

# A Visual mining based framework for classification accuracy estimation

**Pattathal Vijayakumar Arun**

National Institute of Technology  
Madhya Pradesh, India  
e-mail: arunpv2601@gmail.com

Received: 17 April 2013 / Accepted: 08 July 2013

**Abstract:** Classification techniques have been widely used in different remote sensing applications and correct classification of mixed pixels is a tedious task. Traditional approaches adopt various statistical parameters, however does not facilitate effective visualisation. Data mining tools are proving very helpful in the classification process. We propose a visual mining based frame work for accuracy assessment of classification techniques using open source tools such as WEKA and PREFUSE. These tools in integration can provide an efficient approach for getting information about improvements in the classification accuracy and helps in refining training data set. We have illustrated framework for investigating the effects of various resampling methods on classification accuracy and found that bilinear (BL) is best suited for preserving radiometric characteristics. We have also investigated the optimal number of folds required for effective analysis of LISS-IV images.

**Keywords:** Data mining, Remote sensing, Decision tree, Image classification, Visualization, WEKA, PREFUSE

---

## 1. Introduction

Remote sensing techniques are used extensively in land use- land cover analyses in a way that will greatly improve our predictive capability. The increased global availability of spatial data has enhanced the applicability of semiautomatic pre-processing techniques. Classification, which involves assigning of pixels to respective classes, is the most commonly used operation for extracting land use information (Zhang et al, 2009). The accuracy of pixel based classification approaches are affected by the increase in resolution of images and literature suggests a great deal of advanced methodologies for the purpose (Vapnik et al., 1998). Supervised classification techniques have improved accuracy when compared to its unsupervised counterparts and involves the use of training sets to classify the actual dataset (Nghi Dang Huu and Chi-Mai Luong, 2008). The efficiency of supervised classification approaches depend on the training samples hence it is important to measure the effectiveness of training samples.

Effective analysis of signature separability and mixed pixel effects are required for accurate classification. Literature reveals many statistical measures in this regard among which error matrix is the most commonly used strategy. It expresses several characteristics about classification accuracy like omission (exclusion) and commission (inclusion) errors, over all accuracy etc (Lillesand et al, 2004). However these methods do not help to visualize the data and also do not reveal the problematic classes and its pixel wise measures effectively. In recent years, despite continuous inventions in establishing and testing new classification methods, classification performance is still not showing demonstrable improvement (Witten and Frank, 2005; Wilkinson, 2005). Thus, visualization of classification process seems to be the only way to minimize misclassification effects and visual data mining tools are found to be effective in this context (Lu and Weng, 2007).

In this paper we investigate the integration of visual data mining techniques with classification processes to identify signature separability and mixed pixel problems of training sets. The suggested methodology helps to explore classification steps leading to misclassifications and facilitate effective visualisation. It also facilitates to assess local changes to the classification process and helps to mould the classifier to a particular image dataset. The framework has been implemented in java using WEKA and PREFUSE interfaces. We have used the strategy to investigate the effects of resampling techniques on classification accuracy.

## **2. Experiment**

In this work, the WEKA (Bouckaert, 2010), an open source data mining package, and the PREFUSE, a tree visualization package (Keim et al, 2003; Durbha, 2005) plug-in, have been used for investigations. The LISS-IV image of Bhopal city in India, acquired in the month of November 2011 was used for the analyses. The FCC of this image is shown in Figure 1(a). The investigations have been done for the area near MANIT campus located within latitude: 23° 07 to 23° 54 N, longitude: 77° 12 to 77° 40 E and is shown in Figure 1(b).

## **3. Research Methods**

Framework for effective analysis and visualisation of training sets has been developed in Java using WEKA and PREFUSE. WEKA has been used for analysing classification results and PREFUSE for effective visualisation. The subset images are split in to three constituent bands in TIFF format and pixel values are exported to WEKA compatible (.ARFF) format. PREFUSE visualization tree plug-in is integrated with WEKA and the size of JVM is adjusted automatically to handle large datasets. The J48 decision tree classifier of PREFUSE tool is adopted for analysis as it provides effective visualisation. The data stored in .ARFF format is processed using WEKA to

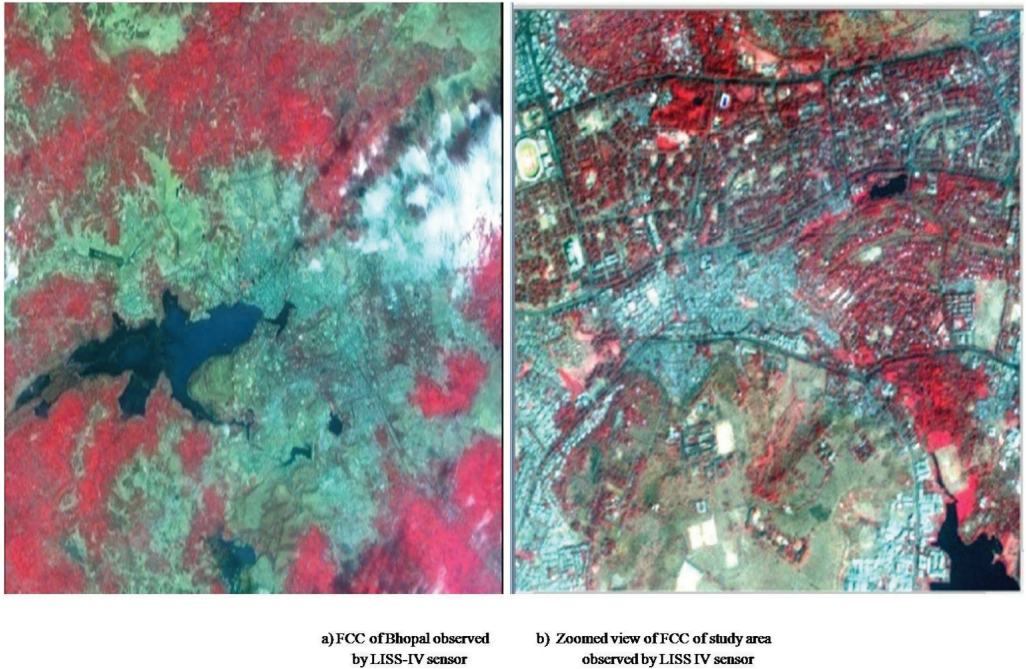


Fig. 1. Study area

obtain different classification accuracy parameters such as confusion matrix, producer accuracy, over all accuracy etc. The PREFUSE interface is then used to generate decision tree and incorrectly classified pixels are highlighted to provide effective visualisation.

The frame work has been used to investigate the effect of different resampling techniques on classification accuracies. Ten different land use classes namely built-up, water, vegetation, agriculture, grass, wetland, playground, swimming-pool, road, and unclassified were decided on the basis of site visit. The LISS-IV image subset of study area was classified in to these land use classes using supervised and unsupervised classifiers without georeferencing the image. Effect of different resampling techniques such Nearest Neighbour (NN), Bi Linear (BL) and Cubic Convolution (CC) methods on the classification accuracy was investigated. Statistical parameters as well as decision tree of all datasets were compared for the investigation. Schematic representation of the methodology adopted is summarized in Figure 2.

#### 4. Results and Discussions

Proposed frame work has been used to investigate the effect of resampling techniques on classification accuracy. The unregistered image False Color Composite (FCC) and its constituent bands used for the classifier training are shown in Figure 3(a) to (d).

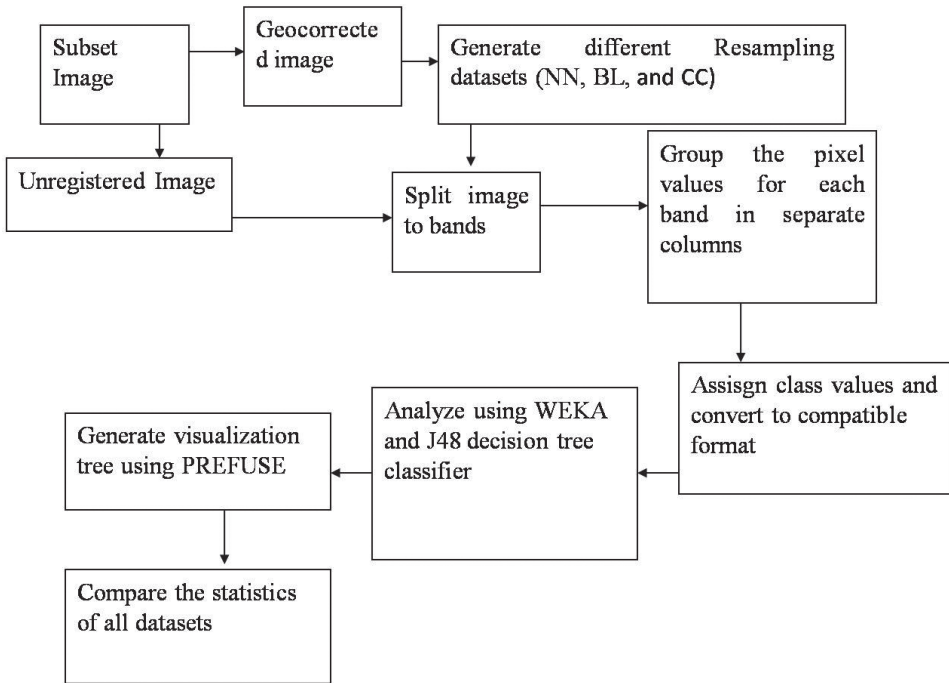


Fig. 2. Methodology flow chart

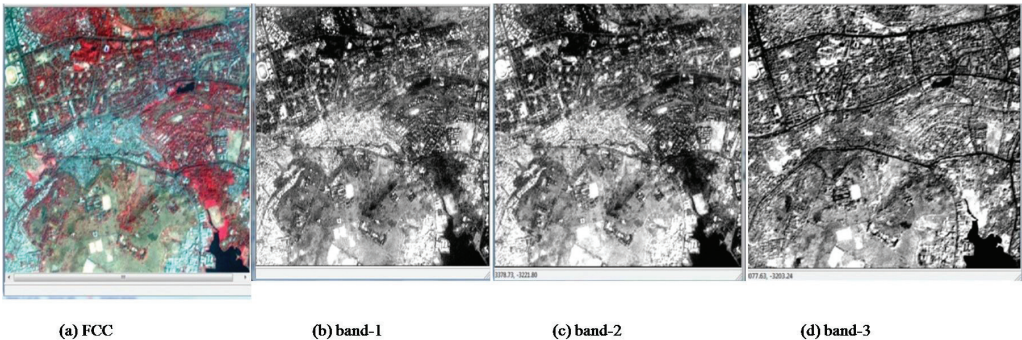
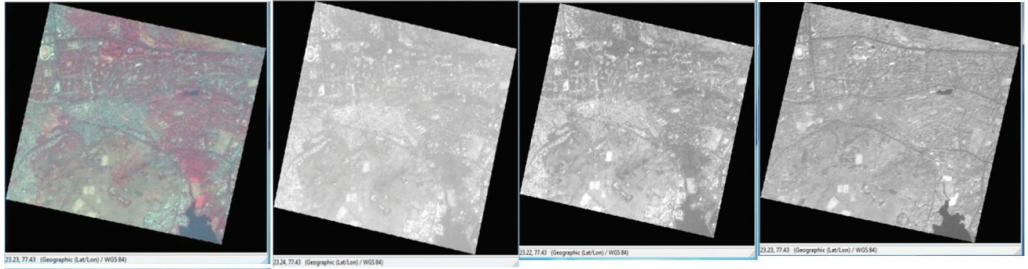


Fig. 3. Geo-coded image data set using NN resampling

The visualization facilitated by decision trees provides a better understanding of the classification context and help to better identify misclassifications. It provides effective evaluation of signature separability and mixed pixels. Tree also provides an idea of the band that better distinguishes specific classes. Figures 4 and 5 respectively show path and zooming facilities of PREFUSE tree visualizer.



(a) FCC

(b) band-1

(c) band-2

(d) band-3

Fig. 4. PREFUSE tree visualizer showing path

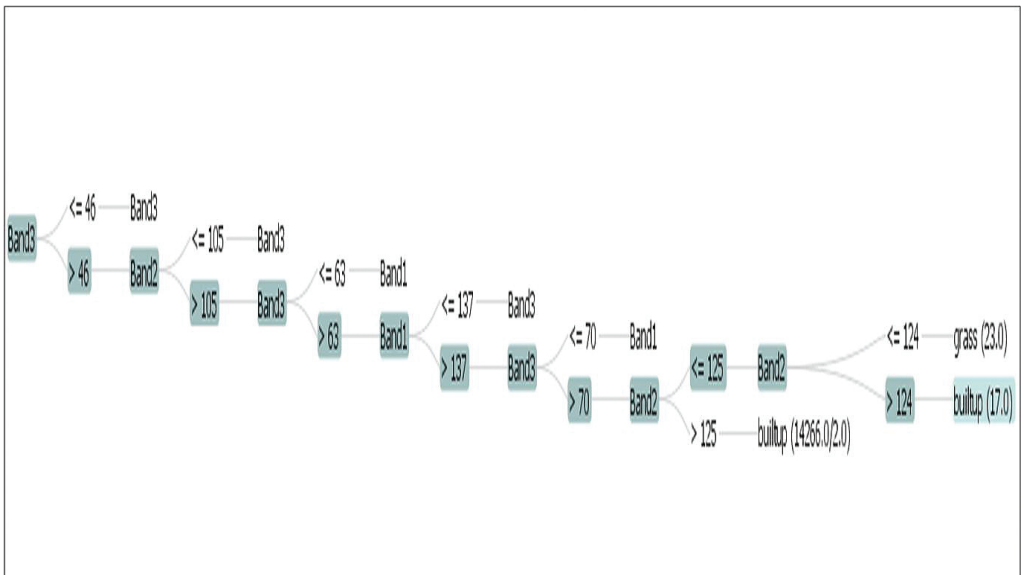


Fig. 5. PREFUSE tree visualizer showing facility of zooming

Investigation revealed that the geocorrection process affects classification accuracy. Different statistical parameters shows highest accuracy in the case of BL method resampled image dataset as compared the others (NN and CC). Kappa statistic measures the agreement of prediction with the true class, and a kappa value of unity signifies complete agreement. Kappa statistic also shows highest value for BL datasets. Accuracy of different approaches in terms of statistical parameters from WEKA software has been summarized in Table 1.

Table 1. Comparison among results of all the four datasets

Parameters	Unregistered		NN		BL		CC	
	No:	%	No:	%	No:	%	No:	%
Correctly classified instances	401364	94.9731	414857	97.981	417366	97.9957	418471	96.9673
Incorrectly classified instances	108	0.0269	79	0.019	18	0.0043	137	0.0327
Total no: of instances	401472		414936		417384		418608	
Kappa statistic	0.9796		0.9897		0.9799		0.9796	
Mean absolute error	0.0001		0		0		0.0001	
Root mean square error	0.007		0.006		0.0028		0.0079	
Relative absolute error	0.0517%		0.0335%		0.0086%		0.0635%	
Root relative squared error	2.6916%		2.2111%		1.1207%		2.8793%	
Coverage of Cases (0.95 level)	99.9786%		99.9853%		99.9964%		99.9742%	
Mean relative region size (0.95 level)	10.0019%		10.0025%		10.0004%		10.0029%	

The results have shown that resampling processes change the pixel value and classification statistics considerably. It is advisable that geo-coding of image should be done after classification. Bi linear method can be adopted if the geocoding is essential before classification. The NN resampling method which was expected to perform better has been out performed by BL. This is because NN resampling method when applied to a scene with abrupt radiometric variation and narrow features introduces radiometric more alternations compared to BL. CC as expected smoothens the image and hence severely affect the pixel value and hence classification accuracy. Effective analysis has to be optimized by selecting appropriate number of folds for the given dataset (we took unregistered dataset), without sacrificing the processing accuracy. Different performance parameters as given in Table 2 were evaluated for 2 to 10 folds and it has been observed that 5 folds are appropriate for the dataset used in this investigation.

Table 2. Parameters of cross validation for 2 to 10 folds

Cross validation Parameter	2 folds		3 folds		4 folds		5 folds		6 folds		7 folds		8 folds		9 folds		10 folds	
	No:	%	No:	%	No:	%	No:	%	No:	%	No:	%	No:	%	No:	%	No:	%
Correctly classified instances	401305	97.96	401337	97.96	401349	97.97	401356	97.97	401354	97.97	401356	97.97	401363	97.97	401362	97.97	401364	97.97
Incorrectly classified instances	167	0.04	135	0.03	123	0.03	116	0.03	118	0.03	116	0.03	109	0.03	110	0.03	108	0.03
Total no: of instances	401472		401472		401472		401472		401472		401472		401472		401472		401472	
Kappa coefficient	0.9894		0.9895		0.9776		0.9786		0.9791		0.9790		0.9599		0.9721		0.9796	
Mean absolute error	0.0001		0.0001		0.0001		0.0001		0.0001		0.0001		0.0001		0.0001		0.0001	
Root mean square error	0.0087		0.0080		0.0075		0.0074		0.0074		0.0073		0.0071		0.0071		0.007	
Relative absolute error	0.08		0.07		0.06		0.06		0.06		0.06		0.05		0.05		0.05	
Root relative squared error	3.33		3.05		2.88		2.82		2.83		2.80		2.70		2.70		2.69	
Coverage of Cases (0.95 level)	99.97%		99.97		99.97		99.98		99.98		99.98		99.98		99.98		99.98	
Mean relative region size (0.95 level)	10.00%		10.00%		10.00%		10.00%		10.00%		10.00%		10.00%		10.00%		10.00%	

## 5. Conclusion

We have discussed an efficient framework for analyzing the classification accuracy of training data sets. Proposed approach integrates image mining techniques with classification for effective visualization using WEKA and PREFUSE open source tools. Framework has been adopted to investigate the effects of resampling on classification which revealed that geo-coding drastically changes the pixel radiometry and this may affect the accuracy. Results suggest that it is not advisable to perform the image geo-coding operations before classification. However BL resampling technique can be adopted where prior geo coding is critical. The efficiency of analysis using the suggested framework depends on the selection of appropriate number of folds for a given dataset. Number of fold should be optimally chosen without sacrificing the processing accuracy.

## 6. Acknowledgement

This research work has been carried out at Maulana Azad National Institute of Technology-Bhopal, India. Authors wish to acknowledge Dr. Appukuttan K K along with other faculties of the institute for their helpful discussions and support towards this research work.

## References

- Bouckaert, R. R., Frank E., Hall M. A., Holmes G., Pfahringer B., Reutemann P. & Witten I. H. (2010). WEKA—Experiences with a Java Open-Source Project. *Journal of Machine Learning Research*, vol.11, 2533-2541.
- Durbha, S. S. & King R. L. (2005). Semantics-enabled framework for knowledge discovery from Earth observation data archives. *IEEE Transaction on Geoscience and Remote Sensing*, vol. 43(11), 2563-2572.
- Keim, D. A., Panse C. & Sips M. (2003). *PixelMaps: A New Visual Data Mining Approach for Analyzing Large Spatial Data Sets*. In Proceedings of Third IEEE International Conference on Data Mining (ICDM'03), 19-22 November, 2003, IEEE CS Press, (pp. 565-568). Melbourne, Florida, USA: Institute of Electrical and Electronics Engineers (IEEE). DOI: [ieeecomputersociety.org/10.1109/ICDM.2003.1250978](https://doi.org/10.1109/ICDM.2003.1250978).
- Lillesand, T. M, Kiefer R. W & Chipman J. W. (2004). *Remote Sensing and Image Interpretation*. Singapore: John Wiley & Sons.
- Liu, Y. & Salvendy G. (2007). Design and evaluation of visualization support to facilitate decision trees classification. *International Journal of Human- Computer Studies*, vol. 65(2), 95-110. DOI: [10.1016/j.ijhcs.2006.07.005](https://doi.org/10.1016/j.ijhcs.2006.07.005).
- Lu, D. & Weng Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, vol. 28(5), 823-870.
- Nghi, Dang Huu & Chi – Mai Luong. (2008). An object-oriented classification techniques for high resolution satellite imagery. *GeoInformatics for Spatial-Infrastructure Development in Earth and Allied Sciences (GIS-IDEAS)*, 230-240.



- Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley Publishers Inc.
- Wilkinson, G. G. (2005). Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing* vol. 43, 433-440.
- Witten, I. H. & Frank E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann. 120-134.
- Zhang, J., Gruenwald L. & Gertz M. (2009). VDM-RS: A visual data mining system for exploring and classifying remotely sensed images. *Journal of Computers and Geosciences*, 1188-1192.

## Podstawy wizualnej eksploracji do szacowania dokładności klasyfikacji

Pattathal Vijayakumar Arun

National Institute of Technology  
Madhya Pradesh, India  
e-mail: arunpv2601@gmail.com

### Streszczenie

Techniki klasyfikacji są szeroko wykorzystywane w różnych aplikacjach teledetekcyjnych, w których poprawna klasyfikacja pikseli stanowi poważne wyzwanie. Podejście tradycyjne wykorzystujące różnego rodzaju parametry statystyczne nie zapewnia efektywnej wizualizacji. Wielce obiecujące wydaje się zastosowanie do klasyfikacji narzędzi do eksploracji danych. W artykule zaproponowano podejście bazujące na wizualnej analizie eksploracyjnej, wykorzystujące takie narzędzia typu open source jak WEKA i PREFUSE. Wymienione narzędzia ułatwiają korektę pól treningowych i efektywnie wspomagają poprawę dokładności klasyfikacji. Działanie metody sprawdzono wykorzystując wpływ różnych metod resampling na zachowanie dokładności radiometrycznej i uzyskując najlepsze wyniki dla metody bilinearnej (BL).