ROAD ACCIDENT PREDICTION BY SUPPORT VECTOR MACHINE USING POLYNOMIAL KERNEL AND SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOT)

Neetu Panwar[1] , Ritu Rani[2]
Ngfcet, Palwal
(E-mail:Neetupanwar26@gmail.com)[1]
(E-mail:ritugarg101@gmail.com)[2]

Abstract: Traffic mishaps are a big problem nowadays in many continents and countries that, necessitate staid action because of the huge figure of fatalities ensuing in human casualties and possessions losses. However, machine learning holds the mechanism to predict and forecast such problems to analyze and provide the data results for further conservative management to protect human lives and to stop irreplaceable injuries in terms of costs to lives. This scheme using the supervised machine learning model based on classification classifies the severity of traffic accidents. The scheme applies Support Vector Machine incorporating Polynomial Kernel along with Synthetic Minority Over-Sampling Technique (SMOT) for the prediction, performance, and accuracy. Therefore, the scheme can fetch the Accuracy of 67.30%, Recall of 67.30%, Precision of 72.50%, and F1 score of 65.35% over the 1796 records of data provided by NHAI, India.

Keywords*: Road Accidents, Machine Learning,* Natural Language Processing, Synthetic Minority Over-Sampling Technique, *Support Vector Machine.*

1. INTRODUCTION

A traffic accident is an unexpected and unintentional event on the Road involving vehicles with or without other road users resulting in human casualties and/or property losses . Various accidents include victims who died, serious and minor injuries, as well as material losses that were not small in number . In general, traffic accidents that occur are caused by several factors, such as human negligence, road conditions, vehicle eligibility, and not yet optimal traffic law enforcement.

There are around 752 pedestrians and 786 motorized vehicle users killed in the world per day (WHO, 2019). Several countries still have fatalities due to traffic accidents with high mortality rates as reported by the Global Status Report on Road Safety in 2015[2], including China, India, Nigeria, Brazil, and Malaysia. Traffic accidents in India by WHO are considered to be the third biggest killers, under coronary heart disease and tuberculosis (TB) . Investigation of road accident circumstances at state and city points demonstrates that there is an enormous deviation in casualty risk crosswise states and cities. Fatality risk in 16 out of 35 states and union territories is higher than the all-India average. Although the burden of road accidents in India is marginally lower in its metropolitan cities, almost 50% of the cities face higher fatality risk than their counterparts. In general, while in many developed and developing countries including China, the road safety situation is generally improving, India faces a worsening situation. Without increased efforts and new initiatives, the total number of road traffic deaths in India is likely to cross the mark of 250,000 by the year 2025 .

Accident data will be very meaningful if dug properly so that it can find knowledge from the data and is used to obtain hidden information . By applying the Data Mining technique, it will solve the problem by analyzing data owned by the National Highway Authority of India.

Data Mining is the discovery of new information by looking for certain patterns or rules from a very large amount of data. There are three Data Mining techniques, namely Association, Clustering, and Classification. In this study, the authors used classification techniques. Classification is the process of finding a model or function that explains or distinguishes data concepts or classes intending to estimate the unknown class of an object. One classification technique is the Support Vector Machine (SVM) .

Support Vector Machine (SVM) was developed by Boser, Guyon, Vapnik and was first presented in 1992 at the Annual Workshop on Computational Learning Theory . This method is a machine learning method (Learning Machine) to find the best separating function [8, 10] that separates two classes in the input space. Today SVM has been successfully applied to real-world problems (real-world problems), and in general, provides better solutions than conventional methods such as Artificial Neural Networks . SVM also works well with high-dimensional data sets.

When proposing solutions for road safety, the security forces security of the different countries, as well as their traffic agencies, are usually focused on solving problems related to the basic triangle driver-vehicle infrastructure, where the use of regulatory measures such as control radars speed, drug and alcohol controls, or inspections compliance controls techniques are considered sufficient actions to reduce accidents traffic on

your part. However, there are other types of actions that, without a doubt, would take advantage of the work in a more efficient way than this type of public organization performs . The collection of all the circumstances that involve a traffic accident, therefore it is a painstaking job. This study, however, ends in commissioning of relatively obvious and sometimes dubious measures to safeguard the vehicle occupant safety. The palliative measures approach on security based on artificial intelligence solutions and related with inferring knowledge and transferring it to infrastructure or automobiles, until Now, it has not been posed optimally, or has not wanted to assume for public bodies with competencies in the matter, despite being the agencies they have more information about it . Therefore, in times when the vehicle autonomous already begins to be a reality, under this scheme it is necessary to assume a proactive attitude in this line. In this work, we propose to process two massive sets of real accidents to recognize the causes that determine the severity of it and also are associated with each victim. The initial approach is to search for a function or classifier that allows us to predict the outcome of the occupants of a vehicle in an accident based on a series of characteristics. As a result of the study, statistical regression models and classification of the severity of the injuries produced will be applied to all occupants of a vehicle, in the event of a traffic accident. For this, each accident that occurred in INDIA will be analyzed based on a set of variables involved in it, where statistical models will allow us to infer the relationships between accidents and their contributing factors, thus allowing us to extract information that may be very useful in planning traffic accident reduction policies, proposals for improving existing infrastructures, signaling and communications with drivers, among others .

2. RELATED RESEARCH

2.1 CONVENTIONS OF TERMS ASSOCIATED WITH ACCIDENTS TRAFFIC

It is believed necessary to include a series of terms that clear any doubts that may arise concerning the types of vehicles, people involved in an accident, meteorology, and infrastructure, among others, that are necessary to understand exactly their relationship to the accident, thus as the accuracy of the data they represent.

1. Accident: refers to the one that produces personal injuries that occurred on public roads (including sidewalks) in which at least one road vehicle is involved. The police are informed within 30 days after their occurrence. An accident can lead to multiple casualties.

2. Vehicles Involved In Accidents: Vehicles whose drivers or passengers are injured, hitting and damaging a pedestrian or other vehicle whose driver or passengers are injured, or who contribute to the accident. Vehicles that collide after the initial accident that caused the injury are not included unless they aggravate the degree of injury or cause more casualties. Includes track-mounted pedal cycles. The severity or severity of accidents Based on the notes, definitions, and conventions that the Governments of India, the severity or severity of victims in a traffic accident in these countries are classified into three levels .

a. Fatal Accidents (Fatal): The usual international definition of an accident with a fatal consequence, as adopted by the Vienna Convention , is: ' A human victim who dies within 30 days after the collision due to injuries received in the accident '.

b. Serious Injury: The definition of a serious accident is less clear and may vary over time and depending on the location. For example, the definition in India refers to the hospitalization of a person due to injuries caused by the accident: fractures, contusions, internal injuries, burns (excluding friction burns), severe cuts, severe general shock that requires medical treatment, even if it does not result in hospitalization.

2.2 SUPPORT VECTOR MACHINE (SVM)

SVM is a collection of learning methods used for classification and regression which are included in the general linear classification section. The special property of SVM is minimizing empirical misclassification and maximizing geometric margins. So SVM is called Maximum Margin Classifiers. SVM is based on Structural Risk Minimization (SRM). Vector input of SVM map to a higher dimensional space where maximum hyperplane separation field is created. Two parallel hyperplanes are built on each side of the hyperplane that separates data. It is assumed that the greater the margin or distance between hyperplanes, the better the generalization of classification errors [15-19].

2.2.1. SVM in Linear Separate

Data The concept of SVM [15-19] can be explained simply as an attempt to find the best hyperplane that functions as a separator of two classes in the input space.

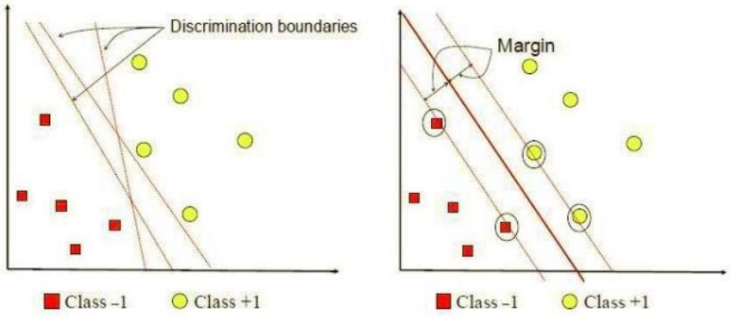

Figure 2.1. The SVM illustration finds the best hyperplane that separates two classes -1 and +1

Figure2.1 shows the members of two classes: +1 and -1. In class -1, it is symbolized by a red square while class +1 is symbolized by a yellow circle. "The classification problem can be translated by trying to find a hyperplane that separates the two groups. The best hyperplane separator between the two classes can be found by measuring the hyperplane margin to the nearest point of each class. The closest point is called a support vector. The solid line in Figure 2.1 to the right shows the best hyperplane, which is located in the middle of the two classes, while the red and yellow dots in the black circle are support vectors. Label of each $y_i \in \{-1,+1\}$ $i = 1,2,3, ..., l$. SVM linear classification hyperplane as in equation 2.1:-

$$\vec{w}.\vec{x} + b = 0 \qquad \text{(eq.2.1)}$$

Information:

$\vec{w}$ = weight vector

$\vec{x}$ = attribute input value

b = bias

$\vec{x}_i$ which belongs to class -1 (negative sample) can be formulated to meet inequality:

$$\vec{w}.\vec{x} \leq -1 \quad \text{(eq.2.2)}$$

Whereas $\vec{x}_i$ which belongs to class +1 (positive sample)

$$\vec{w}.\vec{x} \geq +1 \quad \text{(eq.2.3)}$$

Margin the largest can be found by maximizing the value of the distance between the closest distance and point, i.e. $1/\|\vec{w}\|$. This can be formulated as Quadratic Programming (QP) problem, i.e. look for the minimum point of equation 2.4, taking into account the constraints of equation 2.5.

$$\min_{\vec{w}} \tau(w) = \frac{1}{2} \|\vec{w}\|^2 \quad \text{(eq.2.4)}$$

$$y_i(\vec{x_i}.\vec{w} + b) - 1 \geq 0, \forall \quad \text{(eq. 2.5)}$$

x_i is input data, is output from data x_i, w, b is a parameter-the parameter in search for value. In the formulation above, want to minimize objective function $(w)=\frac{1}{2}$ or maximize quantity $\|\vec{w}\|^2$ concerning the barrier $y_i(wx_i + b) \geq 1$ 1. When output data $y_i=+1$, then delimiter to be $(wx_i + b) \geq 1$ delimiter as $wx_i + b) \leq -1$. In cases that are not feasible (infeasible) where some data may not be classified correctly, the mathematical formulation is as follows.

$$\min_{\vec{w}} \tau(w) = \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{l} t_i \quad \text{(eq. 2.6)}$$

$$y_i(\vec{x_i}.\vec{w} + b) + t_i \geq 1, \forall \quad \text{(eq. 2.7)}$$

$$t_i \geq 0, i = 1, \ldots, l$$

This problem can be solved by various computational techniques, including Lagrange Multiplier.

$$L(w, b, \alpha) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i((\vec{x_i}.\vec{w} + b) - 1)) \quad \text{(eq. 2.8)}$$

$$L(w, b, \alpha) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i(\vec{x_i}.\vec{w} - b \sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i \quad \text{(eq. 2.9)}$$

With the addition of the cosmetics, $\alpha_i \geq 0$ (value of the Lagrange coefficient). With minimizing L concerning w and b.

$$\frac{\partial}{\partial b} L(w, b, \alpha) = 0 \quad \text{(eq. 2.10)}$$

$$\frac{\partial}{\partial w} L(w, b, \alpha) = 0 \quad \text{(eq. 2.11)}$$

From equation 2.8 and equation 2.9 the following equation is obtained:

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \quad \text{(eq.2.12)}$$

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \quad \text{(eq.2.13)}$$

2.2.2. Kernel Trick and Non-Linear Classification in SVM

In general, problems in the real-world domain (real-world problem) are rarely linear. Most are non-linear. To solve non-linear problems, SVM is modified by entering the Kernel function. In non-linear SVM, the data x is first mapped by the function Φ (x) to a vector with a higher dimension ". In this new vector space, a hyperplane separating the two classes can be constructed. This is in line with the Cover theory which states. If a transformation is non-linear and the dimensions of the feature space are high enough, then the data in the input space can be mapped to a new feature space, where those patterns at high probability can be linearly separated. An illustration of this concept can be seen in figure 2.2. Figure 2.2 shows the data in the yellow class and the data in the red class in the two-dimensional input space that cannot be linearly separated. Furthermore, the function Φ maps each data in the input space to a new vector space with a higher dimension (dimension 3), where the two classes can be linearly separated by a hyperplane.

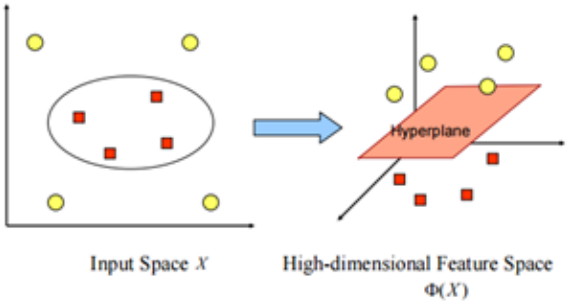

Figure 2.2: High Dimension Vector spaces using SVM

Figure 2.1. The function Φ maps data to vector spaces of higher dimensions so that the two classes can be linearly separated by a hyperplane. In solving non-linear problems in using the concept of the kernel in the workspace dimension height, by finding a hyper plane that can maximize margins between data classes. The hyperplane is useful in separating 2 class +1 and class –1 groups where each class has each pattern. In making decisions with the SVM method is used the kernel function $K(x_i, x^d)$. The kernel to use with the research shown in Equation below:

$$K(\ , x_d) = (X_i^T X_j + C)^d, \gamma > 0 \text{ (eq.2.14)} \quad \text{(eq. 2.14)}$$

Processing is done on training data Sequential training algorithm is used because it is a simple algorithm without takes a lot of time with calculation stages:

1. Initialization of various parameters,
   like α_i, γ, C, and ε.
   αi = alpha, to find support vector
   γ = gamma constant to control the speed
   C = slack variable
   ε = epsilon is used to find value error
2. Calculate the Hessian matrix obtained from multiplication between polynomial kernels and y
is vector 1 and -1.
The equation from the Hessian matrix is: $D_{ij} = y_i y^j (K(x_i, xj) + \lambda^2)$
3. Perform the following calculations until the interaction Data i to j:
   a. $Ei = \sum_j^i aj\ Dij$
   b. $\delta\alpha_i = \min(\max[\ \gamma(1 - E_i), \alpha_i]\ , C - \alpha_i$
   c. $\alpha i = \alpha i + \delta\alpha i$
4. Perform the three steps above in a manner repeat until it reaches the maximum limit
Iteration
5. Sequential learning process from stage 1 up to 4 will get value from support vector (SV), where the value SV = $(\alpha i > th\ resh\ oldSV)$. After that, it needs to be done the calculation of the value of bias b obtained from eq below:-

$$b = -\frac{1}{2} \sum_{i=0}^{n} \alpha_i y_i (x_i, x-) + \sum_{i=0}^{n} \alpha_i y_i (x_i, x+)) \quad \text{(eq. 2.15)}$$

Support vector classifier and support vector machine are two different things, although the two terms are often used in similar contexts [20]. Support vector classifier is a simpler concept. Although this chapter touches on the support of vector machine skins, we do not discuss in detail. However, we hope that the reader will be able to get intuition. You can think of this chapter as a continuation of the story.

2.3 SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

SMOTE is a good and effective oversampling technique to deal with over fitting in the oversampling process in dealing with imbalances in the module classes that are defective in the minority (positive) class. The SMOTE technique was chosen because of its effectiveness in handling class imbalance problems in software defect datasets . The SMOTE method increases the amount of minor class data to be equivalent to the major class by generating artificial data. The purpose of adding this data so that the amount of minor data is equivalent to major data. Artificial data or synthesis is made based on k-nearest neighbor (k- nearest neighbor). The number of k-nearest neighbors is determined by considering the ease of implementing it. Numerical scale artificial data generation differs from categorical. Numerical data are measured by their proximity to Euclidean distance, while categorical data is simpler by mode value. Calculation of distances between examples of minor classes with variable scale variables is carried out using the Value Difference Metric (VDM) formula [22-24], which is:

$$\Delta(X, Y) = W_x W_y \sum_{i=1}^{N} \delta(x_i, y_i)^r \quad (\text{eq. } 2.16)$$

Where :
Δ (X, Y) = the distance between observations X and Y
W_x, W_y = observed weight (negligible)
N = number of explanatory variables
R = 1 (Manhattan distance) or 2 (Euclidean distance)
$\Delta (x_i, y_i)^r$ = distance between categories, with the formula:

$$\Delta(V_1, V_2) = \sum_{i=1}^{N} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right| \quad (\text{eq. } 2.17)$$

Where:-
Δ (V1, V2) = distance between values V1 and V2
C_{1i} = number of V1 included in class i
C_{2i} = number of V2 included in class i
I = number of classes; i = 1,2, ..., m
C_1 = the number of values of 1 occurs
C_2 = number of 2 values occurred
N = number category
K = constant (usually 1)
Procedure for generating artificial data for:
1. Numeric Data
a. Calculate the difference between the main vector and its nearest k-neighbor.
b. Multiply the difference by the random number between 0 and 1.
c. Add the difference to the main value in the original main vector so that you get a new main vector.
2. Categorical Data
a. Choose the majority between the major vectors considered and their closest k-neighbors for face value. If the same value occurs then choose randomly.
b. Make the value of the sample data for the new class.
Illustration of data distribution after applying the SMOTE method can be seen in Figure 2.3.

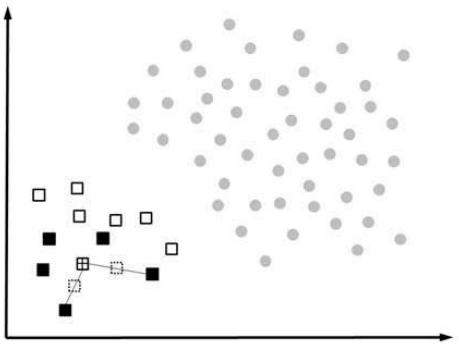

Figure 2.3 Illustration of SMOTE

Although SMOTE is quite effective in increasing the classification accuracy of minority data, there are still problems, among others, namely the occurrence of overgeneralization. The SMOTE synthesis data is still possible to spread to the minority and majority data so that it will reduce the classification performance .

2.4 LITERATURE REFERENCES

Unlike keyword-spotting as well as lexical affinity established by strictly ruled/predetermined basic rules, machine learning is chosen because it can search for patterns of rules/rules independently. Machine learning techniques are also preferred over concept-based techniques because the data to be processed has sufficient quantities and lower business costs. Some machine learning algorithms that are often used and proven optimal for SA are Naive Bayes, Logistic Regression, and Support Vector Machines. The application of the Naïve Bayes algorithm in SA was done by McCallum & Nigam and produced an accuracy of 87%, Kaur & Mohana with an accuracy of 68.15%, and Preety & Dahiya with an accuracy of 89%. Whereas the application of Logistics Regression in SA was carried out by Al-Tahrawi with an F1-Measure value of 86.5%, and Kaur & Mohana with an accuracy of 82, 15% And for Vector Machine Support.

3. PROPOSED ALGORITHM

Below is the model using a machine learning scheme for road accident prediction with high accuracy.

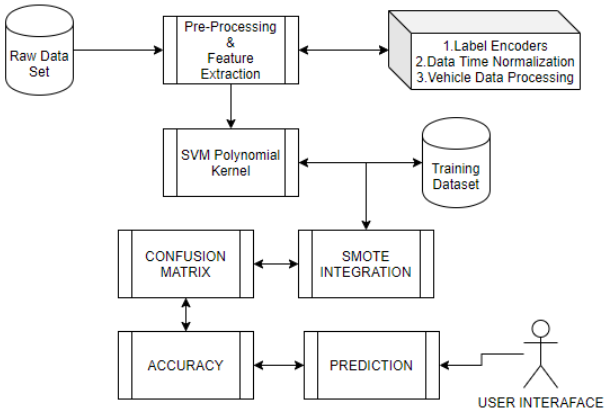

Figure 3.1: Proposed Scheme

3.1 PREPROCESSING

Usually, the data we get from flat files or databases is raw data. Machine learning classification algorithm works with data that will be formatted in a certain way before they start the training process. To prepare data for consumption by the machine learning algorithm, we must process it first and convert it to the right format.

1. Binarization: the binarization process is when we want to convert numeric variables into boolean values (0 and 1).
2. Mean Removal: mean removal is a common method in preprocessing techniques used in machine learning, eliminating the average is usually very useful from a variable, so the variable is in the middle at 0. We do it to eliminate the bias of the variable.
3. Scaling: usually in a dataset that is still raw, some variables have very variable and random values, so it is very important to scale these feature features, in my perspective these features are very large or small because of the nature of measurements.
4. Normalization: usually in preprocessing, the normalization process is to modify the values in a variable so that we can measure it on a general scale. In machine learning, we use various forms of normalization. Some of the most common forms of normalization aim to change the values so that the number becomes 1. Normalization of Level1 which refers to the Smallest Absolute Deviation works by ensuring that the sum of absolute values is 1 in each row. Normalization Level 2, which refers to the smallest square, works by ensuring that the sum of squares is 1. In general, the Level 1 normalization technique is considered to be stronger than the Level 2 normalization technique. The Level 1 normalization technique is strong because it is resistant to outliers in the data. Often, data tends to contain outliers and we cannot do anything about it. We want to use techniques that can safely and effectively ignore them during calculations. If we solve problems where outliers are important, then maybe normalizing Level 2 would be a better choice.
5. Label Encoding: when we classify, we usually deal with a lot of labels. These labels can be in the form of words, numbers, or something else. The machine learning function expects them to be numbered. So if they have become a number, then we can use it directly to start training. But this is not usually the case. In the real world, labels are made in the form of words, because words can be read by humans. we label training data with words so that mapping can be traced. To convert word labels to numbers, we need to use an encoding label maker. Label encoding refers to the process of transforming word labels into numerical forms. In the case of regression, if it contains variable categories and their values cannot be factored into levels, a dummy process is performed, each value in that variable becomes another variable.

3.2 PSEUDO CODE SVM WITH POLYNOMIAL KERNEL FUNCTION

The SVM method provides optimal solutions and speeds up the iteration process rather than using conventional solutions. The steps of the polynomial kernel method are as follows:-

1. Initialize parameter λ (lambda), γ (learning rate), C complexity), ε (epsilon), and maximum iteration.
2. Initialize αi = 0, and compute matrix D for i, j=1,2,…,n

$D_{ij} = y_i(K(x_1, x_j) + \lambda^2)$ #polynomial kernel call

3. Then, for each pattern, i=1,2, n, compute:

Step a. $E_i = \sum_{j=1}^{n} \alpha_j D_{ij}$

Step b. $\delta\alpha_i = \min\{\max[\gamma(1 - E_i), -\alpha_i] , C - \alpha_i\}$

Step c. $\alpha_i = \alpha_i + \delta\alpha$

4. The iteration will be stopped, if it is achieved maximum iteration or $M(|\delta\alpha|) < \varepsilon$, else go to step b.

After the above process is finished, then it will be obtained the α value and Support Vector. So that, the formula of road accident analysis in this research is as Equation is.

$$(x) = \sum_{i=0}^{n} \alpha_j y_i K(x, x_i) + b$$

Where b, bias value is

$$b = \frac{1}{2} \sum_{i=0}^{n} \alpha_i y_i K(x_i, x-) + \sum_{i=0}^{n} \alpha_i y_i K(x_i, x^+))$$

3.3 SMOTE ALGORITHMS

Minority data will be used as a model for making synthetic data with SMOTE. Data will get additional synthetic data to get the same ratio between minority data with majority data dataset has the same ratio (balance) to conclude with a confusion matrix.

The steps are as under:-

1. S-maj = Majority class sample
2. S-min = Minority class sample
3. N = synthetic sample
4. k1 = Predicts noise in the minority class
5. k2 = Major class neighbor that is used as a minor class boundary
6. k3 = Minor class neighborhood used to create synthetic data

Phase 1

1. S-min is the nearest neighbor search based on Euclidean distance.
2. S-min is the maker of minor classes from sample data and removes class samples major.
3. k2 to find the closest neighbors in the major class according to Euclidean distance.
4. S-bmaj to determine the major class borderline.
5. k3 determines the minor class that will be used to create synthetic data according to Euclidean distance.
6. S-imin determines a collection of informative minor classes.

Phase 2

1. I-w to determine the weighting of S-imin members.
2. S-w to select members from class I-w to be synthetic candidates.
3. S-p converts S-w members into probabilities.

Phase 3

1. Lm looking for clusters from the S-min sample.
2. S-min = S-min to make initials.
3. Performing repetitions to select samples from S-imin according to probability distribution (S-p)

for example L-k cluster members. Choose another sample, randomly from members of the L-k cluster, and generate synthetic data.

4. RESULTS AND SIMULATION

Analysis using traffic accident data in India between 2014-2018 with a total of 1756 data victims. On research, it uses non-linear SVM with the Polynomial Kernel function. Classification using the polynomial Kernel SVM with the polynomial kernel SVM function uses values cost ie 1. Where the value of these fees is applied to 1756 original data before using the SMOTE method. Furthermore, the value of these costs is applied to get a confusion matrix. The following is the output confusion matrix from dataset classification:-

```
def polynomial_kernel(x, y, p=3):
    return (1 + np.dot(x, y)) * p

class SVM(object):

    def __init__(self, kernel=polynomial_kernel, C=None):
        self.kernel = kernel
        self.C = C
        if self.C is not None: self.C = float(self.C)

    def fit(self, X, y):
        Encoded_Vehicle, n_features = X.shape

        # Gram matrix
        K = np.zeros((Encoded_Vehicle, Encoded_Vehicle))
        for i in range(Encoded_Vehicle):
            for j in range(Encoded_Vehicle):
                K[i,j] = self.kernel(X[i], X[j])

        P = cvxopt.matrix(np.outer(y,y) * K)
        q = cvxopt.matrix(np.ones(Encoded_Vehicle) * -1)
        A = cvxopt.matrix(y, (1,Encoded_Vehicle))
        b = cvxopt.matrix(0.0)
```

```
[[232   0   0   0]
 [  0 153   0   0]
 [  0   0 173   0]
 [  0   0   0  76]]
```

Confusion Matrix by SVM

```
Accuracy score is 1.0
Recall score is 1.0
Precision score is 1.0
F1 score is 1.0
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       232
           2       1.00      1.00      1.00       153
           3       1.00      1.00      1.00       173
           4       1.00      1.00      1.00        76

   micro avg       1.00      1.00      1.00       634
   macro avg       1.00      1.00      1.00       634
weighted avg       1.00      1.00      1.00       634
```

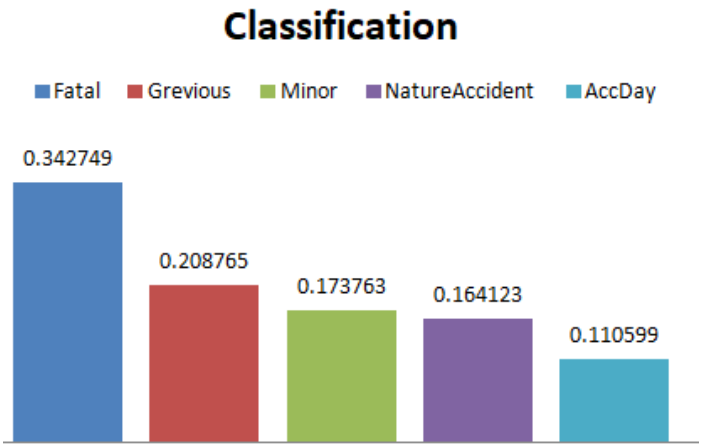

Figure 4.2 Bar Chart Classification using SVM Polynomial Kernel Function

4.1 DATA ANALYSIS WITH SMOT INCORPORATING SVM MODEL

A problem with imbalanced classification is that there of a minority class for a model to effectively learn the decision boundary therefore the SMOT is incorporated for appropriate results.

```
Feature Importance Derived from SVM:-
Fatal              0.342749
Grievous           0.208765
Minor              0.173763
NatureAccident     0.164123
AccDay             0.110599

Out of box features score is 1.0
```

```
[[133  19   0   0]
 [ 39 252  39   0]
 [  0  78 176   0]
 [ 19  77  20  38]]
```

Table: 4.4 Confusion Matrix of SMOT

```
Accuracy score is 0.6730337078651686
Recall score is 0.6730337078651686
Precision score is 0.7250381873897181
F1 score is 0.6535666631028231
```

```
              1       0.70      0.88      0.78       152
              2       0.59      0.76      0.67       330
              3       0.75      0.69      0.72       254
              4       1.00      0.25      0.40       154

    micro avg       0.67      0.67      0.67       890
    macro avg       0.76      0.64      0.64       890
 weighted avg       0.73      0.67      0.65       890
```

Figure 4.3: Results Derived From SMOT with SVM

Based on figure 4.3 it can be seen that Kernel functions used as Polynomial Kernel have 100% of accuracy without modeling the imbalanced data and classes. When viewed from the level of precision the kernel functions, obtained equation of the level of precision in the class Minor Injuries that are not evaluated by SVM there using SWOT the imbalance data and class is incorporated to classify the appropriate results at the level of precision of grade minor injuries class.

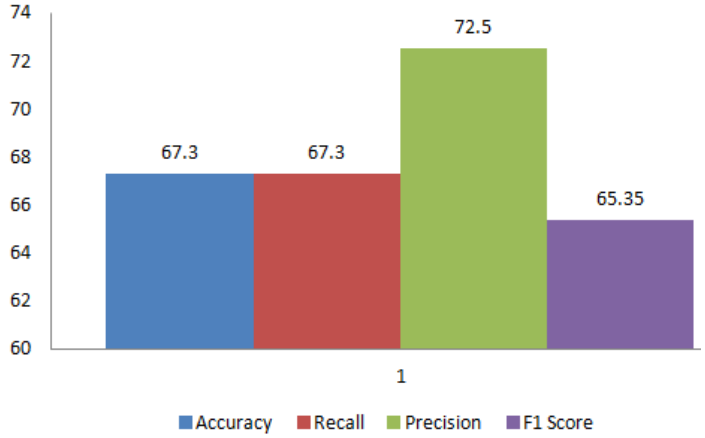

Figure 4.4 Graph Representing Results Derived from SMOT with SVM

5. CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Based on the discussion, it can be concluded that from the results of the analysis conducted, there is some information, namely:

1. General description of the accident data in India between 2014-2018, namely the total number of traffic accident cases are 1756 with the fatal accidents victims are 615, grievous 374, Minor accidents are 311, accidents due to nature are 294 and due to roads 198.
2. The results of the classification of the severity of injuries of traffic accident victims in India in 2014-2018 using the SVM (Support Vector Machine) Polynomial Kernel and SMOTE method are as follows:

a. Training data from 1796 on traffic accidents using SVM over Polynomial Kernel Function the accuracy of fatal accidents are 34.27%, grievous is 20.87, Minor accidents are 17.37%, an accident caused by nature is 16.12% and accident due to bad roads are 11.00% which produces the confusion matrix of 4 folds with the dimension of 232,153,173 and 76 respectively.

b. Training data vide confusion matrix using SVM Polynomial Kernel is passed to SMOT model with SVM for further analysis which originates the confusion matrix based on tested data having 4 fold of dimension 133,252,176 and 38 for further analyses thus resulting Accuracy of 67.30%, precision of 72.50% percent and recall of 67.30%.

5.2 FUTURE SCOPE

Suggestions Based on the conclusions above, several suggestions can be made, namely:

1.It is expected that the traffic police of India should optimize the traffic accident data recapitulation activities over High ways so that there is no missing data in further research.

2. For further research so that similar studies are needed to add variables that affect traffic accident data to get a more accurate pattern and classification on accident data the following year.

3. Future studies are expected to use better-oversampling methods to overcome the imbalanced data.

REFERENCES

[1] Gururaj G, Gautham M S. Advancing Road Safety in India-Implementation is the Key, Bengaluru, 2017. National Institute of Mental Health & Neuro Sciences; 2017. Publication Number :136.

[2] Laura Sminkey, Communications OfficeThe Global status report on road safety 2015 was superseded by the Global status report on road safety 2015 which was launched on 7 December 2018.

[3] Shantajit, Thokchom & Kumar, Chirom & Quazi Syed, Zahiruddin. (2018). ROAD TRAFFIC ACCIDENTS IN INDIA: AN OVERVIEW. International Journal of Clinical and Biomedical Research. 4. 36-38. 10.31878/ijcbr.2018.44.08.

[4] Singh, Sanjay. (2017). Road Traffic Accidents in India: Issues and Challenges. Transportation Research Procedia. 25. 4712-4723. 10.1016/j.trpro.2017.05.484.

[5] Sachin Kumar and Durga Toshniwal, A data mining framework to analyze road accident data, Journal of Big Data (2015) 2:26, Springer, DOI 10.1186/s40537-015-0035-y

[6] (2020). Support Vector Machines. 10.1007/978-981-15-2770-8_8.

[7] Ahmed, Hosameldin & Nandi, Asoke. (2019). Support Vector Machines (SVMs). 10.1002/9781119544678.ch13.

[8] Blanco, Víctor & Japón, Alberto & Puerto, Justo. (2019). Optimal arrangements of hyperplanes for SVM-based multiclass classification. Advances in Data Analysis and Classification. 10.1007/s11634-019-00367-6.

[9] Chang, Mark. (2020). Artificial Neural Networks. 10.1201/9780429345159-5.

[10] Vinge, Rikard & Mckelvey, Tomas. (2019). Understanding Support Vector Machines with Polynomial Kernels. 1-5. 10.23919/EUSIPCO.2019.8903042.

[11] RS in 10 countries,https://www.grsproadsafety.org/wp-content/uploads/RS-10-factsheet-V4-web.pdf

[12] Anna Roy ,Advisor (Industry), NITI Aayog, Discussion Paper National Strategy for Artificial Intelligence June 2018, India..

[13] GOVERNMENT OF INDIA MINISTRY OF ROAD TRANSPORT & HIGHWAYS NEW DELHI, Annual Report 2019

[14] Zhang, Xian-Da. (2020). Support Vector Machines. 10.1007/978-981-15-2770-8_8.

[15] Ahmed, Hosameldin & Nandi, Asoke. (2019). Support Vector Machines (SVMs). 10.1002/9781119544678.ch13.

[16] Žižka, Jan & Dařena, František & Svoboda, Arnošt. (2019). Support Vector Machines. 10.1201/9780429469275-10.

[17] Zhang, Dengsheng. (2019). Support Vector Machine. 10.1007/978-3-030-17989-2_8.

[18] Yang, Xin-She. (2019). Support vector machine and regression. 10.1016/B978-0-12-817216-2.00014-4.

[19] State, Luminita & Cocianu, Catalina. (2011). A New Learning Algorithm of SVM from Linear Separable Samples. Applied Mechanics and Materials. 58-60. 10.4028/www.scientific.net/AMM.58-60.983.

[20] Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.

[21] Gosain, Anjana & Sardana, Saanchi. (2019). Farthest SMOTE: A Modified SMOTE Approach. 10.1007/978-981-10-8055-5_28.

[22] Majzoub, Hisham & Elgedawy, Islam. (2020). AB-SMOTE: An Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification. International Journal of Machine Learning and Computing. 10. 31-37. 10.18178/ijmlc.2020.10.1.894.

[23] Torres, Fredy & Carrasco-Ochoa, Jesús & Martínez-Trinidad, José Francisco. (2019). Deterministic oversampling methods based on SMOTE. Journal of Intelligent & Fuzzy Systems. 36. 4945-4955. 10.3233/JIFS-179041.

[24] Gulowaty, Bogdan & Ksieniewicz, Paweł. (2019). SMOTE Algorithm Variations in Balancing Data Streams. 10.1007/978-3-030-33617-2_31.

[25] McCallum, A. and Nigam, K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. Proceedings in Workshop on Learning for Text Categorization, AAAI'98, 41-48.

[26] Sukhnandan Kaur, Rajni Mohana, International Conference on Emerging Research in Computing, Information, Communication and Applications, 2016/7/29, Springer, Singapore

[27] Preety and Sunny Dahiya. "SENTIMENT ANALYSIS USING SVM AND NAÏVE BAYES ALGORITHM." (2015).

[28] Mayy M Al-Tahrawi, Sumaya N Al-Khatib, Journal of King Saud University-Computer and Information Sciences, 2015/10/1

[29] H He, Y Ma. John Wiley & Sons, 2013. 332, 2013. A self-organizing learning array system ... YX Yang, JL Li, AB Wang, JY Xu, HB He, HR Guo, JF Shen, X Dai.