Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 136 (2018) 472–478

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Young Scientist Conference on Computational Science

Crime rate prediction in the urban environment using social factors

Varvara Ingilevich[a*], Sergey Ivanov[b]

[a,b]*ITMO University, Saint-Petersburg, Russian Federation*

Abstract

The aim of this study is to compare different approaches to the problem of forecasting the number of crimes in different areas of the city. During this research we studied three types of predictive models: linear regression, logistic regression and gradient boosting. The predictive factors used in these models have been selected using the feature selection techniques. This approach allowed us to increase the accuracy of predictions and to avoid the model's overfitting. The obtained models were tested on criminal data of the city of Saint-Petersburg. We compared the results of model predictions and determined that gradient boosting is the most appropriate method for the problem of crime rate prediction in certain urban area.

Keywords: Predictive analysis; Crime analysis; Regression models; Forecasting methods; Feature Selection

1. Introduction

Urbanization creates a lot of social problems. One of these problems inherent in all cities of the world is crime. Police databases accumulate a large amount of data that could be analyzed in order to reduce crime rates. The analysis of criminal activity and prediction of number of crimes remains one of the most interesting problem for researchers.

Several research studies show that the criminal world organizes a complex network with its own rules [1]. In order to make a prediction of crime rate with high accuracy it is important to understand the "nature" of a crime.

[*] Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: v_ingilevich@niuitmo.ru

10.1016/j.procs.2018.08.261

Previous research on the spatial analysis of criminal activity has established that crimes are not randomly organized in space [2]. In [3], [4], [5], [6] researchers note that there are a lot of external factors that influence the formation and displacement of crime hotspots in a city. The main factors identified by scientists are Gross domestic product (GDP), population density, unemployment and number of homeless in the streets. In addition to social factors, there are also different spatial indicators. For example, in [7] Yanqing Xu et al. have postulated a link between street lights and spatial criminal patterns. In [8] the researchers have determined the relationship between traffic and crime. In [9], [10], [11] the authors determined that the weather also influence the criminal activity.

Several approaches to crime prediction have been proposed by researchers earlier. In [12] Hyeon-Woo Kang and Hang-Bong Kang proposed the predictive method based on deep neural network. With their model they achieved the accuracy of 74.35%. In [13] the scientists used the algorithm of random forest regressor to measure the impact of urban factors on homicides and predict the future number of crimes of this type. However, scientists note that the model developed in [13] works well only with small datasets. Renjie et al. [14] apply the Bayesian Learning Theory to implement the model, that could predict serial crime. Anneleen et al. [15] explored 3 approaches to predictive modeling: logistic regression, neural network, and ensemble model. The obtained models were tested on 3 types of crime: home burglary, street robbery and battery. The results with the highest accuracy have been achieved using the logistic regression. Several models, developed not only in the environmental but also in temporal context, were proposed by researchers [16], [17]. Cheng et al. [17] have implemented ARIMA (autoregressive integrated moving average), SES (simple exponential smoothing) and HES (Holt-Winters Exponential Smoothing) models to make a short-term forecasting of property crime. The comparison of prediction results shows that the ARIMA model has the highest accuracy.

In the existing models the authors pay a little attention to the stage of the selection of factors that influence the criminal rate. However, this stage is one of the main ones in the procedure of modeling. The careful selection of features and their filtering improve the performance of predictive models and avoid their complexity and overfitting, is it was shown in the modeling tasks in other thematic areas [18], [19], [20], [21]. For example, in [19] authors use the feature selection method for prediction of solar radiation, Nicole et al. [20] implement this approach for handwritten character recognition. In the both of these studies the authors highlight that the use of this technique greatly improves the accuracy of the models.

In the present study, we compare 3 approaches to crime rate prediction: linear regression, logistic regression and gradient boosting. To avoid the overfitting of the models, we apply the clustering and feature selection technique.

2. Problem Statement

The aim of this essay is to determine the model that could make the crime rate prediction with the best accuracy. The three following models were considered in this work:

- Linear regression
- Logistic regression
- Gradient boosting (GB)

The structure of the research is presented in Fig. 1

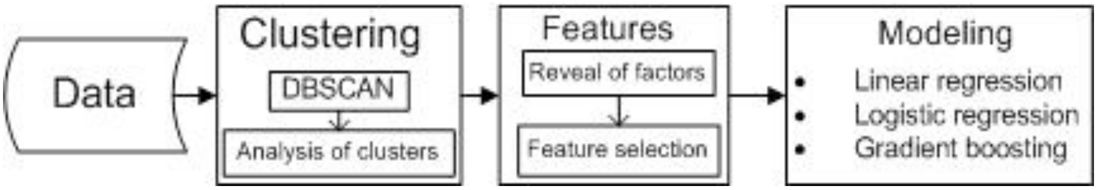

Fig. 1. Structure of the research

3. Data

The data used in this research are provided by the Ministry of Internal Affairs of the Russian Federation. It includes records of crimes in Saint-Petersburg that were registered from 1/1/2014 to 2/28/2017. The dataset contains the following information:

- Date of crime
- Coordinates
- Description

Example of one record is presented in Table 1.

Table 1. Example of record.

Date	Longitude	Latitude	Description
01.01.2016 0:05:00	59.875348	30.312553	Knife wound
02.01.2016 0:11:00	59.939208	30.279163	Brawl

For this study street crimes were selected from the total number of data. The selected records were divided into three main groups according to the keywords in the description:

- banditry (e.g. 'shooting on the street', 'assault')
- massacre (e.g. 'beating in apartment', 'fight in the street')
- robbery (e.g. 'theft', 'theft from car')

The category "robbery" is the most numerous of the selected types and contains 142452 records.

4. Methods

4.1. Clustering

In order to understand better the nature of the data, it is important to remove the noise from the dataset. One of the most effective method for this task is clustering. The dataset used in this research represents the high-density data, that is why for clustering we have chosen the method DBSCAN (Density-based spatial clustering of applications with noise). The major advantage of DBSCAN is that this algorithm finds arbitrarily shaped clusters.

The clustering was performed using the Python package *scikit-learn* [22]. As a result of the procedure we obtained the set of separated individual clusters for each type of crime. The numbers of clusters are presented in table 2. The visualization of clusters is presented in Fig. 2.

Table 2. Number of clusters

Type of crime	Number of clusters	Number of points in clusters
banditry	30	3347
massacre	33	86200
robbery	14	3133

The resulting clusters allowed us to study the spatial patterns of each type of crime in more details. We studied the various spatial "factors" that are located inside the clusters and could probably influence the criminal activity. Using the correlation analysis and crime mapping techniques we revealed the determinant spatial indicators that influence the formation of clusters. These are the following factors: number of population, police stations, schools, malls, churches, alcohol shops, number of buildings and bars.

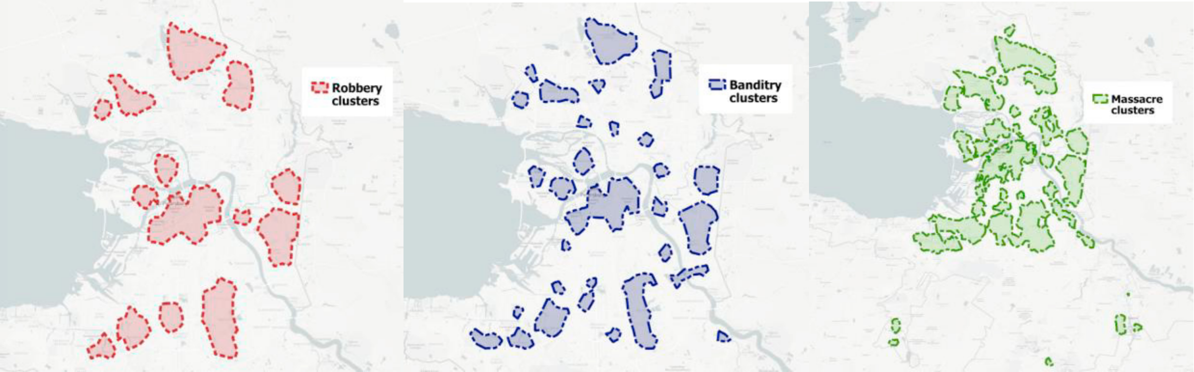

Fig. 2. Clusters for different types of crime

4.2. Feature selection

We defined 8 factors that influence the criminal activity, but the use of all factors in predictive model leads to it complexity and overfitting. The feature selection technique is the key to solving this problem, and it plays a vital role in creating of an effective predictive model.

In our research we applied the method of chi-squared test to the set of selected features. The method of chi-squared test has several advantages such as ease of computation and simple interpretation of results. Since each type of crime has its own "nature", the feature selection method was applied to every crime category separately. However, the resulting set of features turned out to be the same for all types of crimes.

The feature selection technique allowed us to get the insights about the relative importance of the features with the target variable. As a result of filtering, the original number of features was reduced to 4 and the final features that were used for modeling are population number, bars, churches and schools. The weight coefficients for each feature are respectively the following: 0.6, 0.3, 0.1, 0.2. As it can be seen, the factor of availability of bars in a certain area is the most decisive for the criminal activity.

All selected features are from numerical type.

4.3. Modeling

We attempt to estimate the number of crimes, using 3 different methods: linear regression, logistic regression and gradient boosting. As the training data we consider the points from clusters and the independent variables of models are the features defined in section 4.2.

The results of modeling were evaluated using the Mean Absolute Error (MAE) and the coefficient of determination (R^2). The R^2 coefficient ranges from 0 to 1 and measures, how well the model predicts the values.

Several preliminary settings have been made before the modeling. As the logistic regression model output ranges from 0 to 1, the values, on which this type of the model was fitted, were normalized relatively to the value corresponding to the cluster, that has the largest number of crimes.

As a training data we used the 70% of the number of clusters for each type, as the test data we considered the remaining clusters. To evaluate the results of the models, we implemented the cross-validation and determined the average values of the metrics MAE and R^2, that are presented in Table 3.

Table 3. Evaluation of prediction

Type of crime	Method	MAE (Cross-validation)	R² (Cross-validation)	MAE of prediction for grid
banditry	Linear regression	17	0.9	18
	Logistic regression	216	0	156
	Gradient boosting	17	0.9	31
massacre	Linear regression	231	0.9	110
	Logistic regression	5021	0	3516
	Gradient boosting	56	0.9	264
robbery	Linear regression	28	0.9	364
	Logistic regression	189	0.3	5861
	Gradient boosting	34	0.9	470

From the data in Table 3, it is apparent that the logistic regression model works worst, while the other two models make predictions with high accuracy. The R² parameter shows that the linear regression and gradient boosting models are the most relevant to the data.

To assess the performance of the models from the more practical point of view, we tested the models for the whole city of Saint-Petersburg. For this purpose, we have created the hexagonal grid and calculated the number of features of each type for every polygon. The grid contains 130 polygons and the area of one polygon is equal to the average areas of all clusters, 21 km².

The results of the model prediction for the whole city are presented in Fig.3 The linear regression and the gradient boosting models still give the best results. The least mean absolute error, that could be found in the Table 3 in the column “MAE of prediction for grid”, relates to the linear regression model. However, the pitfall of the linear regression method is that it produces the negative predicted values, that is theoretically impossible for the value that indicates the number of crimes. These predicted negative values can be clearly seen in the Fig.3 for the massacre crime type.

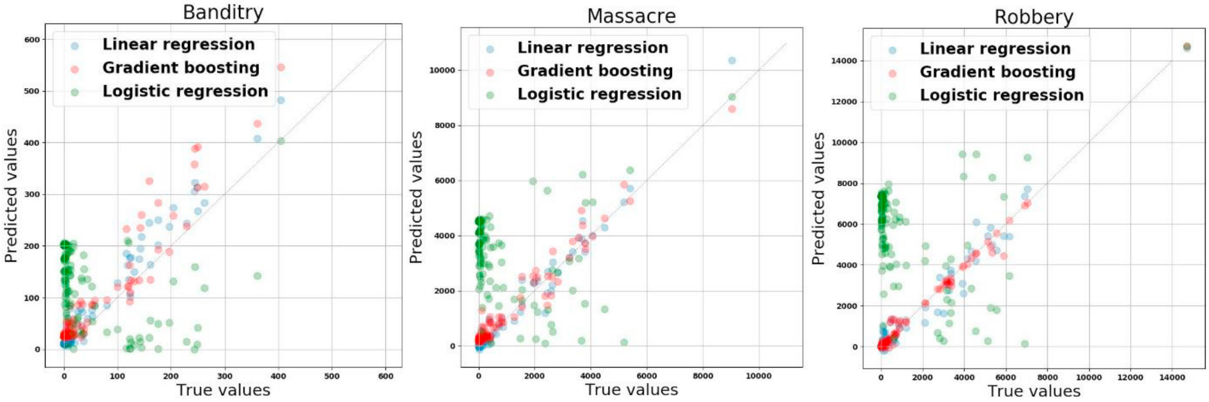

Fig. 3. Predicted number of crimes with different types of models.

Therefore, the gradient boosting method is the most appropriate for the task of crime rate prediction. The visualization of the results of the crime rate prediction using the GB model is presented in Fig.4.

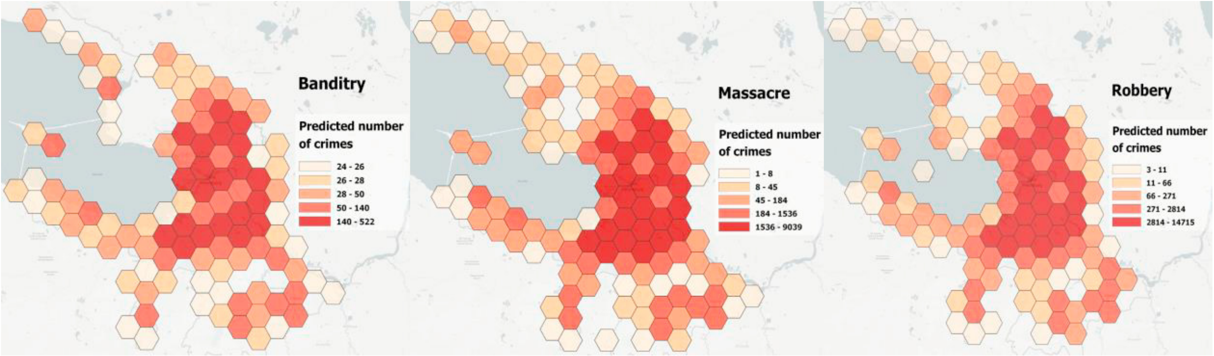

Fig. 4. Predicted number of crimes with GB model.

5. Conclusion and future work

The main goal of this research work was to determine how different statistical tools can be used to predict the number of crimes in a specific urban area. In the process of predictive model's implementation, we used the clustering technique to determine the spatial patterns of the crime and reveal the factors that influence the criminal activity. We also applied the feature selection method to select the most important factors and to avoid the overfitting of future model. Three types of models were used to predict the crime rate. During the study we revealed some disadvantages of the models, such as the prediction of negative values, that were made by the linear regression model. The best accuracy was achieved by the gradient boosting model. The results of our research can help the police to find the most optimal places for the location of police stations in a city and to determine the strategies to reduce the crime rate.

A natural progression of this work is to perform the temporal analysis of criminal data, and to implement the model that could predict not only the number of crimes, but also the time of the crime

Acknowledgements

This research is financially supported by The Russian Foundation for Basic Research, Agreement #15-29-07034, The Russian Science Foundation, Agreement #17-71-30029 with co-financing of Bank Saint Petersburg.

References

[1] P. A. C. Duijn, V. Kashirin, and P. M. A. Sloot, "The relative ineffectiveness of criminal network disruption," *Sci. Rep.*, vol. 4, 2014.

[2] S. Curtis-ham and D. Walton, "Mapping crime harm and priority locations in New Zealand : A comparison of spatial analysis methods," *Appl. Geogr.*, vol. 86, pp. 245–254, 2017.

[3] O. K. Ha and M. A. Andresen, "Journal of Criminal Justice Unemployment and the specialization of criminal activity : A neighborhood analysis," vol. 48, pp. 1–8, 2017.

[4] J. Phillips and K. C. Land, "The link between unemployment and crime rate fluctuations: An analysis at the county, state, and national levels," *Soc. Sci. Res.*, vol. 41, no. 3, pp. 681–694, 2012.

[5] J. J. Allen, C. A. Anderson, and B. J. Bushman, "The General Aggression Model," *Curr. Opin. Psychol.*, vol. 19, pp. 75–80, 2018.

[6] M. Coccia, "A Theory of general causes of violent crime: Homicides, income inequality and deficiencies of the heat hypothesis and of the model of CLASH," *Aggress. Violent Behav.*, vol. 37, no. November 2016, pp. 190–200, 2017.

[7] Y. Xu, C. Fu, E. Kennedy, S. Jiang, and S. Owusu-Agyemang, "The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan," *Cities*, no. October 2017, pp. 0–1, 2018.

[8] L. P. Beland and D. A. Brent, "Traffic and crime," *J. Public Econ.*, vol. 160, no. March, pp. 96–116, 2018.

[9] S. J. Michel *et al.*, "Investigating the relationship between weather and violence in Baltimore, Maryland, USA," *Injury*, 2016.

[10] S. A. Salleh, N. S. Mansor, Z. Yusoff, and R. A. Nasir, "The Crime Ecology: Ambient Temperature vs. Spatial Setting of Crime (Burglary)," *Procedia - Soc. Behav. Sci.*, vol. 42, no. July 2010, pp. 212–222, 2012.

[11] J. Tiihonen, P. Halonen, L. Tiihonen, H. Kautiainen, M. Storvik, and J. Callaway, "The Association of Ambient Temperature and Violent Crime," *Sci. Rep.*, vol. 7, no. 1, pp. 1–7, 2017.

[12] H.-W. Kang and H.-B. Kang, “Prediction of crime occurrence from multi-modal data using deep learning,” *PLoS One*, vol. 12, no. 4, p. e0176244, 2017.

[13] L. G. A. Alves, H. V Ribeiro, and F. A. Rodrigues, “Crime prediction through urban metrics and statistical learning,” *Physica A*, vol. 505, pp. 435–443, 2018.

[14] R. Liao, X. Wang, L. Li, and Z. Qin, “A novel serial crime prediction model based on Bayesian learning theory,” *2010 Int. Conf. Mach. Learn. Cybern.*, no. July, pp. 1757–1762, 2010.

[15] A. Rummens, W. Hardyns, and L. Pauwels, “The use of predictive analysis in spatiotemporal crime forecasting : Building and testing a model in an urban context,” *Appl. Geogr.*, vol. 86, pp. 255–261, 2017.

[16] W. Gorr, A. Olligschlaeger, and Y. Thompson, “Short-term forecasting of crime,” *Int. J. Forecast.*, vol. 19, no. 4, pp. 579–594, 2003.

[17] P. Chen, H. Yuan, and X. Shu, “Forecasting crime using the ARIMA model,” *Proc. - 5th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2008*, vol. 5, no. January 2017, pp. 627–630, 2008.

[18] J. D. López-cabrera and J. V Lorenzo-ginori, “Feature selection for the classification of traced neurons,” *J. Neurosci. Methods*, vol. 303, pp. 41–54, 2018.

[19] M. Almaraashi, “Investigating the impact of feature selection on the prediction of solar radiation in different locations in Saudi Arabia ଝ,” *Appl. Soft Comput. J.*, vol. 66, pp. 250–263, 2018.

[20] N. D. Cilia, C. De Stefano, F. Fontanella, and A. Scotto, “A ranking-based feature selection approach for handwritten character recognition,” *Pattern Recognit. Lett.*, vol. 0, pp. 1–10, 2018.

[21] P. Sciencedirect, M. Masila, A. Jalil, F. Mohd, N. Maizura, and M. Noor, “ScienceDirect ScienceDirect A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection,” *Procedia Comput. Sci.*, vol. 116, pp. 113–120, 2017.

[22] É. D. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, “Scikit-learn: machine learning in python,” *Journal of machine learning research*, 2011. [Online].