An Improved Flexible Similarity Function for Clustering-Based Crime Analysis

Hiram Calvo, Salvador Godoy-Calderón, Marco A. Moreno-Armendáriz

Centro de Investigación en Computación, Instituto Politécnico Nacional,
Av. Juan de Dios Bátiz e/M.O. Mendizábal s/n, Nva. Ind. Vallejo, 07738
{hcalvo, sgodoyc, mam_armendariz}@cic.ipn.mx

Abstract. In this paper, a novel similarity function is used to identify hot-spots of criminal activity in large crime-datasets. This function considers the space and times when each crime was committed, as well as some elements of the perceived *modus operandi* of the perpetrator, in order to compare specific crime patterns and then cluster them using a density-based clustering algorithm. The clusters so formed are then graphically shown to the crime analyst using diverse GIS-tools, in order to provide him/her with high quality information about the current state of criminal activity. Several experiments performed, as well as a case-based comparison with previously published similar proposals, yield significant advantages of the proposed function over classical Euclidean-distance comparisons and other space-time similarity functions.

Keywords. Crime Analysis, Clustering, Spatio-Temporal Similarity Function.

1 Introduction

Recently, crime analysis has become one of the major concerns of public security departments around the world. Crime analysis is defined as a set of processes followed in order to gather, identify, organize, analyze and process data corresponding to the criminal activity in the area under study, to obtain useful information for preventing and fighting crime. Presently there are many analysis techniques (McCue, 2006; Osborne & Wernicke, 2003; Mena, 2003; Liu & Eck, 2008), and several commercial systems have been proposed in literature to address this problem. For more information, see (Levine, 1996, Chen *et al*., 2002) and BAIR's tools[1]. In practice, all of the commercially available crime-analysis systems are based on probability distribution analysis techniques, as well as on probabilistic forecast tools such as the Naïve-Bayes classifier. Only very few of these systems perform spatial and temporal analysis by using clustering techniques and time series analysis.

In Mexico, the legal system allows two types of police institutions: preventive and investigative. The task of the first one is focused on crime prevention; therefore, spatio-temporal analysis turns out to be of the utmost importance, since it allows the study of urban areas with high concentration of criminal activities. On the other hand, the main goal of the investigating police is to identify and capture individuals or groups responsible for committing specific crimes within a delimited jurisdiction. For this kind of police institution a careful analysis of each

[1] www.bairanalytics.com

Advances in Computer Science
Research in computing Science 83, 2014 pp. 25-34

crime's *modus operandi* (henceforth, MO) is crucial. The aforementioned differences among police institutions are relevant to this research because it is focused on managing both contexts. Both types of institutions operate on exactly the same datasets; the main difference is the weighting of the criminal features to be analyzed. If preventive analysis is our main concern, space-time features are the most important ones, followed by the type of crime and its features. If investigation analysis is our focus, some characteristics of crime contribute more information, followed by space-time features. Our proposed methodology may be adapted to the work of both: preventive and investigative police forces yielding, in both cases, a more suitable interpretation for the created clusters, in each of the analysis contexts.

1.1 Related work

Related to clustering-based crime-analysis, Nath (2006) presents a similar work to us. We proceed to describe it for comparison purposes. Nath uses a *k*-means algorithm is used for identifying crime patterns—a crime pattern is described as a specific group of criminal actions with similar MO characteristics. He calls a group or cluster of crimes, a pattern. Experiments are made on a small sample of data, shown in Figure 1.

Crime Type	Suspect's Race	Suspect's Sex	Suspect's Age gr	Victim's Age gr	Weapon
Robbery	B	M	Middle	Elderly	Knife
Robbery	W	M	Young	Middle	Bat
Robbery	B	M	?	Elderly	Knife
Robbery	B	F	Middle	Young	Piston

Figure 1. Example of criminal data (Nath, 2006)

By applying a *k*-means clustering algorithm to the datasets, the author groups crimes with similar MO. He explains that in the robbery sample (Figure 1) pattern behavior may be observed in rows 1 and 3, where the suspect's description matches, as well as the victim's profile. However, no explanation is given about the intra-class diversity of the crimes, and why the author used *k*-means as the clustering algorithm of choice. Figure 2 summarizes the results published by S.V. Nath.

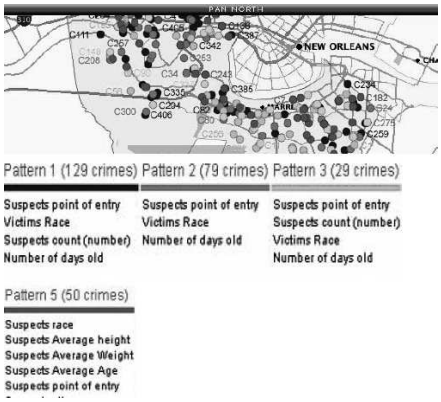


Figure 2. Experiments results (Nath, 2006)

In this paper we propose the use of a clustering technique based on pattern density, together with a space-time similarity function to identify areas with high concentration of crime (hot-spots). Then we compare the results obtained with our similarity function with those obtained by the proposed similarity function used in the original paper of the ST-DBSCAN algorithm (Birant and Kut, 1996). The comparison criteria used by the space-time similarity function and its specific use to cluster criminal activities are the main contributions of this paper.

2 Clustering-Based Crime Analysis

The purpose of using density-based clustering techniques in the context of crime-analysis is to achieve a non-statistical identification of the observed spatial and temporal trends in the commission of crimes, as well as to isolate exception cases that do not fit into those trends. This information is useful for the crime-analyst in order to develop specific strategies, both to fight and prevent delinquency, in the middle and long terms.

2.1 Pattern representation

Let us define a crime-pattern as the abstract representation of a single criminal phenomenon. This crime-pattern consists of three main components: crime specifics, space, and time; all three related to the perpetration of a specific crime. This allows the specialist to identify periods of high criminal activity and their geographic location. Therefore, the components of a criminal pattern are the following:

1) *Crime specifics*: It indicates the specific type of crime committed as well as many of its characteristics, such as the level of violence, number of persons implicated, types of weapons, *modus* operandi, etc.
2) *Space Location*: Geographic area where the crime was committed. This feature may be observed at different levels of detail: state level, patrolling sector, residential development or even street and number.
3) *Time Location*: Time when the crime was perpetrated. In this component, as well as in the one mentioned above, several levels of detail are possible: year, month, date or even the time of the day.

Each one of these components may consist of one or more variables that will be called pattern features that provide a higher level of detail to the pattern. A crime-pattern (D) has the following structure:

$$D = \left(\langle crime\ specifics \rangle, \langle space\ location \rangle, \langle time\ location \rangle \right)$$

The trend-identification process starts by analyzing a set of crime-patterns, each one with the same level of detail and within a limited geographic location, occurred within a given time interval. Table 1 contains a sample of burglary patterns obtained from the Cuautitlán Izcalli area, in the State of Mexico.

Table 1. Data sample (out of 80 patterns) of the *burglary* dataset.

Weapon	Location	Date
Firearm	Ensueños	22/08
Not specified	Cumbria	21/08
Sharp instrument	Arcos de la Hacienda	13/09
Banned weapon	San Isidro Labrador	01/03

In Figure 3 can be seen that such patterns are plotted in the map of the corresponding area divided into surveillance sectors.

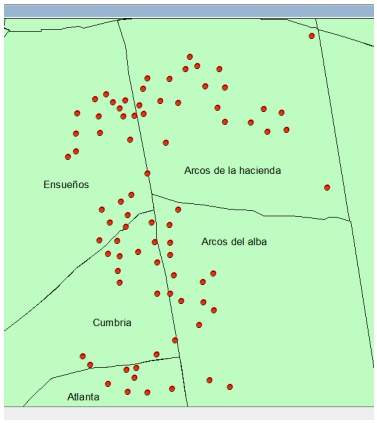


Figure 3. Criminal data referred to the geographic area where they were committed. Each red spot represents a crime incident of the "burglary" type.

In Table 2 we show another sample of the same dataset with a different level of detail from the one shown in the former sample. The differences between the two sets can be observed on the temporal and crime specific components of the patterns. This second set contains robbery-patterns and their location is another surveillance center within the same district of Cuautitlán Izcalli, Mexico.

Table 2. Data sample (out of 126 patterns) from the *robbery* data set.

Robbery type	Weapon	No. of Members	Location	Month
Break-in	Sharp instrument	2	El Rosario	DEC
Auto-parts theft	Without weapons	4	Adolfo López Mateos	FEB
Robbery to passer-by	Firearm	3	El Rosario	FEB
Robbery to passer-by	Sharp instrument	1	Cofradía-III	DEC

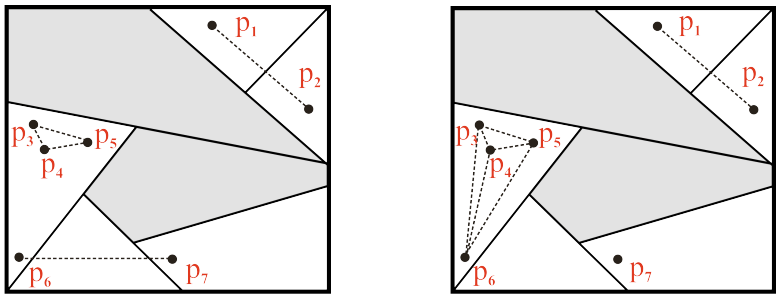
The complete first dataset (burglary) contains 80 patterns, while the second one (robbery) contains 126 patterns.

2.2 Space Comparison Criterion and the Similarity Function

The comparison between patterns is achieved by a similarity function formed by the weighted sum of a set of comparison criteria (Cc) normalized according to the number of features that form the pattern. Each of these comparison criteria

measures the similarity of each feature that makes-up the pattern. The way in which the similarity of the space component is measured is shown below.

Figure 4 shows an area divided into four surveillance sectors: A, B, C and D. A surveillance sector is a geographic zone made up by location (residential developments) and it is defined by the police department for the allocation of financial resources. According to the space comparison criterion based on Euclidean distance, patterns belonging to the same sector but separated by a long distance will not be similar (See figure 4a, patterns p4 and p6), while patterns which are geographically close to each other, even if they belong to different sectors, will be. See Figure 4a, patterns: p1 and p2.



4a) Euclidean comparison **4b)** Comparison by sector division

Figure 4. Surveillance Sectors

The spatial comparison criterion based on regions divided by surveillance sectors, strongly suggests that the maximum space-similarity should be achieved by patterns belonging to the same sector (See Figure 4b, patterns: p3 to p6), followed by patterns belonging to contiguous sectors (See Figure 4b, patterns: p1 and p2). This kind of clustering yields much more useful clusters because patrolling routines, as well as investigative teams usually schedule their operations by sector. Of course, different comparison criteria can be considered depending on the needs by each crime-analyst.

Our proposed similarity function is defined by the following equation:

$$f(o_i,o_j)=\frac{1}{r}\sum_{s=1}^{r}\left(\alpha_s Cc_s(o_i,o_j)\right) \quad (1)$$

Where:
- r is the number of features that make up the pattern.
- $o_i o_j$ are the patterns being compared.
- α_s is the weighting factor of feature s.
- $Cc_s(\,)$ is the space-time and attribute comparison criteria for feature s.

Experiments shown in Section 3 will show that this similarity function based on our space comparison criterion produce better results than the space comparison criterion based on Euclidean distance.

2.3 Density-Based Clustering

Density-based clustering algorithms can group patterns with a high spatial concentration within a delimited region isolating in the processes those patterns with a low degree of spatial similarity, which are regarded as non-related appearances of criminal activity. In this paper, we use the ST-DBSCAN algorithm, an extension of DBSCAN to work with space-time components (Birant and Kut, 2007). This algorithm was implemented as originally described with the exception of the similarity function. Our proposed spatial comparition criteria and similarity function (section 2.2, equation 1) were substituted in place of the original ones based on Euclidean distance and sector division.

2.3.1 The ST-DBSCAN algorithm.

Briant and Kut (2007) describe an extension of the DBSCAN algorithm for spatio-temporal clustering, called ST-DBSCAN. This extension proposes the separate calculation of space and time similarities between patterns.

The ST-DBSCAN algorithm requires, besides the dataset to be processed, the following two parameters: the minimum number of neighbors around an object to consider a high density situation, which will be called ***MinPts***; and the radius of the neighborhood which will be called ***Eps***. For more information about the ST-DBSCAN algorithm refer to (Briant and Kut, 2007).

3 Experimental Results

The values of ***Eps*** and ***MinPts*** in our implementation of ST-DBSCAN were calculated with the following equations:

$$Eps = 1 - \min(f(o_i, o_j)) \tag{2}$$

where

N is the number of patterns

$f(o_i, o_j)$ is the similarity function : $i = 1, 2, \ldots, n;\ j \neq i$

$$MinPts = |O| + 1 \tag{3}$$

where: $|O|$ is the cardinality of the pattern.

The first experiment was conducted over the *burglary* dataset. For this experiment the feature called *weapon* is the most relevant attribute, so the best result is deemed to be the one that groups criminal incidents committed with the same weapon.

A very important aspect to be considered is the clarification of the difference between the patterns identified as noise and the outliers in the dataset. Noise refers to those patterns that may be spatially similar to other patterns or groups, but do not share other similar characteristics, while an outlier is a pattern geometrically separated from other patterns or groups. This clarification is necessary because the ST-DBSCAN identifies noise.

The results generated by the Euclidean space similarity function vs. our proposed space competition criteria are shown in Figure 6. In both figures (5a and

5b) the patterns labeled by question signs represent noise patterns (elements that do not have characteristics similar to others).

Cluster ***A*** (see Figure 5a) represents burglary acts perpetrated with a firearm, while the ***n2*** noise pattern is another burglary crime perpetrated with a "non-specified" weapon. Despite this, not all of the patterns in cluster ***A*** were perpetrated with a "firearm". The four triangular patterns surrounding the ***n2*** pattern were committed "without weapons" (see Figure 5a). That is where the Euclidean space similarity function fails, because it groups them in the same cluster given their geographic similarity, although they are not closely related.

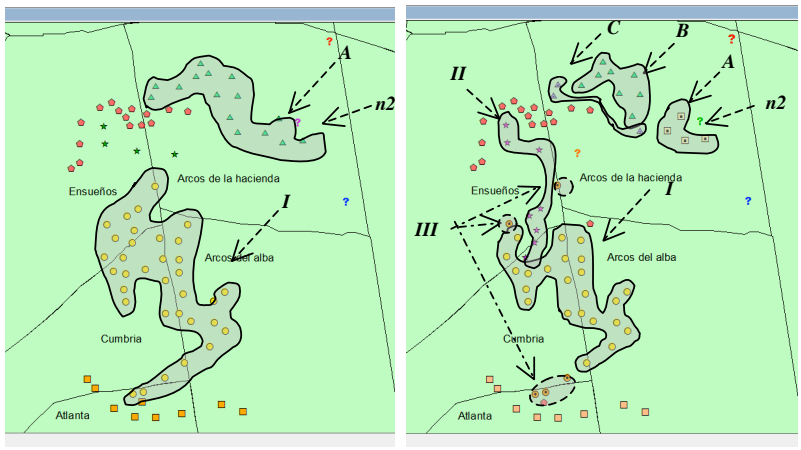


Figure 5. ST-DBSCAN results: 5(a) with Euclidean space similarity function (left), 5(b) with space similarity function based on sector division (right).

Figure 5b shows this difference, the noise pattern identified by ***n2*** is the same as the noise pattern identified in Figure 5a. Cluster ***A*** (see Figure 5b) contains the four criminal patterns perpetrated using a firearm, while the patterns that make-up Cluster ***B*** were committed without weapons. This result turns out to be very important because, following this path, crimes that were probably perpetrated by the same aggressors can be semantically identified.

In the second experiment we worked with patterns that have a higher level of detail, which means more descriptive features. Also, each component (set of related features) in the crime-patterns, were weighted as follows: 60% space-time, 30% crime-specifics and 10% crime features, due to the fact that some features are more important than others.

This experiment is more related with the work performed by the preventive police, due to the fact that the family of crimes related to robbery is the one under study. This crime-family is made up by: robbery to passerby, break-in, and auto-parts theft.

The weighting may be obtained through a criminology expert. The objective is to identify trends by taking advantage of the expert's knowledge in criminology. Figure 6 shows the results achieved.

Table 3 shows the results of our latest experiment. Of the two residential areas studied in the North areas, the one containing a higher amount of crimes from the robbery family is the Santa Barbara residential area, which belongs to the sector with the same name. Besides, we found that those months of the year with the

highest incidence of crime are July and September, so it is necessary to undertake programs and campaigns in such sector, and in that season of the year have prevention programs and campaigns to fight this type of crime.

Table 3. Results of clustering of the "robbery" type of crime data.

Cluster	Type of crime	Month	Year
●	Robbery to passer-by	July, August & September	2006, 2007 & 2008
■	Auto-parts theft	July & September	2007 & 2008
▲	Break-in	January & April	2008
⬟	Auto-parts theft	February & September	2006 & 2007
? *(noise)*	Break-in	December	2007

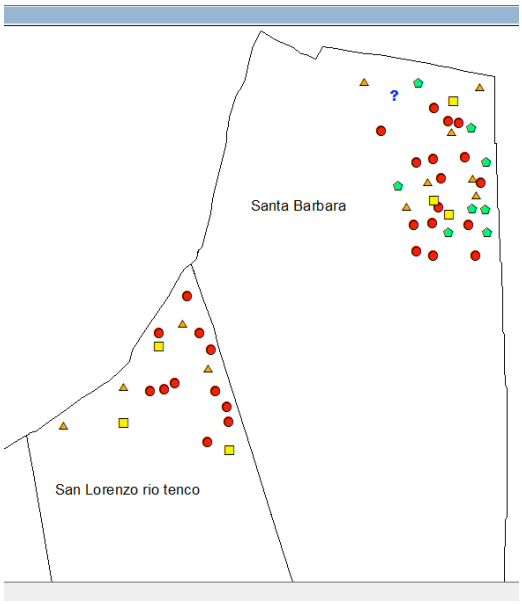


Figure 6. Clusters identified with ST-DBSCAN with a space similarity function based on sector division (Northern area).

4 Conclusions

This paper has shown that the ST-DBSCAN algorithm implemented with the space similarity function based on sector division generates better results than the one based on Euclidean distance, taking the following aspects into account: (1) The semantics adapt better to reality under the context of the type of analysis made and (2) Higher percentage of noise identification contributes to the reduction of elements for the analysis.

Acknowledgements: The authors thank the support of the Mexican Government (CONACYT, SNI, SIP-IPN, COFAA-IPN, and PIFI-IPN), and especially to the Cuautitlán Izcalli local government.

References:

1. Birant D., A. Kut,(2007) "ST-DBSCAN: An algorithm for clustering spatial-temporal data", Data & Knowledge Engineering 60: 208-221.
2. Chen, H., J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, A. W. Clements, (2002) "COPLINK Connect: information and knowledge management for law enforcement", Decisions Support Systems, 34: 271-285.
3. Ester M., H. P. Kriegel, J. Sander, X. Xu, (1996) "A density-based for discovering clusters in large spatial databases with noise", Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 226- 231.
4. Levine, N. (1996) "Spatial statistics and GIS: Software tools to quantify spatial patterns", Journal of the American Planning Association, 62 (3): 381-392.
5. Liu, Lin and John Eck (2008), "Artificial Crime Analysis Systems. Using Computer Simulations and Geographic Information Systems", IGI Global.
6. McCue, C. (2007), "Data mining and predictive analysis: intelligence gathering and crime analysis", Butterworth-Heinemann ELSEVIER.
7. McCue, Colleen (2006), "Data Mining and Predictive Analysis. Intelligence gathering and Crime Analysis", Elsevier.
8. Mena, Jesus (2003), "Investigative Data Mining for Security and Criminal Detection", Butterworth Heinemann.
9. Nath, S. V. (2006), "Crime Pattern Detection Using Data Mining", Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, 41-44.
10. Osborne, Deborah A. and Susan C. Wernicke (2003), "Introduction to Crime Analysis. Basic Resources for Criminal Justice Practice", The Haworth Press, Rutledge Taylor & Francis Group.