IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. II (Jan - Feb 2020), PP 37-41
www.iosrjournals.org

Use of the Logistic Regression Model for the analysis of mafia infiltration in Italian municipalities

Giacomo Abbattista[1], Vito Nicola Convertini[2], Vincenzo Gattulli[3], Lucia Sarcinella[4]

[1]*(Department of Computer Science, University of Bari Aldo Moro, Italy)*
[2]*(Department of Computer Science, University of Bari Aldo Moro, Italy)*
[3]*(Department of Computer Science, University of Bari Aldo Moro, Italy)*
[4]*(Department of Computer Science, University of Bari Aldo Moro, Italy)*

Abstract: *Illegal and criminal systems have become a widespread reality today, so as to favor the establishment of criminal mafia infiltrations within the Municipalities.*

In this study, a predictive algorithm on mafia criminal infiltrations is proposed that can lead to the dissolution of a municipal administration. Some variables are considered two types of variables: numeric and boolean. The first case includes variables such as the number of inhabitants, bars, earth moving companies, banks, supermarkets and more. The second case includes variables such as cases of murder, attacks, drug dealing, extortion. The prediction is realized through Machine Learning techniques, in particular through a logistic regression model.

Key Word*: Criminal mafia infiltrations; Machine Learning; Logistic Regression Model.*

Date of Submission: 28-01-2020 Date of Acceptance: 13-02-2020

I. Introduction

Illegal and criminal systems find their cultural and social legitimacy in the sets of values shared by the community within which they are born and grow.

The illegal and criminal system bases its existence on an 'alegality' widespread within the territory.

This cultural carpet produces the popular legitimation of behaviour extraneous to civism and criminal rootedness.

The measurement of the rootedness takes place, normally, ex post, thanks to the interventions of the investigative (judiciary) and investigative (law enforcement) forces.

There are certainly behaviors that can allude to the presence of a moral and cultural carpet favorable to the criminal and mafia settlement and infiltration, such as, for example, the illegal dumping of waste, poor respect for public spaces, excessive consumption of drugs and alcohol or money at betting centers.

Many other examples can be given to understand what this alegality is.

To define it, it is the cultural humus that legitimizes socially deviant, illegal and/or criminal behaviour.

It is a system of values that is not always born in the local, but that finds in the local spaces where to practice. In short, alegality is the chromosomal set of values and opinions of common sense that makes the illegal tendencies of a territory and a society clearly emerge.

Measuring it to build a predictive system is the subject of this study.

II. Description of the problem

There is an increasing tendency to confuse illegality with crime, forgetting that the two phenomena are not the same and that the second is contained in the first.

Illegality is a container concept, with precise characteristics, which can vary from place to place, from society to society. Illegality is what lies - voluntarily or not - outside the confines of the Law.

It is therefore illegal what goes beyond the limits of the written rules to contravene them. This contravention may not be limited in time and space.

If it extends into space/time dimensions and if it is constructed by organized social groups, it necessarily becomes crime and/or organized crime.

The concept of crime refers more to social norms shared by a territory.

An act that is not considered criminal in a society other than the one in which it is committed may be criminal.

The present study, consequently, intends to produce a predictive algorithm on the determination of the mafia/criminal infiltration up to the maximum degree, that which leads to the possible dissolution of a municipal administration, starting from data of different character and nature extracted from the Municipalities dissolved for Mafia over the years.

Once the variables (qualitative/quantitative) which imply the penetration and/or the criminal genesis of a territory have been identified, an algorithm is designed which predicts scenarios of the mafia-based criminal risk in the Italian Municipalities.

We can try to predict - therefore, to prevent - certain criminal trends thanks to their collocation within the alegal framework. As will be said later, machine learning techniques and a progressive increase in the quantity of data sources allow a continuous improvement of the output.

In the following figure the iterative development model of the algorithm is presented.

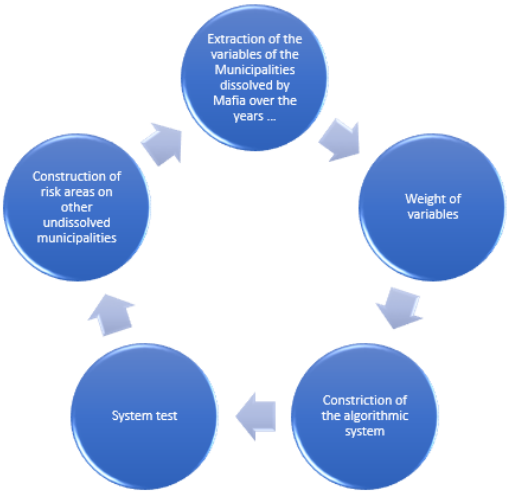
Extraction of the variables of the Municipalities dissolved by Mafia over the years ...
Weight of variables
Constriction of the algorithmic system
System test
Construction of risk areas on other undissolved municipalities

Figure 1: *Iterative algorithm development process*

III. The quantitative variables chosen and their measurement

The choice of the variables used is conditioned by the existing Databases, of different nature. The selection has followed a social logic.

Below is an example of a general value: the diffusion of betting centers. It is a national data that has been gaining interesting weight and gravity since the 1980s (F. Calderini , S. Caneppele, 2009). The criminal geography of contracts. The infiltration of organized crime into public contracts in the South of Italy... There are mafia families specialized in the imposition of slot machines in the different territories. The volume of business is enormous and the system has become a means to launder large amounts of dirty money (fig.2).

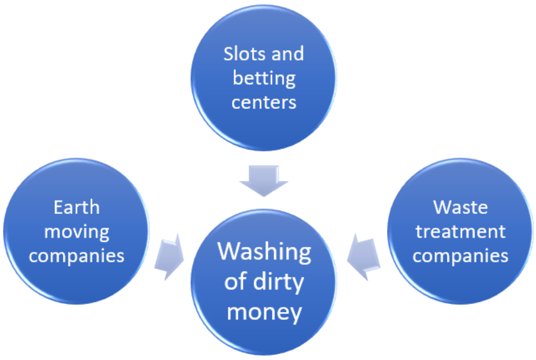

Figure 2 *Money laundering*

This activity, while reducing the proceeds of the clans, legalizes a substantial part of the money produced with the drug dealing and the racket.

At the same time, the imposition of the slot machines establishes a consistent system of control of the territory. These relations of force damage the livability of the territories and favour the mafia roots because they determine the increase of insecurity and become an instrument of power in the hands of the clans.

This, like the other variables, duly weighed, represents the quantitative basis for the determination of the predictive algorithm of the scenarios.

IV. Proposed solution

The monitoring of the mafia presence can be carried out through the collection, and subsequent analysis, of certain variables, selected on the basis of knowledge of the territorial and criminal history that criminal organizations put in place on a territory to launder money or to demonstrate their hegemony.

To this end, it was decided to create a computer system that, through the collection of publicly accessible data, would be able to predict mafia infiltration through Machine Learning techniques.

In the first phase, representative variables were chosen, which were freely accessible.

The parameters

The parameters have been divided into two types: numerical and boolean.

To the first group belong: number of inhabitants, number of bars, number of earthmoving companies, number of garages and car parks, number of car dealers, number of building companies, number of banks and financial companies, number of jewellers, number of supermarkets, number of petrol stations, number of car wrecks, number of pizzerias/restaurants, number of surveillance institutions, number of betting shops, number of waste treatment companies.

The Boolean parameters are: murders (not only of mafia), bomb attacks/burning, drug dealing, presence of extortion phenomena.

The numerical parameters have been collected through Web Scraping techniques on the portal http://www.misterimprese.it, a database of Italian companies divided by category.

Since it was not possible to acquire in a numerical way the data present in the Boolean parameters, a manual collection of each event, represented in "True/False" form, was carried out through research on the main press organs.

For the classification of the data, the Logistic Regression model was chosen, using the optimization algorithm of the Gradient Binder et all. (1997). Lauritzen, S. (1995). Russell et all (1995).

Training and Test Datasets

The representative sample, usedas a Training Set, iscomposed of the 42 cities whoseadministrationwasdissolved by the mafia in the two-yearperiod 2017-2018, as per the decree ex art. 143: Bompensiere, Borgetto, Bova Marina, Brancaleone, Briatico, Caivano, Calvizzano, Camastra, Canolo, Casabona, Casavatore, Cassano allo Jonio, Castelvetrano, Cirò Marina, Crispano, Cropani, Crucoli, Gioia Tauro, Isola di Capo Rizzuto, Lamezia Terme, Laureana di Borrello, Lavagna, Limbadi, Manduria, Marina di Gioiosa Jonica, Mattinata, Parabita, Petronà CZ, Platì, San Biagio Platani, San Felice a Cancello, San Gennaro Vesuviano, San Gregorio d'Ippona, Scafati, Scilla, Siderno, Sogliano Cavour, Sorbo San Basile, Strongoli, Surbo, Trecastagni, Vittoria (Ragusa).

There are also 42 municipalities never dissolved for mafia:

Assisi, Bibbiena, Biella, Borgo San Lorenzo, Bosco Marengo, Caorle, Casalpusterlengo, Castel di Sangro, Castiglion Fiorentino, Ceva, Collecchio, Colletorto, Desenzano del Garda, Fiuggi, Guglionesi, Ladispoli, Larino, Matera, Melendugno, Melfi, Montecatini-Terme, Norcia, Orvieto, Pescasseroli, Pietrelcina, Predappio, Recanati, Rocca di mezzo, Ruvo di Puglia, San Donà di Piave, San Giovanni Rotondo, San Giuliano Terme, Santa Croce di Magliano, Scanno, Schio, Siena, Stornara, Termoli, Trento, Tuscania, Udine, Volpiano.

The municipalitiestested are: Argenta, Cerignola, Barletta, Trani, Melfi, Lanzo Torinese, Saronno, Brescello (alreadydissolved for mafia in 2015), Rivoli, Cellino San Marco (alreadydissolved for mafia in 2014), Lido di Ostia (alreadydissolved for mafia in 2015), Rivarolo Canavese (alreadydissolved for mafia in 2012).

Logic

The logic of the system uses machine learning techniques for the analysis of the parameters present in the dataset.

Classification is an important aspect in the application of supervised learning.

Of all the available classification algorithms, logistic regression is very effective to conduct regression analysis when the target variable (i.e. the dependent variable) is dichotomous (binary).

Logistic regression is a method of classification within the family of supervised learning algorithms and allows the generation of a result that, in fact, represents a probability that a given input value belongs to a given class.

In binomial logistic regression problems, the probability that the output belongs to a class will be P, while the probability that it belongs to the other class will be 1-P (where P is a number between 0 and 1 probability expression).

K-Fold Validation

Cross-validation is a statistical technique to measure the accuracy of a machine learning model, which can be used in the presence of a good number of the observed learning sample (training set).

K-Fold Cross Validation is a resampling procedure, and is performed by dividing the total dataset into k-part of equal numerosity and, at each step, the k-third part of the dataset is used as Validation Set, while the remaining part is used as Training Set.

For each of the k-part you train the model, avoiding problems of overfitting and asymmetric sampling of the training dataset, typical of the division into two parts only (i.e. Training and Validation Set).

Once the sample is divided into groups of equal size, iteratively exclude one group at a time and try to predict it with the groups not excluded.

A generic algorithm that describes the model is the following:

1. mix the dataset randomly;
2. divide the dataset into k groups;
3. for each group:
    - extract the group to use as Validation Set;
    - move the remaining groups to the Training Set;
4. train the model on the Traning set, and make predictions on the Validation Set;
5. memorize the accuracy percentage and start again with the next k;
6. calculate the score based on the number of correct predictions for all k ratings.

The k-value must be chosen accurately according to the dataset: an incorrect choice of the k-value could cause an incorrect calculation of the model accuracy, such as overestimation.

The most common techniques for choosing the k-value are the following:

- Representative: the k value is chosen so that each training/test group is large enough to be statistically significant.

- k=10: the value assigned to k is 10, a value that generally produces an estimate with low bias and modest variance.

- k=n: the value assigned to k is n, where n is the size of the dataset, so that each sample can be used in the Validation Set.

It is preferable to choose a value for k that allows the sample to be divided into k groups, so that the same number of examples can be obtained in order to obtain equivalent accuracy.

The studied dataset is composed of 84 elements:

- 42 loose municipalities by mafia;
- 42 municipalities that have never been dissolved.

The value chosen for k is 6, so as to divide the dataset into 6 folds of 14 elements each.

The accuracy of the calculated model is 83.90%.

V. Application scenario

To test the functioning of the system, an application scenario was simulated.

As anticipated in the previous chapters, the training set is composed of 84 municipalities, 42 of which are loose due to the presence of mafia infiltration.

The analysis of the training set calculated the following weights for the parameters:

- Number of inhabitants: -1.9729639828452
- Number of bars: 1,523576163546
- Number of earthmoving enterprises: 0.89451494263746
- Number of garages and car parks: 0.09055432747986
- Number of car dealers: 1,5178567040533
- Number of construction companies: -3,4258733856075
- Number of banks and financial companies: -3,9466029239561
- Number of jewellers: -2.1741529502053
- Number of supermarkets: -1.1903983954102
- Number of petrol stations: 0.56998406352458
- Number of self-demolitions: 0.27725245617654
- Number of pizzerias/restaurants: -7,5241611651057
- Number of supervisory institutions: 0.46499096734627
- Number of betting rooms: 3,4154250816639
- Number of waste treatment companies: -0.28533848845667
- Homicides (not only mafia): 3,8369948615976

- Bomb attacks/firearms: -0.94176958069549
- Drug dealing: -0.089102601193584
- Presence of extortionary phenomena: 1,9489615881

The values represent the weight, and consequently the extent to which the phenomenon contributes to the risk of dissolution. A negative value contributes negatively to the dissolution hypothesis. For the test, 12 cities were chosen according to random parameters: however, to test the functioning of the system, three municipalities dissolved by mafia in a period before 2017 (and therefore not present in the training set) were chosen: Cellino San Marco, RivaroloCanavese, Brescello, Lido di Ostia.

The system has correctly recognized all four municipalities already dissolved by mafia infiltration, confirming the goodness of the forecast.

The municipalities of Cerignola, Trani and Melfi have also been recognized as at risk of mafia infiltration.

The analysis of the weights attributed by the model to the current dataset has highlighted the extreme importance of the parameter "Homicides": an instrument notoriously used by the clans to incite terror and impose their dominion.

The second parameter of importance is the number of betting rooms, an activity notoriously tied to the money laundering of the clans, followed by the "Presence of extortionary phenomena".

Finally, the parameters "Number of bars", "Number of car dealers" and "Number of earthmoving enterprises" contribute to the Mafia classification (listed in order of importance), although with less relevance.

VI. Conclusion

The system has proved to be surprisingly effective in identifying municipalities at risk of Mafia infiltration, but remains strongly limited by the scarcity of parameters currently present, and by the small number of elements in the training set.

The current analysis, although carried out with a small number of parameters, has already made it possible to identify some significant parameters for the research of the mafia phenomenon, and should be extended by carrying out a study of possible additional indicators (e.g.: Threats to public figures, number of revoked contracts, political guidelines in the process of dissolution).

Like all learning models, the reliability of forecasts increases proportionally to the number of data present in the training set.

At present, the collection of the necessary parameters is limited by the lack of publicly accessible information and the need for semi-manual input.

Having access to databases that contain the required data, it would be possible in the future to implement a mechanism that automatically imports data into the training set, and that is capable of performing automated evaluations.

The system will be constantly used to analyze the events that will occur in the near future, in order to monitor the mafia infiltration and improve the prediction.

In addition, with the collaboration of the authorities, this algorithm could be used in many other areas such as: road safety, health and hydrogeological risk.

Bibliography

[1]. Binder, J., Koller, D., Russell, S. J., and Kanazawa, K. (1997a). Adaptive probabilistic networks with hidden variables. Machine Learning, 29, 213-244.

[2]. Calderoni, F., Caneppele, S. (2009). The criminal geography of procurement. The infiltration of organized crime in public procurement in the South of Italy. Franco Angeli Publisher

[3]. Lauritzen, S. (1995). The EM algorithm for graphi- cal association models with missing data. Computa- tional Statistics and Data Analysis, 19, 191-201.

[4]. Russell, S. J., Binder, J., Koller, D., and Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables. In IJCAI-95, pp. 1146-52.

Giacomo Abbattista, etal. “Use of the Logistic Regression Model for the analysis of mafia infiltration in Italian municipalities." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 37-41.