Land Cover Classification Using Supervised and Unsupervised Learning Techniques

Rahul Nijhawan,
Department of Earthquake Engineering, Indian Institute of Technology Roorkee,
rahulnijhawan2010@gmail.com,

Ishita Srivastava,
College of Engineering Roorkee, Roorkee,
srivastava.ishita@yahoo.com

Pushkar Shukla,
College of Engineering Roorkee, Roorkee,
pushkar_shukla@yahoo.com

***Abstract*— The aim of this study is to propose a suitable methodology for accurate Land cover classification in the Joshimath district, India. The study proposes K-mean clustering algorithm approach for accurate mapping of land cover. We tried several combinations of parameters and opted for the one which gave the best classification results. The results were also compared with current state of art machine learning algorithms, artificial neural network, maximum likelihood classifier and iso-cluster algorithm. Accuracy assessment was performed by means of confusion matrix. It was observed that the proposed approach gave the highest classification accuracy (93.5%) with value of kappa coefficient 0.91. While the lowest accuracy (77.8%) was achieved by iso-cluster algorithm.**

***Keywords*— *Machine learning, artificial neural network, Iso-cluster algorithm, K-mean clustering algorithm, Maximum-likelihood classifier, supervised learning, Unsupervised learning*.**

I. INTRODUCTION

Over the past few years, land cover is found to have close association with our environments and ecosystems. Land cover is the physical and biological cover of earth. Because of its robustness and cost-effectiveness the technique of remote sensing is widely used in land cover classification [2].In the process of classification ,by judging the character of the pixels, labels are attached to it. On the basis of this character the pixels show different behavior in spectral ranges. Pixels are grouped in image classification techniques to represent land cover features. In the past years, a number of methods have been developed for image classification of remotely sensed data. In the recent years ground surveying has increased and so data volumes have increased which have improved the rate of remote sensing in the past years [2]. In image classification, there are two types of techniques-supervised learning and unsupervised learning.

A. Supervised Learning

In Supervised classification techniques the analyst has the task of deciding the training areas so that the characteristics and properties of each category can be determined. Discriminating information is extracted which is used to assign categories to each pixel in the image [1]. When the user has known outcomes that the data samples are to represent, supervised learning turns out to be more useful. The analyst plays an important role in supervised learning as the analyst locates training sites which represent the spectral characteristics of these known areas and are used to train the classifier. The analyst defines the training data in such a way that it has close similarity with the classes and features used during the analysis of the classifier. The properties of each individual class are obtained from the extracted training data. Supervised classification is performed in two stages-training phase and testing phase. In the training phase, the training of the classifier is done, and in the testing phase, the classifier's performance is tested on unknown pixels. In the training stage, the analyst has the task of defining regions; these regions are then used to extract information from the training data. This information helps in estimating the data properties of the classification task [3]. After the classification stage, labelling of unknown pixel is done on the basis of their spectral similarity in the training data. An unknown class is assigned to the pixels which do not show spectral similarity with any class. In the result produced, every pixel of the image gets label of a class. The performance of the classifier depends largely on the properties of the training data which were decided by the analyst [1].

B. Unsupervised Learning

In unsupervised classification, assessment of relative positions of pixels in the image helps in searching for clusters present within the data. Each cluster is assumed to represent unique characteristics. Reflectance property of pixels is used to form these clusters [3]. It is the user's task to identify the number of clusters to be generated and bands to be chosen. On the basis of this information, the software used in image classification forms clusters. Based on the classes formed, the user identifies the classes manually. If multiple classes result in representing the same class, it is the user's task to merge such classes together. Thus unsupervised learning plays a major role when the subdivisions into which the data samples should be divided is unknown to the user.In such cases, unsupervised learning is used to generate distinct classes also called spectrally in-separate classes also known as clusters. The variance between and within the clusters is computed to obtain these clusters [4]. According to the land cover features present in the classification task, labelling of pixels is done after determination of clusters. Thus, the image is classified.

978-1-5090-5595-1/17/$31.00 ©2017 IEEE

II. STUDY AREA

In the Chamoli district of Uttarakhand (India), lies the city of Joshimath.It is located at a height of 6150 feet.The Latitude and Longitude of Joshimath is 30.55 and 79.55 respectively. Joshimath has a very rugged topography and formidable physical feature. It lies in the Nandadevi biosphere reserve. It embeds itself in a mega mountain and is a hill station. The study area is shown in figure 1.

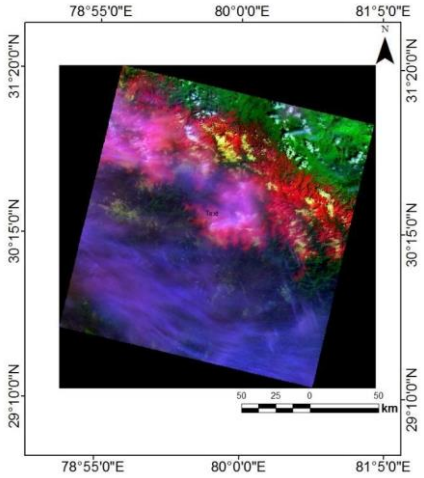


Fig.1 Landsat imagery of Joshimath.

The Landsat multi spectral image used here was acquired on 20th March, 2016.

III. DATA AND SOFTWARE USED

Cartosat multi-spectral data for the area Joshimath with 24m resolution was used. The satellite imagery was first geo-referenced with the help of ArcMAP 10.2.1 Software. Machine learning algorithms were implemented in MATLAB 7 and the results obtained were accessed for accuracy in ArcMAP 10.2.1.

IV. RESEARCH METHODOLOGY

A. *Land cover classification*

To represent the distribution of land cover in the classification task, the following classes were made.

TABLE I. LAND COVER CLASSIFICATION DONE IN THE PAPER.

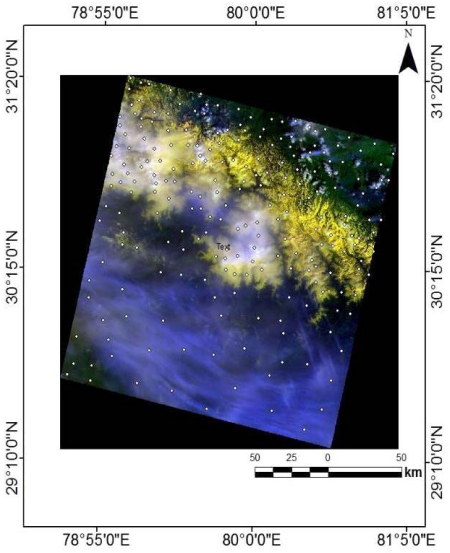
Classes formed	Description
Snow and hills.	Snowcapped mountains
Cloud	cloudy regions
Vegetation grassland.	natural forest/cropland/
Water body	streams, river, lake and reservoir

B. *Training sample*

To see the effect of the choice of training samples on the algorithm, we did uniform selection of training samples such that each subclass has 60 samples. Previous researches done show that the training sample for each class should have a size greater than or equal to 10-30 times the number of bands [5-7]. Figure 2 represents the distribution of the test samples.

C. *Test sample*

For this research, collection of 370 pixels as test data on ArcMAP 10.2.1 was done. For each land class the test size was so selected that it is greater than 40 pixels. Kappa coefficient was used for evaluation.



` Fig.2. Figure represents the distribution of test samples.

D. *Classification*

Four machine learning algorithms were used for the classification of the Landsat imagery. Two supervised learning algorithms namely, Artificial Neural Network and Maximum Likelihood Field Classifier and two unsupervised learning algorithms namely K-mean clustering and Iso-cluster algorithms were implemented in MATLAB 7.The algorithms used are selected because they are openly accessible. Table II shows the documentation and the source of code references.

TABLE II. SET UP OF PARAMETERS USED IN THE ALGORITHMS AND THEIR SOURCE OF CODES.

Name of of Algorithm used	Type of parameter used	Parameter Set	Source code
Maximum-Likelihood Classification	Mean and covariance matrix	estimated for training samples	OpenCV
K-mean clustering	K weight	1,3,5,7,11 no weighting 1/distance	Weka
Artificial Neural Network	hidden layers	1,2,3,4,5,6,7,8	Weka
Iso-cluster	clusters	for 6 bands-1,2,3,4,5,6 For 4 bands-1,2,3,4	Weka

V. RESULTS

The results obtained by the algorithms were analyzed and the confusion matrix is obtained for each classification algorithm. The confusion matrix is a specific table layout that allows visualization of the performance of an algorithm.

Kappa coefficient describes the percentage of similarity between the validation sites and classification results [5].

A. Supervised learning algorithm

1. Maximum likelihood classifier

In the Maximum Likelihood Classifier, while assigning cell to one of the classes which are represented in the signature file, the variances and covariance of the class signatures were considered. The covariance matrix and mean vector can be used to characterize a class if the class sample comes out to be normal. With these two characteristics, the membership of each cell to a particular class is determined by computing the statistical probability for each cell. Classification of cells is done in such a way that they are given label of the class which they have the highest probability of being a member.

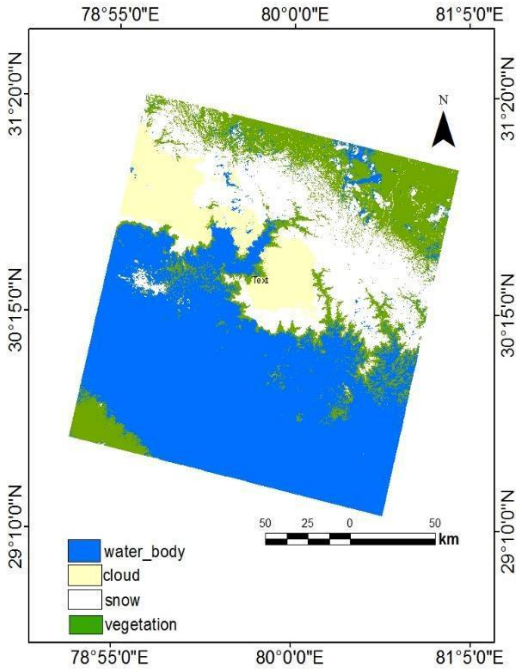


Fig. 3. Classified image obtained by maximum likelihood

TABLE 3. CONFUSION MATRIX FOR MAXIMUM LIKELIHOOD

CLASS_NAME	PREDICT	TRUTH_1	TRUTH_2	TRUTH_3	TRUTH_4	PERCENT	PREDICTIONS
SNOW	1	**106**	20	9	5	75.7 %	140
CLOUD	2	3	**50**	0	2	90.9 %	55
VEGETATION	3	4	0	**88**	0	95.6 %	92
WATER_BODY	4	1	8	10	**90**	82.5 %	109
	PERCENT	92.9 %	64.1 %	82.2 %	92.7 %	**90.2 %**	
	COUNT_TRUTH	114	78	107	97		396 TOTAL POINT
Kappa coefficient : 0.788						**overall accuracy:90.2%**	

Maximum Likelihood classification resulted in an overall classification accuracy and Kappa coefficient of 90.2% and 0.788, respectively. The accuracies for sub-sectional classes using this classifier were following: 95.6% for Vegetation, 82.5% for water body, 90.9% for cloud and 75.7% for snow. Figure 3 represents the classification results by MLC algorithm.

2. K-mean clustering (proposed approach)

This algorithm attempts to splits a given data set into a fixed number of clusters. Initially a fixed number of centroids are chosen which is defined as a data point at the center of a cluster. In practice a centroid is selected in a way that all centroids have unique values. These centroids are used to train a KNN classifier which classifies the data and produces a randomized set of centroids.The arithmetic mean of the cluster is computed and the centroid is assigned its value. This process of classification and adjustment is iterated until the values of the centroids stabilize. The stabilized centroids thus obtained are used to produce the final classification/clustering of the input data. Thus, all the data points in the data set get a class identity. Figure 4 shows the result obtained by this approach.

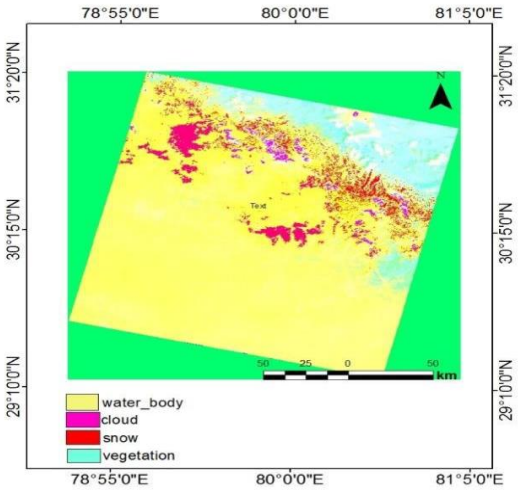


Fig. 4. Classified image obtained by k-mean clustering

TABLE 4. CONFUSION MATRIX FOR K-MEAN CLUSTERING

CLASS_NAME	PREDICT	TRUTH_1	TRUTH_2	TRUTH_3	TRUTH_4	PERCENT	PREDICTIONS
SNOW	1	**123**	15	0	2	87.8 %	140
CLOUD	2	3	**45**	1	0	91.8 %	49
VEGETATION	3	0	1	**90**	1	97.8 %	92
WATER_BODY	4	1	0	0	**88**	98.8 %	89
	PERCENT	96.8	73.7 %	98.9	96.7 %	**93.5 %**	
	COUNT_TRUTH	127	61	91	91		370 TOTAL POINT
Kappa coefficient :0.91						**overall accuracy:93.5%**	

K-mean clustering algorithm resulted in an overall classification accuracy and Kappa coefficient 93.5% and 0.91, respectively. The accuracies for sub-sectional classes using this classifier were following: 97.8% for Vegetation, 98.8% for water body, 91.8% for cloud and 87.8% for snow.

B. Unsupervised learning algorithm

1. Iso cluster algorithm

In this algorithm, minimum Euclidean distance is calculated to assign each candidate cell to a cluster by using an iterative process. The closest of means is computed and each cell is assigned to it. After each iteration, means are calculated based on the attribute distances of the cells that belong to the cluster. After the process is completed, each cell is assigned to the closest mean in multidimensional attribute space. Figure 5 shows the classified image.

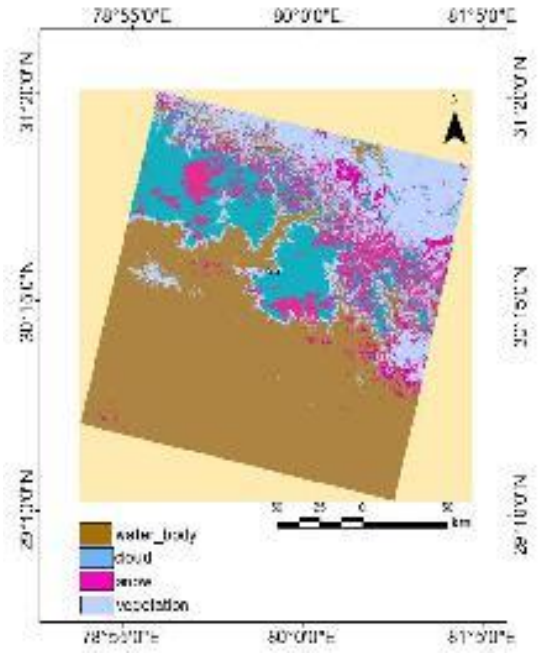


Fig. 5. Image obtained after iso-cluster classification

2. Unsupervised artificial neural network

In this classifier there are three types of connected nodes namely input, output and hidden and therefore it is called hidden network. The input layer represents each input feature. The output layer represents a node for each class to be predicted. In between input and output layers lie the hidden layer. Connections exist between nodes of input layer and hidden layer and similarly exist between output layer and hidden layer. The parameters in this model are the weights assigned to each connection. During the training phase, these parameters are learnt. The sum of the weighted values from the input nodes is provided as input to each node of the hidden layer. The activation function is then provided with these values as input. The training data is adjusted in such a way

that they have a standard deviation of 1 and mean value of 0 so that they are considered equally in the training process. The class probabilities are obtained from the output values.

TABLE 5. CONFUSION MATRIX FOR ISO-CLUSTER ALGORITHM

CLASS_NAME	PREDICT	TRUTH_1	TRUTH_2	TRUTH_3	TRUTH_4	PERCENT	PREDICTIONS
SNOW	1	**90**	41	7	2	64.3 %	140
CLOUD	2	7	**38**	10	0	69.0 %	55
VEGETATION	3	2	14	**74**	2	80.4 %	92
WATER_BODY	4	1	2	0	**86**	96.6 %	89
	PERCENT	90.0 %	40.0 %	81.3 %	95.5 %	**77.8 %**	
	COUNT_TRUTH	100	95	91	90		376 TOTAL POINT
Kappa coefficient : 0.68						**overall accuracy:77.8%**	

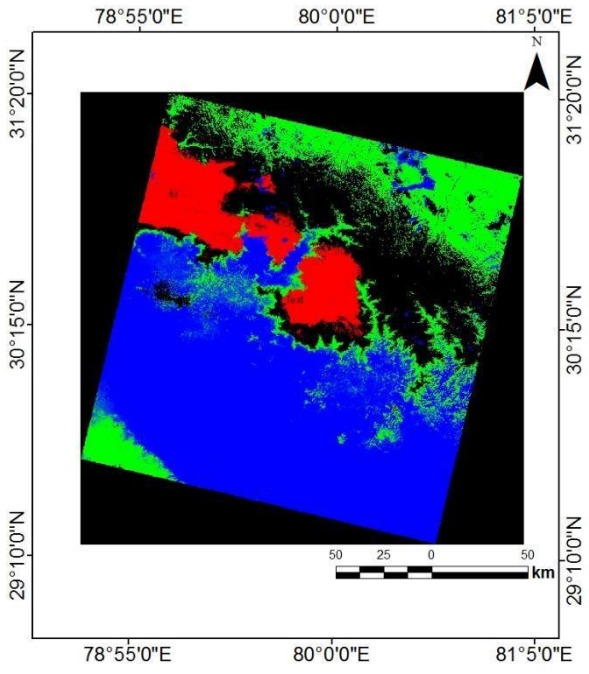


Fig. 6. Image obtained after artificial neural network classification

Artificial neural network algorithm resulted in an overall classification accuracy and Kappa coefficient 85.6% and 0.80, respectively. The accuracies for sub-sectional classes using this classifier were following: 95.5 for Vegetation, 96.6 for water body, 91.8% for cloud and 71.4% for snow. Figure 6 shows the classified image.

TABLE 6. CONFUSION MATRIX FOR ARTIFICIAL NEURAL NETWORK

CLASS_NAME	PREDICT	TRUTH_1	TRUTH_2	TRUTH_3	TRUTH_4	PERCENT	PREDICTIONS
SNOW	1	**100**	38	0	2	71.4 %	140
CLOUD	2	2	**45**	2	0	91.8 %	49
VEGETATION	3	2	2	**86**	2	95.5 %	92
WATER_BODY	4	1	2	0	**86**	96.6 %	89
	PERCENT	95.2 %	51.7 %	97.7 %	95.5 %	**85.6 %**	
	COUNT_TRUTH	105	87	88	90		370 TOTAL POINT
Kappa coefficient :0.80						**overall accuracy:85.6%**	

VI. CONCLUSION

In this study we compared supervised and unsupervised learning algorithms and came to the conclusion that supervised learning algorithms produce better results (as seen by the kappa values and overall accuracy of each algorithm) in land cover classification. For all classifiers the overall accuracy and kappa values were calculated as shown in tables 3, 4, 5 and 6. The results obtained can be summarized as follows:

The k-mean clustering algorithm shows the highest overall accuracy of 93.5% followed by maximum likelihood classifier, artificial neural network classifier and lastly the iso-cluster classifier.

The k-mean clustering also gives the highest kappa value of 0.91 followed by artificial neural network classifier, maximum likelihood classifier and lastly the iso-cluster classifier.

The areas labelled as snow were best classified by k-mean clustering, followed by maximum likelihood classifier, artificial neural network and lastly iso-cluster classifier.

The areas labelled as cloud gave equal prediction with k-mean clustering and artificial neural network followed by maximum likelihood classifier and lastly iso cluster.

Vegetation area was best classified by k-mean clustering followed by maximum likelihood classifier, artificial neural network and lastly iso cluster classifier.

Conclusions can be made on seeing these results that our proposed algorithm gave best performance in land cover classification for the joshimath region.

Considering the processing time of these algorithms, the k-mean clustering and maximum likelihood classifier were slower than iso-cluster and artificial neural network in the training phase. But after completion of the training phase they were very fast to classify the images.

In terms of training data, Neural Network did not require many training sets for the classification process as this classifier is not a statistical approach.

REFERENCES

[1] A. Tzotsos and D. Argialas, Support Vector Machine Classification forObject-Based Image Analysis, LNGC, 2008, 663–667.

[2] Dhawan, A. P., Chitre, Y., Kaiser, C., and Moskowitz, Ms, (1996). Analysis of mammographic microcalcifications using gray-level image structure features, IEEE Trans. Medical Imaging, 246–259.

[3] Hartman, E. J., Keeler, J. D., and Kowalski, J. M., (1990). Layered neural networks with Gaussian hidden units as universal approximations. Neural Computation, 210–215.

[4] Ritter, G. X. and Wilson, J. N., (1996). Handbook of Computer Vision Algorithms in Image Algebra. Boca Raton: CRC Press.

[5] Mather, P.M. Computer Processing of Remotely-Sensed Images, 3rd ed.; John Wiley & Sons, Ltd Chichester, UK, 2004.

[6] Piper, J. Variability and bias in experimentally measured classifier error rates. Pattern Recognition. Lett.1992, 13, 685– 692.

[7] Van Niel, T.; McVicar, T.; Datt, B. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. Remote Sensing of Environment. 2005, 98, 468-480