

Application of Machine Learning in Land Use Land Cover in Ranchi District, Jharkhand State of India

Swagata Ghosh

Ph.D. Scholar, Department of Computer Science & Engineering,
Sarala Birla University, Ranchi,
swagataghosh2k7@gmail.com

Dr. P. Paul

partha.paul@sbu.ac.in
Associate Professor, Department of Computer Science & Engineering,
Sarala Birla University, Ranchi.

Dr. Neeraj Kumar Sharma

sharmank@rediffmail.com
Scientist, Jharkhand Space Application Centre, Ranchi

Article Info

Page Number: 1849 – 1856

Publication Issue:

Vol 72 No. 1 (2023)

Abstract

The increasing population coupled with the modernization is affecting the urban areas. These lead to the depletion in the natural water bodies. The present paper is based on the study of application of machine learning technique to detect the differences in the extent of the land use land cover the past two years from the Sentinel 2A satellite data. The supervised method of machine learning the maximum likelihood classification is used to categorize various features in the images on the basis of colour bands. Finally, the principal component analysis provides the result for the classification process of a reference raster images to the trained image. The covariance matrix of the bands represent the variation of the data of obtained from different processes of classification.

Article History

Article Received: 15 October 2022

Revised: 24 November 2022

Accepted: 18 December 2022

Keywords: - Remote sensing data, Sentinel 2A, machine learning, maximum likelihood classification, random forest, principal component analysis algorithm.

1. Introduction

The natural water bodies get affected by modernization increasing population and other anthropogenic activities. The remote sensing data obtained from the satellites of past two years are used to study the change in the availability of the resources implementing the machine learning methods[1].The supervised method of machine learning learns from the surrounding behaviour of a system from a set of training data. The remote sensing images are the significant tools in the application of image classification [2]. The machine learning is used in the remote sensing data due to its voluminous size. The supervised classification is used for quantitative analysis of the remote sensing image data [3, 4].

Various computations on the raster images are applied with the statistical methods that provide a candid visualization of the data. The state of Jharkhand, popularly known for its lush green undulating terrains is affected by the urbanization activities. The remote sensing

data of November, 2022 collected to study the land use land cover classification of in the area.

2. Study Area

The remote sensing data in the raster form are acquired from sentinel 2A data provided by the European Satellite Agency (ESA) [5]. The study area is in the north east of Ranchi district, state capital of Jharkhand, in India. The images range from $23^{\circ}42'$ N to $23^{\circ}62'$ N latitude and $85^{\circ}26'$ E to $85^{\circ}70'$ E longitude. The images are geo referenced to WGS-84 datum and Universal Transverse Mercator Zone 35 North coordinate system. There are twelve bands in the Sentinel 2A imagery of 10 m, 20 m and 60 m resolutions [6]. The bands with their characteristics are represented (Table1).

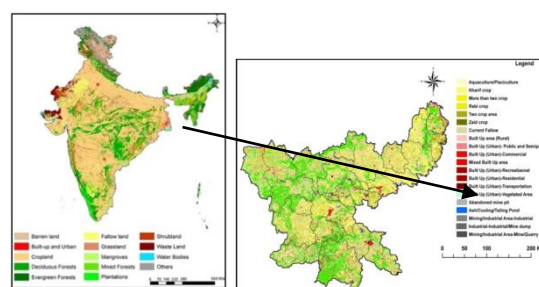


Fig. 1.a Map of India

Fig.1b. Map of Jharkhand state



Fig. 1 Map of the study area

3. Methods

The data of Sentinel 2A are obtained from European Satellite Agency (E.S.A) and with collective bands ranging from 10 meters, 20 meters and 60 meters. The open source software QGIS 3.28 version is used for this purpose. The Semi Automatic Classification plug in installed in the QGIS software is used to execute the features of Machine Learning algorithms. A band set is created with all the bands provided. The secondary colour bands for vegetations are chosen [7]. These wide range of colours help to distinguish the images as the various classes. The composite colour indices of the land use classes are verified from the different combinations of colour indices (Table I).

Data is pre-processed by taking a part of image by clipping the raster data. Atmospheric correction is done by the solar radiance as the raster layer of reflectance is obtained [8]. The classification method is a supervised machine learning methods which are applied on the previously trained categories as training input data to a given image. These are based upon a particular layer of a land form type [9]. One or more area is used to represent a particular

class. The pixels are categorized from the known region with their spatial extent as training area, as supervised. The multi-label image classifiers as water, vegetation, soil or a built up in the form of buildings, roads, query generate a pattern to date back to the previous form [10]. In virtual layer bands 2, 3 and 4 namely, blue, green and red change to the bands 1, 2 and 3 in the virtual raster layer. Total twelve separate bands with numbers for each bands is taken. Each band of the satellite file in tiff format is used for post processing of the images. The classification of raster bands is done to detect any anomalies present (Fig. 2).

Training areas are created by drawing polygons as a region of a particular type as a class is defined by the region of interest (ROI) and the signature. The non parametric rules are applied in the classification process. Supervised classification is easily quantified and frequently used in the study of remote sensing image data [11].

Bands	Colour	Wavelength (nm)
B2	Blue	490
B3	Green	560
B4	Red	665
B5	Red edge1	705
B6	Red edge2	749
B7	Red edge3	783
B8	Near Infra Red	842
B8A	Red edge4	865
B11	Shortwave Infra Red1	1610
B12	Shortwave Infra Red2	2190

Table 1

Every pixel in the image is assigned to cover the type to which its signature is most comparable. The nomenclature of the separate images is matched with the appropriate colours. The conversion of the raster images to that of the vector, confirms to that with the clear boundaries. The signature classes are created from the data. In case of supervised classification, fewer numbers of classes are formed.

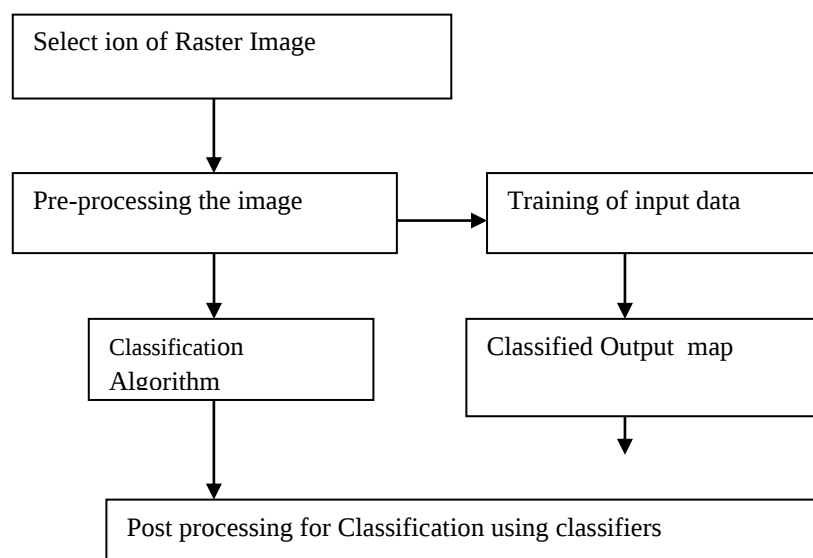


Fig.2. Flow chart of the method

3.1 Minimum Distance – The Minimum distance is calculated by the distance or feature space between the pixel and centre of all the pixels belong to each class is assigned to the closest one. The standard Euclidean distance is calculated.

3.2 Maximum Likelihood -Maximum likelihood estimation for Machine Learning is used to calculate the conditional probability of sample data from the observed probability distribution and the distribution parameters. First a guess is made, and then a predicted probability distribution for the data is made. This is achieved by maximizing a likelihood function. These are unbiased, consistent and efficient [12].

3.3 Random Forest -Random Forest classification algorithm uses a classifier that contains a number of decision trees from the subsets of the data sets used from the raster data. In the random forest, the average data is taken out from the entire data set which is used to improve the predictive accuracy of the datasets. The final output is obtained from each tree and based on the majority votes of predictions. The number of tree in this algorithm is increased to get a higher accurate output. The individual tree does not give an accurate prediction, have low correlation. But in cumulatively the prediction is more accurate. The output of the Random Forest for three classes and one thousand samples were used as training samples. For individual tree, the error rates ranged from 0.0232 in class 0.0, 0.0696 in class 1.0 and class 2.0 to 0.0928.

3.3 Principal Component Analysis- Principal Component Analysis method is used to reduce the dimensions of measured variables in the form of the raster bands. The individual components are uncorrelated and each component has variance less than the previous bands (Table 4.) The spectral bands are obtained by matrix calculation. The covariance matrices for the nine bands are correlated. The pattern analysis is followed from the bands [13, 14].

The eigenvectors and eigenvalues of covariance matrix are identified from two bands. Eigen values show the Principal Component Analysis, the direction of the data.

Results

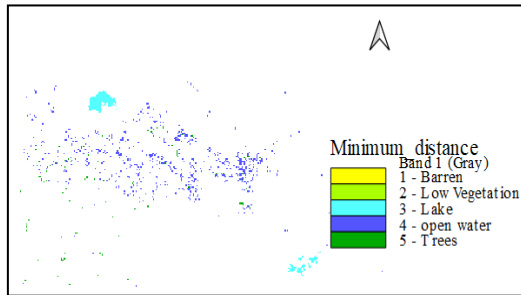
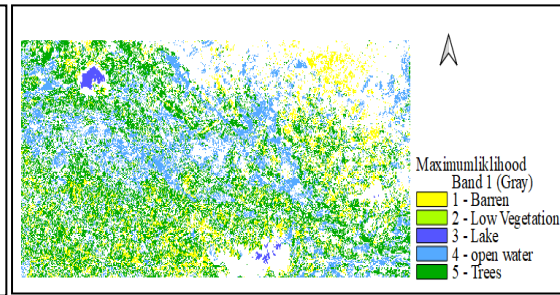
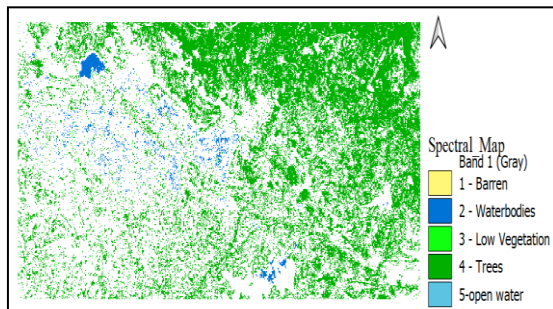
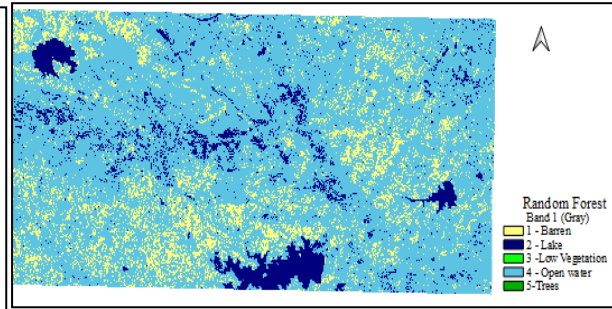
In all the classification process, in the raster, the pixel values correspond to each class id and each colour represent a particular land cover class.

The distance of the blue pixels appear as water bodies range from 0.28 for water bodies and for lake is 0.31 where as the vegetation is 0.39 to 0.43. Water bodies and the lakes are prominent (Fig.3).

The Maximum likelihood yield more barren land as well as the low vegetative covers and trees than the Minimum distance (Fig.4).

The Spectral angle mapping depicts the lake as the water body reflects it clearly. The lush vegetation is scattered with some error in classification (Fig. 5). The spectral angle for dam is 15 to 18 degrees and vegetation contains comparatively less 10 to 13 degrees.

Random forest classification output the water bodies, lake and the barren lands (Fig.6). The vegetation is least represented.

**Fig.3. Minimum distance****Fig.4. Maximum Likelihood****Fig.5. Spectral Angle Mapping****Fig.6. Random Forest**

Three classes class 0.0, with 97 percentage class 1.0 with 93 percentage, and class 3.0 with 90 percentage of accuracy with error rates 0.02, 0.06 and 0.09 respectively is obtained (Table. 2). There were 864 samples and the root mean square error is 0.69 and bias -0.009. Class 0.0 has a false negative of 5.0 class 1.0 14.0 and class 3.0 has 21.0.

```
RandomForest classifier scp20230122_212535856090

Cross Validation
Number of classes = 3
class 0.0: 20230122_212535570856324_temp
accuracy = 0.9768 precision = 0.8837 correlation = 0.8738 errorRate = 0.0232
TruePositives = 38.0000 FalsePositives = 5.0000 TrueNegatives = 383.0000 FalseNegatives = 5.0000
class 1.0: 20230122_21253574024599_temp
accuracy = 0.9304 precision = 0.9435 correlation = 0.8568 errorRate = 0.0696
TruePositives = 267.0000 FalsePositives = 16.0000 TrueNegatives = 134.0000 FalseNegatives = 14.0000
class 3.0: 20230122_21253582483116_temp
accuracy = 0.9072 precision = 0.8190 correlation = 0.7730 errorRate = 0.0928
TruePositives = 86.0000 FalsePositives = 19.0000 TrueNegatives = 305.0000 FalseNegatives = 21.0000

Using Testing dataset, % correct predictions = 90.7193
Total samples = 864
RMSE = 0.6980250562252973
Bias = -0.009280742459396807

Distribution:
class 0.0: 20230122_212535570856324_temp 86 (9.9537%)
class 1.0: 20230122_21253574024599_temp 563 (65.1620%)
class 3.0: 20230122_21253582483116_temp 215 (24.8843%)
```

Table 2. Output of the Random Forest Classification

Principal Component Analysis The covariance for the nine spectral bands with each band is the variance. The values in are in vector transpose. It ranges from 0.0002112342660427771 in band 1 to 1.053510148440174e-05 in band 9 along X vector (Table 3). Each component has variance less than the previous component.

The main diagonal in the correlation matrix is the cumulative of X and Y values of the matrix. It is 1.0 for the main diagonal. The positive values indicate that the variables are correlated.

The positive eigen values show the principal component and the direction of the data. Each spectral band produces a separate band. The statistical report produces mean, maximum, minimum, standard deviation and statistics valid percentage of the two bands.

Covariance matrix

Bands	1	2	3	4	5	6	7	8	9
1	0.00021	0.000 2373	0.000 343	0.0003 16	8.82E- 05	2.31E- 05	0.0004 7	0.0004 92	1.05E-05
2	0.00024	0.0003 273	0.000 445	0.0005 29	0.0003 98	0.0003 55	0.0007 4	0.0006 69	0.000379
3	0.00034	0.0004 452	0.000 715	0.0007 23	0.0003 26	0.0002 21	0.0011 51	0.0011 16	0.000234
4	0.00032	0.0005 289	0.000 723	0.0010 35	0.0011 08	0.0011 18	0.0015 24	0.0012 05	0.00124
5	8.82E- 05	0.0003 976	0.000 326	0.0011 08	0.0029 37	0.0033 91	0.0021 77	0.0010 96	0.003749
6	2.31E- 05	0.0003 552	0.000 221	0.0011 18	0.0033 91	0.0039 74	0.0023 32	0.0010 64	0.004397
7	0.00047	0.0007 396	0.001 151	0.0015 24	0.0021 77	0.0023 32	0.0034 28	0.0025 5	0.002607
8	0.00049	0.0006 685	0.001 116	0.0012 05	0.0010 96	0.0010 64	0.0025 5	0.0021 86	0.00118
9	1.05E- 05	0.0003 785	0.000 234	0.0012 4	0.0037 49	0.0043 97	0.0026 07	0.0011 8	0.004891

Correlation matrix

Bands	1	2	3	4	5	6	7	8	9
1	1	0.9023 922	0.882094	0.6765 59	0.1119 15	0.0252 55	0.5524 59	0.7247 59	0.010 364
2	0.90239	1	0.920366	0.9088 76	0.4055 77	0.3114 3	0.6982 16	0.7904 24	0.299 193
3	0.88209	0.9203 663	1	0.8402 18	0.2252	0.1309 66	0.7351 59	0.8926 26	0.125 069
4	0.67656	0.9088 756	0.840218	1	0.6357 47	0.5514 31	0.8091 6	0.8010 2	0.550 921
5	0.11192	0.4055 77	0.2252	0.6357 47	1	0.9924 02	0.6860 53	0.4325 98	0.989 01
6	0.02525	0.3114 296	0.130966	0.5514 31	0.9924 02	1	0.6318 47	0.3608 64	0.997 262
7	0.55246	0.6982 159	0.735159	0.8091 6	0.6860 53	0.6318 47	1	0.9315 63	0.636 574
8	0.72476	0.7904 241	0.892626	0.8010 2	0.4325 98	0.3608 64	0.9315 63	1	0.360 995
9	0.01036	0.2991 928	0.125069	0.5509 21	0.9890 1	0.9972 62	0.6365 74	0.3609 95	1

Eigen vectors		
Bands	Vector_1	Vector_2
1	-0.0327	0.173196
2	-0.0804	0.1830462

Eigen values	Accounted variance	Cumulative variance
0.015	75.281621	75.28162
0.004	21.503195	96.78482

Table. 3. Output of the values of Principal Component Analysis

4. Conclusion

The Machine Learning method increases the efficiency of the interpretation. It is useful in the analysis of raster data of voluminous size. The process of classification provides various land form types from the monochromatic raster data by the machine learning algorithm. In land cover classification, if a pixel is assigned to a wrong land cover class, because of the spectral similarity of the classes, leads to the discrepancy in the classification. A wrong class definition during the selection of region of interest also results into a wrong classification. The various Machine Learning algorithms used in the study of raster data produced different classification output. The accuracy assessment of the present raster layer with a standard layer is verified.

References

- [1] A.M. Abdi, "Land cover and Land use classification SVM performance of Machine Learning algorithms RF, in a boreal landscape using Sentinel-2 data", *GIScience & Remote Sensing*, 2020, vol. 57 no.4, pp 1-20.
- [2] S. S. Rwanga and J. Ndambuki, "Accuracy Assessment of Land use /Land cover classification using remote sensing and GIS" *International Journal of GeoSciences*, vol. 8, April, 2017, pp 611-622. <http://doi.org/10.4236/ijg.2017.84033>.
- [3] D. J. Lary, A. H Alavi, A. H. Gandomii, A.L. Walker, "Machine Learning in geosciences and remote sensing," *Geo Science Frontiers*, 2015, pp. 3–10. <http://doi.org/10.1016/j.gsf.2015.07.003>.
- [4] A Karpatne, I Ebert-Uphoff, S Ravela, H. A Babaie, V Kumar, "Machine learning for the geosciences: Challenges and opportunities", *IEEE Transactions on Knowledge and Data Engineering*. 2018 Jul, vol.31 no. 8, pp. 1544-54.<http://doi.10.1109/TKDE.2018.2861006>.
- [5] [Copernicus Open Access Hub. Available online <https://scihub.copernicus.eu>]
- [6] P. Yousili, H. A. Jalab et al., "Water body segmentation in satellite imagery applying modified kernel K-mean," *Malaysian Journal of Computer Science*, vol. 31, no.2, pp. 143-154, 2018.<http://doi.10.1016/j.tfp.2020.100018>

- [7] T.K. Thakur et. Al, “Land use land cover change detection through geospatial analysis in an Indian Biosphere Reserve, Trees, Forests & People”, 2020,<http://doi.org/10.1016/j.tfp.2020.10.100018>
- [8] M. Main-Knorn, B. Pflug, J. Louis, V Debaecker , U.M Wilm, F. Gason, “Sen2Cor for Sentinel-2”, Proc SPIE 10427, Image and Signal Processing for Remote Sensing XXIII,1042704, 2017.<https://doi.10.1117/12.2278218>.
- [9] A Roy,A.B.Inamdarl, “Multi-temporal landuse and land cover LULC change analysis of a dry and semi arid river basin in western India following a robust multisensory satellite image calibration strategy”, Heliyon, 2019.<http://doi.org/10.1016/j.heliyon.2019.e10478>.
- [10] X Yang, S. Zhao, X. Qin, N. Zhao, L Lian, “Mapping of urban surface water bodies from Sentinel-2 MSI imagery at 10m resolution via NDWI based image sharpening”, Remote Sensing, 2017, vol. 9, pp.596.
- [11] V.J. Berrocal, Y. Guan, A. Muyskens, H. Wang, B.J. Reich, J.A. Mulholland, H.H. Chang, “A comparison of statistical and machine learning methods for creating national daily maps of ambient PM_{2.5} concentration”, Atmospheric Environment, 2020 Feb 1: vol. 222:117130.<http://doi10.1016/j.atmosenv.2019.117130>.
- [12] G.S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, “DOTA: A large-scale dataset for object detection in aerial images”, Proceedings of the IEEE conference on computer vision and pattern recognition 2018, pp. 3974-3983.<http://doi.10.48550/arxiv.1711.10398>.
- [13] Q Xie, M Zhou, Q Zhao, D Meng, W Zuo, Z. Xu, “Multispectral and hyper spectral image fusion by MS/HS fusion net”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 1585-1594.
- [14] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, H. Huang, “Hyper spectral image super-resolution with optimized RGB guidance”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, pp. 11661-1167.