Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 04, April 2021: 2265-2275

Research Article

Improved Delinquency Detection Using Machine Learning

S Vijayakumar[1], Praveena M[2], Renukadevi J[3], Reshma L[4]

[1]Associate Professor, Department of Computer Science & Engineering, RMK Engineering College, Chennai, India
[2,3,4]Departmentof Computer Science & Engineering, RMKEngineering College, Chennai, India

Abstract

Delinquencies are due to the repressed desire for aesthetic expression. To fulfil one's desire that person involves in creating danger to the entire society. As delinquency is an unlawful activity, the person behind it must be penalized by law. No person have faith with others instead they have fear with others. In the blink of an eye countless delinquencies are happening around us. It is impossible to completely stop the delinquencies but it can be prevented. The aim of this paper is to perform analysis and prediction of delinquency using machine learning algorithms. Identifying delinquency patterns will allow us to tackle problems with unique approaches in specific delinquency category regions and improve more security measures in society. Even though it is not feasible to stop the delinquencies instantly, the tactics used in this paper will assuredly reduce the delinquencies and make an effort to create a delinquency unbound society.

Keywords: *Delinquency; Support Vector Machine; Logistic regression; MLAD*

Introduction

Delinquency is a serious and growing problem in most societies. They can be seen as the plague of humankind. Thousands of delinquencies are committed every day, and probably hundreds are occurring right now in the world. Delinquencies like robbery, murder, rape, false imprisonment, kidnapping, homicide and many more, can happen anywhere anytime from rural to urban areas. Many of the delinquencies that took place in India remained as headlines of the news for the very long time and the cases regarding those also continued for a very long time; sentiments of crores of people of India were attached to the delinquency in sympathy. Many people are too scared to leave their home because of a fear. Delinquencies terrifically affect not only the victims but also the people of the country. The number of delinquencies will only get increased in near future, if we fail to control them. It is the responsibility of police department. As there are enormous amount of data that exist, prediction and classification are the major problems to the police department. There is a need of technology that makes it easy to solve a case and to prevent delinquencies beforehand.

The aim of this paper is to detect the delinquencies using the features present in the dataset. With the help of machine learning algorithm, using python as core we can predict which sort of delinquency will occur in a specific area. This system uses the support vector machine and logistic regression of machine learning domain to implement the detection and classification. SVM can take a large data collection of even millions of data and bring the required output accurately whereas the existing system using KNN classifiers collapses if it has to process large amount of data. The output is displayed in dynamic fashion with the help of graph and chart for better understanding of results. The proposed idea gives a accurate prediction analysis report. This report includes the information that in a given particular place, that suggests in a specific place, the average number of delinquencies happening in a particular day, what sort of delinquencies have taken place in past and what fraction these delinquencies have taken place, what sort of delinquencies that has occurred in most fraction.

The dataset is extracted from the official sites. These data sets are trained using machine learning techniques. By the action of pre-processing, the unwanted data sets and duplicate fields are erased from the data sets to bring stability to the input. After pre-processing step, the input dataset is ready for implementation. The next step is the feature extraction. This is done to extract fields of the dataset that are very much necessary to bring out the results accurately. The input data is subsided to point and it is precise now. Now, the algorithm is implemented for classifying the delinquencies. The reference model is then created and graphs are modelled with different view points with the algorithms result. The accuracy of the dataset considered for implementation can be tested by officials using various testing dataset. This paper will help the officials to know exactly what kind of delinquencies occur at small proportion, which steadily increases. They can focus more inclined towards the problem which arise repeatedly and suitable security measures can be taken to prevent them.

System Configuration

Support Vector Machine (SVM) proposed by vapnik and cortes have been successfully applied for gender classification problems by many researchers. An SVM classifier is a linear classifier where the separating hyper plane is chosen to minimize the expected classification error of the unseen test patterns. SVM is a strong classifier which can identify two classes. SVM classifies the test image to the class which has the maximum distance to the closest point in the training. SVM training algorithm built a model that predict whether the test image fall into this class or another. SVM require a huge amount of training data to select an affective decision boundary and computational cost is very high even if we restrict ourselves to single pose (frontal) detection. The SVM is a learning algorithmfor classification. It tries to find the optimal separating hyper plane such that the expected classification error for unseen patterns is minimized.

For linearly non-separable data the input is mapped to high dimensional feature space where they can be separated by a hyper plane. This projection into high- dimensional feature space is efficiently performed by using kernels. More precisely, given a set of training samples and the corresponding decision values{-1, 1} the SVM aims to find the best separating hyper plane given by the equation WT x+b that maximizes the distance between the two classes.

SVM (Support vector machine) are supervised learning methods that study data needed for classification and regression analysis. According to training examples, each marked for belonging to any of two analyses, a SVM training algorithm builds a model that assigns new examples into either of two analyses and makes it non probabilistic binary linear classifier. An SVM model is a depiction of the examples like points in space, so that the examples of the different analysis are divided by a relevant gap that is as broad as possible. New examples are then mapped into the same existed space and predicted to belong to a category based on which side of the gap they fall.

Logistic Regression*:* Logistic Regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc. Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories.

Literature Survey

3.1 Machine Learning Based Anomaly Detection for Load Forecasting Under Cyberattacks

Author-Mingjian Cui, Jianhui Wang, MengYue, 2019.

Accurate load forecasting can make both economic and reliability benefits for power system operators. However, the cyberattack on load forecasting may mislead operators to make unsuitable operational decisions for the electricity delivery. To accurately and effectively detect the cyberattack, this paper develops a machine learning based anomaly detection (MLAD) methodology.

3.2 Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model

Author-Al Amin Biswas, Sarnali Basak,2019.

It is essential to minimize the crime rate to speed up the development of a country. Rate of crime has increased rapidly within the last few years in Bangladesh which is very much alarming for the advancement of a country like Bangladesh. Hence, it is necessary to analyze the crime data to forecast the crime patterns and trends so that the law enforcement authorities can take some outstanding action to minimize the crime. This paper uses polynomial and randomforest regression of machine learning.

3.3 Analysis of Counterfeit Currency Detection a Techniques for Classification Model

Author-Akanksha Upadhyaya, Dr.Vinod Shokeen, Dr.Garima Srivastava, 2018.

The growth of counterfeited currency is becoming a great threat to world wide by impacting each country thoroughly. The rate of counterfeiting is widely increasing due to fleeting acquisition of technology. The reason of swift adoption is cost, availability and efficiency of technological equipment. From past many years the race is going on between the counterfeiters and the banks. To resolve the issue various researchers came across with variety of techniques and proposed solutions from the area of Machine learning and Image processing for this serious issue.

3.4 A Survey on Machine Learning Techniques for Cyber Security in the Last Decade

Author-Kamran Shaukat, Suhuai Luo, Vijay Varadharajan, IbrahimA. Hameed, Min Xu, 2020.

The cyberspace has become more vulnerable to automated and prolonged cyberattacks. Cyber security techniques provide enhancements in security measures to detect and react against cyberattacks. The usage of machine learning and artificial intelligence techniques is getting expanded rapidly in different area of life. This paper discusses the challenges of using ML techniques in cyber security, also provides the latest extensive bibliography and the current trends of MLin cyber security.

3.5 Framework For Image Forgery Detection And Classification Using Machine Learning

Author–Shruti Ranjan, Prayati Garhwal, Anupama Bhan, Monika Arora, AnuMehra, 2018.
The rates of cybercrimes has been surging prodigiously. It has been proven incredibly easy to create fake documents with powerful photo editing software being as pervasive as ever. Documents can be scanned and forged within minutes with the help of these software that have tools readily available just to do that. While photo manipulations software is handy and ubiquitous, there are also means to deftly investigate these documents. This paper lays a foundation on investigation.

Proposed System

The proposed system analyses and classifies the delinquencies happening around us, it gives a wide view to acknowledge about the delinquency region. The system uses the Support Vector Machine and Logistic Regression algorithms. Based on predicting the type of attribute and finding a hyperplane of data that best disparates the features into different domains, alike delinquencies can be effortlessly encountered. The algorithms considers the past 12 years dataset to forecast the delinquencies. In preparatory implementations, the training dataset is preprocessed and the features are extracted. Now the algorithms are applied to the extracted data which sort and scrutinize the data. From the large data set, the system evaluates and lay the required output accurately. The output generated is in the form of graphical format for better understanding of outcome.

System architecture:

A system architecture is the conceptual model that defines the structure, behaviour, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. An architecture diagram is a graphical representation of a set of concepts, that are part of an architechture, including their principles, elements and components.

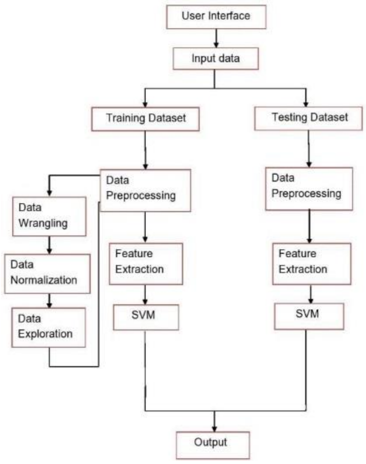

Figure 1 Architechture diagram

The elements include Training dataset, preprocessing and feature extraction, SVM, reference model. This architechture is the representation of Delinquency detection in SVM processing. The user will input the data which is stored in database and detected by checking past delinquency data in which SVM technique is also applied.

Existing System

Delinquencies are not stored or recorded well. Even the output stored are of text description and so it is difficult to get correct overview of the records. The main drawback in prediction is that it has several key parameters that need to be set correctly to achieve the best classification results for any given problem. Parameters that may result in an excellent classification accuracy for problem, may result in a poor classification accuracy for other problem that detect by the system. It does not perform well then the data set has more noise, that is, target classes overlapping. As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probablisitic explanation for the classification. Some of the problems that are faced in the systems are fast report generation is not possible

and the information about delinquency and common people is not properly maintained.

Implementation

User Interface:

User interface is the GUI of the delinquency detetction system where User interface (UI) design is the process of making interfaces in software or computerized devices with a focus on looks or style. Designers aim to create designs users will find easy to use and pleasurable. UI design is more concerned with the surface and overall feel of a design, whereas the latter covers the entire spectrum of the user experience. One analogy is to picture UX design as a vehicle with UI design as the driving console. In GUIs, you should create pleasing aesthetics and animations that convey your delinquency detection values and maximize usability.

Data Analysis:

During this step we performed some descriptive analysis and determined the target variable. We also explored how many classes were in the target and a selection of other possibly problematic (high cardinality) variables. I also Vis ualized the target variable in a histogram which is a good technique for understanding the distribution of the data to assist in parameter tuning.

Data Collection

Dataset is obtained from kaggle, has delinquency dataset of city Chennai in TamilNadu. This dataset is used in CSV format and has the collection of past 10 years of delinquency data.In this data set consists like dates, description, days that the delinquency held and the week of the delinquency.Accurate data collection is essential to maintaining the integrity of research, making informed businessdecisions and ensuring quality assurance.

Data Pre-processing

Data Preprocessing enhance the quality of data to promote the extraction of meaningful insights from the data which increases the accuracy and efficiency of a system. This process organize the selected data by formatting, cleaning and sampling to make it suitable for a building and training the system. Formatting is the process to make the data it suitable for structured format. Cleaning is the process to remove the incomplete variables based on insufficient data, non-representative data, substandard data. Sampling is the process on the data further to reduce running times for algorithms and memory requirements. In order to define the distance metrics for categorical variables, the first step of preprocessing of the dataset is to use dummy variables to represent the categorical variables. Secondly, due to the distinct natures of categorical and numerical data, we usually need to standardize the numerical variables, such as the contributions to the euclidean distances from a numerical variable and a categorical variable are basically on the same level.

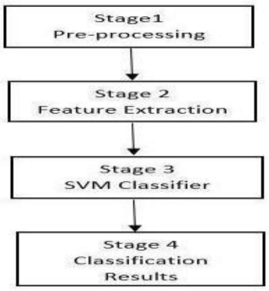

Figure 2. Stages of Delinquency Detection

Data Partition and Transformation*:*

High cardinality variables are dropped during this step as a precursor to the pre-processing step. The pre-processed data partitioned into a training and testing dataset modelling. In training, 70% or 80% of the data goes here and model will be developed on this dataset. In testing, 30% or 20% of the data goes here and model performances will be evaluated on this dataset.

Delinquency Similarities Identification*:*

Feature selection is a subset of the input features for training the model, and ignoring the irrelevant or redundant ones. The attributes used for feature selection are Block, Location, District, Community area, Delinquency description, X coordinate, Y coordinate. Here X, Y denotes the Longitude and Latitude value.

Delinquency Detection*:*

This module we implement the algorithm which one give best accuracy value. Based on that applying the algorithm it should calculate and give the accuracy. Here we are applying the algorithm and get the result on the form of graphical format. Detecting perpetrators is based not only on the conduct of simple investigative activities, but requires specialized knowledge of the motive behind the offender and the circumstances of the offense. The growing role of forensic techniques for the effectiveness of investigations and investigations creates a demand on the labor market for specialized knowledge in this field. Due to the increasing importance of forensic techniques, also achieved through the use of new technologies, including research on digital printers delinquency detection is almost always associated with a series of complicated activities that reveal evidence.

Result And Discussion

This segment covers the analysis done on the dataset and representing them in the form of graphs. Analysis done are - the number of times a particular category of delinquency took place, details of major delinquencies committed and average number of times the delinquencies committed over a week.

In the below graph, y-axis denotes the categories of delinquencies and x-axis denotes their fraction of occurrence, that is, how frequently they take place.

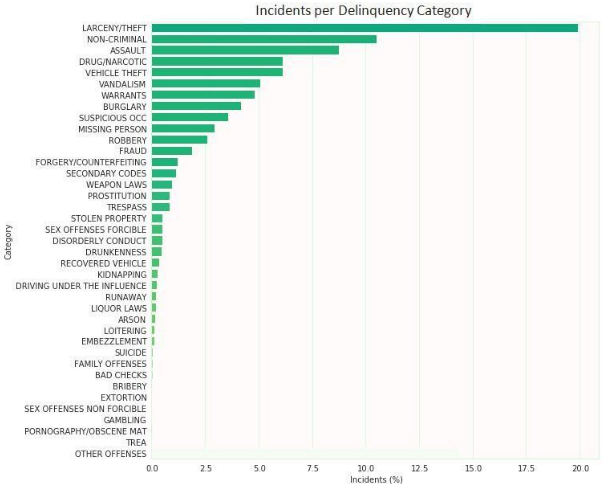

Figure 3 Category vs Fraction of Incident

The graph below shows crime occurrence over particular week. The x-axis denotes weekdays and y- axis denotes the rate of delinquencies.

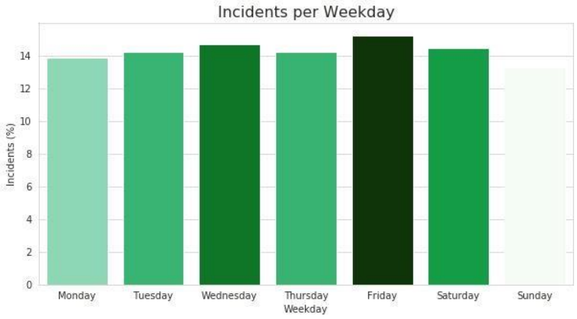

Figure 4. Delinquencies committed over a week

Future Work

This project can also be improved to provide police patrol routes, rather than an optimal position for a police car to locate. This can be done by considering different time periods, seasons, and also special occasions. Using a GIS plan, the GIS data can be integrated with the delinquency dataset. In this way, we can significantly improve the precision and there call of the prediction model trained by the CDAP. Also binding data from back end to front end can be optimized further and it will improve the user experience significantly.

Conclusion

With the help of machine learning algorithms, finding the relation and patterns among various data’s has become simple. The intent of this paper mainly revolves around the types of delinquencies that have taken place, what proportion these delinquencies have taken place and predicting the type of delinquency which may happen if we know it's whereabouts. The model predicts with accuracy of 0.789. Support Vector Machine and Logistic regression Algorithm concepts were used to build this model using training data set that have undergone data cleaning and data conversion. Different sets of data have been interpreted by means of different graphs that gives peculiar information and features. This helps in analysing large datasets in less time with better understanding and helps in preventing delinquencies before they happen to create safer and more secure society.

References

[1] Biswas, A. A., & Basak, S. (2019, September). Forecasting the Trends and Pattems of Crime in Bangladesh using Machine Learning Model. In 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT) (pp. 114-118). IEEE.

[2] Butt, U. M., Letchmunan, S., Hassan, F. H., Ali, M., Baqir, A., & Sherazi, H. H. R. (2020). Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review. IEEE Access, 8, 166553-166574.

[3] Cui, M., Wang, J., & Yue, M. (2019). Machine learning-based anomaly detection for load forecasting under cyberattacks. IEEE Transactions on Smart Grid, 10(5), 5724-5734.

[4] Huang, D., Mu, D., Yang, L., & Cai, X. (2018).CoDetect: financial fraud detection with anomaly feature detection. IEEE Access, 6, 19161-19174

[5] Pastor, A., Mozo, A., Vakaruk, S., Canavese, D.,López, D. R., Regano, L., ... & Lioy, A. (2020). Detection of encrypted cryptomining malware connections with machine and deep learning. IEEE Access, 8, 158036-158055.

[6] Pavithra, R., & Suresh, K. V. (2019, April). Fingerprint image identification for crime detection. In 2019 International Conference on Communication and Signal Processing (ICCSP) (pp. 0797-0800). IEEE.

[7] Ranjan, S., Garhwal, P., Bhan, A., Arora, M., & Mehra, A. (2018, May). Framework for Image Forgery Detection and Classification Using Machine Learning. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1-9). IEEE.

[8] Sandagiri, S. P. C. W., Kumara, B. T. G. S., & Kuhaneswaran, B. (2020, October). ANN Based Crime Detection and Prediction using Twitter Posts and Weather Data. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-5). IEEE.

[9] Shaukat, K., Luo, S., Varadharajan, V., Hameed, A., & Xu, M. (2020). A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. IEEE Access, 8, 222310-222354.

[10] Thota, L. S., Alalyan, M., Khalid, A. O. A., Fathima, F., Changalasetty, S. B., & Shiblee, M. (2017, March). Cluster based zoning of crime info. In 2017 2nd International Conference on Anti-Cyber Crimes (ICACC) (pp. 87-92). IEEE.

[11] Upadhyaya, A., Shokeen, V., & Srivastava, G. (2018, December). Analysis of Counterfeit Currency Detection Techniques for Classification Model. In 2018 4th International Conference on Computing Communication and Automation (ICCCA) (pp. 1-6). IEEE.

[12] Wang, S., Wang, X., Ye, P., Yuan, Y., Liu, S., & Wang, F. Y. (2018). Parallel crime scene analysis based on ACP approach. IEEE Transactions on Computational Social Systems, 5(1), 244-255.

[13] Yadav, S., Timbadia, M., Yadav, A.,Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 1, pp. 225-230). IEEE

[14] Yue, M., Hong, T., & Wang, J. (2019). Descriptive analytics-based anomaly detection for cybersecure load forecasting. IEEE Transactions on Smart Grid, 10(6), 5964-5974.

[15] Zhao, X., & Tang, J. (2017, November). Exploring transfer learning for crime prediction. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 1158-1159). IEEE.