Comparison between the Naïve Bayes and Hierarchical Clustering to Classify The Global Landslide Catalog for the Prediction of the Landslide.

Poonam Verma, Charu Negi, Nisha Chandran, N. S. Bohra

Abstract: ***Machine Learning has been used since long to identify the features of a given datasets that are important for the prediction. Landslides are complex events taking place in the various regions of the world. It is the movement of the debris, soil or rocks from an upper plane in downward direction. Identification of the features that are used for the Landslide involves consideration of various categories of parameters. Present paper studies about the performance comparison between a supervised algorithm Naïve Bayes and unsupervised algorithm Hierarchical Clustering. Naïve Bayes is a non parametric supervised algorithm that can be used for the forecasting purposes in the field of Agriculture, Economics, Aviation etc, whereas Hierarchical Clustering is used to partition the available instances of a dataset into optimal homogeneous groups on the basis of the similarities between the datapoints. The present paper draws a comparison between the accuracy of the Naïve Bayes and Hierarchical Clustering for the prediction of the Landslide dataset. The dataset used is the Global Landslide Catalog that has important parameters like date, location coordinates, country, trigger of the event, continent etc. Before the implementation of both the algorithms, reduction of the parameters is carried out using subset evaluation of the parameters and considering only the most important.***

Keywords: Landslide prediction, GLC, Hierarchical Clustering, Naïve Bayes, Multinomial Text, Machine Learning

I. INTRODUCTION

Landslide the downward movement of the debris, land and rocks. Landslides have been a reason of many causalities and property loss. Prediction about the landslides can help the disaster management response and mitigation. Rainfall is generally the most common trigger event for the landslides apart from other parameters. NASA has developed LHASA (Landslide Hazard Assessment for Situational Awareness) model to identify where and when the landslide hazards are possibly developing, which relies heavily on the susceptibility map[19]. Landslide probability prediction has been carried out in this paper that combines various parameters in combination with the machine learning classifiers. This paper is specifically considering the 13 different indicators separated from the 35 listed parameters.

Revised Manuscript Received on June 30, 2020.

Poonam Verma, Graphic Era Hill University, Clement Town, Dehradun. Email: poonamddn18@gmail.com

Charu Negi , Graphic Era Hill University, Clement Town, Dehradun.

Nisha Chandran, Graphic Era Hill University, Clement Town, Dehradun.

N. S. Bohra, Department of Management Studies, Graphic Era Deemed to be University, Dehradun, Uttarakhand

The first part of the paper is used to find out the important indicators that can help in the prediction of the landslide. The problem has been formulated as a task of binary classification of landslide occurrence in the given region depending on the trend or pattern of the previous 5 years in the dataset.[16] Simpler models provide only slightly inferior predictions to more complex models, and should guide the way for a more widespread application of data mining in regional landslide prediction[17]. Future research may want to develop: better prediction model with less time complexity and better prediction of temporal forecasts of landslides.

II. NAÏVE BAYES:

Naïve Bayes is considered to be a probabilistic supervised algorithm which is commonly used to classify the string datatype datasets with the availability of the two variants of Naïve Bayes namely Naïve Bayes Multivariate Bernoulli Model and Naïve Bayes Multinomial Model. The basic reason behind the popularity of Naïve Bayes being used for the string datatype dataset is that each document is represented by a vector of the binary attributes indicating the frequency of each word captured in the document. Each document is considered to be an event. Each vector consists of words. Given a vocabulary, each space set has the vector consisting of the space of words.

- *Multinomial Naive Bayes*: Feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This is the event model typically use for document classification.
- *Bernoulli Naive Bayes:* In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, however limiting the response to 0 or 1, indicating the presence or absence of the word in the document.

A Bayesian network consists of a directed acyclic graph and a set of conditional probabilities. The graph consists of nodes that represent attributes and arcs represent attribute dependencies. Attribute dependencies are quantified by conditional probabilities for each node. Learning Bayesian networks from data has two elements: structure learning and parameter learning. Bayesian networks are often used for classification problems, in which a learner attempts to

construct a classifier from a collection of labeled training instances. The resulting classifiers are called Bayesian network classifiers. Naive Bayes is the simplest one of Bayesian network classifiers. It assumes that all attributes are independent of each other given the context of the class. This is the so-called attribute independence assumption.

III. HIERARCHICAL CLUSTERING

Hierarchical Clustering requires the creation of the clusters that has ordering from top to bottom. The advantage of the Hierarchical clustering is sometimes there will be no target variable except the independent variables. Two variants commonly used of Hierarchical Clustering are : Agglomerative and Divisive Clustering. The entire data is divided into a set of groups and these groups are known as clusters and the procedure of creating clusters is known as clustering. Steps taken for carrying out Hierarchical Clustering are given as below :

If N data points are required to be clustered, and an N*N distance matrix is available, then following steps are required to be taken:

1. Create N clusters with 1 item belonging to each one of the cluster. Let the distance between the clusters be the same as the similarities between the items consisting in each cluster.
2. Find the most similar pair of clusters that can be merged into a single cluster, thereby reducing the cluster by 1.
3. Now the distances between the freshly merged cluster is computed with the old clusters.
4. Repeat steps 2 and 3, until all the data point are clustered into a single cluster with a size of N.

Two variants of the Hierarchical Clustering are: Divisive Method and Agglomerative Method.

3.1 Divisive Method:

It is also known as the Top Down Clustering where all the observations are assigned to one cluster and then the cluster is divided in two clusters with alike data points. Each cluster is recursively divided into two clusters till there remains only one observation for one cluster. Divisive methods are considered to be more efficient than the Agglomerative Algorithms.

3.2 Agglomerative Methods:

This method is also known as Bottom Up clustering method where each data point is assigned to one cluster and then recursively the clusters are merged to form one cluster. Agglomerative Method requires the proximity matrix that consists of the distance between each point using a distance function. At each step, the matrix is updated.

Hierarchical Algorithms further have single linkage, complete linkage and average linkage. In the Single Linkage Hierarchical algorithms, the distance between any two clusters is the minimum distance between the two data points in the cluster. In the Complete Linkage Hierarchical Algorithms, the distance between the two data points is the maximum possible distance between two data points in each cluster. In Average Linkage Hierarchical Algorithms, the distance between two clusters is defined as the average distance between each data point in one cluster to the other data point in the other cluster.

IV. RELATED WORK

Machine learning is one of the most sought after methods for the classification process in the datasets. Chen,W. et al.[9] studied the prediction of the landslides in the 171 locations selected in china and compared Kernel Linear regression , naïve Bayes and RBF network using chi-squared attribute conditioning. Tien Bui, D. et al. [10] devised a novel ensemble method combining a functional algorithm, stochastic gradient descent and Adaboost were used to predict the landslides in the 98 locations selected by the authors. Pham, B.T.et.al.[11] proposed a novel algorithm using Bagging and Naïve Bayes for the landslide prediction in the north Vietnam. Additionally the comparison was carried out between the Rotation Forest-based Naïve Bayes Trees (RFNBT), single Naïve Bayes Trees (NBT), and Support Vector Machines (SVM). C. N. Madawala et.al.[12] proposed an ensemble method which was based on SVM and Naïve bayes for the prediction of the landslide in Ratnapur area, India. Husam A. H. et.al. [13] compared Random Forest (RF), Naive Bayes (NB), and Boosted Logistic Regression (LogitBoost) on the datasets for predicting the landslide suspecitibility. Dang, V.et al.[14] used Random Forest Classifier to generalize the classification boundary that separates the input information of ten landslide conditioning factors . Basu, T. et. Al.[15] utilized 6 different spatial parameter for 6 different clusters and implemented clustering to predict the landslide map susceptibility. Tang R., [20] made use of probabilistic methods, cluster analysis and artificial neural networks to predict the landslide in Hubei, China.

V. METHODOLOGY

In the present paper, the dataset of the landslides has been obtained as GLC (Global Landslide Catalog)from the official website of NASA. The global landslide catalog was developed to identify the landslide events triggered around the globe independent of their location or impact on the losses incurred. GLC has been compiled since the year 2007 from various sources of media, news articles, disaster databases or scientific papers. The GLC has been compiled by scientists, interns and other colleagues at NASA . The GLC has been cited more than 75 times in peer-reviewed articles. The GLC is a part of the Cooperative Open Online Landslide Repository (COOLR), and can be downloaded on Landslide Viewer.

This datasets consists of 35 columns with the following headings:

Geometric id, Object id, Date, Time, Country, Nearest point, Hazardtype , landslide, trigger, storm name, fatalities, injuries, source_name, Source_link , location, landslide_severity, photo_link , Source, sourceid , country_name, Near Location, distance, admin1, admin2, country, countrycode ,continent code, keycode, version, user_id, changeset, latitude and longitude.

In this paper, we have implemented two different classification algorithms namely Naïve Bayes Multinomial Classifier and Hierarchical Clustering on the

datasets. Naïve Bayes Multinomial text classifier is used to popularly handle text classification problems. It is computationally efficient and easy to be implemented. Two variants of the Naïve Bayes used commonly are the multivariate Bernoulli event model and the multinomial event model. Multinomial Naive Bayes computes class probabilities for each word in the given document. A document is treated as a sequence of words and it is assumed that each word position is generated independently of every other. For classification, we assume that there are a fixed number of classes, c = {1, 2, . . . , m}, each with a fixed set of multinomial parameters.

This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class. Naïve Bayesian classifiers must estimate the probabilities of a feature having a certain feature value. Continuous features can have a large (possibly infinite) number of values and the probability cannot be estimated from the frequency distribution. This can be addressed by modelling features with a continuous probability distribution or by using discretisation. We evaluate Naive Bayes using both discretisation (NBD) and kernel density estimation (NBK). Discretisation transforms the continuous features into discrete features, and a distribution model is not required. Kernel density estimation models features using multiple (Gaussian) distributions, and is generally more effective than using a single (Gaussian) distribution.

VI. EXPERIMENTAL RESULTS

Machine Learning Algorithms perform better on the selected attributes rather than the redundant attributes or noisy features. Hence many Feature Selection algorithms have been also developed to help the algorithms perform better. Instance based methods are more prone to the irrelevant attributes and it has been proved well in many experiments that the number of irrelevant attributes in the dataset effects the result.

In the present paper , we have used the wrapper method of the classifier evaluator that Evaluates the worth of an attribute by using a user-specified classifier. We have used the ZeroR learning scheme for the attribute classifier where ZeroR always predicts the most frequent category value in the training data for problems with a nominal class value, or the average class value for numeric prediction problems.

Number of folds for accuracy estimation: 5

Ranked attributes:

1. the_geom
2. id
3. date_
4. time_
5. country
6. nearest_pl
7. hazard_typ
8. landslide_
9. trigger
10. storm_name
11. injuries
12. source_nam
13. location_a
14. landslide1
15. countrynam
16. distance
17. continentc
18. latitude
19. longitude

Dependent on the results obtained from attribute selectors , out of the defined 35 attributes, 19 important attributes based on the ranking of the attributes have been selected.

VII. MODEL AND EVALUATION SET RESULTS

Class attribute: landslide
Method Used: Hierarchical Clustering
Classes to Clusters:

```
   0    1  <-- assigned to cluster
1536    1 | Medium
 249    0 | Small
 176    0 | Large
  18    0 | Very_large
   9    0 | unknown
   3    0 | medium
   5    0 | large
   1    0 | small
   1    0 | Extra Large
```

Model and Evaluation Results:

Class attribute: landslide
Method used: Naïve Bayes

Confusion Matrix of Naïve Bayes

```
   a   b   c   d   e   f   g   h   i   <-- classified as
1081 305 114  23   5   0   9   0   0 |   a = Medium
  62 172   7   3   3   1   1   0   0 |   b = Small
 102  20  42   7   4   0   1   0   0 |   c = Large
  11   3   3   0   1   0   0   0   0 |   d = Very_large
   2   2   2   0   2   0   1   0   0 |   e = unknown
   0   2   0   0   0   1   0   0   0 |   f = medium
   0   3   1   0   0   0   1   0   0 |   g = large
   0   0   0   1   0   0   0   0   0 |   h = small
   0   0   0   1   0   0   0   0   0 |   i = Extra Large
```

Table 1: Comparison Table between Naïve Bayes and Hierarchical Clustering

Scheme:	**NaiveBayes**	**Hierarchical Clustering**
Instances:	1999	1999
Attributes:	19	19
Time taken	0.05 seconds	29.65 seconds
Correctly Clustered Instances	1299 = 64.9825 %	1536 = 76.8384%
Incorrectly clustered instances	700 = 35.0175 %	463.0 = 23.1616 %

In the above shown table, we have compared the two machine learning algorithms of two different categories , Naïve Bayes , a popular Supervised learning algorithm and Hierarchical Clustering , a popular Unsupervised Learning Algorithm. Since the selected 19 attributes have been taken into consideration for the classification purpose, the results indicate that the unsupervised clustering tends to fare better than the supervised algorithm , although there is a huge compromise on the time complexity of both them.

VIII. FUTURE SCOPE

The present data consists of Nominal and Numeric data. Many classifiers cannot be implemented due to the present format of the dataset. We can preprocess the data and the format of the dataset can be made such that the other classifiers can be implemented on the landslide prediction datasets, thereby improving the accuracy of landslide. From the above listed results, it can be understood that the Hierarchical clustering shows promising results if selected attributes of only high weight are selected. In the future, we would like to compare more clustering algorithms with the popular Naïve Bayes Algorithm.

REFERENCES

1. Kirschbaum, D. B., Adler, R., Hong, Y., Hill, S., & Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: method, results, and limitations. Natural Hazards, 52(3), 561–575. doi:10.1007/s11069-009-9401-4.
2. Kirschbaum, D.B., T. Stanley, Y. Zhou (In press, 2015).
3. Spatial and Temporal Analysis of a Global Landslide Catalog. Geomorphology. doi:10.1016/j.geomorph.2015.03.016.
4. Kirschbaum, D. and Stanley, T. (2018). Satellite-Based Assessment of Rainfall-Triggered Landslide Hazard for Situational Awareness. Earth's Future. doi:10.1002/2017EF000715.
5. "New NASA Model Finds Landslide Threats in Near Real-Time During Heavy Rains", NASA, Greenbelt, MD (March 22, 2018).
6. Stanley, T., & Kirschbaum, D. B. (2017). A heuristic approach to global landslide susceptibility mapping. Natural Hazards, 87(1), 145-164. doi:10.1007/s11069-017-2757-y
7. Kirschbaum, D., Stanley, T., & Yatheendradas, S. (2016). Modeling landslide susceptibility over large regions with fuzzy overlay. Landslides, 13(3), 485-496. doi:10.1007/s10346-015-0577-2
8. "A Global View of Landslide Susceptibility", NASA Earth Observatory, Greenbelt, MD (March 30, 2017)
9. Juang.C.S, et.al., “Using citizen science to expand the global map of landslides: Introducing the Cooperative Open Online Landslide Repository (COOLR)”, PLOS.
10. Chen,W., Yan,X., Zhao,Z. et al. Spatial prediction of landslide susceptibility using data mining based kernel logistic regression, naïve Bayes and RBF Network models for the Long County area (China).Bull Eng Geol Environ 78, 247-266 (2019)
11. Tien Bui, D.; Shahabi, H.; Omidvar, E.; Shirzadi, A.; Geertsema, M.; Clague, J.J.; Khosravi, K.; Pradhan, B.; Pham, B.T.; Chapi, K.; Barati, Z.; Bin Ahmad, B.; Rahmani, H.; Gróf, G.; Lee, S. Shallow Landslide Prediction Using a Novel Hybrid Functional Machine Learning Algorithm. *Remote Sens.* 2019, *11*, 931.
12. Pham, B.T., Prakash, I. A novel hybrid model of Bagging-based Naïve Bayes Trees for landslide susceptibility assessment. Bull Eng Geol Environ 78, 1911–1925 (2019).
13. C. N. Madawala, B. T. G. S. Kumara and L. Indrathilaka, "Novel machine learning ensemble approach for landslide prediction," *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo, Sri Lanka, 2019, pp. 78-84.
14. Husam A. H. Al-Najjar, Bahareh Kalantar, Biswajeet Pradhan, and Vahideh Saeidi "Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms", Proc. SPIE 11156, Earth Resources and Environmental Remote Sensing/GIS Applications X, 111560K (3 October 2019);
15. Dang, V., Dieu, T.B., Tran, X. et al. Enhancing the accuracy of rainfall-induced landslide prediction along mountain roads with a GIS-based random forest classifier. Bull Eng Geol Environ 78, 2835–2849 (2019).
16. Basu, T., Pal, S. A GIS-based factor clustering and landslide susceptibility analysis using AHP for Gish River Basin, India. Environ Dev Sustain (2019).
17. Melchiorre C, Matteucci M, Azzoni A, Zanchi A (2008) Artificial neural networks and cluster analysis in landslide susceptibility zonation. Geomorphology 94:379–400.
18. Jakob M (2000) The impacts of logging on landslide activity at Clayoquot Sound, British Columbia. Catena 38:279–300.
19. Pachauri AK, Pant M (1992) Landslide hazard mapping based on geological attributes. Eng Geol 32:81–100.
20. Pérez-Peña JV, Azañón JM, Azor A, Delgado J, González-Lodeiro F (2009) Spatial analysis of stream power using GIS: SLk anomaly maps. Earth Surf Process Landf 34:16–25.
21. Tang, R., Kulatilake, P.H.S.W., Yan, E. et al. Evaluating landslide susceptibility based on cluster analysis, probabilistic methods, and artificial neural networks. Bull Eng Geol Environ (2020).