Integration of Classification and Clustering for the Analysis of Spatial Data

M. Nachiket Kumar#1, Dr. Venkatesan*2, D. Manoj Prabhu#3

#Computer Science Department, School of Computing Science and Engineering
VIT University, Vellore, Tamil Nadu, India

*Associate Professor, School of Computing Science and Engineering
VIT University, Vellore, Tamil Nadu, India

Abstract—This paper presents the use of classification based clustering approach for landslide(LS) susceptibility analysis. The case is analyzing LS in Coonoor region of Ooty in Tamil Nadu, India. In the study area landslide locations were recognized by analyzing GIS information. Landslide conditioning factors such as Geology, Geomorphology, Soil type, slope, land use and land cover, and rainfall were considered for analysis. These factors are analyzed using Bayes Classification and then k means to classify the landforms into different classes/Zones according to their probability of landslides into zones ranging from “Very High” to “Low”. Also integration of Bayesian/kmeans approach to determine these probabilities corresponding to the region so that it will be easier to analyze the landforms and take decisions accordingly.

Keywords: Landslide, classification, clustering, K-means, Baye's theorem.

I. INTRODUCTION

Consistently a great many lives and billions of dollars are lost because of harms done via avalanches everywhere throughout the world. A yearly evaluated normal harm expense of nearly the harm cost for every capita of 339 rupee with total more than 88 billion rupees in United States alone [13]. Being a harm and influence individuals, commercial ventures, associations, and the nature [13]. Human exercises like deforestation, urban development, overgrazing, mining stacking upslope and street building speedup the landslide [14].

Improperly designed land use like harvesting forest and constructing road add to the risk of landslide. Road intercepting water bodies may degrade quality of water, loss natural spawning habitat of fish and can cause debris jam leading to destruction of riparian vegetation. Road intercepting forest may decrease stability of soil and amplify hazard of landslide in hilly regions. Hence, it is very important to determine susceptibility of landslide for the management of land usage.

In this paper Coonoor area of Tamil Nadu is considered for analysis. Due to the frequent occurrences of landslides in this area landslide analysis has always been a concern here. As per media report a landslide triggered by torrential rain occurred in the Coonoor Ooty region of Tamilnadu killing at least 39 people in November 2009. The landslide demolished nearly 300 tinned roof mud huts. Ketti and its suburbs, about 7 km away from Ooty, received record rainfall of 820mm in 24 hours while Ooty recorded 170mm. Many parts of the Nilgiris continued to remain cut off on Wednesday (11th Nov. 2009) due to landslips. As per another media report as many as 543 landslips has occured in just two days (10-11) in the Nilgiris, and 816 houses razed to debris. Besides, 600 hectares of crops has been devastated and road revetments damaged in 145 places. Above all, 43 precious lives lost and over 1,100 people has been left homeless. Furthermore road accidents and traffic is common problem on Coonoor Ooty National Highway due to the recurrent landslips. In this paper the we try to analyze the given region of land in order to predict its susceptibility to landslide using Naive Bayes therom and then clustering these data sets in to various clusters according to the predicted zone for further analysis.

II. LITERATURE SURVEY

There have been numerous studies did on landslide vulnerability assessment utilizing geographic data framework (geographic information system); for
instance, [14] many landslide risk assessment studies based on geomorphologic relationships between pattern and landslide types and based on the morphological, lithologic and structural settings were summarized. Lately, there have been studies on landslide susceptibility assessment using geographic information system, and probabilistic models have been applied to many of these studies.

Incidence of landslide are broadly because of internal and external reasons. A large number of investigation materials of landslide, the growth inducements of landslide are analyzed.

Late advances in satellite remote sensing innovation and expanding accessibility of higher arrangement geospatial items around the globe have given an uncommon chance to such a study. A system for creating a test continuous forecast framework to distinguish where precipitation activated landslides will happen is proposed by joining two vital segments: surface landslide helplessness (LS) and an ongoing space-based precipitation investigation framework. Initially, a worldwide landslide guide is gotten from a mixture of semi static worldwide surface qualities (soil composition, area spread order advanced height geology, incline, soil sorts, and so forth.) utilizing a GIS weighted straight combo approach. Furthermore, a balanced empiric relationship between landslide event and precipitation power length of time is utilized to survey landslide dangers at territories with high weakness. A study is to concentrate trash source territories by logistic relapse examinations focused around the information from the inclines was carried out on the off chance that study from Kapi, Besparmak and Barla mountains [6][7].

In 1998 Montgomery and Dietrich developed a SHALSTAB model[12] which is another technique used for evaluating landslide risk. Eq. (1) shows basic SHALSTAB evaluation formula

\[
\frac{q}{T} = \frac{\rho_s}{\rho_w} \left(1 - \frac{\tan \Theta}{\tan \Phi}\right)
\]

\[\text{Eq. (1)}\]

where \(a\) is emptying region, \(b\) is limit length, \(\Theta\) is hillslope edge, \(\rho_s\) is mass thickness, \(\rho_w\) is mass thickness, \(\Phi\) is point of interior contact and the proportion \(q/T\) of relentless state powerful precipitation to transmissivity. In the degree \(q/T\) bigger the \(q\) with respect to \(T\) the more probable the ground is to be soaked, and the more prominent the danger of incline disappointment. In any case, yields are hard to look at, the technique does not consolidates instability connected with the parameters utilized for the displaying. An approval investigation of SHALSTAB, a straightforward robotic model for characterizing the relative potential for shallow area sliding over the scene has been accounted for. Landslide disintegration has a constituted history in New Zealand[12].rough Set Theory was utilized as a part of in north focal Idaho zone's western slants of the Rocky Mountains inside Clearwater National Forest[1]. In a Landslide Prone Area neuro-Fuzzy Approach is utilized to foresee landslide in Malaysia[4]. An alternate model was actualized close to Three Gorges, China that proposed utilization of Rough Set and Back-Propagation Neural Networks. Be that as it may there are sure disadvantages of Traditional BP calculation, such as getting stuck effortlessly in neighborhood minima and moderate velocity of merging.

Likelihood hypothesis or fuzzy set hypothesis may be utilized for displaying instabilities. In likelihood hypothesis probabilistic models are utilized to measure vulnerabilities connected with expectation and measure of fragmented learning by destination demonstrating. Then again in fluffy set hypothesis instabilities are displayed focused around master learning and measure of inadequate information by subjective modeling. fuzzy Model, for example, fluffy k-implies, which is like bunch examination however permits class cover, was actualized in soil-scene analysis [16][17]. The relationship between different earth surface procedures soil improvement ,topography were investigated in these studies. A non specific strategy for consequently portioning landforms into landform components utilizing Dems(digital Elevation Model), fluffy rationale and heuristic principles .[18]. By utilization of hypotheses in smudge math, request game plan of the instigations in agreement their importance, set forward obscure judgment tenet of landslide estimation was done by Fuzzy logic in order to analyze and Predict Landslip [15] Weighted Bayesian Classification was utilized focused around Support Vector Machines[8] , Weighted Naive Bayesian Classifier are desribed [9] and Locally weighted gullible bayes[10] attempting to enhance the aftereffects of Bayes Therom.

Other methods, like Baysian classification, facilitate a probable way of transforming fuzzy classification's analytical correlations of landforms with
III. METHODOLOGY AND IMPLEMENTATION

In this paper, a part of the Coonoor - Ooty in Tamil Nadu India, was used for the applying landslip vulnerability analysis because of the repeated incidence of landslips in that area. The Methodology is based on creating classes of continuous landform using kmeans methods bayesian classification to deal with uncertainties. Data set consider consists of mainly six parameters Geology, Geomorphology, Land use and land cover, rainfall, slope and soil. Based on these six parameters label is assigned to each set called zone. Zone may be High, Very High, Moderate or low which indicates the susceptibility of that region for landslide.

A. Bayes Theorem

The Bayes classification is a technique utilized for choice making under vulnerability. This Strategy is system for joining relative likelihood (genuine or not) with contingent likelihood .subjective likelihood is a representation of the level of confidence in an occasion happening focused around an individual's experience, partialities, idealism, and so forth. Contingent likelihood is the information about the probability of the theory to be genuine given a bit of coordinating a harsh set k-implies order proof. Case in point, one can't be sure whether landslides dependably happen in regions of topographic union. The learning may be communicated as the client being 90% sure (i.e., likelihood of 0.9) that landslides will happen in territories of topographic convergence.

Bayesian Classification was utilized for datamining with the assistance of components that triggers landslides for EWLSM to anticipate the landslide risks in Nilgiris area of TN,India[11].

Here $P(h|D)$ follows the Bayesian classification

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$$P(C_j|D) \propto P(C_j) \prod_{i=1}^{n} P(d_i|C_j)$$

In our data set land slide information of Coonor-Ooty is used. Here first Model is created using training data in Rapid Miner where based on the six parameters Geology, Geomorphology, Land use and land cover, rainfall, slope and soil based on which zone is predicted using Naive Bayes Therom which helps in cross validation. After creating the model it is applied to testing data with blank zone which is predicted. Output of this step is then fed to k-means clustering for further cluster anaylsis.
Once the model is generated testing data is applied to it. In testing data zone parameter is intentionally left blank and prediction is done using the generated model.
B. k means Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

$k$-means clustering aims to partition $n$ observations into $k$ clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

In standard algorithm randomly $k$ centroids are selected. Then distance of each set from centroids is calculated. Here $d_{ii}$ is the distance formula,

$$d_{ii} = \sqrt{\sum (x_i - c_i)^2}$$

Membership $\mu$ is determined $m$ is fuzzification factor

By membership function we determine is to which cluster a particular data item belongs, based on the distance of that point from the centroid.

$$\mu_1(x_1) = \frac{\left(\frac{1}{d_{ii}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p} \left(\frac{1}{d_{ii}}\right)^{\frac{1}{m-1}}}$$

Ensuring

$$\sum_{j=1}^{p} \mu_j(x_j) = 1$$

And at each iteration centroids are updated as...
In this paper k-means is applied to cluster the various zones of land. When testing data is applied to Bayes Classifier zone of every data item is predicted. Now the Output of Bayes classifier is given to k-means clustering Algorithm. In k-means clustering various data items are classified according zones.

After application of k-means all the data items with similar zone come under one cluster hence making it easier to understand and analyze.

This methodology is proceeded till stable group focuses (centroids) are found. Different systems can be used to focus ideal number of bunches[2]. Rapid Miner is used for implementation of k-means as shown in Fig 4

Figure 3 shows the centroid movement for one iteration.
IV. RESULT

On applying Bayes theorem for the zoning of landslides prediction zone of the testing data is predicted (with 6 attributes and 1100 tuples).

Cross validation gives the following result: if the number of parameters used in model creation is increased, i.e., more analysis data is available, the accuracy model will increase. In given data set, more the data belonging to a particular zone more will be the accuracy of prediction of that zone. According to the confidence calculation of every data set in testing data set, it is assigned a zone as shown in figure 7.

<table>
<thead>
<tr>
<th>accuracy</th>
<th>70.79% ± 4.42% (micro: 70.80%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>true High</td>
</tr>
<tr>
<td>pred. High</td>
<td>471</td>
</tr>
<tr>
<td>pred. Moderate</td>
<td>25</td>
</tr>
<tr>
<td>pred. Low</td>
<td>0</td>
</tr>
<tr>
<td>pred. Very High</td>
<td>35</td>
</tr>
<tr>
<td>class recall</td>
<td>97.05%</td>
</tr>
</tbody>
</table>

Fig 5: Bayes Classification
K-means clustering is applied on zones to classify the data set into 4 clusters corresponding to each zone viz low, moderate, high, and very high.
Fig 8: k-means Clustering: clustering according to zone

Fig 9: k-means Clustering chart
Fig 10 shows how clusters are created corresponding to various predicted zone of the testing data. Hence it can be seen that certain factors of data like rainfall are more dominant than Geology. Furthermore accuracy of Bayes model will increase with considering more predicting parameters like water content of soil and hence improve the zone prediction.

**V. CONCLUSION**

The paper demonstrates use of Bayesian classification on the training data available to predict the susceptibility landslide for a specific land region which is then clustered using k-means to analyze that region of landslip susceptibility. Thus, a wide range of illustrative variables important for landslip risk forecast were incorporated by continuous classification using k-means. Hence by means of using clustering and classification (Bayes/kmeans) prediction of landslide is much better compared to earlier approaches like SHALSTAB. As a future work dynamic cluster formation can be used to decide best possible number of clusters and using Bayes theorem to determine cluster of required testing data. Best possible number of clusters can be decided by using Modified partition entropy, Fuzzy Performance Index etc. Once the best possible number of clusters is determined as per their landslide susceptibility Bayes theorem can be used for prediction.

**REFERENCES:**


[4] Landslide Susceptibility Mapping by Neuro-Fuzzy Approach in a Landslide-Prone Area (Cameron Highlands, Malaysia) Biswajeet

ISSN: 2278 – 1323 All Rights Reserved © 2014 IJARCET


[7] An artificial neural network application to produce debris source areas of Barla, Besparmak, and Kapi Mountains (NW Taurids, Turkey) M. C. Tunusluoglu1, C. Gokceoglu1, H. Sonmez1, and H. A. Nefeslioglu2.


[9] Weighted Naive Bayesian Classifier Alhammady H.

[10] Locally weighted naïve bayes Eibe Frank, Mark Hall, Bernhard Pföringer.

[11] Improved Bayesian Classification Data mining for early warning landslide susceptibility model using GIS.

[12] A digital terrain model for mapping shallow landslide potential (SHALSTAB) Dietrich WE, Montgomery DR.


[15] Prediction and Analysis of Landslide Based on Fuzzy Theory Chenguang JIANG, Jian-guo PENG Chun-qiao YUAN, Guo-hui WANG, Yong HE, Bo LIU.

