Criminal Hot Spot Detection using Formal Concept Analysis and Clustering Algorithms

Adriana M. G. Faria, Marcos E. Cintra, Angélica F. de Castro, Danniel C. Lopes
Federal Rural University of the Semi-Arid
Department of Exact and Natural Sciences
E-mail: {adrianaguimaraes, cintra, angelica, danniel}@ufersa.edu.br

Abstract—Public safety strategy planning can benefit from artificial intelligence techniques by extracting information from raw data. Crime hot spot detection, based on crime data, supports the distribution of resources, such as police men, patrolling cars, and surveillance cameras, as well as the definition of strategies for crime combat and prevention. In this paper, we present an approach for criminal hot spot detection using the k-means clustering algorithm in order to optimize the distribution of police resources in the city of Mossoró, Brazil. In order to induce the hot spot models, we used a database with 7,486 real records containing criminal information on the city of Mossoró for the first quarter of 2013. We also used the theory of formal concept analysis with the same data to generate graphical representations (lattices) of different types of crimes in different periods of a day to complement the hot spot models. The resulting hot spots and lattices have been evaluated by police experts and received positive feedback. Keywords—criminal hot spot detection, clustering algorithms, formal concept analysis, criminal hot spot detection, clustering algorithms, formal concept analysis

I. INTRODUCTION

Until the popularization of computers, criminal analysis used to be carried out manually, for instance, by means of city maps with pins in locations with large occurrence of crime events. These maps were analysed to evaluate the distribution of resources and to propose strategies for crime prevention. Nowadays, public security agencies collect and store large amounts of digital information on crime which can be processed by computational tools in order to optimize public safety strategies.

Public records of crimes are useful since we know that crime is not distributed randomly [1]. In fact, crime usually has spatial and temporal characteristics related to the population, environment, economic factors, politics, and social events. One of the most effective temporal-spatial analysis for the understanding of the implicit relationships among crime events is the analysis of criminal hot spots. The analysis of criminal hot spots contributes to the combat and prevention of crimes by allowing the planning of strategies that optimize the distribution of police resources. As police resources can be quite limited in some areas, the planning of such distribution becomes a highly relevant task.

In this sense, in this work we present an approach to generate criminal hot spots based on crime information which uses clustering algorithms. We also generated lattices using Formal Concept Analysis (FCA) to provide graphical information on crime type and time. These models aim at providing formal tools to understand the characteristics of the crimes in Mossoró and properly allocate resources for crime combat and prevention. Such resources include the distribution of policemen and patrol cars, as well as surveillance cameras and definition of blitzes.

This paper is organized as follows. Section II introduces the concepts of crime analysis and criminal hot spot detection, as well as a bibliographical review. Section III presents the basic concepts of clustering and formal concept analysis. Section IV presents the proposed methodology used to generated the supporting models, the experiments and a discussion of the results. The conclusions and future work are in Section VI.

II. CRIMINAL ANALYSES AND CRIME HOT SPOTS

The large amounts of data recorded daily by public safety agencies makes the use of computational tools necessary for the process of such data in order to obtain more precise and faster results. In fact, the task of knowledge acquisition based on electronic documents has become essential for police work [2]. The Federal Police of Brazil, for instance, apprehends more than 6,000 hard drives during police investigation yearly. The volume of data stored in such material reaches 720 terabytes, which corresponds to 36 times the volumes of the largest library of the world [2].

Regarding the existing techniques for crime analysis, in [3], the authors propose a computational tool based on FCA for crime analysis based on crime records aiming at finding crime patterns. The approach proposed in [3] provides a graph with relationships among objects (crime records) that share characteristics (attributes) in a hierarchical style.

In [4], the authors use attributes in data that characterize crimes, such as the sex, age, and economic background of criminals, to produce graphical summaries of the patterns that have some correlation.

The idea of computational criminal area is studied in [5]. The proposed strategy, named TOPO, uses an algorithm that segregates neighbourhoods in small areas based on their inhabitants physical and social characteristics. The authors use psychological, social and cognitive characteristics of the inhabitants, as well as a description of the areas where crimes take place. The combination of such techniques is able to define criminal hot spots based on predefined targets.

The authors of [6] use clustering algorithms to define the Modus Operandi, i.e., the method of operation, of criminals. The idea is to cluster criminal records with similar characteristics in order to find similar criminals. The proposal is named...
Segmented Multiple-Metric Similarity Measurement and uses the following attributes: type of crime, sex and age of the criminal, scope of the criminal activity, place, and crime tool.

In this work, we aim at defining Criminal Hot Spots (CHS) for the distribution of public resources. CHSs are defined by a geographical area with high incidence of crimes. In other words, CHSs are areas where the number of crimes is higher than their neighbouring areas and represent areas where people are more likely to be victimized [2].

CHSs can be characterized by the following criteria.
- Total number of events in the area.
- Spatial information on the area.
- Time of events.
- Type of events.

Since such criteria can suffer high variations from place to place, the analysis of hot spots supports police work in the identification of high criminality areas, as well as the types of crimes occurring in such areas [7].

### III. Theoretical Reference

In this section, we detail the problem of crime hot spot detection, the basic concepts of the employed techniques, clustering algorithms and formal concept analysis.

#### A. Clustering

Clustering algorithms [8] are part of non-supervised learning for the separation of objects with similar characteristics. Clustering is a fundamental task in data mining and can be applied in several areas.

The process of clustering objects, also referred to as patterns, evaluates the similarities of such objects according to different strategies in order to separate them into \( k \) groups or clusters. An open problem for clustering algorithms is the definition of the number of clusters. Specifically for this work, this number can be defined according to the resources available, such as number of surveillance cameras or cars for patrolling.

One of the most well-known clustering algorithms is K-MEANS. Next, we present the basic steps of K-MEANS.

1) Define the number of clusters \( (k) \).
2) Define the initial position of the clusters (K-MEANS can handle the random definition of these parameters).
3) Associate each object to the nearest cluster according to the adopted similarity measure.
4) Recalculate the centre of each group by analysing the position of each object.
5) Repeat steps 2 and 3 until no element is reallocated to another group or until a predefined number of iterations.

The basic similarity measures used by K-MEANS is the Euclidean distance and Jaccard measure [9].

#### B. Formal Concept Analysis

FCA [10], [11], is a mathematical technique for extracting concepts and structures from data introduced in the 1980s which, since then, has become increasingly popular. The basic data structure in FCA is the formal context, normally represented in as a table where the columns represent the attributes and the rows represent the objects (see Table II). In other words, a formal context is a representation of the relation between objects and their attributes. Notice that attributes in a formal context are binary. The table representing the formal context contains 1 (true) if object \( i \) has attribute \( j \), and 0 (false) otherwise. It is possible to generate a conceptual lattice from the formal context. A conceptual lattice is basically a graph whose vertices correspond to the formal concepts represented by sets of examples or attributes.

Formally, a context is a triple \( (G, M, I) \), where \( G \) is a set of objects, \( M \) is a set of attributes, and \( I \) is a binary relation \( I \subseteq G \times M \). Given a set of objects \( A \subseteq G \), the shared image of \( A \) in \( M \) is defined as:

\[
A^\uparrow := \{ m \in M | (g, m) \in I \forall g \in A \}
\]

While the shared image of \( B \) in \( G \), given a set of attributes \( B \subseteq M \), is:

\[
B^\downarrow := \{ g \in G | (g, m) \in I \forall m \in B \}
\]

The pair \( (A, B) \) constitutes a formal concept of \( (G, M, I) \) if and only if \( A \subseteq G, B \subseteq M \) and \( A = B^\downarrow, B = A^\uparrow \). \( A \) is called the extent of the concept and \( B \) is called the intent of the concept [10]. In other words, Equation 1 defines the collection of all attributes shared by all objects in \( A \), while Equation 2 defines the collection of all objects sharing all the attributes in \( B \).

As previously stated, in traditional FCA, the relation is binary. For attributes that can take a range of values, we can use idea of “conceptual scaling”, which transforms a many-valued attribute (e.g. a number) into a symbolic attribute. For instance, the attribute “height in centimetres”, whose domain ranges from 0 and 200, can be represented by the following binary attributes “height-less-than-50”, “height-from-50-to-100”, and “height-more-than-100”. As the derived attributes have true/false values, they can be treated within the FCA framework as binary ones.

Using the formal context, it is possible to generate a conceptual lattice that presents the information in a nice visual way. It follows a toy example based on the attribute-value table shown in Table I, which shows the Name, Age, Sex, and Hair Colour of six people [12].

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Sex</th>
<th>Hair Colour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy</td>
<td>48</td>
<td>M</td>
<td>Black</td>
</tr>
<tr>
<td>Lina</td>
<td>29</td>
<td>F</td>
<td>Black</td>
</tr>
<tr>
<td>Mark</td>
<td>23</td>
<td>M</td>
<td>Brown</td>
</tr>
<tr>
<td>Martina</td>
<td>46</td>
<td>F</td>
<td>Blond</td>
</tr>
<tr>
<td>Mike</td>
<td>18</td>
<td>M</td>
<td>Brown</td>
</tr>
<tr>
<td>Suzy</td>
<td>17</td>
<td>F</td>
<td>Blond</td>
</tr>
</tbody>
</table>

Table I: Attribute \( \times \) value table - toy example.
The first step to define the formal context is to create binary attributes representing Age, Sex, and Hair Colour (Table I), as shown in Table II.

Table II: Formal Context based on Table I.

<table>
<thead>
<tr>
<th>Name</th>
<th>Age ≤ 20</th>
<th>&gt;20 &amp; ≤ 30</th>
<th>&gt;30</th>
<th>Sex M</th>
<th>F</th>
<th>Hair Colour Blond Brown Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andy</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Lina</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mark</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Martina</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mike</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Suzy</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

In Table II, the continuous attribute age was discretized, generating three binary attributes. The multi-valued attributes sex and hair colour generated a binary attribute for each of its values. Notice that FCA works with datasets that have a class attribute and with datasets that do not have a class, as in the case of this toy example. Figure 1 shows the generated conceptual lattice based on the formal context shown in Table II.

In the lattice structure, formal concepts are represented by the nodes. Attributes are noted above the nodes while objects are noted under the nodes. To retrieve extensions, one must simply trace all paths leading down from the node representing the object of interest. Similarly, it is possible to retrieve intentions by tracing all paths leading up from the node representing the object of interest. For instance, the intention of the formal concept represented by the node named Mark is the set of attributes \( \{ >20 \text{ & } \leq 30, M, Brown \} \). The extension of the formal concept represented by the node named Blond is \( \{ Martina, Suzy \} \).

Another important aspect of conceptual lattices is the fact that the nodes, i.e., the formal concepts, can be represented with different sizes, according to their frequency in the formal context. This way, more frequent formal concepts are represented by bigger nodes, while less frequent formal concepts are represented as smaller nodes. This characteristic gives an extra information which can be quite valuable in the analysis of the concept lattice.

IV. PROPOSED METHODOLOGY AND EXPERIMENTS

In this section, we detail the data collection and preprocessing methodologies used to generate the hot spot maps and lattices, as well as the experiments and a discussion of the obtained results.

A. Data Collection

Usually, crime records registered by the Brazilian Federal Police include: (i) crime type, (ii) crime place, (iii) crime time, and (iv) information on the involved people.

Other information that can also be collected, but are optional, are: (i) the number of people involved in the crime, (ii) the motivation for the crime, (iii) ages of the involved people, (iv) genre of the involved people, (v) type of place where the crime took place, and (vi) reincidence.

In fact, the criminal records of the Brazilian Federal Police database are defined using only the place and time of the occurrence, as well as information on the involved people, when available. The safety of the cities in Brazil is the responsibility of local police departments whose databases do not follow a pattern. Thus, databases from different police stations in Brazil usually present several differences. This lack of standardization for the crime databases associated to the exponential growth of data collection characterizes a problem for crime analysis. Such differences in crime records also compromise the generation of statistics that represent the actual scenery of criminality in the country.

In fact, a national collection system for crimes was implanted in 2001 in Brazil. This system presents independent, but relational data modules, which makes it easy to implement future improvements. Nevertheless, this system has not been adopted by all local police stations yet.

A national project to

The hot spot maps and lattices were generated using a total of 7,486 records of crimes were used. These records were collected during the first quarter of 2013 in Mossoró. Such records were separated according to the time of occurrence (morning, afternoon, evening, and night) and contain occurrences of the following four crimes: (i) burglary, (ii) theft, (iii) homicide, and (iv) drug traffic.

Each record contains the following attributes: (i) crime type and (ii) crime place. The place of the crimes are defined by the name of the neighbourhood where it occurred. We transformed that information using the latitude and longitude of the centre of each neighbourhood. Although this strategy allows the use of clustering algorithms by providing coordinates for each occurrence, the area of the neighbourhoods directly affects the generated hot spot maps and lattices: larger neighbourhoods tend to concentrate more occurrences and produce more generalized hot spot maps and lattices.

Besides the insertion of the geographical information of the crime place, another preprocessing task was carried out to eliminate records with missing attributes or crimes different from the four listed previously.

Also, the records were separated according to the time of occurrence in order to generate different maps with criminal hot spots for different periods of the days. The number of hot spots were defined from 8 to 20 for each period of the day. These values correspond to an average number of patrol cars available in all police stations of Mossoró (minimum of 8 and maximum of 20).
V. EXPERIMENTS AND RESULTS

The implementation of WEKA [13] was used for the K-MEANS algorithm, considering the following parameters: Euclidean distance with 500 iterations (the number of iterations was defined empirically) and from 8 to 20 clusters.

A total of 652 crime hot spot maps were generated, segmented in groups of 8 to 20 clusters according to four periods of the day: morning (from 6 a.m. to 12 a.m.), afternoon (from 12 a.m. to 6 p.m.), evening (from 6 p.m. to 12 p.m.), and night (from 12 p.m. to 6 a.m.). The generation of the hot spots for each time group made it possible to further analyse the patterns of criminality in the city.

Each hot spot map has geographical coordinates (latitude and longitude) to define their localization, i.e. and the places of high incidence of crimes. The maps were plotted using an API of Google Earth [14].

In order to illustrate the plotted hot spots, Figures 2, 3, 4, and 5 show maps for morning, afternoon, evening, and night, respectively. Notice that these maps were generated defining 8 hot spots, each one represented by a coloured pin.

![Figure 2: Hot spots for the morning period.](image)

![Figure 3: Hot spots for the afternoon period.](image)

From the maps, it is possible to:

- Define the distribution of the available resources for patrolling the city according to the different shifts through the working days, including patrolling cars and surveillance cameras;
- The planning of strategies for the combat and prevention of crimes;
- A better understanding of the crime scene in the city.

It was possible to observe that the increase of hot spots tend to generate maps with close hot spots, some located two or three blocks from each other.

For the generation of the lattices, we used the Conexp tool [15] with default parameters. The lattices were generated using the same data of the CHSs. The time of the crime attribute was scaled to a binary attribute with the following values: morning (from 06:00 am to 12:00 am), afternoon (from 12:00 am to 06:00 pm), evening (from 06:00 pm to 12:00 pm), and night (from 12:00 pm to 06:00 am).

The visual information presented in the lattices are related to the size of the nodes, i.e., formal concepts, and are useful for basically two analysis:

1) The distribution of the different types of crimes according to the time of day.
2) The incidence of each type of crime through the day.

The obtained lattices are shown in Figures 6 and 7.

Such lattices are particularly useful for the scheduling of specialized staff.

The hot spot maps and lattices were analysed by police officers of Mossoró and received positive feedback.

VI. CONCLUSION AND FUTURE WORK

The use of computational techniques to process criminal records are necessary to support the definition of strategies to
combat and prevent crime due to the large amounts of data collected daily.

Criminal hot spot maps are becoming increasingly important for the preventive police operation as they support the process of criminal behaviour investigation.

Formal concept analysis has also been used to process criminal information registered in police stations and provide lattices with information on the incidence of crimes in different periods of the day.

In this paper, we propose the use of the K-MEANS algorithm to detect criminal hot spots by adding geographical coordinates to local police records in Mossoró, Brazil.

The hot spot maps and lattices can be easily analysed. They can be used to plan the distribution of resources, such as patrolling cars and surveillance cameras, and to define prevention and combat strategies to crime.

The generated models were evaluated by local police experts and received positive feedback.

As future work, we intend to propose different levels of dangerousness to the types of crimes in order to generate more specialized hot spots based on such information. We also intend to explore the use of fuzzy clustering algorithms.

ACKNOWLEDGEMENTS

We kindly thank F. P. F. Neto for the support with collecting and processing the data. We also thank the financial support
References


