ROAD ACCIDENT ANALYSIS USING MACHINE LEARNING

Mr. S. Munusamy,
Assistant Professor ,Department of MCA, Sri Balaji Chockalingam Engineering College, Arni,muns.samy@gmail.com

ABSTRACT

There are many inventories in automobile industries to design and build safety measures for automobiles, but traffic accidents are unavoidable. There are a huge number of accidents prevailing in all urban and rural areas. Patterns involved with different circumstances can be detected by developing an accurate prediction models which will be capable of automatic separation of various accidental scenarios. These clusters will be useful to prevent accidents and develop safety measures. We assume that by using certain experimental methods we achieve optimum possibilities for minimizing injuries using low budget capital. This paper summarizes the performance of three machine learning paradigms applied to modeling the severity of injury that occurred during traffic accidents. We considered neural networks trained using hybrid learning approaches, support vector machines, decision trees and a concurrent hybrid model involving decision trees and neural networks. Experiment results reveal that among the machine learning paradigms considered the hybrid decision tree-neural network approach outperformed the individual approaches.

1.INTRODUCTION

Today, traffic safety is one of the main priorities of governments. Considering the importance of topic, identifying the factors of road accidents has become the main aim to reduce the damage caused by traffic accidents. Consider the issue of providing a safety travelling measures on the road network within the urban and suburban one of the fundamental principles governing the engineering, traffic and transportation planning. Nearly 3,500 people die on the world's roads every day lots of people are injured or disabled every year.There are several problems with current practices for prevention of the accidents occurred in the localities. The database we will use is available officially by many institutes and government websites. The data collected will be analyzed, integrated and grouped together based on different constraints using the best suited algorithm. This estimation will be useful in evaluating and determining the incident fault and causes.It will also be helpful while

making roads and bridges as a reference to avoid the same problems faced before. The predictions made will be very much useful to plan the management of such problems.

2.PROPOSED WORK

In Machine Learning Algorithms there are many types of learning systems such as the Supervised learning, Semi supervised learning, unsupervised learning and reinforcement learning. Out of all the kind of machine learning techniques and approached, we use the Supervised learning approach which is dynamic in processing for the Road Accident Analysis. From the Supervised learning approaches, we select the three major compatible techniques KNN, Decision Tree and Naive Bayes.

3. DATASET PREPARATION

The important part of implementing an algorithm is only possible by preparing the dataset for the algorithm to process and to produce result. Effective and complete data records of incidents are the most significant, and by applying machine learning approaches the primary need to achieve better results. For such dataset to be prepared we have to perform certain data processing methods such Collecting data, Pre-Processing the Dataset and Features Selection.

3.1 Collecting Data

There are many road accidents that is been accruing in India, all of in different places and location. For the machine learning to predict the severity accurately and effectively we need to gather a large amount of dataset of the road accident records. To collect such number of datasets we have collected the datasets from the OGD platform India, where it consists a large amount of accident records of 62,000+ that have been occurred in India from the year 2005-2017. These data records are been used for training and testing the machine learning algorithms. Using the python library, we use to train the algorithm.

3.2 Pre-Processing the Dataset

The Process of pre-processing the dataset is to organize the dataset so that the irrelevant data’s do not affect the accuracy and performance of the road accident analysis predictions. The supervised learning algorithm might throw some false result due to the irrelevant data present int the dataset. To remove such data, we pre-process the dataset to organize and fill in the correct features in the dataset. This process is iterative until the dataset features are of correct datatype and persistent.

3.3 Features Selection

Features refers to the data's present in the dataset that is gathered. To handle large number of features we use some algorithms that handles the features in the dataset such as Recursive Feature Elimination and Tree-Based feature selection. The key of using the Features Selection methods is to get more accurate prediction, feature selection to avoid any critical factor while training set. In python we have Sklearn machine learning library which handles the feature selection methods essentially.

A) Recursive Feature Elimination:

The Recursive Feature Elimination (RFE) method is a feature selection approach. It works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

The objective of recursive component end (RFE) is to choose includes by recursively thinking about littler and littler arrangements of highlights. In the first place, the estimator is prepared on the underlying arrangement of highlights and the significance of each element is gotten either through a coef_ property or through a feature_importances_ characteristic. At that point, the least significant highlights are pruned from current arrangement of highlights. That technique is recursively rehashed on the pruned set until the ideal number of highlights to choose is in the end come to.

a. Setting $F = \{1, \ldots, n\}$

b. Where we get w^* as the solution on an SVM on the data set restricted to features in F (Minimize estimation of $R(\alpha, \sigma)$ *wrt.* α).

c. If we select the top features as ranked by the $|w_i^*|$*'s* (Minimize the estimate $R(\alpha, \sigma)$ *wrt.* α and under a constraint that only limited number of features must be selected).

d. Then the process gets iterated from the second stage.

B) Feature Importance:

Feature Importance gives you a score for each component of your information, the higher the score increasingly significant or applicable is the element towards your yield variable. Feature Importance is an inbuilt class that accompanies Tree Based Classifiers, we will utilize Extra Tree Classifier for removing the best 10 features for the dataset.

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

- *ni sub(j)* = the importance of node j.
- *w sub(j)* = weighted number of samples reaching node j.
- *C sub(j)* = the impurity value of node j.
- *left(j)* = child node from left split on node j.
- *right(j)* = child node from right split on node j.

5.THE PROPOSED METHODOLOGY

The proposed systemis a middleware that uses techniques involving data slicing, data analysis and pre-processing of data for secure and optimized results. Data pre-processing is an important and mandatory step for any machine learning model because it involves steps like feature scaling to get exact values. As machine learning models deals with values with close proximity, splitting the dataset in training is necessary as the dataset contains huge amount of features within which various unwanted features are also present that are not required. After pre-processing the

important features are extracted for the problem and based on the dataset they are analysed graphically. Further the model is trained with three algorithms one by one using the important features and at last accuracy is compared for the three algorithms based on true and predicted results.

- k-NN
- Decision tree
- Naïve Bayes

i. k-NN

The k-nearest neighbor algorithm (k-NN) is a non-parametric approach used for classification and regression. In both cases, the input consists of the closest feature space training examples k. The output depends upon the dataset that is induced to the kNN algorithm.

In k-NN characterization, the yield is a class enrollment. An item is ordered by a majority vote of its neighbors, with the article being appointed to the class generally normal among its k closest neighbors (k is a positive whole number, commonly little). On the off chance that k = 1, at that point the item is just doled out to the class of that solitary closest neighbor.

k-NN is a type of instance-based learning, or lazy learning, in which the function is only approximated locally and all computations are deferred until classification.

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

a.The 1-nearest neighbor classifier

The most intuitive nearest neighbor type classifier is the one nearest neighbor classifier that assigns a point x to the class of its closest neighbor in the feature space, that is

$$C_n^{1nn}(x) = Y_{(1)}.$$

As the size of training data set approaches infinity, the one nearest neighbor classifier guarantees an error rate of no worse than twice the Bayes error rate (the minimum achievable error rate given the distribution of the data).

b.The weighted nearest neighbor classifier

The k-nearest neighbor classifier can be viewed as assigning the k nearest neighbors a weight *1/k* and all other *0 weight*. This can be generalized to weighted nearest neighbor classifiers. That is, where thei^{th}nearest neighbor is assigned a weight $w_{ni,}$ with

$$\sum_{i=1}^{n} w_{ni} = 1$$

An analogous result on the strong consistency of weighted nearest neighbor classifiers also holds. Subject to regularity conditions on the class distributions the excess risk has the following asymptotic expansion.

Let C_n^{wnn} denote the weighted nearest classifier with weights $\{w_{ni}\}_{i=1}^{n}$. Subject to regularity conditions on the class distributions the excess risk has the following asymptotic expansion.

$$\mathcal{R}_{\mathcal{R}}(C_n^{wnn}) - \mathcal{R}_{\mathcal{R}}(C^{Bayes}) = \left(B_1 s_n^2 + B_2 t_n^2\right)\{1 + o(1)\},$$

for constants B_1 and B_2 where

$$s_n^2 = \sum_{i=1}^{n} w_{ni}^2 \quad \text{and} \quad t_n = n^{-2/d} \sum_{i=1}^{n} w_{ni}\{i^{1+2/d} - (i-1)^{1+2/d}\}.$$

The optimal weighting scheme $\{w_{ni}\}_{i=1}^{n}$ that balances the two terms in the display above, is given as follows: set $k^* = \lfloor Bn^{\frac{4}{d+4}} \rfloor$,

$$w_{ni}^* = \frac{1}{k^*}\left[1 + \frac{d}{2} - \frac{d}{2k^{*2/d}}\{i^{1+2/d} - (i-1)^{1+2/d}\}\right] \text{for } i = 1, 2, \ldots, k^* \text{ and } w_{ni}^* = 0 \text{ for } i = k^*+1, \ldots, n.$$

With optimal weights the dominant term in the asymptotic expansion of the excess risk is $\mathcal{O}(n^{-\frac{4}{d+4}})$. Similar results are true when using a bagged nearest neighbor classifier.

ii. Decision Tree

The general rationale of utilizing Decision Tree is to make a preparation model which can be utilized to anticipate class or estimation of target factors by taking in the choice standards from preparing information.

The choice tree calculation attempts to take care of the issue by utilizing a tree portrayal. Each interior hub of the tree compares to a property while each leaf hub speaks to a class mark. Property Selection assumes a significant job in Decision tree. Property choice is finished by considering factors like Information Increase, Gini Index, and so on.

The qualities for each trait are determined dependent on these criteria and put away. The characteristic with the high worth will be the base of the tree. Gini Index is a measurement to quantify how frequently an arbitrarily picked component would be mistakenly distinguished. Henceforth, a characteristic with a lower Gini Index would be liked.

A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogeneous). ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is equally divided then it has entropy of one.

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

i. Naïve Bayes Classification:

The Bayes theory is simpler to structure and it is appropriate for applications including enormous informational collections. It takes a shot at the Bayes hypothesis of likelihood to foresee the class of an obscure informational index. A Naive Bayes classifier expects that the nearness of a specific component in a class is irrelevant to the nearness of some other component. Guileless Bayes model is simpler to fabricate and valuable for working with huge informational collections. The given mishap informational collection is first prepared, and afterward a model is made from which expectation should be possible. Expectation should be possible for required conditions and the criticality of the mishap can be anticipated.

For a class variable Y, and feature variables (X1, X2, …,Xn), Bayeshypothesis gives the following relation:

$P(Y|X1\ldots, Xn) =$

$P(Y)\ P(X1\ldots, Xn|Y) / P(X1\ldots, Xn)$

Using Naïve Bayes assumption,

$P(Xi|Y, X1\ldots, Xi\text{-}1\ldots, Xn) = P(Xi|Y)$

For all values of i, the relation can be given by,

$P(Y|X1\ldots, Xn) = P(Y)\ P(X1\ldots, Xn|Y) / P(X1\ldots, Xn) \propto P(Y)\ \Pi P(Xi|Y)$

$Y = arg\ max = P(Y)\ \Pi\ P(Xi|Y)$

4. CONCLUSION

The rules revealed different factors associated with road accidents at different locations with varying accident frequencies. High frequency accident locations mostly involved certain regions. In moderate-frequency accident locations, colonies near local roads and intersection on highway roads are highly dangerous for pedestrian. Low frequency accident locations are scattered throughout the district and the most of the accidents at these locations were not critical. Our approach extracted some useful hidden information from the data which can be utilized to take some preventive effects in these locations.