IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Crime Hotspots Using Machine Learning Techniques

1. POTNURI GAYATRI (ASSISTANT PROFESSOR),

2. CH. NIKHILA, 3. K. SAI KIRAN, 4. B. AKHILESWAR REDDY, 5. A. SURESH

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING,

SANKETIKA VIDYA PARISHAD ENGINEERING COLLEGE, VISAKHAPATNAM, INDIA

Abstract:

This research delves into the application of machine learning algorithms for forecasting crime hotspots by leveraging historical data of public property crime in a major coastal city in southeast China. The study conducts a comparative analysis, emphasizing the predictive efficacy of various machine learning models. Results indicate that the LSTM model surpasses other methods including KNN, random forest, support vector machine, naive Bayes, and convolutional neural networks when utilizing solely historical crime data. Moreover, integrating built environment data such as points of interest (POIs) and urban road network density as covariates into the LSTM model enhances predictive accuracy. These findings bear significance for shaping policing strategies and implementing measures for crime prevention and control.

INDEX TERMS: LSTM model, Comparative analysis, Predictive power, Points of interest (POIs), Urban road network density, Policing strategies, Crime prevention, Control measures, Comparative evaluation.

INTRODUCTION

In recent years, there has been an exponential increase in spatiotemporal data related to public security, yet its effective utilization remains a challenge. Within the realm of crime prevention, numerous scholars have devised models for predicting crime, often relying solely on historical crime data to fine-tune predictive models.

Current research on crime prediction primarily revolves around two main areas: crime risk area prediction and crime hotspot prediction. Crime risk area prediction, rooted in the "routine activity theory," examines the correlation between criminal activities and the physical environment, typically utilizing traditional crime risk estimation methods to identify crime hotspots from historical crime cases.

The terrain risk model, considering the proximity and aggregation of crime elements, tends to incorporate crime-related environmental factors and historical crime data, proving effective for stable, long-term crime hotspot prediction. Empirical research has combined demographic, economic, land use, mobile phone, and crime history data for crime prediction, aiming to forecast the likely location and concentration of future crime events.

Machine learning algorithms, including K-Nearest Neighbour (KNN), random forest, support vector machine (SVM), neural networks, and Bayesian models, have gained popularity for crime trend prediction. Studies have compared linear methods, Bayesian models, neural networks, and spatiotemporal kernel density methods in different crime prediction scenarios.

Among these algorithms, KNN is recognized for its efficiency, SVM for its versatility in classification and regression tasks, and random forest for its strong non-linear data processing capabilities. Naive Bayes is praised for its simplicity and insensitivity to missing data. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) neural networks are known for their robustness in handling intricate classification problems and time-series data, respectively.

RELATED WORK

PRINCIPLES OF THEORETICAL CRIMINOLOGY IN THE PREDICTION OF CRIME HOTSPOTS

Crime hotspot prediction revolves around forecasting the future concentration of criminal events in a specific geographic area, drawing insights from principles of theoretical criminology. Various criminological theories provide crucial guidance, shedding light on the significant influence of location factors on the formation and aggregation of criminal events. These theories establish a fundamental mechanism for law enforcement to utilize crime hotspot information for prevention or control, primarily rooted in routine activity theory, rational choice theory, and crime patterns theory, acknowledged as the theoretical underpinnings of situational crime prevention.

Routine activity theory, jointly proposed by Cohen and Felson in 1979, has evolved through integration with other theories. It posits that most crimes, especially predatory crimes, require the convergence of three elements: motivated offenders, suitable targets, and a lack of ability to defend in time and space.

Rational choice theory, proposed by Cornish and Clarke, suggests that offenders' location, goals, and methods choices can be explained by rational assessments of effort, risk, and reward.

Crime pattern theory integrates routine activities theory and rational choice theory, offering a nuanced explanation of the spatial distribution of criminal events. Individuals create "cognitive maps" and "activity spaces" through daily routines, which potential offenders use to select crime locations in relatively familiar areas. Offenders tend to avoid unfamiliar places and choose locations where criminal opportunities align with cognitive spaces based on rational decision-making. Identifying crime hotspots involves not only analyzing historical crime data but also considering the environmental factors of these locations, which exhibit distinct characteristics conducive to crime "production" or "attraction."

Built Environment Data:

Numerous studies underscore the significant influence of the urban built environment on criminal behavior, shaping crime opportunities and playing a vital role in crime reduction and prevention efforts. The 2007 Global Habitat Report emphasized the pivotal role of built environment elements in the occurrence of criminal acts. Point of Interest (POI) data and road network density data are utilized as covariates in the crime prediction model.

CRIME PREDICTION WITH MACHINE LEARNING ALGORITHMS

Traditional methods typically identify crime hotspot areas based on the historical distribution of crime cases, assuming that past patterns will recur in the future. This assumption holds validity for predicting long-term stable crime hotspots. The widely employed Kernel Density Estimation (KDE) method effectively pinpoints such stable hotspot areas. Particularly, the KDE method, tailored to account for temporal autocorrelation, tends to outperform the general KDE method. Liu et al. conducted a comparison between the random forest algorithm and the spatiotemporal KDE method, revealing the random forest's superior efficiency, especially in smaller time scales and grid space units. Gabriel et al. utilized the Gated Localized Diffusion Network for crime prediction at the street segment level, showcasing a substantial increase in prediction accuracy compared to the traditional Network-time KDE method. The effectiveness of machine learning algorithms in processing non-linear relational data, as corroborated in various domains including crime prediction, is characterized by faster training speeds, adeptness in handling high-dimensional data, and the extraction of key data characteristics.

PREDICTION MODEL

This paper employs the random forest algorithm, KNN algorithm, SVM algorithm, and LSTM algorithm for crime prediction. Initially, historical crime data serve as the sole input for model calibration, facilitating comparison to determine the most effective model. Subsequently, built environment data, such as road network density and POI, are introduced as covariates into the predictive model to assess if prediction accuracy can be further enhanced.

A. KNN:

KNN, or k-nearest neighbor, utilizes the feature vector of the instance as input, calculates the distance between the training set and the new data feature value, and selects the nearest K classifications. The classification decision rule involves majority voting or weighted voting based on distance. The category of the input instance is determined by the majority of K neighboring training instances.

B. RANDOM FOREST:

The random forest comprises a set of tree classifiers $\{h(x, \beta k), k = 1... \}$, where the meta classifier $h(x, \beta k)$ is an uncut regression tree constructed by the CART algorithm. The output is obtained through voting, with randomness introduced by randomly selecting the training sample set using the bagging algorithm and randomly selecting the split attribute set. The final classification result is determined by the vote of tree classifiers.

C. SVM:

SVM, grounded in statistical learning theory, is a versatile data mining method successful in addressing regression, time series analysis, pattern recognition, and classification problems. SVM aims to find a superior classification hyperplane that ensures accuracy and maximizes the blank area on both sides, achieving optimal classification for linearly separable data.

D. NB:

In the field of probability and statistics, Bayesian theory predicts the occurrence probability of an event based on evidence knowledge. The naïve Bayes (NB) classifier, within machine learning, is based on Bayesian theory and assumes the independence of each feature. This classifier leverages conditional probability to determine the likelihood of a given entity belonging to a certain class.

E. CNN:

CNN utilizes one-dimensional convolution for sequence prediction, involving the convolution sum of discrete sequences. The network employs a window size of Kernel size to convolve the sequence, followed by a pooling operation to filter and extract the most useful features.

CONCLUSION

This study applies six machine learning algorithms to predict crime hotspots in a town in the southeast coastal city of China, yielding the following conclusions:

The LSTM model demonstrates superior prediction accuracy compared to other models, showcasing its capability to extract patterns and regularities from historical crime data.

Incorporating urban built environment covariates enhances the prediction accuracies of the LSTM model, surpassing results achieved using historical crime data alone.

The models developed in this study exhibit improved prediction accuracies compared to previous empirical research on crime hotspot prediction. For instance, the LSTM model outperforms prior research, achieving a case hit rate of 59.9% and an average grid hit rate of 57.6%.

For future research, enhancements can be made in several areas:

Temporal Resolution: Exploring finer temporal resolutions to capture changes in crime levels over shorter time intervals, such as days or hours.

Spatial Resolution: Assessing the impact of varying grid sizes on prediction accuracy to determine the optimal spatial resolution.

Robustness and Generality: Testing the robustness and generality of the findings in other study areas to validate the applicability of the research outcomes beyond the current study size.

Despite persistent challenges, the insights gained from this research have proven beneficial in recent hotspot crime prevention experiments conducted by the local police department in the study area.

REFERENCES

[1] Vineet Jain, Yogesh Sharma, Augush Bhatia, Vaibhav Arora. “Crime Prediction using K-means Algorithm”. Global Research and Development journal for engineering Volume 2, issue 5, April 2017.

[2] Shyam Varan Nath, Oracle Corporation, Shyam. Nath, @Oracle.com “Crime Pattern Detection Using Data Mining”.

[3] JERZY Ste FANOUSKI Institute of Computing. Sciences Poznon University at Technology, Polan “Data Mining- Clustering”.

[4] K.S.Arthisree, M.E, A.Jaganraj, M.E, CSE Department. “Identify Crime Detection Using Data Mining Techniques”.

[5] Zijun Zhang, “K-means Algorithm and Cluster Analysis in Data Mining.”

[6] “Clustering”. 15-381 Artificial Intelligence Henry Lin, Modified From Excellent slide of Eamonn Keogh, Ziv Bar-Josept, and Andrew Moore.

[7] “Cluster Analysis: Basic Concepts and Algorithms”.

[8] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal, “Crime Analysis using K-means Clustering”. International Journal of Computer Applications (1975-8887) volume83, No-4, December, 2013.

[9] Hsinchum chen, Wingyan Chug, Yin Qin, Michael Chau, “Crime Data Mining: An Overview and Case Studies”.

[10] Ms.Aruna J.Chamatkar, Dr.PK.Butey “Important of Data Mining with different Types of data Applications and Challenging Areas”.