International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Survey On “Predicting Crime Hotspot Using Machine Learning”

Pawar Aarti Dipak, Mr.Abhale B.A., Kawade Sonali Dagadu

aartipawar1870@gmail.com, atul.abhale@gmail.com, Sonalikawade15@gmail.com

ABSTRACT

In this paper, a detailed study on crime classification and prediction using machine learning is presented. Crime prediction is of great significance to the formulation of policing strategies and the implementation of crime prevention and control. Machine learning is the current mainstream prediction method. However, few studies have systematically compared different machine learning methods for crime prediction. This paper takes the historical data of public property crime from 2015 to 2018 from a section of a large coastal city in the southeast of China as research data to assess the predictive power between several machine learning algorithms. Results based on the historical crime data alone suggest that the LSTM model outperformed KNN, random forest, support vector machine, naive Bayes, and convolutional neural networks. In addition, the built environment data of points of interests (POIs) and urban road network density are input into LSTM model as covariates prediction.

Keywords: Prediction of crime hotspots , machine learning, LSTM, built environment

1.INTRODUCTION

Crimes are the significant threat to the humankind, there are many crimes that happen at regular interval of time, perhaps it is increasing and spreading at a fast rate. With the help of machine learning algorithm, we can predict the type of crime which will occur in a particular area. Spatiotemporal data related to the public security have been growing at an exponential rate during the recent years.

However, not all data have been effectively used to tackle real-world problems.[1] In order to facilitate crime prevention, several scholars have developed models to predict crime . Most used historical crime data alone to calibrate the predictive models.

The research on crime prediction currently focuses on two major aspects: crime risk area prediction and crime hotspot prediction . The crime risk area prediction, based on the relevant influencing factors of criminal activities, refers to the correlation between criminal activities and physical environment, which both derived from the ‘‘routine activity theory’’ . Traditional crime risk estimation methods usually detect crime hotspots from the historical distribution of crime cases, and assume that the pattern will persist in the following time periods[2] . For example, 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 .

2.RELATED WORK

A. PRINCIPLES OF THEORETICAL CRIMINOLOGY IN PREDICTION OF CRIME HOTSPOTS

In [4] Authors state several previous contributions revealing the fact that crime is concentrated at some micro places within the city with high intensity; such places are termed as hotspots. Authors also stated that spatial patterns of crime and use of these features to predict crime require some theoretical framework. So geography of crime that is spatial features can help in certain applications of policing which includes Hotspot policing, predictive policing and geographic profiling. But all these predictions and analytics will be accurate when it is combined with standard crime theories like Social disorganization theory, Rational choice perspective theory, Routine activity theory and Crime pattern theory.

The focus of crime hotspot prediction is to forecast future concentration of criminal events in a geographical space. Theoretical criminology provides the necessary theoretical basis. Specifically, several related criminological theories not only provide guidance for us to understand the important influence of location factors in the formation and aggregation of criminal events, but also provide a basic mechanism for the police to use information of crime hot spots for crime prevention or control. It mainly includes routine activity theory, rational choice theory, and crime patterns theory[5]. These three theories are generally considered as the theoretical basis of situational crime prevention.

Routine activity theory was jointly proposed by Cohen and Felson in 1979, and has now been further developed through integration with other theories. This theory believes that the occurrence of most crimes, especially predatory crimes, needs the convergence of the three elements including motivated offenders, suitable targets, and lack of ability to defend in time and space.

B. BUILT ENVIRONMENT DATA

At present, a large number of studies show that the urban built environment has a significant impact on urban criminal behavior, through the impact of crime opportunities to reduce and prevent crime. In the 2007 Global Habitat Report, it was pointed out that the elements of the built environment have an important impact on the occurrence of criminal acts . Point of interests (POIs) data and road network density data are considered as covariates in the crime prediction model.

1) POI DATA

The urban infrastructure data POI includes the location information and attribute information of various urban facilities . Catering facilities, shopping malls and stores are usually located in places with convenient transportation and large flow of people, gathering a large number of different groups of people to generate the targets for the criminals, while entertainment places attract criminals . These POIs are selected as covariates of the prediction model.

2) ROAD NETWORK DENSITY

The conventional definition of road network density refers to total length of roads divided by the size of an areal unit. The area with a denser road network attracts greater flow of people, including potential victims and criminals. Previous studies have shown that the density of road network has an impact on crime rate, especially in public space .[7]

C. CRIME PREDICTION WITH MACHINE LEARNING ALGORITHMS

The traditional methods usually detect the crime hotspot area from the historical distribution of crime cases, and assume that the past pattern is to be repeated in the future . This assumption tends to be reasonable for predicting long-term stable crime hotspots. The commonly used KDE method can effectively identify such stable hotspot areas . The KDE method based on temporal autocorrelation tends to outperform the general KDE method Liu et al. Compared the random forest and spatiotemporal KDE method, found that the random forest algorithm is more efficient than the traditional spatiotemporal KDE method in the smaller time scale and grid space unit Gabriel et al. used the Gated Localized Diffusion Network for crime prediction at the street segment level . Compared with the traditional Network-time KDE method, the diffusion network approach significantly increased the prediction accuracy. The ability of machine learning algorithm in processing non-linear relational data has been confirmed in many fields, including crime prediction. It has a faster training speed, can handl very high-dimensional data, and can also extract the characteristics of the data[8].

3. PREDICTION MODEL

In this paper, random forest algorithm, KNN algorithm, SVM algorithm and LSTM algorithm are used for crime prediction. First, historical crime data alone are used as input to calibrate the models. Comparison would identify the most effective model. Second, built environment data such as road network density and poi are added to the predictive model as covari ates, to see if prediction accuracy can be further improved[9].

A. KNN

KNN, also known as k-nearest neighbor, takes the feature vector of the instance as the input, calculates the distance between the training set and the new data feature value, and then selects the nearest K classification. If k = 1, the nearest neighbor class is the data to be tested. KNN’s classification decision rule is majority voting or weighted voting based on distance. The majority of k neighboring training instances of the input instance determines the category of the input instance.

B. RANDOM FOREST

The random forest is a set of tree classifiers {h(x, βk),k =1. . . }, in which the meta classifier h(x, βk) is an uncut regression tree constructed by CART algorithm; x is the input vector; βk is an independent random vector with the same distribution, and the output of the forest is obtained by voting. The randomness of random forest is reflected in two aspects: one is to randomly select the training sample set by using bagging algorithm; the other is to randomly select the split attribute set. Assuming that the training sample has M attributes in total, we specify an attribute number F ⩽ M, in each internal node, randomly select F attributes from M attributes as the split attribute set, and take the best split mode of the f attributes Split the nodes. The multi decision tree is made up of random forest, and the final classification result is determined by the vote of tree classifier.

C. SVM

SVM, based on statistical learning theory, is a data mining method that can deal with many problems such as regres sion (time series analysis) and pattern recognition (classi- fication problem, discriminant analysis) very successfully. The mechanism of SVM is to find a superior classification hyperplane that meets the classification requirements, so that the hyperplane can ensure the classification accuracy and can maximize the blank area on both sides of the hyperplane. In theory, SVM can realize the optimal classification of linear separable data

D. CNN

CNN uses one-dimensional convolution for sequence pre diction, which is the convolution sum of discrete sequences. To convolve the sequence, CNN first finds a sequence with a window size of kernel_size, and perform convolution with the original sequence to obtain a new sequence expression. The convolutional network also includes a pooling operation, which is to filter the features extracted by the convolution to get the most useful characteristics.

E. LSTM

LSTM is a kind of deep neural network based on RNN. The core of LSTM is to add a special unit (memory module) to learn the current information and to extract the related information and rules between the data, so as to transfer the information. LSTM is more suitable for deep neural network calculation because of memory module to slow down infor mation loss[12]. Each memory module has three gates, including input gate (it), forget gate (ft), and output gate (ot). They are used to selectively memorize the correction parameters of the feedback error function as the gradient decreases.

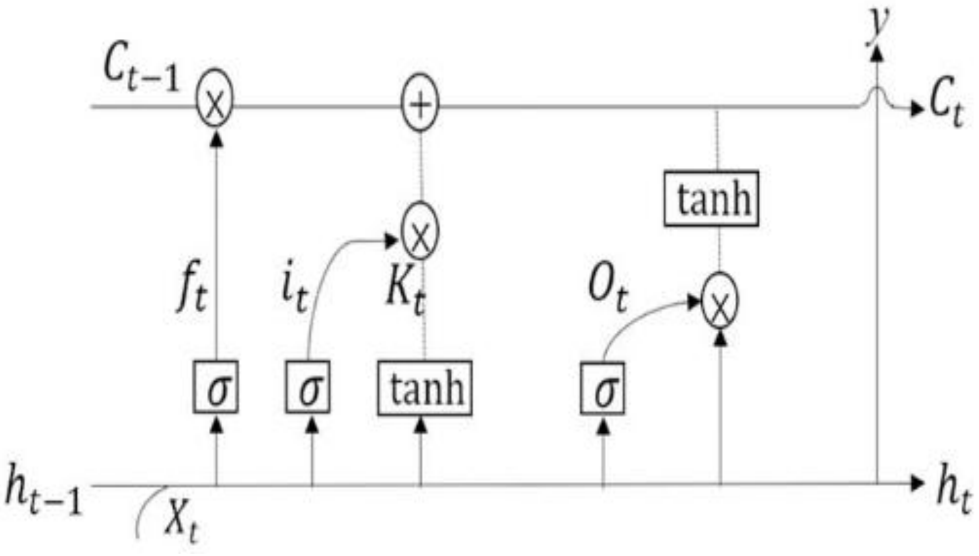

FIGURE 1. The structure chart of LSTM algorithm

Proposed System:

In the proposed system, random forest algorithm, KNN algorithm, SVM algorithm and LSTM algorithm are used for crime prediction. First, historical crime data alone are used as input to calibrate the models. Comparison would identify the most effective model. Second, built environment data such as road network density and poi are added to the predictive model as covariates, to see if prediction accuracy can be further improved[11].

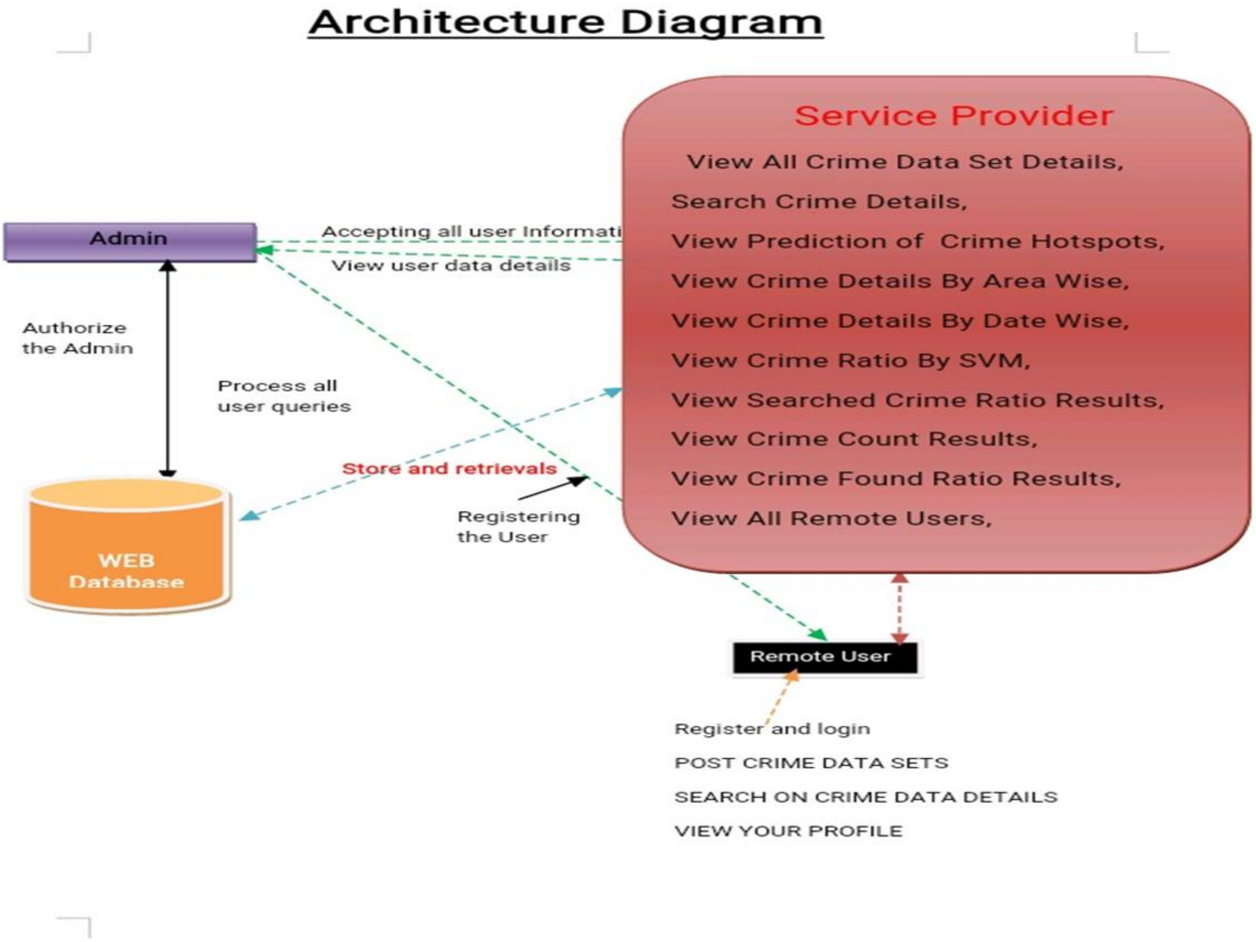

FIGURE 2. Architecture of Predicting crime Hotspot

2.APPLICATION

Security
-Retail
-Media
-Banking
-Prevention[5]

3.CONCLUSION

In this paper, six machine learning algorithms are applied to predict the occurrence of crime hotspots in a town in the southeast coastal city of China. The following conclusions are drawn:1) The prediction accuracies of LSTM model are better than those of the other models. It can better extract the pattern and regularity from historical crime data. 2) The addition of urban built environment covariates further improves the prediction accuracies of the LSTM model. The prediction results are better than those of the original model using historical crime data alone[7].

Our models have improved prediction accuracies, compared with other models. In empirical research on the prediction of crime hotspots, Rummens et al. used historical crime data at a grid unit scale of 200 m×200 m, using three models of logistic regression, neural network, and the combination of logistic regression and neural network . In the biweekly forecast, the highest case hit rate for the two-robbery type is 31.97%, and the highest grid hit rate is 32.95%; Liu et al. Used the random forest model to predict the hot spots in multiple experiments in two weeks under the research scale of 150 m × 150 m . The average case hit rate of the model was 52.3%, and the average grid hit rate was 46.6%. The case hit rate of the LSTM model used in this paper was 59.9%, and the average grid hit rate was 57.6%, which was improved compared with the previous research results,

4.REFERENCES

[1] U. Thongsatapornwatana, ‘‘A survey of data mining techniques for analyzing crime patterns,’’ in Proc. 2nd Asian Conf. Defence Technol. (ACDT),
Jan. 2016, pp. 123–128.
[2] M. Cahill and G. Mulligan, ‘‘Using geographically weighted regression to explore local crime patterns,’’ Social Sci. Comput. Rev., vol. 25, no. 2, pp. 174–193, May 2007.
[3] H. Berestycki and J.-P. Nadal, ‘‘Self-organised critical hot spots of criminal activity,’’ Eur. J. Appl. Math., vol. 21, nos. 4–5, pp. 371–399, Oct. 2010.
[4] W. Gorr and R. Harries, ‘‘Introduction to crime forecasting,’’ Int. J. Forecasting, vol. 19, no. 4, pp. 551–555, Oct. 2003.
[5] W. H. Li, L. Wen, and Y. B. Chen, ‘‘Application of improved GA-BP neural network model in property crime prediction,’’ Geomatics Inf. Sci. Wuhan Univ., vol. 42, no. 8, pp. 1110–1116, 2017.
[6] R. Haining, ‘‘Mapping and analysing crime data: Lessons from research and practice,’’ Int. J. Geogr. Inf. Sci., vol. 16, no. 5, pp. 203–507, 2002.
[7] S. Chainey, L. Tompson, and S. Uhlig, ‘‘The utility of hotspot mapping for predicting spatial patterns of crime,’’ Secur. J., vol. 21, nos. 1–2, pp. 4Feb. 2008.
[8] M. I. Jordan and T. M. Mitchell, ‘‘Machine learning: Trends, perspectives, and prospects,’’ Science, vol. 349, no. 6245, pp. 255–260, Jul. 2015.
[9] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, ‘‘Improved method of classification algorithms for crime prediction,’’ in Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST), 2015, pp. 250–255.
[10] Q. Zhang, P. Yuan, Q. Zhou, and Z. Yang, ‘‘Mixed spatial-temporal characteristics based crime hot spots prediction,’’ in Proc. IEEE 20th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD), May 2016, pp. 97–101.
[11] H. Tyralis and G. Papacharalampous, ‘‘Variable selection in time series forecasting using random forests,’’ Algorithms, vol. 10, no. 4, p. 114,Oct. 2017.