave an accuracy of 98.26%, and the Random forest classifier gave an accuracy of 98.59%. It is observed that Random forests classifier has the highest accuracy in comparison to other classifiers for crop classification. The Random forest classifier is an ensemble classifier and therefore gives better classification accuracy.

In order to avoid bias in classification accuracy due to skewness in number of available training samples in each class, we have carried out uniform sampling of different crop categories and performed the training. In this case we have used Random forest classifier for training and testing. The classification accuracy observed after uniform sampling is 96.8%. Although the overall accuracy comes down, the accuracy across all the class is uniform as seen from Table 8.

ACKNOWLEDGMENT

We are obliged to Dr. Manish Kale, ESEG Group, C-DAC, Pune, for providing input images for our study.

REFERENCES

[1] V. Rodriguez-galiano, B. Ghimire, J. Rogan, M. Chica-olmo, and J. Rigol-sanchez, "An assessment of the eectiveness of a random forest classifier for land-cover classification", ISPRS Journal of Photogrammetry Remote Sensing, vol. 67, pp. 93-104, 2012.

[2] A. Orynbaikyzy, U. Ursula Gessner, M. B, and C. C. C., "Crop type classification using fusion of Sentinel-1 and Sentinel-2 data: Assessing the impact of feature selection, optical data availability, and parcel sizes on the accuracies", Remote Sensing, vol. 12, 2020.

[3] W. Liang, M. Abidi, L. Carrasco, J. McNelis, L. Tran, Y. Li, J. Grant, and W. Liang, "Mapping vegetation at species level with high-resolution multispectral and LIDAR data over a large spatial area: A case study with Kudzu", Remote Sensing, vol. 12, no. 609, 2020.

[4] Z. Yi, L. Jia, and Q. Chen, \Crop classification using multi-temporal Sentinel-2 data in the Shiyang river basin of China," Remote Sensing, vol. 12, 2020.

[5] Neetu and S. S. Ray, "Exploring machine learning classification algorithms for crop classification using Sentinel-2 data", Remote Sensing and Spatial Information Sciences, vol. XLII, no. 3, 2019.

[6] J. Singh, U. Devi, J. Hazra, and S. Kalyanaraman, "Crop-identification using Sentinel-1 and Sentinel-2 data for Indian region", Geoscience and Remote Sensing (IGARSS), pp. 5312-5314, 2018.

[7] R. Saini and S. K. Ghosh, \Crop classification on single date Sentinel-2 imagery using random forest and support vector machine," Remote Sensing and Spatial Information Sciences, vol. XLII, no. 5, 2018.

[8] B. Zheng, S. Myint, P. Thenkabail, and R. Aggarwal, "A support vector machine to identify irrigated crop types using time-series landsat NDVI data," International Journal of Applied Earth Observation and Geoinformation, vol. 34, pp. 103-112, 2015.

[9] F. Waldner, G. Canto, and P. Defourny, "Automated annual cropland mapping using knowledge-based temporal features", ISPRS Journal of Photogrammetry and Remote Sensing, vol. 110, pp. 1-13, 2015.

[10] A. Whyte, K. Ferentinos, and G. Petropoulos, "A new synergistic approach for monitoring wetlands using Sentinels-1 and -2 data with object-based machine learning algorithms", Environmental Modelling and Software, vol. 104, pp. 40-54, 2018.

[11] R. Sonobe, Y. Yamaya, H. Tani, X.Wang, N. Kobayashi, and K. Mochizuki, "Assessing the suitability of data from Sentinel-1A and 2A for crop classification", GIScience and Remote Sensing, vol. 54, pp. 918-938, 2017.

[12] L. Korhonen, P. Packalen, and M. Rautiainen, "Remote sensing of environment comparison of Sentinel-2 and Landsat 8 in the estimation of boreal forest canopy cover and leaf area index", Remote Sensing of Environment, vol. 195, pp. 259-274, 2017.

[13] M. Belgiu and O. Csillik, "Remote sensing of environment Sentinel-2 cropland mapping using pixel-based and object-based time- weighted dynamic time warping analysis", Remote Sensing of Environment, vol. 204, pp. 509-523, 2018.

[14] Q.Wang and P. Atkinson, "Spatio-temporal fusion for daily Sentinel-2 images", Remote Sensing of Environment, vol. 204, pp. 31-42, 2018.

[15] Q.Wang, W. Shi, Z. Li, and P. Atkinson, "Fusion of sentinel-2 images", Remote Sensing of Environment, vol. 187, pp. 241-252, 2016.

[16] P. Hawryo, B. Bednarz, P. Wyk, and M. Szostak, "Estimating defoliation of scots pine stands using machine learning methods and vegetation indices of Sentinel-2", European Journal of Remote Sensing, vol. 51, pp. 194-204, 2018.

[17] Y. Shao and R. Lunetta, "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points", ISPRS Journal of Photogrammetry and Remote Sensing, vol. 70, pp. 78-87, 2012.

[18] C. Man, T. Nguyen, H. Bui, and K. Lasko, "Improvement of land-cover classification over frequently cloud-covered areas using landsat 8 time-series composites and an ensemble of supervised classifiers",

International Journal of Remote Sensing, vol. 39, pp. 1243-1255, 2018.

[19] N.-T. Son, C.-F. Chen, C.-R. Chen, and V.-Q. Minh, "Assessment of Sentinel-1A data for rice crop classification using random forests and support vector machines", Geocarto International, vol. 6049, pp. 1-15, 2017.

[20] M. Li, L. Ma, T. Blaschke, L. Cheng, and D. Tiede, "A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments," International Journal of Applied Earth Observation and Geoinformation, vol. 49, pp. 87-98, 2016.