Vol 12, Issue 07, July/2021
ISSN NO:0377-9254

Comparison of Machine Learning Algorithms for Predicting Crime Hotspots

[1]Bera Srujana , [2]Chaganti B N Lakshmi

[1]M.Tech Scholar, Department of Computer Science and Engineering, TKR college of Engineering and Technology, Hyderabad, India.

[2]Professor, Department of Computer Science and Engineering, TKR college of Engineering and Technology, Hyderabad, India.

1. ABSTRACT

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 Bays, and convolution 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. It is found that the model with built environment covariates has better prediction effect compared with the original model that is based on historical crime data alone. Therefore, future crime prediction should take advantage of both historical crime data and covariates associated with criminological theories. Not all machine learning algorithms are equally effective in crime prediction.

2. INTRODUCTION

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. In order to facilitate crime prevention, several scholars have developed models to predict crime [1]. 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 [2], [3] and crime hotspot prediction [4], [5].

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" [6]. 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 [7]. 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].

3. LITERATURE SURVEY

The information exploration are actually info evaluating methods which utilized to evaluate unlawful task info formerly placed from various means discover developments and designs in illegal things in current times. In added, they could be put on to increase capabilities in solving the criminal activities quicker but also are set on quickly notify those tasks being unlawful. Nevertheless, you will find numerous facts methods which could be mining.

In order to make sure you're in a position to develop productivity of crime growth, it's essential to buy the given info exploration methods suitably. This particular report response the literatures on many info exploration programs, especially software which utilized on re resolve the crimes. Find out in addition throws lightweight in analysis holes as well as problems of criminal specifics exploration.

In additional compared to that, this particular papers supplies understanding related to the information exploration to uncover the patterns and trends in illegal exercise to appropriately be also utilized as well as to be an assist for novices within the exploration of crime facts mining.

The analysis offered right here enjoys 2 goals which are important. The very first will be using chance terrain (RTM) which is actually modeling to the task of yours which is actually criminal of. The chance landscapes maps which had been generated from RTM take advantage of bunch of contextual tips very related to the opportunity structure of shootings to estimated likelihood of possible shootings as they're shipped throughout a geography.

The goal which is actually 2nd to examine the predictive power with this particular threat landscapes maps more than 2 time which is actually six month, in order to evaluate each one of them against the predictive possibility of retrospective area maps.

Outcome declare which possibility landscapes present a forecast that's mathematically large of shootings across a range of cut details as they're substantially much a lot more precise compared to retrospective spot mapping that's sexy. Together with this, choices terrain maps emit data which may be operational zed by police directors effectively and easily, like for pointing authorities patrols to coalesced regions which are high risk.

4. ALGORITHM

In this Project, 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.

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 determine the category of the input instance.

B. RANDOM FOREST:-The random forest is a set of tree classifiers {h(x, eke), k = 1 . . .}, in which the Meta classifier h(x, eke) is an uncut regression tree constructed by CART algorithm; x is the input vector; eke 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 regression (time series analysis) and pattern recognition (classification problem, discriminate analysis) very successfully. The mechanism of SVM is to find a superior classification hyper plane that meets the classification requirements, so that the hyper plane can ensure the classification accuracy and can maximize the blank area on both sides of the hyper plane. In theory, SVM can realize the optimal classification of linear separable data.

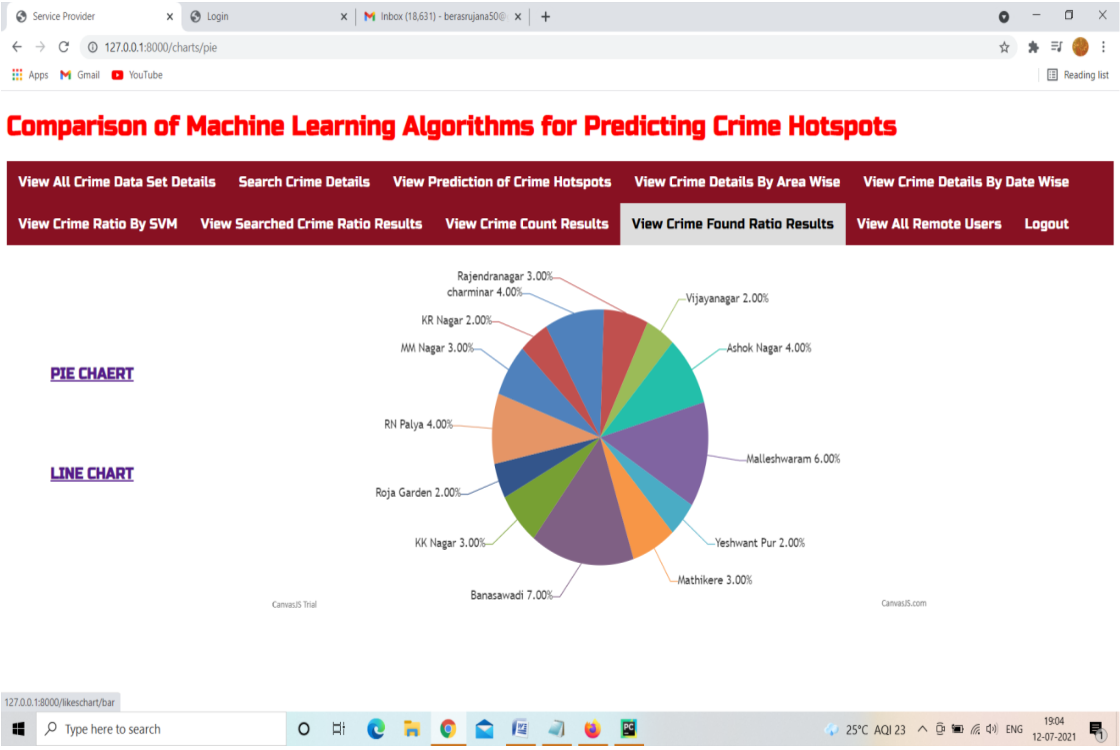
D. NB: - In the field of probability and statistics, Bayesian theory predicts the occurrence probability of an event based on the knowledge of the evidence of an event. In the field of machine learning, the naïve Bays (NB) classifier is a classification method based on Bayesian theory and assuming that each feature is independent of each other. In abstract, NB classifier is based on conditional probability, to solve the probability that a given entity belongs to a certain class.

E. CNN: - CNN uses one-dimensional convolution for sequence prediction, 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 convolution network also includes a pooling operation, which is to filter the features extracted by the convolution to get the most useful characteristics.

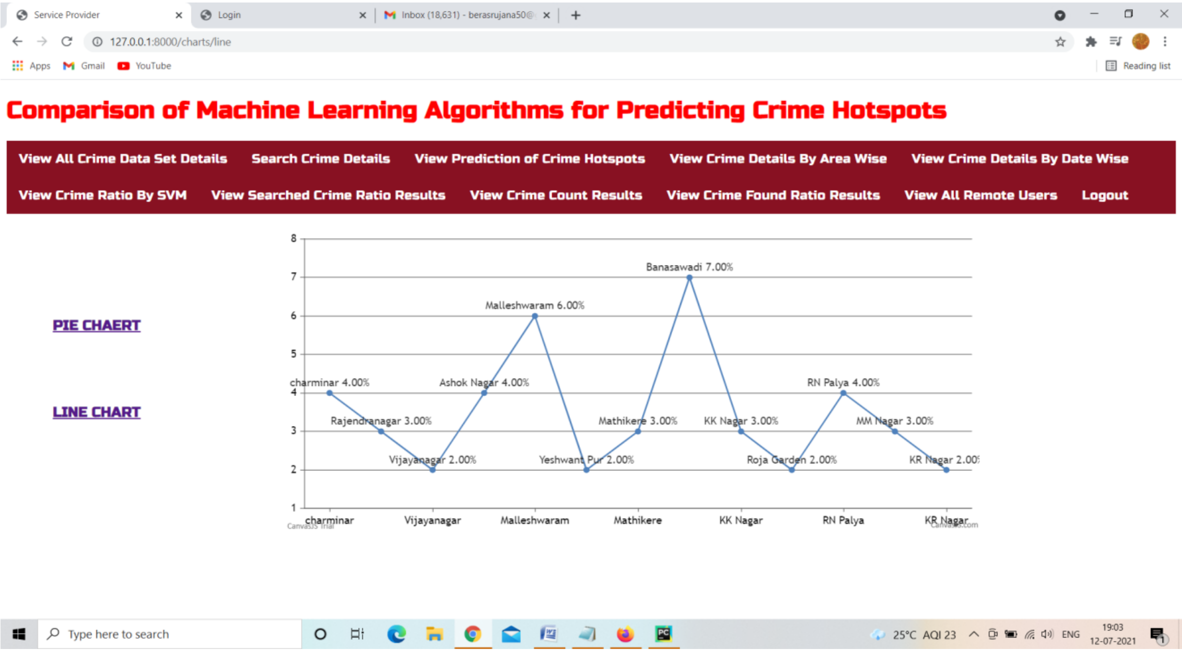
F. LSTM: - LSTM is a kind of deep neural network based on RNN. The core of LSTM is to add a special unit (memory module) to learn the current information and to extract the related information and rules between the data, so as to transfer the information. LSTM is more suitable for deep neural network calculation because of memory module to slow down information loss. 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

5. RESULT ANALYSIS

The experiment carried out for the distinctive machine learning models and KDD dataset, and the results obtained, Intrusion Detection Systems (IDS) are among the ways against these attacks. Furthermore, modern technologies of upcoming generation networks such as for instance Wireless network normally referred to as We If have emerged, which call for a notable comprehension of the key difficulties and constraints that deal with the layout as well as setup of an IDS for such methods. IDS often have to boost the performance of its in conditions of raising the precision and lessening false alarms. In machine learning grounded IDS, integrating effective feature selection as well as attribute dimensionality minimization with intrusion detection has proven to be a booming strategy since it is able to assist in choosing probably the most informative features and minimize the function dimensionality from the full set of characteristics.



Displaying of Crime Count Results through pie Chart



Comparison of Machine Learning Algorithms for Predicting Crime Hotspots

Displaying of Crime Count Results through Line Chart

CONCLUSION

In this project, 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. Our models have improved prediction accuracies, compared C] with other models. In empirical research on the prediction of crime hotspots, Rumens et al. used historical crime data at a grid unit scale of 200 m200 m, using three models of logistic regression, neural network, and the combination of logistic regression and neural network [41]. 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 150m150m [23]. 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, for the future research, there are still some aspects to be improved. The rest is the temporal resolution of the prediction. Felon et al. revealed that the crime level changes with time [43] some studies have shown that it is useful to check the variation of risks during the day [44].We chose two weeks as the prediction window. It does not capture the impact of crime changes within a week, let alone the change within a day. The sparsely of data makes the prediction of crime event difficult if the prediction window is narrowed down to day of a week or hour within a day. There is no viable solution to this challenging problem at this time. The second is the spatial resolution of the grid. In this paper, the grid size is 150m ------150m. Future research will assess the impact of changing grid sizes on prediction accuracy. Third, the robustness and generality of the findings of this paper needs to be tested in other study areas. Nonetheless, the findings of this

REFERNCES

[1] U. Thongsatapornwatana, ``A survey of data mining techniques for analyzingCrime patterns," in Proc. 2nd Asian Conf. Defense Technol. (ACDT),Jan. 2016, pp. 123128.

[2] J. M. Kaplan, L. W. Kennedy, and J. Miller, ``Risk terrain modeling:
Brokering criminological theory and GIS methods for crime forecasting,"
Justice Quart., vol. 28, no. 2, pp. 360381, Apr. 2011.

[3] M. Cahill and G. Mulligan, ``Using geographically weighted regression
To explore local crime patterns," Social Sci. Compute. Rev., vol. 25, no. 2,pp. 174193, May 2007.

[4] A. Almehmadi, Z. Judoka, and R. Alkali, ``Language usage on Twitter
predicts crime rates," in Proc. 10th Int. Conf. Secure. Inf. Newt. (SIN), 2017,pp. 307310.

[5] H. Berestycki and J.-P. Nodal, ``Self-organized critical hot spots of criminal
activity," Eur. J. Appl. Math., vol. 21, nos. 45, pp. 371399, Oct. 2010.

[6] K. C. Baumgartner, S. Ferrari, and C. G. Sulfate, ``Bayesian network
modeling of offender behavior for criminal prolong," in Proc. 44th
IEEE Conf. Decius. Control, Eur. Control Conf. (CDC-ECC), Dec. 2005,pp. 27022709.

[7] W. Garr and R. Harries, ``Introduction to crime forecasting," Int. J. Forecasting, vol. 19, no. 4, pp. 551555, Oct. 2003.

[8] W. H. Li, Lawmen, and Y. B. Chen, ``Application of improved GA-BP neural
network model in property crime prediction," Geometrics Inf.