Journal of Cardiovascular Disease Research

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

Support Vector Machine Algorithm for Analysis of FBI Crime Data

[1]Sujit Kumar Panda, [1]Siddharth Bhusan Neelamani, [1]Satya Ranjan Pattanaik

[1]Professor, [1]Dept. of CSE

[1]Gandhi Institute for Technology, Bhubaneshwar, India

Abstract

Cyber-incidents are a mixture of discrete instances with new illegal acts. Cybercrime incidents occur as separate criminal offences and, according to the national crime statistics and surveys, the instances are increasing. The existing system classifies the cybercrimes and cyber-incidents with less accuracy. These overlapping in its result and the lack of a unique algorithm for classification are the main drawback of the existing model.Thus, to solvethese problems, the paper gives an exclusive way of classifying various crimes depending on the physical factors such as time and date. It gives a solution to users to carry out an easy efficient classification outcome using a cybercrime classifier with support vector machine (SVM). It uses the grouping of the dataset by either decision trees or random forest to build up a model to prepare over a preparation set in order to get the most exact outcomes. It is a modest and productive approach to group cybercrimes with the goal that the affected can identify the kind of occurrence and follow-up correspondingly. Additionally, it examines the information and creates various charts for the correct portrayal of the information. The above model designed to categorize convicted criminals into low, medium and high risk of turning into recidivists helps curb the increasing crime rates in the society, thus ensuring the welfare and well-being of its citizens. The extensive simulation results show that the proposed method gives the outstanding classifier compared to the state of art approaches.

Keywords: Cybercrime data, data analytics, machine learning, support vector machine.

1. Introduction

In an FBI report [1], more than 260,000 complaints were registered of crimes over the internet in 2014, which is 1600% rise compared to the previous report. In accordance with a PwC report, there's been around 50% surge in information wrongdoings in 2014, which comes to be around 117,000 attacks per day. Also, a hacking incident happened in Kudankulam, Tamil Nadu, and India in the month of October, 2019 when the nuclear plant was hacked [2]. Fortunately, this plant had two exclusive infrastructures, one operational and one technical. The hack was made in the organizational infrastructure. Had it been in the technical architecture [3], it would have been fatal. Similarly, cases of cyber malfunctions came up in nuclear plants like the Hatch power plant, near Baxley, Georgia, US, which without a proper emergency system could have proved to be catastrophic. Therefore, nuclear plants pose a grave danger to the security, confidentiality and secrecy of the nation [4]. It's very evident that these crimes pose serious danger to the world economy, its safety, and the overall functioning of society. In recent reports, it had been highlighted that these crimes aren't only increasing quantitatively but also becoming progressively destructive and affect a sizeable information range and vectors. Some reports also suggest that crimes are escalating not only in numbers but also in their gravity [5]. There still isn't much information available of what these cyber-incidents can take the shape of. This also makes dealing with such evils a very tedious task. Hence, a way to classify, detect and counter these crimes is required accordingly.

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

At present, the criminal cases that are pending in India are rapidly increasing with the number of crimes committed are increasing. To solve a case based upon a particular data there should be a thorough investigation and analysis that is to be done internally [6-7]. With the amount of crime data that is present in India currently the analysis and decision making of these criminal cases is too difficult for the officials. Identifying this major problem this paper concentrates on creating a solution for the decision making of crime that is committed. Machine Learning is the branch of science where computers decide without human intervention. In recent times Machine Learning is being used in various domains one of the examples of such cases is automated or self-driving cars. By Machine Learning algorithms there is a way where we can predict certain results based upon our inputs given and provide a solution to solving crime cases in India. The two common types of prediction techniques are classification and regression. This crime data prediction is a domain where classification is applied. Classification is a supervised prediction technique and it has been used in various domains like forecasting stock, medicinal area, etc. [8-9].The main aim of this paper is to consider some algorithms which can be used to predict and analyze the crime data and improve the accuracy of those models by data processing in order to obtain better results. The purpose is to train the required model to predict the data using the training data set by validation of the test data set [9-10]. The models which are being used here are Logistic Regression, Decision Tree classification, Random Forest classification.

The major contributions of the paper as follows:

- Preprocessing of the data has been performed, so errors in the data and malwares are effectively removed in crime dataset.
- The SVM method was implemented classification on public available dataset, the results shows that the proposed SVM classification gives the better performance compared to other approaches.

Rest of the paper is organized as follows; section 2 deals with the various literatures with their drawbacks respectively. Section 3 deals with the detailed analysis of the proposed method with its operation. Section 4 deals with the analysis of the results with the comparison analysis. Section 5 concludes the paper with possible future enhancements.

2. Proposed Method

In the Proposed system, the dataset used is pre-processed, cleaned and transformed using various data mining techniques. Further, the result of this phase gives us a cleaned compatible dataset free of anomalies which can be fed into the next phase where Machine Learning Algorithms would be applied onto the processed dataset. The next phase entails Machine Learning Implementation where various Supervised Classification Algorithms are employed to classify the records of criminals into three categories namely Low, Medium and High. We train our models using classification algorithm namely SVM. Subsequently, Voting takes place to determine which algorithm provides us with the best outcome. This process of Voting is based on the Accuracy Score provided by each algorithm and considering the one with the best accuracy score to be the ideal algorithm for this system. Now, we finalize the model with the highest accuracy and given any tuple for a criminal record the finalized algorithm fetches the best results of classifying the criminal into the respective target categories.

2.1 Data cleaning

The initial dataset consisted of three tuples with null values throughout all attributes. These three tuples were dropped using Pandas and Numpy libraries of Python. Thus, out of 60000 records

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

dropping three tuples full of Null values would negligibly affect the performance of Machine Learning Models. Thus, we drop and eliminate these three tuples. Other than three tuples, the data set consisted of few null values which were replaced by the mean value of that specific column.

Elimination of Null Values The initial dataset consisted of three tuples with null values throughout all attributes. These three tuples were dropped using Pandas and Numpy libraries of Python. Thus, out of 60000 records dropping three tuples full of Null values would negligibly affect the performance of Machine Learning Models. Thus, we drop and eliminate these three tuples. Other than three tuples, the data set consisted of few null values which were replaced by the mean value of that specific column.

Elimination of Duplicate: Data The Dataset contained multiple duplicate values for all tuples. Hence, to avoid over fitting we had to eliminate all repeated records. This scaled down the dataset from 60000 repeated to 18000 unique records.

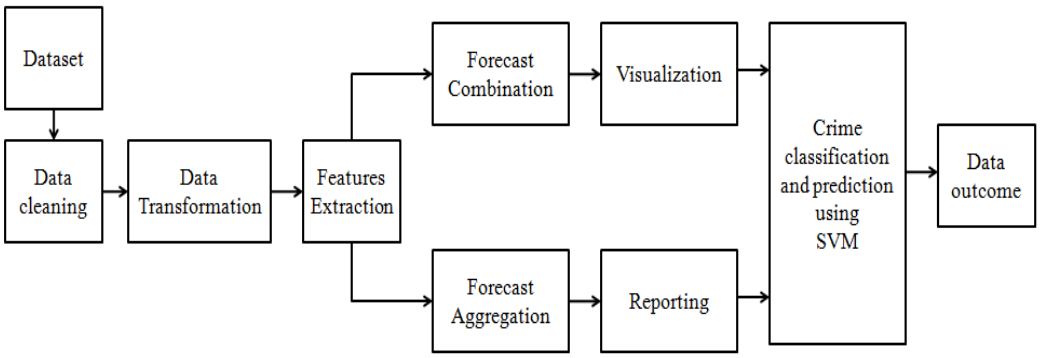

Figure 1: Proposed Method

2.2 Data Transformation:

The dataset contains various attributes with values in the string datatype. The String datatype is not compatible with the Machine Learning Models. To convert the data into compatible format (Integer, Float), we perform Label Encoding Technique. In the associated dataset, label encoding is done manually to reduce the biases on the dataset which is one of the most important factors that needs to be take care of.

2.3 Features Taken into Consideration

The dataset taken into consideration consists of 20 features. In total, 16 important features are considered out of 20 for the implementation of prediction of the models. The features mainly considered are ‘Ethnic code’, ‘Marital Status’, ‘Age’, ‘Sex Code’, ‘Legal Status’ and many more are considered while training the model on the dataset. The target variable is ‘Score_Text’ with values low, medium and high. The table given below describes the integer type attributes of the dataset, providing with statistical values like Mean, Standard Deviation, Minimum Value, and Maximum Value. To explicate this, let’s consider the attribute ‘Age’, we observe that the Mean age of the criminals is between 34 and 35 years. The Standard Deviation is 12.20 years. The Youngest criminal in the dataset is 16 years whereas; the oldest criminal is 84 years old. Likewise, inferences can be made and extrapolated to other such attributes as well.

From the above diagram, the crime data set is collected and a four-part operation called data mining happens. Data mining refers to the practice of inspecting huge pre-existing datasets in order to

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

generate new statistics. Initially, the data goes through a process called data cleansing or data cleaning that detects and corrects corrupt or inaccurate records. After the data is altered and corrected a method called data preprocessing takes place which converts the raw data into a useful and efficient format. Then a process where the raw data is reduced to a more manageable group of processing is referred to as feature extraction. Once the data is grouped, a method called data processing takes place where operations of the data are executed by the computer to retrieve, transform or classify information. The below flowchart describes the process of the prophet model. Initially, the data which is combined from trend, seasonality and holidays are calculated and the data is obtained. The first method that takes place is feature selection where the user automatically or manually selects those features which will give the most to the prediction variable or output. Once the features are selected the data undergoes a four-step process i.e. modeling, forecast evaluation, surface problems and visually inspecting the forecasts. After the process is complete the data sets are analyzed to summarize their main characteristics often with visual methods.

2.4 Classification

To carry out the above-mentioned algorithms, our target variable is the 'ScoreText' attribute of the dataset, which throughout the dataset may have 3 values - Low, Medium and High. This is the measure of the tendency of the criminal to commit recidivism. Thus, we have set the 'ScoreText' as the target attribute, since we need to classify the criminals based on their tendency of committing a crime again. The classification into Low, Medium and High risk provides a cogent perspective to the authorities while processing the given criminal for parole or bail. To explicate this, the criminal with a higher tendency, must not get a bail/parole, as compared to one with a lower risk. This solves the purpose of checking and curbing criminal recidivism in society, hence ensuring the safety of citizens and eschewing a potential crime. The attributes selected as features, for the algorithms, directly or indirectly affected and related to the recidivism tendency of the given criminals. Nearly 17 attributes of the dataset were selected for the training and testing of our machine learning models. For a thorough comparative study among all 3 algorithms used, the features of the model remained the same.

A bias can be encountered in any Machine Learning model, this basically may influence the outcome generated by the machine learning model. Biases in models must be removed since they provide us with an impartial outcome. A bias-free machine learning model cannot exist as it requires a certain amount of bias to model the data and to analyze predictions. However, the aim is to reduce these biases occurring in our model. In the case of training models for Criminal Recidivism, there can be various biases which may be encountered. To explicate this, some of the biases may be against certain races, where people of a particular race may be impartially evaluated for gauging the recidivism score. Another such bias may occur in the gender of any offender, where a person of a particular gender may be more biased/likely to be categorized as a recidivist. There can likely be the following biases –

1) Sample Bias - This is an inevitable type of bias arising due to the randomness and irregularities in the data samples. This occurs during the training phase of the model. This can be an intuitive way of the model to pick up the more frequently occurring values of a particular attribute. For instance, in the dataset used to train our model, the number of cases, where the tuple has value "Single" in the "Marital Status" column is much higher than other values. This inevitably creates an occurrence of sample bias, where the model would be inclined to categorize most of the "single" values to higher recidivism.
2) Algorithmic Bias - This is the type of bias which is introduced by the algorithmic phase of the machine learning model and is not present due to anomalies in data samples. Data Scientists

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

strive to attain a perfect balance between high variance and high bias. Here, in our model, the Random Forest Classifier introduces bias, when training and testing the given dataset. This is an inherent property of the algorithm.

3) Measurement Bias - It occurs when we select the features we wish to incorporate in the model. It may be the way these features/attributes are used in the machine learning phase. A striking example of this is the use of this for criminal recidivism, where any priory committed crimes or crimes committed by relatives/friends may also taint the outcome of the model. Thus, attributes like “Agency Type”, “Custody Status”, “Legal Status”, etc. may create a measurement bias for criminals in evaluating their score text.
4) Prejudice Bias - This type of bias is mainly due to the influence of social stereotypes and orthodox opinions. It mainly occurs on training data, where prejudice against a particular culture, gender, ethnicity or any such factor may make the model biased while generating the output. If the algorithm is exposed to a more even-handed data distribution, then the statistical relationship between such potentially prejudiced attributes can be avoided.

3. Simulation Results

3.1 Understanding the Dataset

The dataset considered for implementation is from Carnegie Mellon University (CMU). In this phase, we understood the features of the dataset and the type of data present in the dataset. The Dataset contains approximately 60000 tulles consisting of around 22 attributes. It contains the records of criminals in the United States of America, with their various traits and characteristics which may help us classify them as recidivists. Moreover, we also checked for the uniformity of the dataset to decide the performance parameter for the Machine Learning Model.

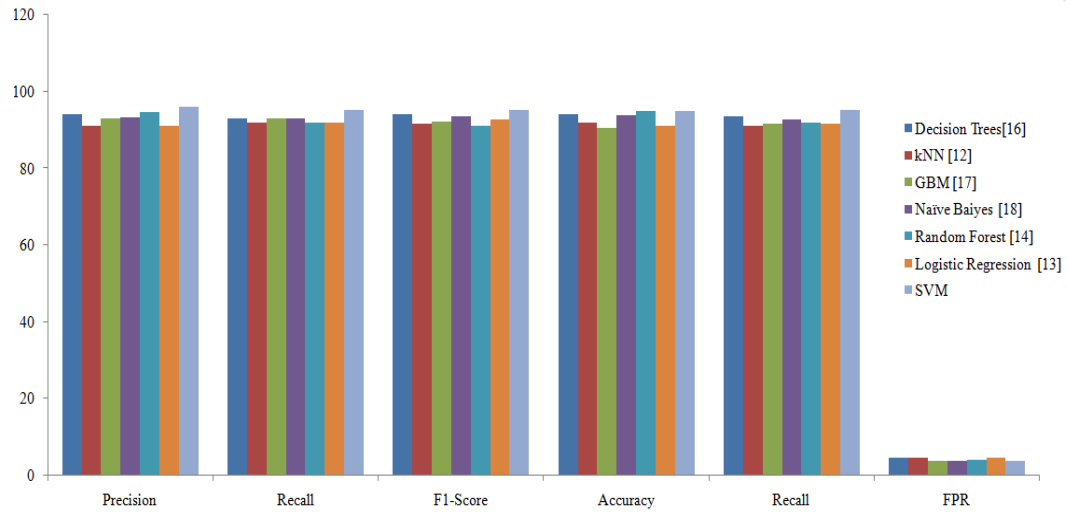

Figure 2: Experimental Results for Various ML Methods

Experimental results from Figure 2 demonstrate the percent of the different assessment parameters for just the credit card fraud dataset for distinct machine learning techniques. Findings indicate that SVM techniques demonstrate an accuracy percentage Decision Trees [16], kNN [12], GBM [17], Naïve Baiyes [18], Random Forest [14] and Logistic Regression[13]demonstrate a precision percentage of machine learning crimes identification. For any machine learning technique, greater values are shown to be accepted as just a higher performance method of precision, accuracy, recall, and F1-score. As we have seen, there are a few algorithms that have surpassed others as well quite significantly. Thus,

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

selecting SVM over all other techniques could be a sensible approach in attaining a greater degree of completeness when decreasing quality just significantly.

4. Conclusion

The above model designed to categorize convicted criminals into low, medium and high risk of turning into recidivists helps curb the increasing crime rates in the society, thus ensuring the welfare and well-being of its citizens. In this way, Machine Learning can be made of paramount importance to perpetuate the security and safety of innocent citizens, who might be potential victims of assaults or any such ordeal. Thus, our Machine Learning model aims at resolving one of the above mentioned causes, by overcoming the shortcomings of law enforcement practices, by enabling the authorities to make an informed, statistically and analytically cogent decision, in matters of granting parole to the criminals, who might be potential recidivists.The above research methodology system should be extrapolated for more localized prisons/facilities so that the results are more specific and pertaining to the respective location. Moreover, the research should be conducted for a longer period of time, thus generating more records of criminals to obtain better accuracy. The proposed model should help in reducing criminal recidivism in the upcoming years. The results of this research should also be merged with the oral statements provided by criminal on day of trial to judge for more authenticity using Natural Language Processing techniques. Extensive analysis should be carried out locally and subsequently on a larger scale to gain insightful trends. This system should help the parole granting authorities to abstain from granting parole to criminals with a medium to high risk of recidivism.

References

[1]. Tamilarasi, P., and R. Uma Rani. "Diagnosis of Crime Rate against Women using k-fold Cross Validation through Machine Learning." *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2020.

[2]. Hossain, Sohrab, et al. "Crime Prediction Using Spatio-Temporal Data." *International Conference on Computing Science, Communication and Security*. Springer, Singapore, 2020.

[3]. Ch, Rupa, et al. "Computational system to classify cyber crime offenses using machine learning." *Sustainability* 12.10 (2020): 4087.

[4]. Saravanan, P., et al. "Survey on Crime Analysis and Prediction Using Data Mining and Machine Learning Techniques." *Advances in Smart Grid Technology*. Springer, Singapore, 2020. 435-448.

[5]. Yang, Bo, et al. "A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery." *International Journal of Geographical Information Science* 34.9 (2020): 1740-1764.

[6]. Wheeler, Andrew P., and Wouter Steenbeek. "Mapping the risk terrain for crime using machine learning." *Journal of Quantitative Criminology* (2020): 1-36.

[7]. ATEŞ, Emre Cihan, Gazi Erkan BOSTANCI, and M. S. G. Serdar. "Big Data, Data Mining, Machine Learning, and Deep Learning Concepts in Crime Data." *Journal of Penal Law and Criminology* 8.2 (2020): 293-319.

[8]. Trisnawarman, Dedi, and Muhammad Choirul Imam. "Prediction Analysis Of Criminal Data Using Machine Learning." *IOP Conference Series: Materials Science and Engineering*. Vol. 852. No. 1. IOP Publishing, 2020.

[9]. Yerpude, Prajakta. "Predictive Modelling of Crime Data Set Using Data Mining." *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol* 7 (2020).

ISSN:0975-3583,0976-2833 VOL12,ISSUE07,2021

[10]. Das, Priyanka, et al. "A framework for crime data analysis using relationship among named entities." *Neural Computing and Applications* 32.12 (2020): 7671-7689.