Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 6, No. 2, December 2018, pp.360-364
Available online at: http://pen.ius.edu.ba

A crime prediction model based on spatial and temporal data

Hicham Aitelbour[1], Soumia Ounacer[1], Yassine Elghomari[1], Houda Jihal[1], Mohamed Azzouazi[1]
[1] Information Technology and Modeling Laboratory, Hassan II University, Casablanca, Morocco

Article Info

Received Jun 29, 2018

Keyword:

Big data
Crime prediction
Machine learning
Deep learning

ABSTRACT

In a world where data has become precious thanks to what we can do with it like forecasting the future, the fight against crime can also benefit from this technological trend. In this work, we propose a crime prediction model based on historical data that we prepare and transform into spatiotemporal data by crime type, for use in machine learning algorithms and then predict, with maximum accuracy, the risk of having crimes in a spatiotemporal point in the city. And in order to have a general model not related to a specific type of crime, we have described our risk by a vector of n values that represent the risks by type of crime.

Corresponding Author:

Hicham Aitelbour,
Information Technology and Modeling Laboratory,
Faculty of sciences Ben M'sik
Hassan II University, Casablanca, Morocco,
Email: bourhicham@yahoo.fr

1. Introduction

The classic approach of computing was to create models from inputs and programs to obtain output results. In other words, the growth in data sizes and the evolution of data mining techniques that allow the induction of behaviors and correlations between data observed in nature, has allowed us to generate programs from data and results that describe laws and rules between inputs and outputs (Fig.1).

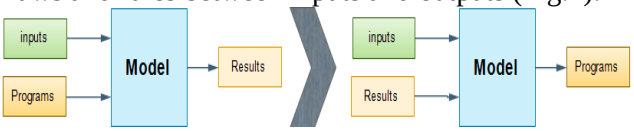

Figure 1: Evolution of computing models.

Prediction being the result of this technological evolution, it consists in studying a phenomenon and modeling it to better understand it in order to estimate the results from new data [1]. Indeed, this is done by training on previous experiments that are presented and forms input and output data. Training consists of understanding the behavior of the system and translating it into applications or programs.
We can say that prediction is a concept based on learning, testing, and evaluation. That is, an intelligent system must train on data and then be tested on data from other situations not encountered in the draft shield [2], [3]. Subsequent evaluation can adjust the results. These learning methods are based on supervised or unsupervised classification algorithms that depend on the nature of the problem.
Crimes are acts that are not acceptable in society. Police deploy significant human, financial and technological resources to conduct patrols and minimize crime rates. Crime prediction is intended as a solution to help police optimize their resources to better plan these patrols and fight crime. In other words, we are trying to know where, who and when with sufficient time lag to approach a goal of "zero crimes" [4].

This is how we can use spatial and temporal information to learn us the trend and behavior of crimes in both dimensions: space and time and, above all, to inform ourselves about the types of crime we exactly expect [5].
In this work, we propose a machine-learning model that will train on past crimes to predict future crimes. We will then experiment this model with several machine-learning algorithms to make a comparison and conclude with which we will continue our research.

2. Research goal.

The objective of this work is to predict crimes on a given day of the year and at a specific point in the city based on historical crime data. That is to say, it consists of determining a model f defined as in (1) where p a point in the city and s a part of the day.

$$y = f(p, s) \quad (1)$$

3. Crime Prediction Model.

Several studies around the world have examined the subject of crime prediction through the open data movement in some countries that displays data to the public. Some of these studies focused on different data fields to characterize a crime. For example in [6], They use place and time. In [7] they used social networks and local geolocation data [8].
These researches involve using several ways of representing data in order to use them in learning algorithms to predict crimes.
This research also used crime type approaches with models that do not describe the complete process from data to prediction.
We propose as a solution a model with several layers based on two types of data, namely historical crime data (HIST DATA) and site information (LOC DATA) as shown in Fig.2. The model includes a data preparation phase that takes these two types of data as input and consists of two layers: cleaning layer and transformation layer. The results of these layers will be stored using the Storage layer. This stored data will be used by the "feature selection layer" which allows choosing the most relevant fields. The selected data will be divided into two parts: the first for model training and the second for post-training testing. A prediction layer then allows the model to be interrogated with test data and evaluated [9].

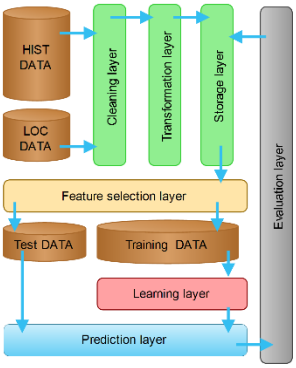

Figure 2: Proposed model

Transformation is the most important layer in this work. We thought of the mathematical expression (2) for the calculation of risk in a spatiotemporal point x.

$$RT(x) = Prob\big(Labelx = T/f1(x), f2(x)\big) \quad (2)$$

Where $x = (p, s)$ / p : point in the city et s : slice in a day, $f1(x)$ crime spatial density in the point p, $f2(x)$ crime temporal density at the s.

For n type of crime, the risk in the spatiotemporal point x is a vector as shown in Eqs (3).

$$R(x) = (RT1(x), \dots, RTn(x)) \quad (3)$$

The transformation consists of calculating for each crime of our dataset a risk vector. These vectors participate in the learning to predict, at each point in slice s, a vector of size n.
After the transformation phase, our dataset will include the prediction fields characterizing time and location as well as the decision vector that contains the n risk values by crime type. Therefore each line of the dataset will be of the form (x,R) where x is the vector of the prediction fields and R is the vector of the risks of n values where n is the number of crime types.
The training phase will be carried out using a learning algorithm and there we can mention a wide range of algorithms of machine-learning such as Regression, Naive Bais, support vector machine (SVM), Decision Tree and Random forest decision [10]. We can also mention deep learning algorithms such as Neural Network [11], Recurrent Neural Network (RNN) [12] and Long short-term memory (LSTM) [13]. In the experimentation stage, we test some examples of these algorithms.

4. Dataset exploration.

In our work, we use as an example a crime dataset from the city of Chicago in the United States of America that we were able to retrieve through the city's website mentioned in [14]. This data set contains 6755055 lines or crimes of 18 years (Fig.3) from 2001 to 2018 and with 22 columns of information.

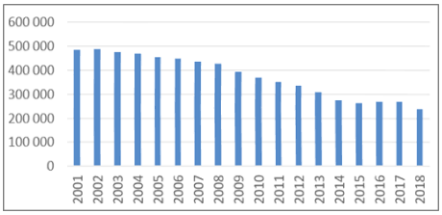

Figure 3: Number of crimes per year.

Among these columns, we have a date and time column for each crime. We also mention the "Primary Type" field, which allows us to see that there are 34 types of crimes; only the first 11 types of crimes have remarkable rates (Fig.4).

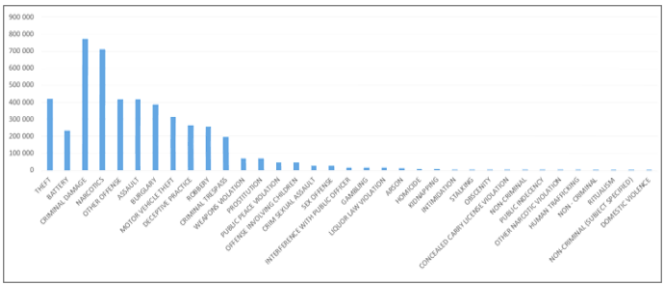

Figure 4: Number of crimes by type.

For this experiment, we chose the "Community Area" field defined in [15], which refers to the 77 geographical divisions of Chicago to characterize the spatial aspect of our model before using more geospatial features in future work that can enrich the experience such as place category or place images.

5. Transformation.

We started by transforming the dates, times into day, month, year, and slice number. The slice number is the result of dividing the day into four six-hour periods.
We then transformed our dataset into a table that contains the columns: day, month, year, slice number and community area number as well as the 34 columns with crime type headings that contain the risk of each type of crime in the rental and slice (The risk is the number of crimes in this location and slices).

6. The environment of the test.

The test takes place in a professional Windows 10 environment with a 4 CPU processor of 2.5 GHz 8GB Ram and 4GB graphics card.

We chose to use python 3.7 in the Anaconda environment for the ease and availability of machine learning algorithm packages such as sci-kit-learn and tensorFlow.

7. Experiments and discuss.

In the experimentation phase, we used several machine-learning algorithms to search for the algorithm that gave good results in terms of accuracy and response time. The PYTHON language offers advantages in the use of learning algorithms. The scikit-learn offers a variety of classifiers for easy use with the appropriate settings. In the majority of our tests, we used 90% of our dataset for learning and 10% for testing as presented in Table 1.
The test is done in several iterations. The exact number of iterations depends on two things: either the algorithm is very slow and we could not use several iterations or after a certain number of iterations, the accuracy no longer changes so we decide to stop learning.

Table 1: Comparison of tests according to learning algorithms

	Naive Bais	**Decision tree**	**Random forest**	**Neural network**	**SVM**
Language	python	python	python	Python	Python
Iterations number	0	20	1	70	0
Test rate	10%	10%	10%	0%	10%
Classifier	Multinomial NB Bernoulli NB Gaussian NB From sklearn	Decision Tree Classifier from sklearn	Random Forest Classifier from sklearn	TensorFlow	SVC from sklearn
Accuracy	---	38%	59%	81%	---

After several tests and with the limitations of the environment. We have found that some algorithms do not work with our data formats such as Naive Bais and SVM. The Random forest takes a long time for a single iteration and gives 59% accuracy. The decision tree gives an accuracy of 38% from the second iteration and remains stable. The neural network algorithm is more suitable for our data format and faster in response and gives an interesting accuracy of 81%, which would have been improved later.

8. Conclusion and future work

In this work, we proposed a two-dimensional crime prediction model based on the spatial and temporal activity of individuals that we were able to model in a mathematical form.
In the next works, we intend to further explain our model by opening up to other types of data such as weather and location images and thus give a more advanced experimental aspect in a more efficient test environment using the best learning algorithm that we will specify after our in-depth comparative study.
In future experiments, we will use algorithms such as the RNN (recurrent neural network) and LSTM for the perceived advantages of using time series, especially since we see that our dataset has a seasonal time series structure.

9. References

[1] M. A. Talhaoui, A. Daif, M. Azzouazi, and Y. Oubrahim, “A Gamified Recommendation Framework,” Int. J. Eng., p. 6.

[2] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Rand Corporation, 2013.

[3] R. A. Flarence, S. Bethu, V. Sowmya, K. Anusha, and B. S. Babu, “Importance of Supervised Learning in Prediction Analysis,” vol. 6, no. 1, p. 14.

[4] A. Newton and M. Felson, “Editorial: crime patterns in time and space: the dynamics of crime opportunities in urban areas,” Crime Sci., vol. 4, no. 1, p. 11, Jun. 2015.

[5] S. Garnier, J. M. Caplan, and L. W. Kennedy, "Predicting Dynamical Crime Distribution From Environmental and Social Influences," Front. Appl. Math. Stat., vol. 4, May 2018.

[6] M. Al Boni and M. S. Gerber, "Area-Specific Crime Prediction Models," in Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on, 2016, pp. 671–676.

[7] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," Decis. Support Syst., vol. 61, pp. 115–125, 2014.

[8] F. K. Bappee, A. S. Junior, and S. Matwin, "Predicting Crime Using Spatial Features," ArXiv180304474 Cs, Mar. 2018.

[9] H. Ait El Bour, M. A. Talhaoui, M. Azzouazi, and R. Moulouki, "Crime Prediction in the Era of Big data," Int. J. Eng. Technol., vol. 7, no. 4.32, pp. 84–86, Dec. 2018.

[10] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," Mach. Learn. Appl. Int. J., vol. 2, no. 1, pp. 1–12, Mar. 2015.

[11] O. Gursoy and Md. H. Sharif, "Parallel Computing for Artificial Neural Network Training," Period. Eng. Nat. Sci. PEN, vol. 6, no. 1, p. 1, Jan. 2018.

[12] B. S. Babu, A. Suneetha, G. C. Babu, Y. J. N. Kumar, and G. Karuna, "Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network," vol. 6, no. 1, p. 12, 2018.

[13] A. Stec and D. Klabjan, "Forecasting Crime with Deep Learning," p. 13, 2017.

[14] "Community areas in Chicago," Wikipedia. 30-Nov-2018.

[15] "City of Chicago | Data Portal | City of Chicago | Data Portal." [Online]. Available: https://data.cityofchicago.org/. [Accessed: 30-Dec-2018].