Mineração de Dados Criminais: Um Estudo de Caso no Observatório Criminal do Tapajós

Criminal Data Mining: A Case Studyin Criminal Observatory Tapajós

Bruno M. de Melo, Jarsen L. C. Guimarães, Adriângela S. de Castro, Clayton A. M. Santos
Instituto de Engenharia e Geociências
Universidade Federal do Oeste do Pará
Santarém – PA – Brasil
brunomachadodemelo@gmail.com

Durbens M. Nascimento
Núcleo dos Altos Estudos Amazônicos
Universidade Federal do Pará
Belém – PA – Brasil
durbens.naea@gmail.com

Adriano Del Pino Lino
Center for Informatics and Systems
Department of Informatics Engineering
University of Coimbra
Polo II, 3030-290, Coimbra, Portugal
adlino@dei.uc.pt

Resumo — **A mineração de dados (*data mining*) permite investigação à procura de padrões, muitas vezes não visíveis pela simples observação dos dados. É um processo de exploração com o intuito de detectar relacionamentos entre as variáveis, buscando inferir ou criar previsões para dados futuros. Na área de segurança pública a mineração de dados ou MD, pode ser utilizada em diversas formas: identificar relação de um tipo de crime com algum bairro, determinar a existência de um padrão para a idade, sexo, dia da semana e horário de quem comete determinado tipo de crime, entre tantas outras possibilidades. O objetivo deste artigo é utilizar técnica de MD no OBCRIT – Observatório Criminal do Tapajós, que é um banco de dados que registra boletins de ocorrências do 3º Batalhão de Polícia Militar do Estado do Pará. Os resultados das simulações por meio da ferramenta Weka propõe traçar um perfil das ocorrências que fazem parte de um mesmo agrupamento, encontrando indícios comuns para os crimes cadastrados e mostrando a importância da utilização da mineração de dados no processo de extração do conhecimento a níveis criminais.**

Palavras Chave - Mineração de dados; Criminalidade; OBCRIT; Segurança Pública.

Abstract — **Data mining allows research to reach patterns, often not visible by the simple observation of data. It is an exploration process in order to detect relations between the variables, seeking infer or create forecasts for future data. In public safety area data mining or MD, can be used in several ways: to identify the relation of a crime type with some neighborhood, determine the existence of a pattern for age, sex, day and the time that somebody commits some type of crime, among many other possibilities. The purpose of this article is to use MD technique in OBCRIT - Criminal Observatory Tapajós, which is a database that records reports of occurrences of the 3rd Battalion of Military Police of Pará State. The simulation results using Weka tool proposes to trace a profile of the occurrences that are part of the same group, finding common indications for the registered crimes and showing the importance of the use of data mining to in the process of extract knowledge to criminal levels.**

Keywords - Data mining; criminality; OBCRIT; Public Security.

I. INTRODUÇÃO

Mineração de dados (MD) ou *Data Mining* é uma alternativa eficaz para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar as instituições nas tomadas de decisões mais rápidas ou até mesmo a atingir um maior grau de confiança [1].

Na segurança pública, a MD tem se tornado uma das técnicas cada vez mais utilizadas na identificação de criminosos, mapeamento de crimes por região, detecção do perfil de criminosos e vítimas, ou até mesmo para a análise específica de certos tipos de crimes, tudo isso somado a possibilidade de previsão de determinado crime acontecer em certo lugar e tempo. Atualmente, o tema criminalidade mobiliza toda a sociedade na discussão de como combatê-la ou minimizá-la em tempos tão complexos, nos quais os valores, a pobreza e o consumo estão à prova. Em 2007, o ministério da justiça do Brasil expôs em um curso nacional de multiplicador de polícia comunitária que as estratégias de policiamento que funcionaram há décadas passadas, não estão sendo mais eficazes, sendo necessária a modernização e incentivos em inovações tecnológicas capazes de acompanhar o crescente avanço da criminalidade [2].

O propósito desta pesquisa é fazer um levantamento das informações que estão cadastradas na base de dados do OBCRIT a fim de verificar as relações entre os crimes e os demais atributos presentes em cada ocorrência. Os dados obtidos foram cedidos a partir dos boletins de ocorrências do 3º Batalhão de Polícia Militar do estado do Pará, por meio de uma parceria entre o OBCRIT, Polícia Militar, Polícia Civil, Centro de Perícias Criminal Renato Chaves. Todos os dados são referentes à cidade de Santarém, localizada na região oeste do Pará, totalizando mais de 5.400 ocorrências cadastradas entre os meses de julho de 2013 a abril de 2014. O banco de informações contém diversos tipos de variáveis, entre elas: data, hora, dia da semana, cidade, bairro, entre outras. Saber qual a relação entre o crime e esses atributos fará com que sejam criadas estratégicas eficazes no combate de possíveis

outros delitos, a partir do momento em que a mineração de dados fornece agrupamento e informações estatísticas da relação entre eles. Por exemplo, saber que os crimes de furto ocorrem em bairros intermediários (bairros que estão entre o centro e a periferia) e que são cometidos por jovens (18 a 24 anos), no final do mês (21 a 31), à noite (19 à 00 hora), em um dia útil, faz com que a polícia tome medidas cautelares de prevenção a esse crime que há bastante tempo vem causando transtornos à população. Ou seja, pode ser feita intervenções por meio de palestras nesses bairros, onde a polícia se aproxima mais da população por meio de programas sociais, ou até mesmo blitz educativas com mais frequências. Tudo isso com o intuito de inibir a prática do delito. As informações produzidas pela análise criminal desta pesquisa poderão ser instrumentos válidos para aplicação eficiente por parte do administrador de segurança pública ao gerir recursos que lhe são disponibilizados pelo Estado e ser efetivo no proposito de controlar e neutralizar as manifestações criminosas e violentas.

O presente artigo encontra-se estruturado da seguinte forma: na seção II, é apresentada a fundamentação teórica; na seção III, apresenta-se os trabalhos correlatos; na seção IV é conduzida a metodologia para a descoberta de padrões; a seção V descreve-se os resultados obtidos; e, na seção VI, anuncia-se as conclusões e trabalhos futuros decorrentes desta pesquisa.

II. FUNDAMENTAÇÃO TEÓRICA

A. *Mineração de Dados (Data Mining)*

Segundo [3] a mineração de dados (MD) se classifica como uma ferramenta de interface entre a estatística e outras áreas do conhecimento. Este método destaca-se por sua capacidade de descobrir padrões com alguns significados práticos para o usuário. O autor revela ainda que o usuário deve se ater ao fato de que a má utilização desta ferramenta e a não observância de seus critérios de aplicação e análise pode levar ao viés dos resultados e, consequentemente, a uma interpretação errônea dos mesmos.

Dentre os processos criados para a análise de grandes bancos de dados, destaca-se o KDD (*Knowledge Discovery in Databases*), definido [4], como um método para extrair conhecimento útil a partir de volume de dados. Esse é um processo global com o objetivo de extrair informações de grandes bancos de dados, utilizando algum procedimento automático (matemático ou computacional).

O processo KDD exige diálogo direto com o conhecimento prévio do pesquisador na tomada de decisão sobre os resultados encontrados pelo método. A "Fig. 1" destaca as etapas de KDD necessárias ao processo de descoberta de padrões sobre o banco de dados.

Este processo mais abrangente possui uma metodologia própria para preparação e exploração de dados, interpretação de seus resultados e assimilação dos conhecimentos minerados. Mas a etapa de mineração de dados ficou mais conhecida, pois é nela que os algoritmos são aplicados. Abaixo faz-se uma breve apresentação a respeito do algoritmo de mineração dedados que é aplicado nesta pesquisa.

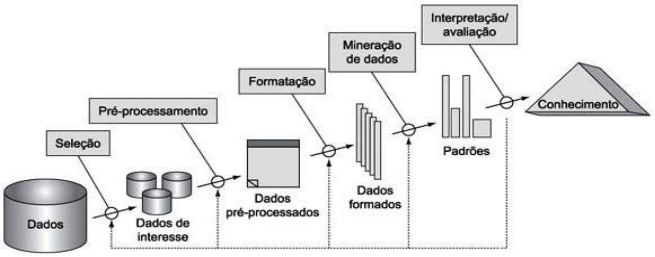


Figura 1. Visão geral das etapas que constituem o processo [4]

B. *Agrupamento (Clusters)*

A tarefa de agrupamento implementa algoritmo não supervisionado em que o rótulo da classe de cada amostra de treinamento não é conhecido, e o número ou conjunto de classes a ser treinado pode não ser notório a priori. No agrupamento os objetos são incorporados de modo que a semelhança seja máxima dentro de cada *cluster* e mínima entre instâncias de *clusters* diferentes [5]. Segundo [6] esta funcionalidade visa segmentar um conjunto de dados num número de subgrupos homogêneos ou *clustering*. Seu objetivo é formar grupos baseados no princípio de que os agrupamentos devem ser o mais homogêneo em si e mais heterogêneo entre si. A diferença fundamental entre a formação de agrupamentos e a classificação é que no *cluster* não existem classes predefinidas para classificar os registros em estudo. Os registros são incorporados em função de suas similaridades básicas, ou seja, quando se deseja formar agrupamentos, seleciona-se um conjunto de atributos (variáveis) e em função da similaridade dessas características são formados os grupos.

Os critérios para a escolha de agrupamentos ao invés de outras técnicas como classificação e associação, se deve ao fato de que outros trabalhos envolvendo dados criminais já foram desenvolvidos utilizando a técnica de *cluster*, gerando resultados bastante interessantes. Outro critério é o fato da técnica de agrupamento ser não supervisionada, ou seja, os rótulos das classes no conjunto de treinamento são desconhecidos, o objetivo então é estabelecer a existência de classes ou grupos nos dados. Por exemplo, observando os crimes contra o patrimônio, quais foram às características agrupadas em relação ao crime de furto? Como resposta obteve-se que eles ocorrem em determinados bairros (intermediários) e a característica de quem os comete são de jovens (18 a 24 anos) do sexo masculino, no final do mês, à noite, em um dia útil da semana. Então, todos esses grupos formados a partir dos *clusters* faz com que seja bem visualizada a forma de intervenção posterior por parte do administrador de segurança pública.

Em um primeiro momento buscou-se a utilização de apenas uma técnica para gerar resultados iniciais, os autores pretendem incluir novas técnicas, até mesmo para aprimorar e comparar os resultados no decorrer dos novos trabalhos.

C. *Algoritmo de Agrupamento K-means (K-médias)*

É um algoritmo de partição que busca de forma direta pela divisão ótima (ou aproximadamente ótima) dos *n* elementos sem a necessidade de associações hierárquicas [7]. Servem exclusivamente para agrupar os *n* registros em *k clusters* [8]. Proposto por J. MacQueen em 1967, este é um dos algoritmos

mais conhecidos e utilizados, além de ser o que possui o maior número de variações [7]. O algoritmo inicia com a escolha dos *k* elementos que formaram as sementes iniciais. Esta escolha pode ser feita de muitas formas, entre elas: (1) Selecionando as *k* primeiras observações; (2) Selecionando *k* observações aleatoriamente; (3) Escolhendo *k* observações de modo que seus valores sejam bastante diferentes. Por exemplo, ao se agrupar uma população em três grupos de acordo com a altura dos indivíduos, poderia se escolher um indivíduo de baixa estatura, um de estatura mediana e um alto.

O critério da escolha de *k-means* na pesquisa ocorre a partir de outros trabalhos envolvendo mineração de dados criminais já utilizarem esse algoritmo para geração de agrupamentos. Em [14], os estudos realizados originaram resultados significativos, o que gerou expectativa nos autores do OBCRIT em gerar efeitos semelhantes. Outro critério significativo é a eficiência do método em tratar grandes volumes de dados, além de sua simplicidade computacional, sendo um algoritmo de fácil aplicação.

D. Distância Euclidiana

Dentre as medidas de distância, a Euclidiana é a mais conhecida. Para calcular a distância entre dois objetos (*i* e *j*) com dimensão *p* (número de variáveis), ela é definida por:

$$d(i,j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{il})^2} \quad (1)$$

Conforme [9], uma derivação muito usada é Distância Euclidiana Média (DEM), onde a soma das diferenças ao quadrado é divida pelo número de elementos envolvidos:

$$d(i,j) = \sqrt{\frac{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}{p}} \quad (2)$$

E. Weka

O *Weka*[1] (*Waikato Environment for Knowledge Analysis*) [10] - *software* de código aberto - é formado por um conjunto de implementaçõesde algoritmos de diversas técnicas de Mineração de Dados. O *Weka* está implementado na linguagem *Java*, que fornece a característica de ser portável, desta forma pode rodar nas mais variadas plataformas e aproveitando os benefícios de uma linguagemorientada a objetos como modularidade, polimorfismo, encapsulamento, reutilização de código, dentre outros. As principais interfaces desta ferramenta, segundo [14] são:

- *Simple Client*- O usuário opera por meio de linhas de comando, sendo muito aproveitado por usuário avançados;
- *Explorer* - Visualização mais usada, que aborda o pré-processamento dos dados, mineração de dados (associação, classificação, agrupamento) e pós-processamento dos dados (apresentação de resultados);
- *Experimenter*- Ambiente para testes estatísticos entre algoritmos utilizados;
- *Knowledge Flow* - Ferramenta para planejamento de ações.

Na pesquisa de [14] e [15] expõe que a ferramenta possui um formato próprio para o arquivo de entrada, o Arff [11]. Nele, descrevem-se os dados que são usados no processo de mineração. Antes de aplicar o algoritmo de mineração aos dados na ferramenta *Weka*, estes devem ser convertidos para o formato Arff. Esse processo pode ser feito da seguinte forma:

1) Criar um arquivo de texto composto de duas partes. A primeira contém uma lista de todos os atributos, onde devemos definir o tipo do atributo ou os valores que ele pode representar. A segunda parte consiste em instâncias, os registros a serem minerados;

2) Criar o arquivo Arff implicitamente, onde cada atributo é definido via código e as instâncias são carregadas diretamente do banco de dados.

III. TRABALHOS CORRELATOS

Segundo [12] o uso de mineração de dados como apoio à tomada de decisão ocorre em diversas áreas e setores, tais como: governo, marketing, medicina, economia, engenharia, administração, etc. Tratando-se de governo, especificamente no setor de segurança pública, alguns trabalhos têm sido feitos ressaltando-se a utilização da mineração de dados como apoio ao planejamento estratégico adotado pela polícia.

Em [13], mostram quais tipos de crimes ocorriam em determinadas horas e locais, servindo assim como grande apoio ao planejamento estratégico policial, onde poderiam ser alocados efetivos naqueles locais entre as horas determinadas após o resultado da mineração.

Nos estudos de [14] foram utilizadas técnicas de mineração de dados sobre o sistema SISGOP, um banco de dados da Polícia Militar do Estado de Alagoas para descobrir informações que auxiliam ações estratégicas da polícia, baseadas no comportamento dos criminosos e vítimas. Na pesquisa, foram registradas diversas variáveis relacionadas ao crime (tipo, arma empregada, localidade, data, horário e etc.) A aplicação teve como base o emprego da tarefa de agrupamento, utilizando o algoritmo *Simple Kmeans*, nas condutas dos criminosos e na descrição das vítimas, procurando potencializar indicações de diretrizes para policiamento em termos de distribuição do efetivo e das razões pelas quais ocorreu o delito. Foram obtidos resultados relevantes, gerando grupos de características comuns de condutas criminosas e de situações de risco das vítimas. Dentre os resultados encontrados, o mais interessante foram os casos das vítimas de roubo e ameaça que ocorria no bairro Jacintinho, no período das 14hs às 20hs, nos dias de domingo, o que segundo os autores poderia ser feita uma distribuição policial para orientar as vítimas sobre os riscos desses crimes, além de tentar descobrir as razões para tantos delitos dessa natureza no local especificado.

Assim como em [13] e [14], este artigo objetiva a aplicação de mineração de dados no problema envolvendo a

[1] http://www.cs.waikato.ac.nz/ml/weka

criminalidade, a fim de evidenciar as ocorrências por meio de agrupamentos entre as variáveis cadastradas na base de dados do OBCRIT. Ao relacionar os delitos cometidos aos dados do cotidiano como dia da semana, horário, início, meio e fim do mês, é possível agrupar os crimes cadastrados, com o intuito de traçar perfis, considerando aqueles que fazem parte do mesmo montante. Portanto, esses trabalhos motivaram a utilização da técnica de agrupamentos na base de dados do OBCRIT com intuito de, a partir dos resultados encontrados sejam criadas medidas preventivas a crimes futuros ao utilizar dados do passado.

IV. METODOLOGIA

A. *Escolha da base de dados*

A escolha da base de dados ocorre a partir dos boletins de ocorrência do 3º Batalhão de Polícia Militar do Estado do Pará cadastrados até o momento em um período de 10 meses (julho de 2013 a abril de 2014) com mais de 5.400 ocorrências. Desde o ano de 2014 as informações dos chamados BO/PM (Boletim de Ocorrência da Polícia Militar) estão sendo tabuladas e cadastradas mensalmente no OBCRIT a partir da liberação por parte do órgão competente descrito anteriormente. Depois de cadastradas essas informações geram indicadores de criminalidade e visualização de crimes no site www.obcrit.com. Devido a grande quantidade de dados cadastrados em um curto período de tempo, faz-se necessária à ampliação do conhecimento para a exploração por meio de mineração de dados por se tratar de uma grande ferramenta para a descoberta de padrões ainda desconhecidos pelos desenvolvedores.

B. *Pré-processamento e seleção das variáveis mais significativas*

Os dados que são processados nessa pesquisa estão armazenados no banco de dados do OBCRIT, divididos em 12 tabelas, sendo a principal delas chamada de "ocorrências". A estrutura dessa tabela é baseada em 16 atributos, são eles: *identificador da ocorrência*, *sexo do denunciante*, *idade do denunciante*, *sexo do denunciado*, *idade do denunciado*, *cidade*, *bairro*, *tipo de crime*, *crime*, *endereço*, *data*, *hora*, *dia da semana*, *tipo de chamada*, *latitude* e *longitude*. Desses atributos apenas 10 são os escolhidos para a fase de pré-processamento, são eles: *sexo* e *idade do denunciante* e *denunciado*, *bairro*, *tipo de crime*, *crime*, *data*, *hora* e *dia da semana*. Foram selecionados apenas 10 atributos, devido à falta de informação que poderia ser extraída por parte dos atributos que ficaram de fora. Por exemplo, o atributo *idOcorrência* apenas informa qual o identificador da ocorrência, sendo ele único para cada caso. O *tipo de crime* já havia sido utilizado para separar os crimes em suas devidas categorias, portanto não havia necessidade de repeti-lo já que todas as ocorrências de crimes contra o patrimônio, por exemplo, seriam do mesmo tipo de crime. O *endereço* gerou pouca relevância já que foram verificadas inconsistências no que diz respeito à rua, alameda, avenida, travessa e etc. A mesma rua, ora aparecia como travessas, ora como avenida e isso gerou bastante "lixo" com informações distorcidas. Sendo assim, optou-se pela exclusão desse atributo. O atributo *tipo de chamada* teve em sua maioria apenas um tipo que foi por "chamada via CIOP (Centro Integrado de Operações)", ou seja, mais de 95% dos resultados da mineração apareceriam com o atributo de *tipo de chamada* via CIOP, então elegeu-se a exclusão. Por último os atributos *latitude* e *longitude* só são necessários para apresentação das ocorrências no mapa do OBCRIT, e para a mineração é aconselhável à utilização de valores categóricos e não reais como é o caso desses atributos. Os dados escolhidos foram separados para um banco de testes chamado "mineração". A partir dessa etapa atributos foram padronizados para valores categóricos conforme o padrão exigido pela ferramenta *Weka*. Abaixo são apresentadas as padronizações dos atributos após a fase de pré-processamento:

- *Sexo do Denunciante* e *Denunciado*: MAS (Masculino) e FEM (Feminino);
- *Idade do Denunciante* e *Denunciado*: Criança (0 a 11 anos), Adolescente (12 a 17 anos), Jovens (18 a 24 anos), Adultos1 (25 a 34 anos), Adultos2 (35 a 64 anos), Idosos (Acima de 65 anos);
- *Bairro*: Central, Intermediario1 (próximos aos bairros centrais), Intermediário2 (próximos aos bairros Intermediários1) e Periférico (afastados da área central);
- *Tipo de Crime*: CCP (Crime contra a pessoa) Ex: Ameaça, Agressão e etc., CP (Crimes contra o patrimônio) Ex: Roubo, Furto, etc., CCDS (Crime contra a Dignidade Sexual) Ex: Estrupo, Assédio Sexual e etc., TD (Tráfico de Drogas) Ex: Tráfico, Associação ao tráfico e etc.
- *Data*: Início (01 a 10), Meio (11 a 20) e Fim (21 a 31) do mês.
- *Hora*: Manhã (06 às 12), Tarde (13 às 18), Noite (19 às 00) e Madrugada (01 às 05).
- *Dia da Semana*: Dias úteis (segunda, terça, quarta, quinta e sexta) e finais de semana (sábado e domingo).
- *Crimes*: Os crimes foram apenas trocados os nomes por siglas para facilitar a identificação. Ex: Ameaça (AME), Furto (FUR), Roubo (ROU) e assim sucessivamente.

C. *Criação do arquivo .Arff*

Após a fase de pré-processamento os dados foram exportados para o formato .CSV (separado por vírgula). Foram criados 4 arquivos, um para cada tipo de crime (contra a pessoa, patrimônio, dignidade sexual e tráfico). A "Fig. 2" ilustra a criação do arquivo, onde foram inseridos o cabeçalho da mineração contendo todos os atributos categóricos anunciados anteriormente (retângulo vermelho). Após o termo "@data" são listados os elementos referentes a cada ocorrência cadastrada no banco de dados (retângulo azul). Depois de padronizado o arquivo .CSV foi convertido para o arquivo .Arff por meio da troca de nome de CSV por Arff.

```
@relation ccp
@attribute sexoDenunciante {MAS,FEM}
@attribute idadeDenunciante {criança,adolescente,jovens,adultos1,adultos2,idosos}
@attribute sexoDenunciado {MAS,FEM}
@attribute idadeDenunciado {criança,adolescente,jovens,adultos1,adultos2,idosos}
@attribute idBairro {central, intermediario1,intermediario2,periferico}
@attribute idCrime {HOM,PTS,CONT,THO,PAB,ASS,CDP,LMP,DAN,MTS,PAF,LEC,ADI,TSU,AME,
@attribute data {inicio,meio,fim}
@attribute hora {manha,tarde,noite,madrugada}
@attribute diaSemana {util,final_de_semana}
@data
FEM,adultos1,FEM,idosos,intermediario2,AME,inicio,noite,util
MAS,idosos,FEM,adultos1,intermediario1,AME,inicio,noite,util
MAS,jovens,MAS,adolescente,intermediario2,ATF,inicio,madrugada,util
```

Figura 2. Arquivo .Arff dos crimes contra a pessoa.

V. SÍNTESE E ANÁLISE DOS RESULTADOS

A. *Resultados dos crimes contra a pessoa*

A "Fig. 3" ilustra os resultados dos agrupamentos dos crimes contra a pessoa. O resultado do *cluster*0 agrupou 70% das ocorrências (2.549) e do *cluster*1 incorporou os outros 30% (1.080). O interessante desse resultado é que em ambos os agrupamentos o *sexo* e *idade* dos denunciantes e denunciados permaneceram os mesmo, podendo assim dizer que os crimes de desordem (DES) são denunciados por pessoas do sexo feminino de idade entre 35 a 64 anos (adultos2) e que os denunciados são homens de 18 a 24 anos (jovens). O crime também é cometido em um *bairro* periférico ou em um intermediário2 podendo avaliar assim a menor incidência de crime dessa natureza em bairros da área central da cidade. Outro fato também visível é que o período emque acontece geralmente esse tipo de delito é no meio ou no fim do mês, de noite ou à tarde, sendo a maioria das ocorrências registradas em um dia útil da semana.

```
Final cluster centroids:
                                          Cluster#
Attribute                 Full Data             0                1
                           (3629.0)        (2549.0)         (1080.0)
=============================================================================
sexoDenunciante                 FEM             FEM              FEM
idadeDenunciante            adultos2        adultos2         adultos2
sexoDenunciado                  MAS             MAS              MAS
idadeDenunciado              jovens          jovens           jovens
idBairro             intermediario2      periferico  intermediario2
idCrime                         DES             DES              DES
data                            fim             fim             meio
hora                          noite           noite            tarde
diaSemana                      util            util  final de semana
-----------------------------------------------------------------------------
Clustered Instances

0       2549 ( 70%)
1       1080 ( 30%)
```

Figura 3. Resultado dos crimes contra a pessoa

B. *Resultados dos crimes contra o patrimônio*

Os resultados dos crimes contra o patrimônio são ilustrados na "Fig. 4". Na imagem nota-se que o *cluster*0 agrupou os crimes de ROU (roubo), responsável por 64% das ocorrências, enquanto que o *cluster*1, incorporou os crimes de FUR (furto). Ambos os crimes são mais denunciados e cometidos por pessoas do sexo masculino, e a idade de quem denuncia fica entre 35 a 64 anos (adultos2) para o crime de roubo e 18 a 24 anos (jovens) para o crime de furto. O interessante dessa análise é que ambos os crimes são cometidos por pessoas de baixa idade, 18 a 24 anos (jovens) para roubo e 12 a 17 anos (adolescentes) para furto. Os crimes comumente são cometidos em um dia útil, porém o roubo ocorre com mais frequência no inicio do mês e no período da tarde, enquanto que o furto ocorre no final do mês e no período da noite.

```
Final cluster centroids:
                                          Cluster#
Attribute                 Full Data             0                1
                           (1638.0)        (1051.0)         (587.0)
=============================================================================
sexoDenunciante                 MAS             MAS              MAS
idadeDenunciante            adultos2        adultos2           jovens
sexoDenunciado                  MAS             MAS              MAS
idadeDenunciado              jovens          jovens       adolescente
idBairro             intermediario2      periferico  intermediario2
idCrime                         ROU             ROU              FUR
data                            fim          inicio              fim
hora                          noite           tarde            noite
diaSemana                      util            util             util
-----------------------------------------------------------------------------
Clustered Instances

0       1051 ( 64%)
1        587 ( 36%)
```

Figura 4. Resultado dos crimes contra o patrimônio

C. *Resultado dos crimes contra a dignidade sexual*

A "Fig. 5" explana os resultados dos agrupamentos dos crimes contra a dignidade sexual. O *cluster*0 apresenta 59% (34) das ocorrências cadastradas, enquanto que o *cluster*1 agrupou 41% (24). Ambos os agrupamentos foram utilizando o crime de ESTU (estupro), onde foram analisados que na maioria das vezes esse crime é realizado em bairros afastados da área central (intermediários1 e periféricos) em dias úteis. A maioria dos denunciantes são pessoas do sexo feminino seja ela jovem (18 a 24 anos) ou crianças (0 a 11 anos). Já os denunciados são jovens (18 a 24 anos) ou adultos1 (25 a 34 anos), que cometem esse crime no período da tarde ou da noite, no meio ou no final do mês.

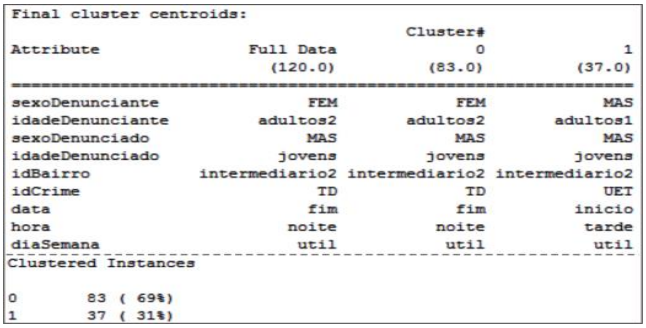
```
Final cluster centroids:
                                          Cluster#
Attribute                 Full Data             0                1
                             (58.0)          (34.0)           (24.0)
=============================================================================
sexoDenunciante                 FEM             FEM              FEM
idadeDenunciante             jovens          jovens          crianca
sexoDenunciado                  MAS             MAS              MAS
idadeDenunciado              jovens          jovens        adultos1
idBairro             intermediario1  intermediario1       periferico
idCrime                         ESTU            ESTU             ESTU
data                            meio             fim             meio
hora                          noite           noite            tarde
diaSemana                      util            util             util
-----------------------------------------------------------------------------
Clustered Instances

0       34 ( 59%)
1       24 ( 41%)
```

Figura 5. Resultado dos crimes contra a dignidade sexual

D. *Resultados dos crimes de tráfico de drogas*

A "Fig. 6" apresenta o último resultado com os crimes relacionados ao tráfico de drogas. O *cluster*0 apresenta 69% (83) instâncias de crimes de TD (tráfico de drogas), enquanto que o *cluster*1 informa 31% (37) instâncias de crime de UET (uso de entorpecente). Ambos os crimes são cometidos por jovens, do sexo masculino, em bairros intermediários2, em dias úteis. Porém, o crime de tráfico de drogas é denunciado por pessoas do sexo feminino de idade entre 35 a 64 anos (adultos2), no final do mês, à noite. Enquanto que o uso de entorpecente ocorre é denunciado por pessoas do sexo masculino, de idade entre 25 a 34 anos, no início do mês, à tarde.



```
Final cluster centroids:
                                        Cluster#
Attribute                Full Data            0            1
                           (120.0)         (83.0)       (37.0)
=============================================================
sexoDenunciante                FEM          FEM          MAS
idadeDenunciante          adultos2     adultos2     adultos1
sexoDenunciado                 MAS          MAS          MAS
idadeDenunciado             jovens       jovens       jovens
idBairro           intermediario2 intermediario2 intermediario2
idCrime                        TD           TD          UET
data                          fim          fim       inicio
hora                        noite        noite        tarde
diaSemana                   util         util         util
Clustered Instances

0       83 ( 69%)
1       37 ( 31%)
```

Figura 6. Resultados dos crimes de tráfico de drogas

Para todos os tipos de crimes separados na metodologia o algoritmo de agrupamento utilizado é o *Simple kmeans*, acompanhado da função de distância Euclidiana, ambos citados no referencial teórico. Embora utilizar apenas um algoritmo de agrupamento, a pesquisa em questão gerou resultados satisfatórios, já que a integração de novas técnicas como associação e classificação exigiria um estudo de caso e aprimoramento das ocorrências para um caráter mais complexo, fato este que pretende-se agregar mais a frente, como trabalhos futuros. Contudo, por meio dos resultados, pode-se demonstrar o potencial das aplicações de mineração de dados na obtenção do conhecimento em base de dados, proporcionando elementos capazes de formar diretrizes de prevenção de crimes futuros e melhor distribuição dos recursos destinados à segurança pública, tanto em termos de efetivo policial, quanto em orientações das vítimas, evidenciando possibilidades de prevenção e repressão a crimes com maiores indícios.

VI. Considerações Finais e Trabalhos Futuros

Nesta pesquisa é apresentada uma abordagem para a descoberta de padrões de conhecimento a partir de dados dos boletins de ocorrências do 3º Batalhão de Polícia Militar do Estado do Pará cadastrados na base de dados do OBCRIT – Observatório Criminal do Tapajós. A proposta apresentada teve o intuito de demonstrar a aplicação de algoritmo de agrupamento (*clusters*) realizada diretamente no banco de dados do sistema, visando benefícios dessas informações para conhecer as relações entre os entre as variáveis cadastradas no banco de dado do OBCRIT. No Brasil e em muitos países do mundo, observa-se um crescente avanço nos índices de criminalidade e da violência o que gera bastante preocupação por parte dos governantes. Por isso a análise dos resultados oriundos desta pesquisa aponta para uma importante forma de prevenção a possíveis outros delitos, pois a partir do momento em que se percebe, por meio da mineração de dados, que determinado crime ocorre em um bairro "x", em data, hora e dia da semana de forma frequente, ou seja, todos os meses, fica mais fácil criar um diagnóstico da criminalidade para subsidiar o planejamento de novas políticas públicas de combate à criminalidade, de forma específica para cada bairro, por meio de palestras, blitz educativas, criação de unidades de polícia pacificadora, rondas ostensivas, programas sociais, entre tantas outras possibilidades. Como trabalhos futuros, pretende-se incluir novos meses e anos na base de dados para serem utilizados novos algoritmos, podendo fornecer resultados superiores aos encontrados.

Agradecimentos

Um agradecimento especial primeiramente a Deus que nos ilumina em todo o decorrer dos trabalhos de pesquisa e elaboração de artigos. Aos professores que estão sempre disseminando o conhecimento com muita dedicação. A equipe do projeto OBCRIT, bolsistas e voluntários, que sedem horas de trabalhos para que o banco de dados esteja sempre atualizado. E a todos que direta ou indiretamente contribuem para o sucesso do projeto.

Referências Bibliográfica

[1] O.N.P Cardoso, R.T.M. Machado, "Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras", RevAdm Pública, pág. 42(3):495-528, 2008.

[2] Ministério da Justiça. "Secretaria Nacional de Segurança Pública. Curso Nacional de Multiplicador de Polícia Comunitária". 2nd ed. Brasília: SENASP, 2007.

[3] D. J. Hand, "Data Mining: Statistics and More?" The American Statistician, vol. 52, nº 2, maio, 1998.

[4] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, v.39, n. 11, p. 27– 34, 1996.

[5] I. H. Witten, and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, 1999.

[6] S. C. Côrtes, R.M. Porcaro, S. Lifschiz. "Mineração de Dados – Funcionalidades, Técnicas e Abordagens", PUC – Rio Inf. MCC 10/12, Maio, 2002.

[7] C. A. Diniz, F. N. Louzada, "Data Mining: uma introdução", São Paulo, ABE, 2000.

[8] JR. A. Johnson, D. W. Wichern, "Applied Multivariate Statistical Analysis", New Jersey, Prentice-Hall, Inc., 1982.

[9] W. O. Bussab, E. S. Miazaki, D. F. Andrade, "Introdução à Análise de Agrupamentos". 9rd Simpósio Nacional de Probabilidade e Estatística, São Paulo: ABE, 1990.

[10] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)", Morgan Kaufmann, 2000.

[11] Arff, "Attribute-relationfileformat", disponível em <http://www.cs.waikato.ac.nz/ml/weka/arff.html>, acesso em 20 de janeiro de 2015.

[12] L. A. V. Carvalho, "A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração", 1st edição, 2005.

[13] L. A. Silva Filho, F. H. M. Santos, "A Utilização de Sistemas de Informação no Apoio à Tomada de Decisão na Segurança Pública do Estado do Pará". Trabalho de Conclusão de Curso, Belém/PA (UFPA), 2007.

[14] L. M. Braz, R. Ferreira, D. Dermeval, D. Véras, M. Lima, W. Tiengo, "Aplicando Mineração de Dados para Apoiar a Tomada de Decisão na Segurança Pública do Estado de Alagoas", WCGE, Bento Gonçalves, Rio Grande do Sul, Julho, 2009.

[15] A. Puurula and S. Myaeng. "Integrated instance- and class-based generative modeling for text classification". In Proc 18th Australasian Document Computing Symposium, pages 66-73. ACM, 2013.