IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analyzing Crime Hotspot Prediction With Machine Learning Techniques

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

5. P. GAYATRI (ASSISTANT PROFESSOR)

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING,

SANKETIKA VIDYA PARISHAD ENGINEERING COLLEGE, VISAKHAPATNAM, INDIA

Abstract:

This study explores the use of machine learning algorithms for crime prediction using historical data of public property crime in a large coastal city in southeast China. The comparative analysis focuses on the predictive power of several machine learning models. The findings indicate that the LSTM model outperformed other methods such as KNN, random forest, support vector machine, naive Bayes, and convolutional neural networks when solely using historical crime data. Furthermore, incorporating built environment data such as points of interest (POIs) and urban road network density into the LSTM model as covariates improved the prediction effect. These results have implications for the formulation of policing strategies and the implementation of crime prevention and control.

INDEX TERMS Prediction of crime hotspots, machine learning, LSTM, built environment.

INTRODUCTION

The volume of spatiotemporal data related to public security has grown exponentially in recent years. However, not all of this data has been effectively utilized to address real-world problems. In the pursuit of crime prevention, numerous scholars have developed models to predict crime, with many relying solely on historical crime data to calibrate the predictive models.

Current research on crime prediction is primarily focused on two main aspects: crime risk area prediction and crime hotspot prediction. Crime risk area prediction, which is based on the relevant influencing factors of criminal activities, examines the correlation between criminal activities and the physical environment, drawing from the "routine activity theory." Traditional crime risk estimation methods typically identify crime hotspots from the historical distribution of crime cases and assume that the pattern will persist in subsequent periods.

Considering the proximity of crime places and the aggregation of crime elements, the terrain risk model tends to use crime-related environmental factors and crime history data and is relatively effective for long-term, stable crime hotspot prediction (reference 2). Many studies have carried out empirical research on crime prediction in different periods, combining demographic and economic statistics data, land use data, mobile phone data, and crime history data. Crime hotspot prediction aims to predict the likely location of future crime events and hotspots where the future events would concentrate (reference 8). A commonly used method is kernel density estimation (references 9–12). A model that considers temporal or spatial autocorrelations of past events performs better than those that fail to account for the autocorrelation (reference 13). Recently, machine learning algorithms have gained popularity. The most popular methods for crime trend prediction include K-Nearest Neighbour (KNN), random forest algorithm, support vector machine (SVM), neural network, and Bayesian model, among others [6]. Some studies have compared linear methods for crime trend prediction [14], while others have compared the Bayesian model and BP neural network [15], [16], and some have compared the spatiotemporal kernel density method with the random forest method in different periods of crime prediction [12].

Among these algorithms, KNN is recognized as an efficient supervised learning method [17], [18]. SVM is widely used in machine learning for its ability to implement classification and regression tasks as well as detect outliers [4], [19]. The random forest algorithm has been shown to have strong non-linear data processing ability and high prediction accuracy in multiple fields [20]–[23]. Naive Bayes (NB) is a classical classification algorithm with few parameters and is not sensitive to missing data.

Convolutional Neural Networks (CNN) exhibit robust expansibility, enhancing their expressive capability by incorporating deep layers to address more intricate classification problems [25], [26]. Long Short-Term Memory (LSTM) neural networks excel at extracting time-series features, proving effective in handling data with pronounced time-series trends [27]–[29]. This study aims to compare six machine learning algorithms, ultimately recommending the most effective one to showcase predictive performance both with and without the inclusion of covariates.

RELATED WORK

PRINCIPLES OF THEORETICAL CRIMINOLOGY IN THE PREDICTION OF CRIME HOTSPOTS

The focal point of crime hotspot prediction revolves around forecasting the future concentration of criminal events in a geographical space, drawing from the principles of theoretical criminology. Various criminological theories provide essential insights, guiding the understanding of location factors' significant influence on the formation and aggregation of criminal events. These theories also establish a fundamental mechanism for law enforcement to leverage crime hotspot information for prevention or control, primarily based on routine activity theory, rational choice theory, and crime patterns theory. These three theories are generally acknowledged as the theoretical underpinnings of situational crime prevention.

Routine activity theory [30], jointly proposed by Cohen and Felson in 1979, has undergone further development through integration with other theories. This theory posits that the occurrence of most crimes, particularly predatory crimes, necessitates the convergence of three elements: motivated offenders, suitable targets, and a lack of ability to defend in time and space.Rational choice theory [31], proposed by Cornish and Clarke, asserts that offenders' choices regarding location, goals, and methods can be explained by the rational balance of effort, risk, and reward.

Crime pattern theory [32] integrates routine activities theory and rational choice theory, providing a more nuanced explanation of the spatial distribution of criminal events. Individuals form "cognitive maps" and "activity space" through daily activities, while potential offenders utilize these cognitive maps to select specific locations for crimes in a relatively familiar space. When committing a crime, offenders tend to avoid unfamiliar places and opt for locations where "criminal opportunity overlaps with cognitive space" based on rational decision-making. The identification of crime hotspots relies not only on historical crime data but also on considering the environmental factors of these locations, as they exhibit distinct characteristics conducive to crime "production" or "attraction."

BUILT ENVIRONMENT DATA

Numerous studies currently emphasize the substantial impact of the urban built environment on criminal behavior, influencing crime opportunities and contributing to crime reduction and prevention. The 2007 Global Habitat Report highlighted the crucial role of built environment elements in the occurrence of criminal acts [33]. Point of interest (POI) data and road network density data serve as covariates in the crime prediction model.

CRIME PREDICTION WITH MACHINE LEARNING ALGORITHMS

Conventional methods typically identify crime hotspot areas based on the historical distribution of crime cases, assuming that past patterns will repeat in the future [7], [2]. This assumption proves reasonable for predicting long-term stable crime hotspots. The widely used Kernel Density Estimation (KDE) method effectively identifies such stable hotspot areas [10], [11]. The KDE method, specifically based on temporal autocorrelation, tends to outperform the general KDE method [38]. Liu et al. compared the random forest algorithm with the spatiotemporal KDE method, revealing the random forest's greater efficiency in smaller time scales and grid space units [12]. Gabriel et al. employed the Gated Localized Diffusion Network for crime prediction at the street segment level, demonstrating a significant increase in prediction accuracy compared to the traditional Network-time KDE method [39]. The efficacy of machine learning algorithms in processing non-linear relational data, as confirmed in various fields, including crime prediction, is characterized by faster training speeds, the ability to handle high-dimensional data, and the extraction of 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 neighbour, 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 neighbouring training instances.

B. RANDOM FOREST

The random forest comprises a set of tree classifiers {h(x, βk), k = 1... }, where the meta classifier h(x, βk) 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. The following conclusions are drawn: The LSTM model outperforms other models in terms of prediction accuracy, showcasing its ability to extract patterns and regularities from historical crime data. The incorporation of urban built environment covariates enhances the prediction accuracies of the LSTM model, surpassing the results of the original model using historical crime data alone. Our models exhibit improved prediction accuracies compared to other models in empirical research on crime hotspot prediction. For instance, the LSTM model in this paper outperforms previous research results, achieving a case hit rate of 59.9% and an average grid hit rate of 57.6%. As for future research, improvements can be made in several aspects:

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

While challenges persist, 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

”.