J. Ind. Geophys. Union (July 2017)
v.21, no.4, pp: 271-276

A Hybrid Prediction Algorithm using Naive Bayes Classifier for Improving Accuracy in Classifying LISS III Data

Kalyan Netti*[1] and Y.Radhika[2]

[1]CSIR-National Geophysical Research Institute, Uppal Road, Hyderabad
[2]Associate Professor, CSE Dept., GITAM University, Visakhapatnam
*Corresponding Author: nettikalyan@gmail.com

ABSTRACT

We propose a hybrid model for improving the accuracy in classifying LISS III Data using Naïve Bayes Classifier. The assumption of Conditional Independence among the predictors is one of the main reasons for loss of accuracy in Naïve Bayes Classifier. The effect of conditional independence on the accuracy varies on the data chosen for analysis. As there are cases where the predictor-outcome has become null, ignoring such results is not advisable as the outcome may affect the accuracy. In this paper, remote sensing data for Land use/Land cover is used as an input to the algorithm for classification. The Linear Imaging Self Scanning Sensor (LISS-III) data of Resourcesat-2 satellite has been used for this study. The Naive Bayes algorithm has been applied to the data, and the results are compared with the standard classification methods such as Maximum likelihood classifier and Mahalanobis classifier. The result of the study shows Naive Bayes classification performs better compared to conventional classifiers such as Maximum likelihood classifier and Mahalanobis classifier.

Key words: Hybrid prediction algorithm**,** Naive Bayes Classifier, Conditional Independence, Supervised Classification, Maximum likelihood classifier, Mahalanobis classifier.

INTRODUCTION

In Data Mining, classification is one of the best techniques available to predict outcomes in data sets. Naïve Bayes Classifier, a traditional supervised classification method, is one such classification techniques used to predict outcomes. In general, Naïve Bayes Classifier performs well when compared to other classifiers due to its simplicity, less computational complexity, less memory requirement and good prediction accuracy (Han et al., 2011; Wang et al., 2014). The better performance of Naive Bayes Classifier is attributed to the assumption of independence among predictors. This hypothesis sometimes leads to loss of accuracy in NBC. This loss can be more when data sets for classification has strong inter-relation among attributes. Thus, improving Naive Bayes classifier with the assumption of Independence among predictors is a challenging task (Wilson et al., 2009; Xi-Zhao et al., 2014). The primary goal of a Classifier is to predict the class value accurately for each instance in a given data set (Han et al., 2011; Haleem et al., 2014). In this paper the authors present a model using Naïve Bayes Classifier to estimate accuracy in LIS-III data for estimating accuracy of various factors associated with Forest cover, area specific water body dynamics, Wasteland status with time, Vegetation cover, Fallow and Built-up land particulars. The main aim of this paper is to show that the accuracy estimation using Naïve Bayes Classifier is better compared to Maximum Likelihood Classifier and Mahalanobis Classifier.

Many researchers have demonstrated that supervised classification is a proven technique for automatic generation of land cover maps (Richards, 1993; Benediktsson et al., 1990; Bruzzone et al., 1999; Bruzzone and Fernández Prieto 1999; Bruzzone, 2000). In supervised classification, analyst supervises the pixel categorization process by specifying the various land cover types present in a scene to the computer algorithm. Supervised classification procedures require substantial interaction with the analyst, who must guide the computer by identifying areas in the image that are known prior to the classification, which belongs to particular land use land cover classes. These areas are referred as training sites. The training sites are known identities, which are used to classify pixels of unknown. The locations of the training site pixels must stem from ground truth or higher quality maps or data sets. The computer uses the spectral characteristics of the training pixels to identify other unknown pixels in the image with similar characteristics (Richards, 1993). The quality of these training pixels decides mostly the success of supervised classification method. Parallelepiped, Maximum Likelihood, Minimum Distance and Mahalanobis Distance are the important classifier methods used widely (Khalid and Shakil, 2014; Zhu et al., 2006).

Maximum Likelihood classification assumes that Statistics for each Class belonging to each band is distributed normally and calculates the probability whether a given pixel belongs to a particular class or not. Each pixel is assigned to the class that has the highest probability. If the highest probability is less than a threshold, the pixel remains unclassified. (Bruzzone, 2000)

The Mahalanobis distance classification is a direction-sensitive distance classifier that uses statistics of each class. It is similar to the maximum likelihood classification. However,as it assumes that all class covariances are equal it is realistically classified as a faster method. All pixels are classified to the closest Region of Interest (ROI) class unless distance threshold is not specified.In such cases, some pixels may be unclassified as they do not meet the threshold.

Specific study particulars, results and conclusions are structured in the manuscript as per section wise details. Section-II explains Naïve Bayes Classifier, Section –III presents Implementation and Section-IV explains Data considered in this paper. Results are presented in Section-V and Section-VI offers conclusions.

Naive Bayes Classifier

Naïve Bayes Classifier, a supervised classification technique based on Bayes' Theorem, is used to predict the class from the attributes of a data set (Han et al., 2011; Dunham, 2006). Bayesian Classifier is stated at (1)

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{1}$$

Where, the posterior probability, *P(C/X)* (the probability of a attribute value X belonging to a class C), is calculated using Class Prior Probability *P(C)* (probability of class), Predictor Prior Probability *P(X)* (probability of attribute value) and Likelihood *P(X/C)* (probability of attribute value X given class C).

In Naïve Bayes Classifier, the assumption that the probabilities of each attribute with respect to a class are independent of all other attribute values is found to be apt. This assumption is made to basically simplify the calculation of probabilities. This assumption is called as Conditional Independence (Srisuan and Hanskunatai 2014; Domingos and Pazzani, 1996). It is explained, in this scenario as: the predictor (X) value of class (C) has little effect on the predictor's values of the other.This inturn leads to the loss of accuracy.

The proposed new method considers numerical attributes as input, and the values are Gaussian distributed. For Gaussian distribution mean and standard deviation need to be computed; the Gaussian distribution function is stated at (2).

$$P(X_j|C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} exp\left(-\frac{\left(X_j - \mu_{ji}\right)^2}{2\sigma_{ji}^2}\right) \tag{2}$$

μ_{ji} : Mean (average) of feature values x_j of examples for which $c{=}c_i$

σ_{ji} : Standard deviation of feature values x_j of examples for which $c{=}c_i$

Accuracy Assessment

The accuracy assessment of classification technique was carried out using confusion matrix. The confusion matrix was plotted with respect to the reference and predicted classes of three classifications. The confusion matrix compares the pixel classified by the classification method against the same site in the field. In confusion matrix, diagonals represent sites that are classified correctly according to the reference data. Off-diagonals are misclassified. The result of the confusion matrix provides the overall accuracy of the each class map. The overall accuracy of the classification is the ratio of the sum of correctly classified pixel to the total number of pixel (Overall Accuracy= No. of Correct Plots / Total No.of plots). However, the problem with overall accuracy is that it does not reveal if (any) error was evenly distributed among classes or not. For example,if some classes are null and others are excellent, the performance of a classification method is measured based on user and producer accuracy.

The producer's accuracy is derived by dividing the number of correct pixels in a particular class divided by the total number of pixels as derived from reference data. It means, the producer's accuracy measure for a given class in reference plots is based on how many pixels on the map are labeled correctly. It includes the error of omission, which refers to the proportion of observed features on the ground that is not classified in the map.

User accuracy measures for a given class are defined based on how many of the pixels on the map are rightly classified. The user's accuracy measures the commission error and indicates the probability that a pixel classified into a given category represents that group on ground.

Producer's Accuracy (%) = 100% - error of omission (%)
User's accuracy (%) = 100% - error of commission (%)

For a given class in reference plots, how many of the pixels on the map are labeled correctly

DATA

Study Area

The study area covers Nagpur and surrounding region (Maharashtra; India) centered at 21°22′N/78°59′E.The study area is one of the worst drought affected areas.

Resourcesat-2 LISS III

ResourceSat-2 carries three electro-optical cameras such as LISS-III, LISS-IV, and AWiFS. Resourcesat-2 provides continuity and increases the observation timeliness (repetitive) in tandem with ResourceSat-1. The Resourcesat data is very useful for agricultural crop discrimination and monitoring, crop acreage/yield estimation, precision farming, water resources, forest mapping, rural infrastructure development, disaster management. In the present study, we have used LISS III data. The Linear Imaging Self Scanning Sensor (LISS-III) is a multi-spectral camera operating in four spectral bands, three in the visible and near to infrared and one in the SWIR region, as in the case of IRS-1C/1D. LISS III sensor has the following configuration (Table 1) with resolution of 24m.

Table 1. LISS III sensor configuration.

Bands	Wavelength (μm)
B2 - Green	0.52 - 0.59
B3 - Red	0.62 - 0.68
B4 - NIR	0.77 - 0.86
B5 - SWIR	1.55 - 1.70

Conversion of Radiance from Digital number

The image acquired is recorded as digital numbers. To convert back to the original object reflectance values, the DN values are processed using Equation-1. It needs the maximum and minimum radiance value for each band, which is unique for each sensor. This information is provided with the header file of the image.

$$L_{rad} = (DN / MaxGray) * (L_{max} - L_{min}) + L_{min} \qquad (3)$$

L_{rad}: Radiance for a given DN value (Table 2), DN: Digital count, MaxGray: 255
L_{min} / L_{max}: Minimum/ Maximum radiance value for a given band available in the header file of the image

Table 2. Conversion table for DN to radiance for LISS-III.

Satellite Image	Band	L_{min}	L_{max}
IRS – 1C	band1 band1 band1	1.76 1.54 1.09	14.4500 17.0300 17.1900

RESULTS AND DISCUSSION

The results of the three classifiers, the proposed model using Naive Bayes, MXL, Mahanoblis are shown in Tables 3, 4, 5. Figure 1 shows the data used for the study i.e., Resoursesat-2 LISS III & figures 2, 3, 4 show the output when classified using Naive Bayes, MXL and Mahanoblis. The proposed model using Naive Bayes classifier produces output with the user accuracy of 0.91,1.0, 0.80, 0.81, 0.82, 0.77 for Forest,Water body, Wasteland, Vegetation,

Table 3. Confusion matrix of Naive Bayes classification.

Naive Bayes	Forest	Water body	Waste land	Vegetation	Fallow	Built-up	Total	User accuracy
Forest	1217	0	0	106	0	2	1325	0.918
Water body	0	169	0	0	0	0	169	1
Wasteland	0	0	179	23	0	21	223	0.802
Vegetation	314	0	0	1452	7	6	1779	0.816
Fallow	1	2	3	34	640	97	777	0.823
Built-up	0	0	25	2	153	612	792	0.772
	1532	171	207	1617	800	738	5065	
Producer accuracy	0.794	0.988	0.864	0.897	0.8	0.82		

Table 4. Confusion matrix of MXL Classification.

MXL	Forest	Water body	Waste land	Vegetation	Fallow	Built-up	Total	User accuracy
Forest	1169	0	0	263	0	0	1432	0.816
Water body	0	171	0	0	5	8	184	0.929
Wasteland	20	0	168	51	3	21	263	0.638
Vegetation	341	0	0	1194	3	3	1541	0.774
Fallow	2	0	2	104	625	67	800	0.781
Built-up	0	0	37	5	164	639	845	0.756
	1532	171	207	1617	800	738	5065	
Producer accuracy	0.763	1	0.811	0.738	0.781	0.86		

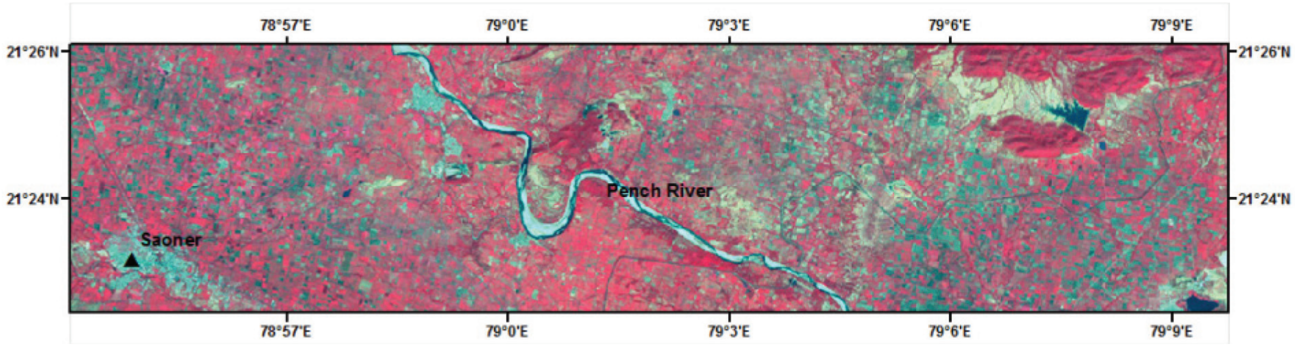

Figure 1. Resoursesat-2 LISS III data.

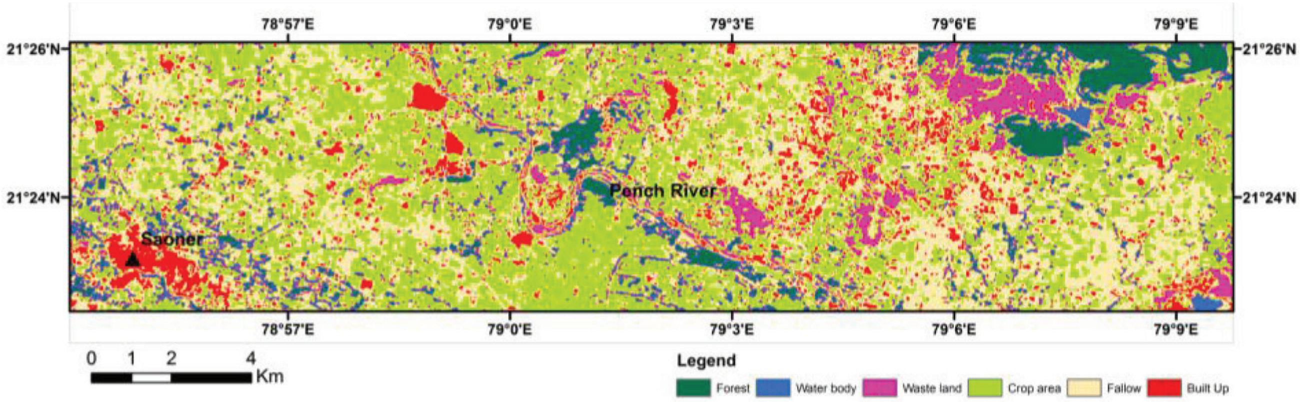

Figure 2. Propsed model using Naïve Bayes Classifier Output.

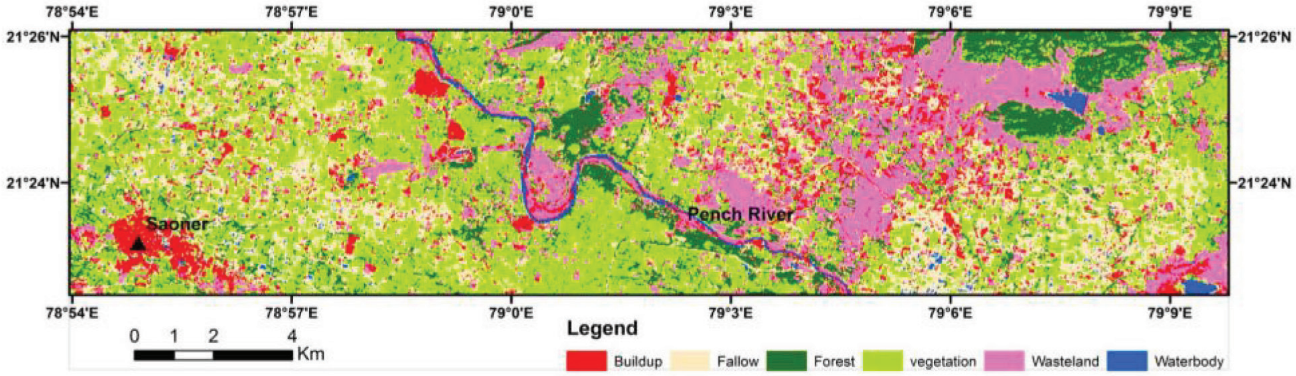

Figure 3. MXL classification Output.

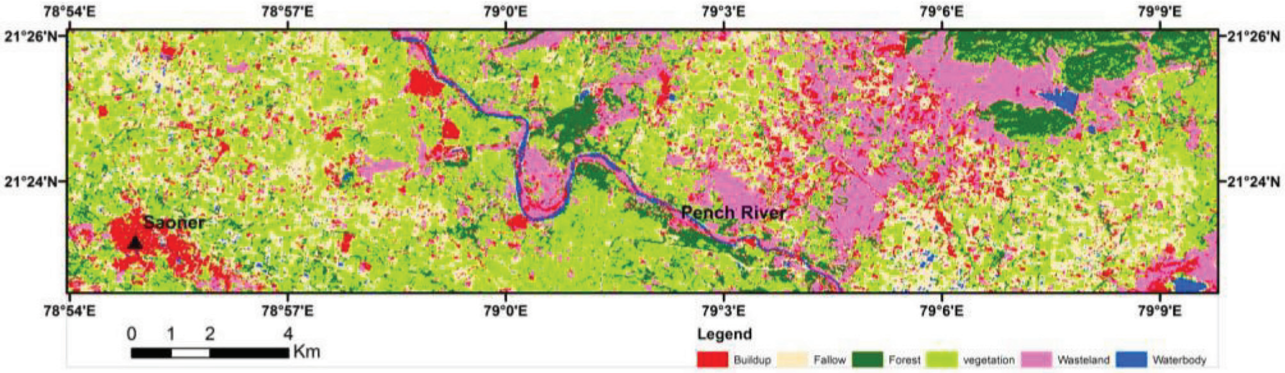

Figure 4. Mahalanobis classification Output.

Table 5. Confusion matrix of Mahanoblis Classification.

Mahanoblis	Forest	Water body	Waste land	Vegetation	Fallow	Built-up	Total	User accuracy
Forest	1169	0	0	250	0	0	1419	0.823
Water body	0	171	0	0	5	6	182	0.939
Wasteland	21	0	166	51	3	19	260	0.638
Vegetation	340	0	0	1198	7	3	1548	0.773
Fallow	2	0	2	113	624	62	803	0.777
Built-up	0	0	39	5	161	648	853	0.759
	1532	171	207	1617	800	738	5065	
Producer accuracy	0.76	1	0.801	0.740	0.78	0.87		

Table 6. User accuracy comparison.

	Proposed Algorithm using NBC	Mahanoblis	MXL
Forest	0.918	0.823	0.816
Water body	1	0.939	0.929
Wasteland	0.802	0.638	0.638
Vegetation	0.816	0.773	0.774
Fallow	0.823	0.777	0.781
Built-up	0.772	0.759	0.756

Fallow and Built-up area, respectively. For the same class, producer accuracy of Naive Bayes is 0.79, 0.98, 0.86, 0.89, 0.80 & 0.82 (Table 3). The user and producer accuracy of MXL classifier are 0.81, 0.92, 0.63, 0.77, 0.78, 0.75 & 0.76, 1.0, 0.81, 0.73, 0.78, 0.86 for Forest, Water body, Wasteland, Vegetation, Fallow, Built-up area, respectively (Table 4). The user and producer accuracy of Mahanoblis are almost same as the MXL (Table 5). The comparison of the three classifiers shows (Table 6) that the accuracy of the proposed algorithm is better when compared with the accuracies of MXL and Mahanoblis (Table 6). For example the user accuracy of Wasteland is 0.802 using the proposed algorithm, whereas the user accuracy is 0.638 for MXL and Mahanoblis. The accuracies for Forest, Water body, Fallow, Vegetation, Built-up area are also better for the proposed algorithm (Table 6). The results indicates that the accuracy has significantly improved by using NBC even with the assumption of Conditional Independence.

CONCLUSION

One of the reasons for loss of accuracy in Naïve Bayes Classifier is Conditional Independence. The Hybrid model proposed in this paper has better accuracy with the assumption of Conditional Independence when applied on LISIII data. The results show that the method used in this article has the highest prediction accuracy when compared with standard classification methods such as Maximum likelihood classifier, Mahalanobis classifier. The conclusion, based on the experimental results is that accuracy of Naïve Bayes classifier can be improved even with the assumption of Conditional Independence. Based on the experimental results, we can assert that addressing the loss of accuracy in Naïve Bayes Classifier due to Conditional Independence proved advantageous for better analysis of data.The proposed model has improved the accuracy when applied on a complex data with more number of attributes in a given data. Our future work includes observing the impact of the complex data on the performance of the proposed model and exploring the possibility of cleaning the data before applying Naïve Bayes Classifier for further improvement of accuracy.

ACKNOWLEDGEMENTS

Authors want to sincerely thank Director, CSIR-NGRI for permitting them to carry out this study and allowing its publication in JIGU. Our thanks are due to Prof.A.D.Sarma for critically evaluating the manuscript and for useful suggestions. We are grateful to Chief editor for his support to our research endeavours and for apt editing.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest and adhere to copyright norms.

REFERENCES

Benediktsson, J.A., Swain, P.H., and Ersoy, O.K., 1990. Neural Networks:Approaches versus statistical methods in classification of multisource remote sensing data,IEEE Trans. Geosci. Remote Sensing, v.28, pp: 540-552.

Bruzzone, L., 2000. An approach to feature selection and classification of remote sensing images based on the Bayes rule for minimum cost. IEEE Trans. Geosci. Remote Sensing, v.38, pp: 429-438.

Bruzzone, L., and Fernández Prieto, D., 1999. A Technique for the selection of kernel function parameters in RBF neural networks for classification of remote-sensing images, IEEE Trans. Geosci. Remote Sensing, v.37, pp: 1179-1184.

Bruzzone, L., Fernández Prieto, D., and Serpico, S.B., 1999. A neural statistical approach to multi-temporal and multisource remotesensing image classification. IEEE Trans. Geosci. Remote Sensing, v.37, pp: 1350-1359.

Domingos, P., and Pazzani, M., 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Proceedings of International Conference of Machine Learning, pp: 105-112.

Dunham, M.H., 2006. Data Mining: Introductory and advanced topics. Pearson Education India.

Haleem, H., Sharma, P.K., and Sufyan Beg, M.M., 2014. Novel frequent sequential patterns based probabilistic model for effective classification of web documents. In Computer and Communication Technology (ICCCT), 2014 International Conference on IEEE., pp: 361-371.

Han, J., Kamber, M., and Pei, J., 2011. Data mining: concepts and techniques. Elsevier.

Khalid Omar Murtaza, and Shakil A. Romshoo, 2014. Determining the Suitability and Accuracy of Various statistical Algorithms for Satellite Data Classification. International journal of geomatics and geosciences, pp: 585-599.

Richards, J.A., 1993. Remote Sensing Digital Image Analysis, 2nd ed. New York: Springer-Verlag.

Srisuan, J., and Hanskunatai, A., 2014. The Ensemble of Naïve Bayes classifiers for hotel searching. In Computer Science and Engineering Conference (ICSEC), 2014 International, IEEE., pp: 168-173.

Wang, X.Z., He, Y.L., and Wang, D.D., 2014. Non-naive Bayesian classifiers for classification problems with continuous attributes.Cybernetics, IEEE Transactions on, v.44, no.1, pp: 21-39.

Wilson, M.L., 2009. Exploring heterogeneous datasets from different searcher perspectives.

Xi-Zhao Wang,Yu-Lin He and Debby D.Wang., 2014. Non-Naïve Bayesian Classifiers for Classification Problems with Continous Attributes. IEEE Transactions on Cybernetics, v.44.

Zhu, G.B., Liu, X.L., and Jia, Z.G., 2006. A multi-resolution hierarchy classification study compared with conservative methods, ISPRS Workshop on Multiple representation and interoperability of spatial data, pp: 79-84.

Received on: 9.1.17; Revised on: 10.2.17; Re-revised on: 24.3.17; Accepted on: 26.3.17

*“Twenty years from now,
You will be more disappointed
By the things you didn’t do,
Than by the ones you did
So
Throw off the bowlines,
Sail away from the safe harbour,
Catch the tradewinds in your sails,
DREAM
EXPLORE
DISCOVER”
Scott J. Fitzgerald – one of the greatest American writers of the 20th century.