Extração de Regras de Associação de Dados Criminais no Município de Goiânia

Sílvio Passos Severino[1], **Nádia Félix F. da Silva**[1]

[1]Instituto de Informática – Universidade Federal de Goiás – Goiânia, GO – Brasil

silviopass@gmail.com, nadia@inf.ufg.br

Abstract. *The present work aims to extract associations in data of criminal occurrences in the city of Goiânia, identifying the neighborhoods where there is a greater concentration of crimes, considering the spatial and temporal distribution, as well as the socioeconomic profile of the place of occurrence. The developed process uses all phases of knowledge discovery - KDD (Knowledge Discovery in Databases), including the selection of attributes, cleaning, standardization, pre-processing and data transformation. The data set used in this study refers to the record of criminal occurrences provided by the Information Analysis Management of the Public Security Secretariat of the State of Goiás. The result obtained through the use of association rules gathers important information about criminal occurrences, such as the identification of the most frequent crimes in a specific place, in a time period, with a specific victim profile. The results are relevant to help decision making and also to inform and alert the population about the places with the most criminal occurrences at certain times in the municipality of Goiânia.*

Resumo. *O presente trabalho tem como objetivo extrair associações em dados de ocorrências criminais no município de Goiânia, identificando os bairros onde há maior concentração de crimes considerando a distribuição espacial e temporal, bem como o perfil socioeconômico do local da ocorrência. O processo desenvolvido utiliza-se de todas as fases da descoberta de conhecimento – KDD (Knowledge Discovery in Databases) compreendendo a seleção de atributos, limpeza, padronização, pré-processamento e transformação dos dados. O conjunto de dados utilizado no presente estudo refere-se ao registro de ocorrências criminais fornecido pela Gerência de Análise de Informações da Secretaria de Segurança Pública do Estado de Goiás. O resultado obtido através do uso de regras de associação reúne informações importantes sobre as ocorrências criminais como, por exemplo, a identificação dos crimes mais frequentes em um determinado local, em um período de horário, com um perfil de vítima específico. Os resultados são relevantes para o auxílio à tomada de decisão, bem como informar e alertar a população sobre os locais com mais ocorrências criminais em determinados horários no município de Goiânia.*

1. Introdução

Não são poucas as notícias e imagens, expondo o sério problema da violência no Brasil, especialmente nas capitais. A alta criminalidade no país tem gerado custos altíssimos. De acordo com o Anuário de Segurança Pública, só em 2016, o governo brasileiro gastou

na área de segurança pública mais de 67 bilhões de reais. No estado de Goiás a situação não é diferente, em 2016 foram gastos mais de 3 bilhões de reais em segurança pública [Fórum 2017].

Devido ao elevado número de crimes, em 2015 Goiânia chegou a ocupar o 13º lugar no ranking das capitais mais violentas do Brasil [Fórum 2017]. Buscando reduzir os índices de crimes, o governo de Goiás investiu no ano 2015, somente em inteligência de informações, o valor de 1,5 milhões reais. Um dos investimentos foi a implantação da Plataforma de Sistemas Integrados (PSI). Lançada em abril de 2016, a PSI é composta pelos programas de Registro de Atendimento Integrado (RAI), Sistema Geográfico de Informação (GisGestão), Mapeamento de Operações Policiais Integradas (MOPI), Mapeamento de Ações Sociais Integradas (MASI) e o Aplicativo de Integração entre Polícia e Cidadão (I9X). Segundo a Secretaria, o RAI foi desenvolvido para que as instituições que compõem o Centro Integrado de Inteligência, Comando e Controle (CIICC) possam ter base de informações sobre os registros das ocorrências policiais integradas e em tempo real.

É importante ressaltar que em análise de dados criminais, fatores relacionados ao espaço geográfico do crime devem ser considerados como, por exemplo, demografia, perfil socioeconômico, entre outros. Abordando sobre o relacionamento entre crime e níveis socioeconômicos [Villarreal and Silva 2006] foi encontrada uma associação entre o nível social e as taxas de criminalidade no Brasil. De um modo geral a violência atinge, sem distinção, todos os segmentos sociais. No entanto, é relevante considerar fatores socioeconômicos para identificar padrões de crimes no município de Goiânia e o conhecimento adquirido a partir de bases de dados criminais unificadas a informações socioeconômicas pode auxiliar na análise sobre os crimes em determinados bairros.

2. Trabalhos Relacionados

Na literatura existem diversos estudos sobre análise de dados criminais, bem como técnicas com a finalidade de auxiliar os gestores da área na tomada de decisão ou para outra finalidade, a seguir são apresentados alguns estudos sobre o tema:

- [Mande et al. 2012] apresenta uma metodologia que usa Modelo Gaussiano de Mistura Generalizada para mapear informações especificadas pela testemunha ocular utilizando o agrupamento k-means com base no tipo de crime.
- No trabalho [Nath 2006] é utilizado o agrupamento K-means para encontrar alguns padrões criminais, implementando uma estrutura que funciona com a informação geoespacial, o tipo de crime, o perfil da vítima e o perfil do criminoso.
- [Keyvanpour et al. 2011] trata da relação dos crimes entre si e os criminosos através de uma abordagem sistemática do uso de Redes Neurais Auto-Organizáveis - SOM e Multicamadas Perceptron - MLP para agrupar e classificar dados do crime.
- [Wang et al. 2013] utiliza o algoritmo representado como GDPatterns para encontrar a informação espacial e o relacionamento dos crimes com outros fatores, apresentando uma ferramenta de otimização de hotspots que considera outros crimes ocorridos nas proximidades.

- Já [Vijayakumar et al. 2014] apresenta os resultados da predição e de padrões criminais focando em como incorporar o fator tempo na análise de dados utilizando o modelo STEM "Modelo de espaço-tempo-evento".
- O trabalho proposto por [Almanie et al. 2015] apresenta dois modelos classificadores e regras de associação para analisar dados criminais considerando a localização, tempo e tipo de crime.

3. Fundamentação Teórica

A tecnologia da informação possui vários recursos que pode auxiliar a análise de dados e a compreender padrões e tendências dos crimes ocorridos a partir de um conjunto de dados. Dentre estes se destaca o processo de Descoberta de Conhecimento em Bases de Dados (KDD, do inglês *Knowledge Discovery in Databases*). Uma definição clássica para o KDD, segundo [Fayyad et al. 1996], *é o processo não trivial de identificar informações válidas, novas, potencialmente úteis e padrões compreensíveis nos dados*, e contém uma série de etapas a saber: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação dos dados.

3.1. Mineração de Dados

A mineração de dados é uma etapa do KDD, compreende os principais algoritmos que permitem obter conhecimento em grandes conjuntos de dados [Zaki and Jr 2014] e tem como finalidade encontrar anomalias, padrões e correlações para depois os apresentarem na forma de representação e visualização [Silva 2004]. Em mineração de dados existem diferentes técnicas e tarefas para diferentes propósitos, cada uma com vantagens e desvantagens. A escolha da técnica está relacionada com o tipo de conhecimento que se deseja extrair ou com o tipo de dado no qual ela será aplicada.

A Classificação é a tarefa de mineração de dados que associa ou classifica objetos a determinadas classes buscando prever um rótulo de classe para um determinado objeto não rotulado [Zaki and Jr 2014].

A Clusterização busca reunir instâncias com características comuns em grupos, que posteriormente podem ser classificados. Exemplos de tarefas de agrupamento são: identificar grupos de clientes para direcionamento de campanhas, identificar fraude ou até mesmo classificar instâncias, quando não houver classe conhecida. [Fayyad et al. 1996] define Clusterização como uma tarefa descritiva comum em que se procura identificar um conjunto finito de categorias ou *clusters* para descrever os dados.

A tarefa de Associação busca encontrar relacionamentos significativos entre os itens de dados armazenados através de regras de associação, identificando associações entre atributos e apresentando padrões frequentes em um conjunto de dados [Kantardzic 2011]. Neste trabalho a técnica utilizada foi a Associação.

3.2. Regras de Associação

Uma Regra de Associação dá-se por meio da utilização de métodos probabilísticos sobre a ocorrência simultânea de determinados eventos em uma base de dados [Zaki and Jr 2014]. Por exemplo, para a regra R assumem-se as seguintes variáveis como binárias *R : (SE A=1 E B=1 ENTÃO C=1)* com probabilidade *P*, Assim, tem-se:

$$P = P\,(C = 1 — A = 1, B = 1) \tag{1}$$

Uma regra de associação é composta de dois conjuntos de itens, um antecedente e um consequente e são representadas na forma: Antecedente (A) $\rightarrow$ Consequente (B), interpretada da seguinte forma: Se A então B, e ambos formam um conjunto de itens (*itemsets*) [Zaki and Jr 2014]. Para um item ser considerado frequente deve satisfazer alguma condição previamente definida, para tanto é necessário definir medidas, entre as mais usadas estão o suporte e a confiança.

O suporte representa o número de transações incluindo todos os itens na posição de antecedente e consequente da regra. Assim, para uma a regra de associação $\{A\} \rightarrow \{B\}$ o suporte mede o número total de registros de transação que contêm os conjuntos de itens A e B. Neste caso o suporte da regra $\{A\} \rightarrow \{B\}$, em que A e B são conjuntos de itens é dado pela seguinte expressão:

$$\textit{Suporte}\ (A \rightarrow B) = (\textit{Frequência de A e B}) / (\textit{Total de T}) \tag{2}$$

A confiança de uma regra é a probabilidade condicional que uma transação contém *A*, dado que contém *B* é representada pela seguinte expressão:

$$\textit{Confiança}\ (A \rightarrow B) = \textit{suporte}\ (A \cup B) / \textit{suporte}\ (A) \tag{3}$$

Existem muitos algoritmos que utilizam regras de associação e entre os mais populares estão o Apriori e o FPGrowth [Amaral 2016].

3.2.1. Algoritmo FPGrowth

O algoritmo FPGrowth leva uma abordagem diferente do Apriori que usa o paradigma "gerar e testar", a abordagem do FPGrowth consiste no desenvolvimento de uma estratégia baseada na técnica de dividir para conquistar, na qual o problema é fracionado em subproblemas [Han et al. 2000]. O FPGrowth codifica o conjunto de dados através de uma estrutura compacta chamada Frequent Pattern tree (FP-tree) e extrai os conjuntos de itens frequentes diretamente desta estrutura. Isso possibilita um melhor desempenho na geração das regras de associação, pois o número de varreduras na base de dados é menor [Tan et al. 2005]. Neste presente trabalho o algoritmo utilizado para descoberta das regras de associação é o FPGrowth.

4. Experimentos

Nesta seção são descritos o conjunto de dados e a metodologia do processo utilizado.

4.1. Conjunto de Dados de Ocorrências Criminais

Fornecido pela Gerência do Observatório de Segurança Pública do Estado de Goiás – GEOSP, o conjunto de dados compreende o registro de ocorrências no Estado de Goiás obtidas através da integração da Polícia Militar, Polícia Civil, Corpo de Bombeiros Militar e da Superintendência de Polícia Técnico-Científica (SPTC). O conjunto é composto por 16 atributos, contém 1.785.488 instâncias e os principais atributos fornecem a natureza do crime, data, hora, localização geográfica e informações do comunicante e/ou

envolvido como, por exemplo, sexo e data de nascimento da vítima. O período referente às ocorrências registradas foi entre os dias 1º de abril de 2016 a 31 de março de 2017, totalizando doze meses.

4.2. Conjunto de Dados Socioeconômicos

O conjunto de dados contém a renda domiciliar dos bairros do município de Goiânia e foram extraídos do Censo Demográfico 2010 - Resultados Agregados por Setor Censitário, os dados compreendem as características dos domicílios particulares e das pessoas que foram investigadas para a totalidade da população e são denominados por convenção resultados do universo [IBGE 2017].

4.3. Pré-processamento de Dados

A seguir são descritos os passos da etapa de pré-processamento dos dados, os quais foram realizados através das ferramentas IBM SPSS Statistics [IBM 2011] e Orange Canvas [Demšar et al. 2018]

O conjunto de dados criminais foi disponibilizado em 12 arquivos no formato CSV. Sendo assim, o primeiro passo foi fazer a integração dos arquivos para facilitar a manipulação dos dados.

Após a integração dos arquivos, verificou-se a existência de valores ausentes nos atributos 'profissão', 'escolaridade', 'orientacão_sexual', 'cor_raça'. Como os índices eram altos, todos estes atributos foram eliminados da base.

Em seguida foram eliminadas as instâncias que não eram do município de Goiânia, em sequência foram eliminadas todas as instâncias que não se caracterizavam como uma ocorrência criminal. Depois de observado que uma ocorrência gerava várias instâncias, ou seja, uma instância para cada envolvido na ocorrência, considerando a abordagem do trabalho que foi com foco na qualificação da vítima foram eliminadas da base todas as instâncias em que o atributo 'qualificação' era diferente de 'vítima' ou 'vítima comunicante'. O passo seguinte foi padronizar o atributo 'natureza', pois em algumas ocorrências havia mais de uma 'natureza'. Os dados foram reestruturados novamente, criando uma instância para cada crime.

Com o objetivo de delimitar o escopo do estudo e levando em consideração os objetivos específicos do trabalho, os crimes foram delimitados nas seguintes categorias: crimes contra a dignidade sexual, crimes contra a pessoa, crimes contra o patrimônio e crimes contra o estatuto da criança e do adolescente. Assim, foi necessário identificar as instâncias que não se enquadravam na categoria desses crimes e excluí-las da base de dados, classificando-as segundo a Tabela de Naturezas - RAI[1]

Terminado o processo de limpeza dos dados, o próximo passo foi obter a idade da vítima, através da subtração do atributo 'data_ocorrencia' e 'data_nascimento'.

Com o intuito de gerar análises por dia do mês, o atributo 'data' foi transformado em três atributos, 'data_ano', 'data_mes' e 'data_dia'.

Ao final deste processo o conjunto de dados contava com 10 atributos: data_dia, data_mes, dia_semana, hora, bairro, natureza, logradouro, qualificacao, sexo e idade.

[1] http://observatorio.ssp.go.gov.br/

Foi observada a necessidade de reduzir a diversidade dos valores de alguns atributos. Assim, foi aplicada a transformação de dados aos atributos, mapeando seus valores para que se enquadrarem em grupos menores. O objetivo foi obter padrões mais frequentes e aumentar a precisão do modelo. Para o atributo 'natureza', foi minimizada a lista de crimes criando uma categoria denominada 'outros crimes', agrupando os crimes com frequência menor que 1%. Ao final, a lista de natureza de crimes ficou com 20 categorias. Para o atributo 'hora' do crime, foram mapeados os valores em intervalos de 4 horas.

Finalmente, utilizando o atributo 'bairro' presente nos dois conjuntos (criminal e socioeconômico) as bases foram unificadas, gerando assim um único conjunto de dados. A figura 1 apresenta o conjunto de dados após o pré-processamento.

	Atributo	Tipo	Categorias
1	id	numérico	...
2	data_dia	numérico	...
3	data_mes	numérico	...
4	dia _semana	nominal	Domingo; Segunda-feira; Terça-feira; Quarta-feira; Quinta-feira, Sexta-feira; Sábado
5	turno	nominal	T1: 01:00 as 04:59; T2: 05:00 as 08:59; T3: 09:00 as 12:59; T4: 13:00 as 16:59; T5: 17:00 as 20:59; T6: 21:00 as 00:59
6	bairro	nominal	(64 categorias)
8	crime	nominal	(20 categorias)
9	sexo	nominal	Masculino; Feminino
10	faixa_etaria	nominal	Criança (<11); Adolescente (11-17); 2 Jovem (18-30); 3 Adulto (31-59); 4 Idoso (>59)
11	renda_domicilio	numérico	...

Figura 1. Conjunto de dados após pré-processamento

4.4. Obtenção de Regras de Associação

Para encontrar relações entre atributos — por exemplo, data, local, gênero — o foco foi aplicar algoritmos de obtenção de regras de associação e avaliar se tais regras obtidas podem auxiliar na tomada de decisões e em estratégias de políticas de segurança pública. Desta forma, a abordagem proposta está focada nos três principais elementos sobre os crimes, que são a natureza do crime, o tempo de ocorrência e a localização do crime. Na tentativa de extrair os possíveis padrões frequentes interessantes baseados nas variáveis do crime, foi aplicada regras de associação.

O Algoritmo FPGrowth é um dos algoritmos mais usados para encontrar padrões frequentes de mineração (3.2.1) e sua estratégia é baseada na técnica de dividir para conquistar, obtendo um melhor desempenho na geração das regras de associação, pois o número de varreduras na base de dados é menor que a do algoritmo APriori. Para FPGrowth encontrar padrões frequentes são necessários dois parâmetros de entrada, o suporte mínimo e a confiança. A implementação deste modelo foi realizada com uma ferramenta de código aberto, Orange [Demšar et al. 2018].

O objetivo inicial foi encontrar todos os possíveis padrões frequentes, independentemente da natureza do crime, criando uma lista de todos os locais com maior ocorrência criminal, juntamente com o dia e o horário mais frequente relacionado.

O algoritmo foi implementado utilizando somente os atributos relacionados à localização e tempo de ocorrência. Além disso, para obter padrões mais frequentes, foi

aplicada a mineração baseada em restrição, restringindo o processo de extração com essa fórmula de três conjuntos de itens específicos (bairro → dia → turno).

O objetivo foi encontrar aproximadamente 50 padrões frequentes. Assim, foram realizados vários experimentos usando diferentes valores mínimos de suporte e de confiança. Finalmente, define-se o valor de suporte mínimo em 0.01 e confiança 0.04, que representou 13.216 instâncias, 12% da base de dados. Com estas definições foram identificados 62 padrões, contemplando 14 bairros.

O segundo objetivo foi identificar que natureza de crime pode estar associada a um local específico, dentro de um tempo particular, e um perfil específico de vitima. Assim, foram realizados outros experimentos, inserindo os atributos anteriormente desconsiderados.

5. Resultados

Nesta seção é apresentado o resumo dos principais resultados obtidos com a aplicação das Regras de Associação utilizando o Algoritmo FPGrowth.

5.1. Análise espacial e temporal

Visando encontrar os pontos criminais espaciais e temporais foram extraídos todos os padrões interessantes com base nos limites predefinidos. A figura 2 mostra os resultados da aplicação das regras de associação, revelando 62 padrões, que contemplaram 13.216 instâncias. É importante mencionar que neste primeiro momento o objetivo foi encontrar associações entre o bairro, dia da semana e horário da ocorrência. Com esses conjuntos de itens frequentes são revelados os locais de crimes juntamente com seu dia e horário de ocorrência associados.

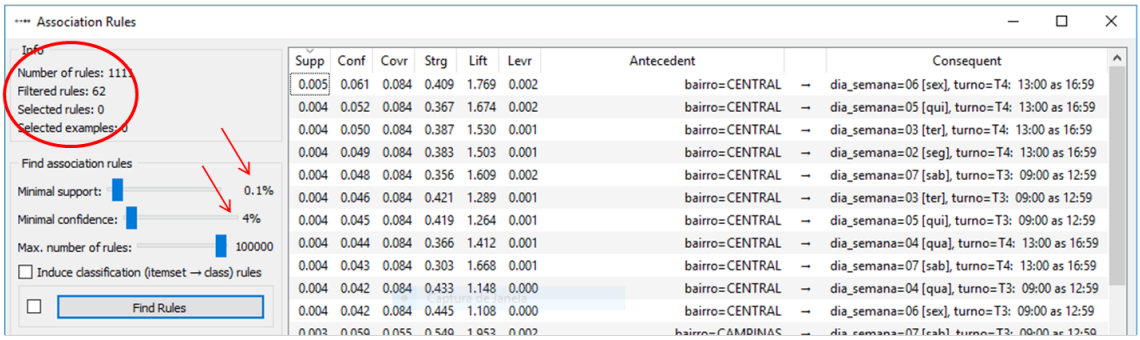

Figura 2. Padrões encontrados

Na sequência são apresentadas as tabelas com os resultados extraídos, onde são listadas a regra (associação de itens), o suporte e a confiança de cada associação. No rodapé de cada tabela são apresentados os parâmetros definidos para extração das referidas regras.

A tabela 1 apresenta os padrões frequentes, suporte mínimo e confiança das regras extraídas entre a associação dos atributos — bairro, dia da semana e turno — indicando que 14 bairros são locais que têm padrões frequentes de crimes. O bairro Central e Campinas são os que têm maior número de padrões. Além disso, nota-se que sexta-feira é o dia da semana com o maior número de padrões. Observa-se também que a maioria

dos padrões estão associados ao período da tarde. Por exemplo, a primeira linha da tabela revela que no bairro Central um padrão de ocorrência é sexta-feira das 13h00min às 16h59min. Na linha 12 observa-se que no bairro Campinas o padrão mais frequente é no sábado das 09h00min às 12h59min. Já no Jardim Goiás, o padrão é no domingo das 17h00min às 20h59min (linha 23).

Regra	Antecedent		Consequent	Supp	Conf
1	CENTRAL	→	sexta-feira,13h00min às 16h59min	0.005	0.061
2	CENTRAL	→	quinta-feira,13h00min às 16h59min	0.004	0.052
3	CENTRAL	→	terça-feira,13h00min às 16h59min	0.004	0.050
4	CENTRAL	→	segunda-feira,13h00min às 16h59min	0.004	0.049
5	CENTRAL	→	sábado, 09h00min às 12h59min	0.004	0.048
6	CENTRAL	→	terça-feira, 09h00min às 12h59min	0.004	0.046
7	CENTRAL	→	quinta-feira, 09h00min às 12h59min	0.004	0.045
8	CENTRAL	→	quarta-feira,13h00min às 16h59min	0.004	0.044
9	CENTRAL	→	sábado,13h00min às 16h59min	0.004	0.043
10	CENTRAL	→	quarta-feira, 09h00min às 12h59min	0.004	0.042
11	CENTRAL	→	sexta-feira, 09h00min às 12h59min	0.004	0.042
12	CAMPINAS	→	sábado, 09h00min às 12h59min	0.003	0.059
13	CAMPINAS	→	sexta-feira,13h00min às 16h59min	0.003	0.053
14	CAMPINAS	→	sexta-feira, 09h00min às 12h59min	0.003	0.053
15	CAMPINAS	→	terça-feira,13h00min às 16h59min	0.003	0.052
16	CAMPINAS	→	quinta-feira, 09h00min às 12h59min	0.003	0.052
17	CAMPINAS	→	terça-feira, 09h00min às 12h59min	0.003	0.052
18	CAMPINAS	→	segunda-feira, 09h00min às 12h59min	0.003	0.050
19	CAMPINAS	→	quinta-feira,13h00min às 16h59min	0.003	0.048
20	CAMPINAS	→	quarta-feira, 09h00min às 12h59min	0.003	0.046
21	CAMPINAS	→	segunda-feira,13h00min às 16h59min	0.002	0.045
22	CAMPINAS	→	quarta-feira,13h00min às 16h59min	0.002	0.043
23	JARDIM GOIÁS	→	domingo,17h00min às 20h59min	0.002	0.065
24	BUENO	→	quinta-feira, 09h00min às 12h59min	0.002	0.044
25	CIDADE JARDIM	→	sexta-feira,17h00min às 20h59min	0.001	0.045
26	LESTE UNIVERSITÁRIO	→	sexta-feira,17h00min às 20h59min	0.001	0.047
27	CIDADE JARDIM	→	sexta-feira, 09h00min às 12h59min	0.001	0.043
28	JARDIM AMÉRICA	→	sexta-feira, 09h00min às 12h59min	0.001	0.048
29	LESTE UNIVERSITÁRIO	→	quinta-feira,13h00min às 16h59min	0.001	0.045
30	JARDIM AMÉRICA	→	quarta-feira, 09h00min às 12h59min	0.001	0.047
31	CIDADE JARDIM	→	segunda-feira, 09h00min às 12h59min	0.001	0.042
32	FINSOCIAL	→	terça-feira,17h00min às 20h59min	0.001	0.041
33	PEDRO LUDOVICO/ ...	→	sexta-feira, 09h00min às 12h59min	0.001	0.050
34	LESTE UNIVERSITÁRIO	→	quinta-feira,17h00min às 20h59min	0.001	0.044
35	LESTE UNIVERSITÁRIO	→	terça-feira,13h00min às 16h59min	0.001	0.043
36	JARDIM GOIÁS	→	domingo,21h00min às 00h59min	0.001	0.051
37	FINSOCIAL	→	quinta-feira,17h00min às 20h59min	0.001	0.040
38	LESTE UNIVERSITÁRIO	→	segunda-feira,13h00min às 16h59min	0.001	0.043
39	FINSOCIAL	→	domingo,17h00min às 20h59min	0.001	0.040
40	JARDIM AMÉRICA	→	segunda-feira, 09h00min às 12h59min	0.001	0.045
41	JARDIM AMÉRICA	→	quinta-feira, 09h00min às 12h59min	0.001	0.043
42	LESTE UNIVERSITÁRIO	→	sexta-feira, 09h00min às 12h59min	0.001	0.041
43	JARDIM AMÉRICA	→	terça-feira, 09h00min às 12h59min	0.001	0.041
44	PEDRO LUDOVICO/ ...	→	terça-feira,13h00min às 16h59min	0.001	0.045
45	CELINA PARK/ ...	→	quarta-feira,17h00min às 20h59min	0.001	0.051
46	JARDIM GOIÁS	→	sábado,21h00min às 00h59min	0.001	0.046
47	OESTE	→	sexta-feira, 09h00min às 12h59min	0.001	0.049
48	OESTE	→	sábado,17h00min às 20h59min	0.001	0.048
49	JARDIM EUROPA	→	quarta-feira, 09h00min às 12h59min	0.001	0.041
50	JARDIM EUROPA	→	sexta-feira, 09h00min às 12h59min	0.001	0.041
51	PEDRO LUDOVICO/ ..	→	sexta-feira,13h00min às 16h59min	0.001	0.042
52	JARDIM EUROPA	→	terça-feira,17h00min às 20h59min	0.001	0.040
53	JARDIM EUROPA	→	quarta-feira,13h00min às 16h59min	0.001	0.040
54	JARDIM EUROPA	→	quinta-feira, 09h00min às 12h59min	0.001	0.040
55	OESTE	→	sexta-feira,13h00min às 16h59min	0.001	0.047
56	NORTE FERROVIÁRIO	→	sábado, 09h00min às 12h59min	0.001	0.085
57	MUTIRÃO E CURITIBA	→	quarta-feira,17h00min às 20h59min	0.001	0.040
58	OESTE	→	segunda-feira, 09h00min às 12h59min	0.001	0.046
59	MUTIRÃO E CURITIBA	→	segunda-feira,17h00min às 20h59min	0.001	0.040
60	OESTE	→	quarta-feira, 09h00min às 12h59min	0.001	0.046
61	CELINA PARK/ ...	→	terça-feira,17h00min às 20h59min	0.001	0.045
62	OESTE	→	terça-feira, 09h00min às 12h59min	0.001	0.044

Settings (Supp min: 0.100; Conf: 4.000; Antecedent min items: 1; Antecedent max items: 99)

Tabela 1. Regras geradas considerando: bairro, data e hora

5.2. Análise espacial e temporal, considerando a natureza do crime e o perfil da vítima

Com o objetivo de identificar qual natureza de crime pode estar associada a um local específico, dentro de um tempo particular e um perfil específico de vítima, foram realizados outros experimentos inserindo atributos do crime anteriormente desconsiderados.

Para geração das regras de associação entre conjunto de atributos (elementos que compõe o crime) e a classe do crime, foi necessário induzir a classificação — conjunto de itens → classe — (SE conjunto de itens ENTÃO classe), onde: o conjunto de itens representa as 'características' da ocorrência e a classe representa a 'natureza do crime'. Desta forma, o algoritmo gerou somente as regras com um valor de classe no lado direito (consequente) da regra, associando os atributos à classe. Por exemplo, a figura 3 3 mostra que: Se o indivíduo for do sexo feminino com idade entre 31 e 59 anos, a natureza do crime é "ameaça" com um suporte de 0.025 e confiança de 0.098.

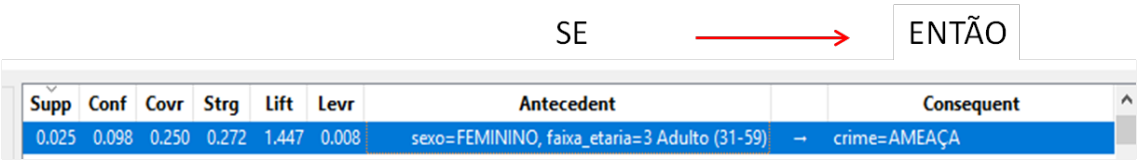

Figura 3. Exemplo de indução de regras de associação

A seguir, são apresentados os resultados obtidos com a indução para associar as características da ocorrência à natureza do crime. Na identificação dos padrões foi feita a filtragem das regras informado o número mínimo de itens do antecedente (lado esquerdo da regra). Sendo assim, inicialmente foi deixado como 1 (um). Com essa configuração é mostrado na tabela 2 que o padrão com o maior suporte, ou seja, ocorre com maior frequência é SE (SEXO = FEMININO ENTÃO CRIME = ROUBO A TRANSEUNTE). Percebe-se que a associação entre 'FEMININO' e 'ROUBO A TRANSEUNTE' é maior que entre 'MASCULINO' e 'ROUBO A TRANSEUNTE' (linha 3), indicando que o crime de roubo a transeunte é mais frequente com a vítima do sexo feminino.

Regra	Antecedent		Consequent	Supp	Conf
1	FEMININO	→	crime=ROUBO A TRANSEUNTE	0.106	0.215
2	Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.098	0.275
3	MASCULINO	→	crime=ROUBO A TRANSEUNTE	0.097	0.191
4	Adulto (31-59)	→	crime=ROUBO A TRANSEUNTE	0.081	0.160
5	17h00min às 20h59min	→	crime=ROUBO A TRANSEUNTE	0.056	0.241
6	MASCULINO, Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.051	0.279
7	FEMININO	→	crime=AMEAÇA	0.050	0.101
8	FEMININO, Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.047	0.271
9	FEMININO, Adulto (31-59)	→	crime=ROUBO A TRANSEUNTE	0.047	0.187
10	MASCULINO	→	crime=OUTROS FURTOS	0.046	0.090
11	05h00min às 08h59min	→	crime=ROUBO A TRANSEUNTE	0.045	0.344
12	Adulto (31-59)	→	crime=OUTROS FURTOS	0.045	0.088
13	Adulto (31-59)	→	crime=AMEAÇA	0.043	0.085
14	MASCULINO	→	crime=FURTO DE DOCUMENTOS	0.042	0.084
15	MASCULINO	→	crime=ROUBO DE VEÍCULO	0.041	0.081
16	Adulto (31-59)	→	crime=ROUBO DE VEÍCULO	0.040	0.080

Settings (Supp mín: 4.000; Conf: 1.000; Antecedent min items: 1; Antecedent max items: 99)

Tabela 2. Regras geradas com a indução da classe (1 item no antecedente)

Na sequência, foram filtradas as regras com no mínimo "dois itens no antecedente", ou seja, serão apresentadas somente associações com no mínimo de dois atributos (tabela 3). Os padrões com maiores suportes são: 'MASCULINO', 'JOVEM' → 'ROUBO A TRANSEUNTE'; 'FEMININO', 'jovem' → 'ROUBO A TRANSEUNTE'; 'FEMININO', 'ADULTO' → 'ROUBO A TRANSEUNTE', indicando que a associação entre vítima "jovem"do sexo masculino e roubo a transeunte é maior que, vítima "jovem"do sexo feminino.

Regra	Antecedent		Consequent	Supp	Conf
1	MASCULINO, Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.051	0.279
2	FEMININO, Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.047	0.271
3	FEMININO, Adulto (31-59)	→	crime=ROUBO A TRANSEUNTE	0.047	0.187
4	MASCULINO, Adulto (31-59)	→	crime=ROUBO A TRANSEUNTE	0.034	0.133
5	17h00min às 20h59min, FEMININO	→	crime=ROUBO A TRANSEUNTE	0.030	0.253
6	17h00min às 20h59min, Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.028	0.307
7	FEMININO, Adulto (31-59)	→	crime=AMEAÇA	0.028	0.110
8	05h00min às 08h59min, FEMININO	→	crime=ROUBO A TRANSEUNTE	0.026	0.407
9	17h00min às 20h59min, MASCULINO	→	crime=ROUBO A TRANSEUNTE	0.026	0.228
10	MASCULINO, Adulto (31-59)	→	crime=ROUBO DE VEÍCULO	0.025	0.100
11	MASCULINO, Adulto (31-59)	→	crime=OUTROS FURTOS	0.025	0.097
12	21h00min às 00h59min, MASCULINO	→	crime=ROUBO A TRANSEUNTE	0.024	0.268
13	21h00min às 00h59min, Jovem (18-30)	→	crime=ROUBO A TRANSEUNTE	0.023	0.329
14	17h00min às 20h59min, Adulto (31-59)	→	crime=ROUBO A TRANSEUNTE	0.022	0.189
15	05h00min às 08h59min, Adulto (31-59)	→	crime=ROUBO A TRANSEUNTE	0.021	0.298

Settings (Supp mín: 1.000; Conf: 4.000; Antecedent min items: 2; Antecedent max items: 99)

Tabela 3. Regras geradas com a indução da classe (2 itens no antecedente)

Deste modo, o número mínimo de itens do antecedente foi alterado progressivamente, até a quantidade de cinco (máximo de atributos considerados na análise, exceto a classe). Com cinco atributos (mais a classe), o padrão com maior suporte é bairro 'CENTRAL', 'sexta-feira','13h00min às 16h59min', 'FEM', 'Adulto (31-59)' → crime='FURTO A USUÁRIOS DE TRANSPORTE COLETIVO', essa associação equi-

vale a 67 ocorrências criminais da base de dados, representando o suporte de 0.001 (tabela 4).

Regra	Antecedent		Consequent	Supp	Conf
1	CENTRAL, sexta-feira,13h00min às 16h59min, FEM, Adulto (31-59)	→	crime=FURTO A USUÁRIOS DE TRANSP	0.001	0.435
2	CENTRAL, quinta-feira,13h00min às 16h59min, FEM, Adulto (31-59)	→	crime=FURTO A USUÁRIOS DE TRANSP	0.001	0.504
3	CENTRAL, segunda-feira,13h00min às 16h59min, FEM, Adulto (31-59)	→	crime=FURTO A USUÁRIOS DE TRANSP	0.001	0.420
4	CENTRAL, terça-feira,13h00min às 16h59min, FEM, Adulto (31-59)	→	crime=FURTO A USUÁRIOS DE TRANSP	0.001	0.413
5	CENTRAL, sexta-feira,13h00min às 16h59min, FEM, 4 Idoso (60 ..)	→	crime=FURTO A USUÁRIOS DE TRANSP	0.001	0.505

Settings (Supp mín: 0.005; Conf: 1.000; Antecedent min items: 5; Antecedent max items: 99)

Tabela 4. Regras geradas com a indução da classe (5 itens no antecedente)

5.3. Análise integrada: ocorrências e perfil socioeconômico do local

A geração de regras de associação envolvendo a informação socioeconômica foi limitada aos seguintes atributos: dia da semana, horário, renda do bairro (renda média por domicílio em salários mínimos). As linhas 1 a 3 da tabela 5 revela a associação entre o crime 'roubo a transeunte', o horário '17h00min às 20h59min' e os bairros com renda domiciliar 'acima de 3,5 salários mínimos' é mais frequente. Porém, observa-se que na linha 4, o mesmo crime (roubo a transeunte) é mais frequente nos bairros com renda 'abaixo de 3,5' no horário de '05h00min às 08h59min'. Deste modo, nota-se que é possível existir a relação entre um tipo específico de crime, um perfil específico de bairro, em um espaço de tempo específico.

Regra	Antecedent		Consequent	Supp	Conf
1	17h00min às 20h59min, renda_domicilio 5.5 - 8.5	→	crime=ROUBO A TRANSEUNTE	0.013	0.238
2	17h00min às 20h59min, renda_domicilio ≥ 8.5	→	crime=ROUBO A TRANSEUNTE	0.012	0.198
3	17h00min às 20h59min, renda_domicilio 3.5 - 4.5	→	crime=ROUBO A TRANSEUNTE	0.012	0.269
4	05h00min às 08h59min, renda_domicilio < 3.5	→	crime=ROUBO A TRANSEUNTE	0.011	0.490
5	13h00min às 16h59min, renda_domicilio ≥ 8.5	→	crime=FURTO A USUÁRIOS DE TRAN	0.011	0.169
6	05h00min às 08h59min, renda_domicilio 3.5 - 4.5	→	crime=ROUBO A TRANSEUNTE	0.011	0.392
7	09h00min às 12h59min, renda_domicilio ≥ 8.5	→	crime=ESTELIONATO	0.011	0.147
8	05h00min às 08h59min, renda_domicilio 5.5 - 8.5	→	crime=ROUBO A TRANSEUNTE	0.010	0.319
9	09h00min às 12h59min, renda_domicilio ≥ 8.5	→	crime=OUTROS FURTOS	0.010	0.135
10	17h00min às 20h59min, renda_domicilio 4.5 - 5.5	→	crime=ROUBO A TRANSEUNTE	0.010	0.244
11	21h00min às 00h59min, renda_domicilio ≥ 8.5	→	crime=ROUBO A TRANSEUNTE	0.010	0.237
12	17h00min às 20h59min, renda_domicilio < 3.5	→	crime=ROUBO A TRANSEUNTE	0.010	0.279

Settings (Supp mín: 0.010; Conf: 1.000; Antecedent min items: 3; Antecedent max items: 99)

Tabela 5. Regras geradas com a inclusão da renda do bairro da ocorrência

Regra	Antecedent		Consequent	Supp	Conf
1	domingo, 17h00min às 20h59min, renda_domicilio 5.5 - 8.5	→	crime=AMEAÇA	0.011	1.000
2	terça-feira, 09h00min às 12h59min, renda_domicilio ≥ 8.5	→	crime=AMEAÇA	0.011	1.000
3	domingo, 17h00min às 20h59min, renda_domicilio < 3.5	→	crime=AMEAÇA	0.010	1.000
4	segunda-feira, 17h00min às 20h59min, renda_domicilio 5.5 - 8.5	→	crime=AMEAÇA	0.010	1.000
5	domingo, 17h00min às 20h59min, renda_domicilio 3.5 - 4.5	→	crime=AMEAÇA	0.010	1.000
6	segunda-feira, 09h00min às 12h59min, renda_domicilio 5.5 - 8.5	→	crime=AMEAÇA	0.010	1.000
7	terça-feira, 09h00min às 12h59min, renda_domicilio 5.5 - 8.5	→	crime=AMEAÇA	0.010	1.000
8	quarta-feira, 09h00min às 12h59min, renda_domicilio ≥ 8.5	→	crime=AMEAÇA	0.010	1.000
9	quinta-feira, 09h00min às 12h59min, renda_domicilio ≥ 8.5	→	crime=AMEAÇA	0.010	1.000

Settings (Supp mín: 0.010; Conf: 1.000; Antecedent min items: 3; Antecedent max items: 99)

Tabela 6. Regras geradas somente para o crime ameaça

Em um segundo momento foi testado a extração de regras de associação isolando dois crimes e tratando-os de formas distintas. Primeiro foi testado o crime de 'ameaça'. Desta forma, a base de dados só conteve as instâncias deste crime. Igualmente, o mesmo procedimento foi feito para o crime 'roubo a residência'. A tabela 6 revela que o crime 'ameaça' é mais frequente em bairros com domicílios de renda média acima de 3,5 salários mínimos. Uma associação interessante na tabela 7 é que o 'roubo em residência'

nos bairros (com renda inferior a 3,5 salários mínimos) tem um mais frequência a partir das 21h00min, enquanto que nos bairros com renda superiores o padrão mais frequente é no período da tarde ou pela manhã.

Regra	Antecedent		Consequent	Supp	Conf
1	quarta-feira,21h00min às 00h59min, renda_domicilio < 3.5	→	crime=ROUBO EM RESIDÊNCIA	0.020	1.000
2	quinta-feira,05h00min às 08h59min, renda_domicilio 5.5 - 8.5	→	crime=ROUBO EM RESIDÊNCIA	0.019	1.000
3	sábado,21h00min às 00h59min, renda_domicilio < 3.5	→	crime=ROUBO EM RESIDÊNCIA	0.017	1.000
4	sexta-feira,21h00min às 00h59min, renda_domicilio 3.5 - 4.5	→	crime=ROUBO EM RESIDÊNCIA	0.016	1.000
5	quarta-feira,17h00min às 20h59min, renda_domicilio 5.5 - 8.5	→	crime=ROUBO EM RESIDÊNCIA	0.016	1.000
6	segunda-feira,17h00min às 20h59min, renda_domicilio < 3.5	→	crime=ROUBO EM RESIDÊNCIA	0.015	1.000
7	quinta-feira,17h00min às 20h59min, renda_domicilio 3.5 - 4.5	→	crime=ROUBO EM RESIDÊNCIA	0.015	1.000
8	quarta-feira,21h00min às 00h59min, renda_domicilio 3.5 - 4.5	→	crime=ROUBO EM RESIDÊNCIA	0.013	1.000
9	terça-feira,17h00min às 20h59min, renda_domicilio 5.5 - 8.5	→	crime=ROUBO EM RESIDÊNCIA	0.013	1.000
10	sexta-feira,17h00min às 20h59min, renda_domicilio < 3.5	→	crime=ROUBO EM RESIDÊNCIA	0.013	1.000
11	sexta-feira,21h00min às 00h59min, renda_domicilio < 3.5	→	crime=ROUBO EM RESIDÊNCIA	0.013	1.000
12	terça-feira,21h00min às 00h59min, renda_domicilio 3.5 - 4.5	→	crime=ROUBO EM RESIDÊNCIA	0.013	1.000
13	segunda-feira,21h00min às 00h59min, renda_domicilio < 3.5	→	crime=ROUBO EM RESIDÊNCIA	0.013	1.000

Settings (Supp mín: 0.010; Conf: 1.000; Antecedent min items: 3; Antecedent max items: 99)

Tabela 7. Regras geradas somente para o crime de roubo a residência

6. Conclusão e Trabalhos Futuros

O trabalho apresentou a construção de um modelo completo para descoberta do conhecimento e utilizou-se de todas as fases da descoberta (da seleção à interpretação). Como fonte de estudo fez uso de uma base criminal do município de Goiânia, visando extrair padrões criminais no que diz respeito à vítima, espaço local e temporal do crime.

Através do estudo foi possível identificar padrões de ocorrências criminais, como por exemplo, os locais onde há mais ocorrências em determinados horários, relacionados a um perfil específico de vítima. Portanto, conclui-se que a pesquisa foi satisfatória alcançando os objetivos propostos.

Em relação à análise dos dados socioeconômicos, de um modo geral não foi um fator identificado como "forte" influência para o "alto"índice de crimes. No entanto, fica a ressalva: a análise de cada natureza criminal deverá ser analisada de forma independente e outros elementos característicos do local do crime deverão ser inseridos.

Finalmente, diante do estudo realizado acerca de um tema tão relevante, que de certa forma todos estão inseridos, espera-se que o conteúdo aqui apresentado, de certa forma possa contribuir na descoberta de conhecimento sobre o referido tema e na decisão de adotar a utilização de técnicas de Mineração de Dados como ferramenta de apoio a tomada de decisão.

Visando a continuidade deste trabalho, no intuito de alcançar novas descobertas que possam contribuir um pouco mais sobre o tema, ficam como sugestão para trabalhos futuros as seguintes linhas de pesquisa: aplicação de outros modelos e técnicas não inseridas neste estudo, como por exemplo, técnicas de visualização de dados, modelos de classificação, entre outros; abordagem de outros aspectos não tratados neste estudo, como por exemplo, associação entre o perfil da vítima e criminoso; expansão da análise considerando o âmbito estadual.

Referências

Almanie, T., Mirza, R., and Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *CoRR*, abs/1508.02050.

Amaral, F. (2016). *Introdução à Ciência de Dados: mineração de dados e big data*. ALTA BOOKS.

Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. (2018). Orange data mining. version 3.13.0. https://orange.biolab.si/.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *Ai Magazine*, 17:37–54.

Fórum, B. d. P. S. (2017). Anuário brasileiro de segurança pública. http://www.forumseguranca.org.br/11o-anuario-brasileiro-de-seguranca-publica/. acesso em: 30 abril 2018.

Han, Pei, and Yin (2000). Mining frequent patterns without candidate generation. *SIGMODREC: ACM SIGMOD Record*, 29.

IBGE, I. B. d. G. e. E. (2017). Estimativa da população. https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/9103-estimativas-de-populacao.html?=t=resultados. acesso em: 02 fev 2018.

IBM (2011). Ibm spss statistic. version 20.0.0. https://www.ibm.com/products/spss-statistics.

Kantardzic, M. (2011). *Data Mining: Concepts, Models, Methods, and Algorithms*. Wiley.

Keyvanpour, M. R., Javideh, M., and Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework. volume 3, pages 872–880. Elsevier.

Mande, U., Srinivas, Y., and Murthy, J. V. R. (2012). Feature specific criminal mapping using data mining techniques and generalized gaussian mixture model. *International Journal of Computer Science and Communication Networks*.

Nath, S. V. (2006). Crime pattern detection using data mining. pages 41–44. IEEE Computer Society.

Silva, M. P. d. S. (2004). Mineração de dados - conceitos, aplicações e experimentos com weka.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.

Vijayakumar, M., Balamurugan, P., and Alhadidi, B. (2014). Crime classification algorithm for mining crime hot spot and cold spot.

Villarreal, A. and Silva, B. F. (2006). Social cohesion, criminal victimization and perceived risk of crime in brazilian neighborhoods. *Social Forces*, 84(3):1725–1753.

Wang, D., 0003, W. D., Lo, H. Z., Stepinski, T. F., Salaza, J., and Morabito, M. (2013). Crime hotspot mapping using the crime related factors - a spatial data mining approach. *Appl. Intell*, 39(4):772–781.

Zaki, M. J. and Jr, W. M. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, NY, USA.