International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-2, Issue-8, August 2014

# Analysis And Prediction Of Landslides in Uttarakhand Using Weka Tool

**Chandradeep Bhatt, Anubhooti Papola**

***Abstract*— Landslides are one of the most popular natural disasters. Cause of landslides damage to infrasructures, blocking of communication, loss of human life and indirect loss of productivity. Therefore, a need is felt to make a landslide prediction and analysis the landslides susceptible area. There are various conventional methods of predict landslides, but application of soft computing technique with information technology has given a n unique direction to this area. In this particular work two different approaches have been proposed for the classifiers of subjects into 4 classes namely: Non-susceptible, Less-susceptible, Moderately-susceptible and High-susceptible. The techniques undertaken are J48 and LMT. Analysis of these classifiers on WEKA tool, then implement the landslide prediction model in MATLAB . The results obtained are very clear and show improvement from the previous works.**

***Index Terms*— landslide , landslide prediction, j-48 classifier, lmt classifier, weka data mining tool, matlab..**

## I. INTRODUCTION

Landslide is a geographical phenomenon in which all varieties of mass movements of hill slope include and it can be defined as the downward movement of slope forming materials composed of rocks, soils, artificial fills along surfaces of separation by falling, sliding and flowing, either slowly or quickly from one place to another. Landslides are one of the major natural hazard that account for hundreds of lives besides enormous damage to properties and blocking the communication links every year in the in the Himalayas. So this become essential to predict landslide before happening. The general process of landslide prediction involves real time disaster information collections, compilations, interpretations, analysis, predictions, illustrations and decision support. It may be observed that advancement of information technology in the form of internet, Geographic information System (GIS), remote sensing and satellite communication can help a great deal in planning and implementation of landslide prediction model. The main goal of our work is to analysis and prediction of Landslides susceptible area and non- susceptible area. We have utilized known data mining techniques to develop a realistic model of Landslide prediction system. Our goals to analysis all classifier performance on Landslide data, and improve the accuracy of prediction. To find out the prime factor is also an objective. In this study we present a report on our attempt at using J48 and LMT (Logistic Model Tree) classifier. First of all our goal is to make Landslide data in ARFF format, by which we can easily analysis the data on WEKA Tool. After that analysis on weka using different classifiers. Implement the classification algorithm which was the best among all in MATLAB.

**Manuscript received August 20, 2014.**

**Chandradeep Bhatt**, Department of Computer Science & Engineering, Uttarakhand Technical University, Dehradun, India, 9634074436.

**Anubhooti Papola**, Assistant Professor in Department of Computer Science & Engineering, Uttarakhand Technical University, Dehradun, India, 9634365084.

## II. DATA MEHODOLOGY

*Data methodology begins with a detailed field study of the selected* site that included *structural and lithological mapping of the area and collection of the necessary geological and geotechnical data. Prior to detailed field investigation a large scale topographic map on 1: 5000 scale was prepared.*

| *Spatial Database* | *Factor* | *Scale Resolution* |
|---|---|---|
| *Landslide* | *Landslide* | *1: 25,000* |
| *Topographic Map* | *Slope* | *1: 25,000* |
| *Geology Map* | *Lithology types* | *1: 63,300* |
| | *Distance From River* | |
| *Land cover Map* | *Land cover* | *30m * 30m* |

**Table 1: Factors and their resolution map scale**

## III. PRIME PARAMETERS

*(i) Lithology:* Lithology of an area combined with various conditions of the rocks such as compaction, deformation, facturing, intusions, ultrations etc. holds good significances as a factor responsible for causing landslides. The stability of a site, thus can be inferred from the lithology / rock type It inhibits and its conditions. The study area is rich in quartizite . Different types of quartizite, such as pink, white and brown having varying degree of weathering and other conditions are included by scatterd pathches of matabolcanics

*(ii) Slope Gradient :* The slope gradient has been assessed for each grid using topographic map of the area. This angle is then classified into four categories as Gentle, Moderate, Step and Very steep.

| *Catagory* | *Gradient* |
|---|---|
| *Gentle* | *< $15^0$* |
| *Moderate* | *>$15^0$ & <$30^0$* |
| *Steep* | *>$30^0$ & <$45^0$* |
| *Very steep* | *>$45^0$* |

**Table 2: categories of slope gradient**

(iii) Land use/ Land cover: Land when citilized efficiently can be saved from erosion and weathering and hence, stabilized. Land cover, thus, need to be studied for potential landslide assessment. Land cover can be inferred as dense forest, Barren Land, Agricultural Land and Industrial or Commercial Land. The slope stability generally increases with better forestation and vegetation.

Barren Land is most susceptible to Landslides whereas a dense forest or highly vegetated land cover is Less susceptible. It has been observed that the slopes which are barren have been utilized for the purpose of road construction and in current study is considered one of the most influencing mthropogenic factors on slope stability based on the land use and vegetation cover.

## IV. LANDSLIDE PREDICTION MODEL

The landslide prediction models are developed using landslide data of Uttarakhand. 80% of the data is used for training the landslide prediction model and 20% of the data used for validation. Fig 1 illustrates the overall methodology of the landslide prediction model.

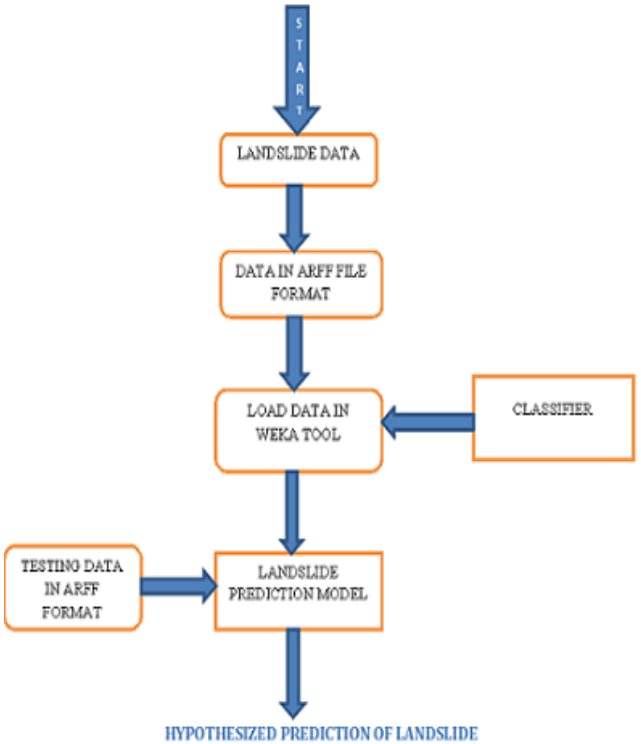


**Figure 1: step for development of landslide prediction model.**

### *A. CLASSIFICATION ALGORITHM*

Classifier is a model which predicts the class value from explanatory attributes. Classifiers can be designed manually, based on expert's knowledge, but nowadays it is more common to learn them from real data. This basic idea is the following: first, we have to choose the classification method, like Decision Trees, Bayesian Networks or Neural Networks. Second, we need a sample of data, where all class values are known. The data is divided into two parts, a training set and a test set. The training set is given to a learning algorithm, which derives a classifier. Then the classifier is tested with the test set, where all class values are hidden. Here in our study we just use two classifiers one is **J48** and other is **LMT**.

**J-48:** J48 is an open source Java implementation of the c 4.5 algorithm in the WEKA data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. This algorithm was developed by Ross Quinlan. The decision tree generated by C4.5 can be used for classification and for this reason; C4.5 is often referred to as a statistical classifier ("C4.5 (J48)",).

C4.5 building a decision trees from a set of training data using the concept of information entropy. The training data is a set $S=S_1, S_2, \ldots..$ of already classified samples. Each sample $S_i$ consists of p-dimensional vector $\{X_{1,i}, X_{2,i}, \ldots\ldots X_{p,i}\}$ where the $X_j$ represent attributes or features of the sample, as well as the class in which $S_i$ falls. At each node of the tree, C4.5 choose the attribute of the data the most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy).

The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then reuses on the smaller sub lists.

The algorithm has a few base cases:

**(i)** All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.

**(ii)** Non of the features provide any information gain. In this case C4.5 creates a decision node higher up the tree using the expected value of the class.

**(iii)** Instance of previously- unseen class encountered. Again C4.5 creates a decision made higher up the tree using the expected value.

**LMT:** A Logistic Model Tree basically consist of a standard decision tree structure with logistic regression functions at the leaves, much like a model trees is a regression tree with regression function at the leaves.

As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with K values, the node has K child nodes and instances are sorted down one of the K branches

depending on their value of the attribute for numeric attribute, the node has two child nodes and the test consists

of comparing the attribute value to a threshold: an instance is sorted down the left branch if its value for that attribute is

smaller than the threshold and sorted down the right branch otherwise.

### B. *WEKA TOOL*

WEKA is a data mining tool, provides implementations of learning algorithms that you can easily apply to your dataset. It also includes a variety of tools for transforming dataset such as the algorithm for discretization. We can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance, all without writing any program code at all.

The workbench includes methods for all the standard data mining problems: regression, classification, clustering, association rule mixing, and attribute selection. Getting to know data is integral part of the work, and many data visualization facilities and data preprocessing tools are provided. All algorithms take their input in the form of single relational table in ARFF format. WEKA would not only provide a toolbox of learning algorithms, but also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation.

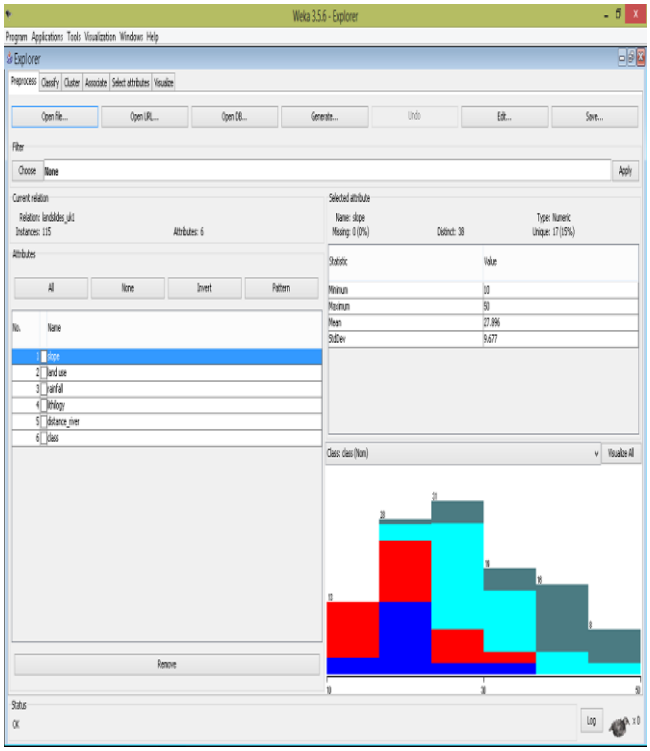

**Figure 2: Weka explorer user interface.**

The classification process involves following step:-

A – Create training data set

B – Identify class attribute and classes.

C – Identify useful attributes for classification (Relevance analysis).

D – Learn a model using training examples in training set

E – Use the model to classify the unknown data.

### C. *PREDICTION USING J48 AND LMT CLASSIFIER IN WEKA*

We choose the j-48 classifier first and apply to our data set. We got a decision tree created by this classifier, then apply trained model to test data. Then, we have 90% accuracy.

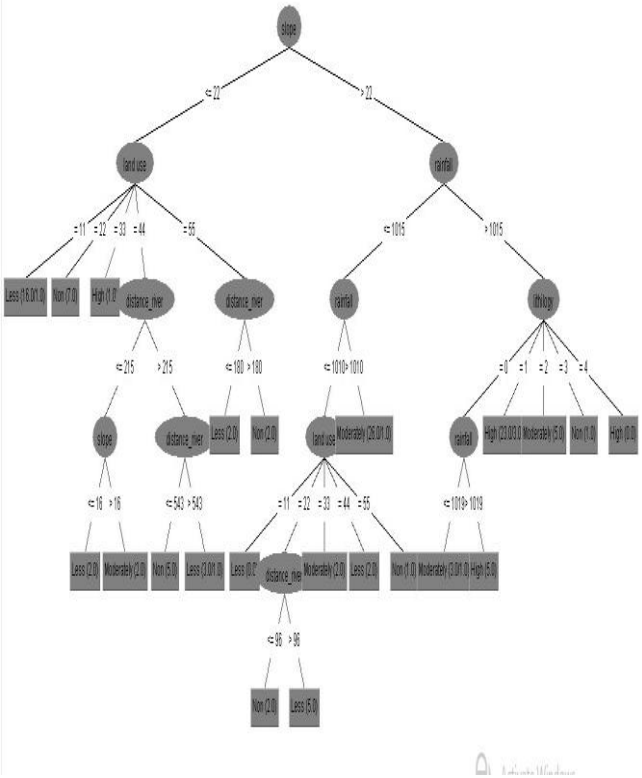

**Figure 3: prediction model tree created by j-48**

The figure 3 shows the tree created by j-48 classifier and it is also a prediction model tree. On the basis of this we can also test the testing data set.

Classifier output

```
=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          27               90      %
Incorrectly Classified Instances         3               10      %
Kappa statistic                          0.8626
Mean absolute error                      0.1048
Root mean squared error                  0.2314
Relative absolute error                 28.3274 %
Root relative squared error             53.696  %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.8       0.08       0.667     0.8       0.727      0.852    Non
  0.8       0.05       0.889     0.8       0.842      0.84     Less
  1         0          1         1         1          1        Moderately
  1         0          1         1         1          1        High

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  4  1  0  0 |  a = Non
  2  8  0  0 |  b = Less
  0  0 10  0 |  c = Moderately
  0  0  0  5 |  d = High
```

**Figure 4: Evaluation on test data using J-48 Classifier.**

Now. We choose LMT classifier and create a trained model using training set. Then apply the trained model to test data. We have 70% accuracy.

```
Classifier output

Time taken to build model: 1.7 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          21               70      %
Incorrectly Classified Instances         9               30      %
Kappa statistic                          0.5878
Mean absolute error                      0.1592
Root mean squared error                  0.298
Relative absolute error                 43.0481 %
Root relative squared error             69.1717 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.6       0.2       0.375     0.6       0.462      0.864    Non
  0.6       0.1       0.75      0.6       0.667      0.91     Less
  0.9       0.1       0.818     0.9       0.857      0.97     Moderately
  0.6       0          1        0.6       0.75       1        High

=== Confusion Matrix ===

 a b c d   <-- classified as
 3 2 0 0 | a = Non
 4 6 0 0 | b = Less
 1 0 9 0 | c = Moderately
 0 0 2 3 | d = High
```

**Figure 5: Evaluation on test data using LMT Classifier.**

We also perform cross validation test and percentage split test.

**5- Fold cross validation:-** For each fold, we randomly assign data points to five sets $d_0$, $d_1$, $d_3$, $d_4$ and $d_5$, so that all sets are equal size (this is usually implemented by shuffling the data array and then splitting it in five). We then train on $d_0$, $d_1$, $d_3$, and $d_4$ and test on $d_5$, followed by training on $d_1$, $d_3$, $d_4$ and $d_5$ and testing on $d_0$ and so on. This has the advantage that our training and test sets are both large, and each data point is used for both training and validation on each fold.

66 percentage split:- It is also known as default percentage split, because in WEKA tool this percentage split already lies in editing textbox. In this split, 66% of data is randomly choose for training dataset and rest for testing dataset.

### *D. IMPLEMENTATION OF PREDICTION MODEL CREATED BY J-48 CLASSIFIER IN MATLAB*

After the analysis of classifiers on landslide dataset using WEKA, we analyzed that the performance and prediction accuracy of J-48 classifier is much better than the logistic model tree (LMT). So we implemented the landslide prediction model created by WEKA in MATLAB. After making source code of prediction model in MATLAB, we will just click the button RUN. Then, we got a application window; in which we have all the parameter of the land slide and front of each of that a blank text box. Figure 6 shows the application window.

A graphical user interface (GUI) is a graphical display in one or more windows containing controls, called components that enable a user to perform interactive tasks. The user of the GUI does not have to create a script or type commands at the command line to accomplish the tasks. Unlike coding programs to accomplish tasks, the user of a GUI need not understand the details of how the tasks are performed.

In this prediction model, we can get the type of landslide may occur by inputting the value of all parameter show in model.


**Figure 7: Landslide prediction model with GUI in MATLAB**

In figure 8, we inputted the value of all parameter like we put the value of slope is 34, value of land use is 44, value of rainfall is 1002, value of lithology is 1 and distance from river is 111. Then, clicking the button submit we have a message that the less susceptible area. This means the data we inputted is belongs to less susceptible area. If we click the button accuracy in GUI prediction model window, we get 93.9. It means the accuracy of trained model with training data is 93.9 %.


**Figure 8: GUI landslide prediction model window in MATLAB with result.**

## V. RESULT ANALYSIS

We inputted the dataset to weka, and then trained the model with training data set. After tha we use the trained model into test data.

| CLASSIFIER | TEST | PREDICTION ACCURACY |
|---|---|---|
| J-48 | TRAINING SET | 93.9 % |
| | TESTING SET | 90.0 % |
| | 5- FOLD CROSS VALIDATION | 74.7 % |
| | 66PERCENTAGE SPLIT | 75.0 % |
| LMT | TRAINING SET | 90.4 % |
| | TESTING SET | 70.0 % |
| | 5-FOLD CROSS VALIDATION | 66.9 % |
| | 66PERCENTAGE SPLIT | 62.5 % |

**Table 3: Accuracy table.**

So, the performance and accuracy for prediction of j-48 classifier is better than MLT classifier for landslide data set.

## VI. CONCLUSION

Remote sensing and GIS have been useful in data preparation and at integration stages. We have analyzed the landslide susceptible area on the basis of data mining tool and techniques.

From the above study it has been observed J48 classifier perform well for landslide susceptible area prediction. Also the accuracy for LMT (Logistic Model Tree) classifier is good but in terms of accuracy J48 performs better than other classifier. The prediction model obtained by training known landslide point data and stable point data possessed high capability and accuracy. The application of landslide prediction on WEKA Tool in mining area is always in exploration, and there are still many things which need to be researched further.

## REFERENCES

**[1]** - B. Pradhan and S. Lee " Regional Landslide susceptibility analysis using back propagation neural network model at Cameron Highlands, Malaysia", Landslides, Volume 7, No. 1, pp 13-30, Mar 2010.

[2] – Weisun, Shuliang Nang, " A new Data Mining Method for Early Warning Landslide based on Parallel Coordinate," 978-1-4244-8351, 2011 IEEE.

[3] – Ruixiang Liao, Shimei Wang, Haifeng, Feifei Xu, " Landslides Hazard Evaluation of Zigui based on GIS Information Model in Three Gorges Reservoir", 978-1-4244-8351, 2011 IEEE.

[4] – S Kundu, D C Sharma, A K Sana, C C Pant, and J Mathew, " GIS based Statistical Landslide Susceptibility Zonation : A case study in Ganeshganga Watershed, The Himalayas", 12th esri India User Conference, 2011.

[5] – Han J., Kamber M., Jian P., Data mining Concepts and Techniques. San Francisco, CA : Morgan Kaufmann Publisher, 2011.

[6] - Pai-Huittsu, Wen-Ray Su, " Hazard Hotspots Analysis from Geospatial Database using Geospatial Data Mining Technology", 978-1-4673-1159, 2012 IEEE.

[7] – Kishor Kumar, Rahul Devrani, Anil Kathait, Neha Aggarwal, "Micro-Hazard Evaluation and Validation of Landslide in a part of North Western Garhwal Lesser Himalaya, India, "International Journel of Geomatics and Geosciences, ISSN – 0976-4380, Volume 2, No 3, 2012.

[8] – Mohammad Onogh, V. K. Kumra and Praveen Kumar Rai, " Landslide Susceptibility Mapping in a part of Uttarakashi District by Multiple Linear Regression Method", ISSN : 2277-2081, 2012 Volume 2 (2) May-Aug, PP. 102-120.

[9] – Mahendra Tiwari, Manu BhaiJha, OmPrakash Yadav, "Performance Analysis of Data Mining Algorithms in WEKA," IOSRJCE, ISSN : 2278-0661, ISBN : 2278-8727, 3 (Sep-Oct 2012), PP. 32-41.

[10] – M. S. Rawat, B. S. Rawat, V. Joshi, M. M. Kimothi, " Statistical analysis of Landslide in South district, Sikkim, India : Using Remote Sensing and GIS", ( IOSR-JESTFT), ISSN : 2319-2402, Volume 2,3 (Nov-Dec) 2012, PP. 47-61.

[11] – Trilok Chand Sharma, Manoj Jain, " WEKA Approach for Comparative Study of Classification Algorithm", IJARCCE, ISSN : 2319-5940, Volume 2,4, April 2013.

[12] – Swati Singhal, Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", IJITEE, ISSN : 2278-3075, Volume 2, Issue-6, May 2013.

[13] – Dr. Garima Krishna, Dr. Anurag Raii, Sachin Saxena, " Landslide Monitoring and Hazard Mapping in Uttarakhand using Reinforcement Learning and Neural Network", Volume 3, 8 Aug 2013, ISSN -2277-128X.

[14] – Tina R. Patil, Mrs. S. S. Sherakar, "Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification", IJCSE, ISSN : 0974-1011, Volume 6,2 April 2013.

[15]- Jay Gholap, "Performance tuning of j48 algorithm for prediction of soil fertility", 2013.

**First Author:**

**Mr. Chandradeep Bhatt,** Chandradeep bhatt is a Graduate in Computer Science Engineering from Hemwati Nandan Bahuguna Garhwal University, Srinagar Garhwal, Uttarakhand. Presently he is pursuing Post Graduate (Final Year) in Computer Science Engineering from Uttarakhand Technical University, Dehradun. His area of interest include computer network and data warehousing & data mining.

**Second Author:**

**Assistant Professor Anubhooti Papola .** nubhooti Papola is a Graduate in Computer Science Engineering from Uttarakhand Technical University, Dehradun, Uttarakhand and a Post Graduate in Computer Science Engineering from Graphic Era University, Dehradun. She was a lecturer in GRD IMT Dehardun and programmer in Anya- Softek, Dehradun. Presently she is Assistant Professor in W.I.T, Uttarakhand Technical University, Dehradun.