

# FERRAMENTA WEB DE SUPORTE À DECISÃO PARA GESTÃO EM SEGURANÇA PÚBLICA

Arlene Santos de França, João Gabriel R. de O. Lima, Ádamo L. de Santana, Carlos Renato Francês

Universidade Federal do Pará – Rua Augusto Corrêa, 01 – 66.075-110 – Belém – PA – Brasil  
{arlene.sf, jgabriel.ufpa}@gmail.com, {adamo, rfrances}@ufpa.br

**Resumo** – No setor de segurança pública, a capacidade de planejamento estratégico é um dos focos principais, no que diz respeito à busca de soluções viáveis no combate ao aumento da criminalidade e pontos críticos existentes, por este motivo a produção do conhecimento a partir de informações existentes depende do potencial do analista. Entretanto, a imensa quantidade de dados pode tornar o trabalho demorado e suscetível a falhas, por isso, o uso de tecnologias que auxiliem o gestor no processo de análise e tomada de decisão em bases de dados é de extrema importância. Com base no nisso, este artigo objetiva apresentar uma ferramenta para realização de análises probabilísticas que envolve o uso de métodos e modelos computacionais a fim de extrair conhecimento de grandes conjuntos de dados e apresentar resultados de forma a facilitar o processo decisório do gestor.

**Palavras-chave** – Redes Bayesianas, Algoritmos Genéticos, Mineração de Dados, Suporte à Decisão, Segurança Pública.

## 1 Introdução

A atual conjuntura da política governamental brasileira, face aos ingentes problemas sociais crônicos, tem compelido as autoridades governamentais a recorrer aos préstimos do pensamento técnico-científico desenvolvido pela Academia. É indubitável que essa tendência dos poderes governamentais em buscar parcerias com as Instituições de Ensino Superior, constitui uma providência estratégica, considerando a complexidade dos problemas enfrentados pelo país [1].

Em vários Estados, a matriz da violência é o tráfico de armas e drogas (o segundo financiando o primeiro e ambos induzindo à expansão e à intensificação da violência envolvida nas práticas criminais). Nas áreas pobres em que a comercialização de entorpecentes se instala, muitas crianças são recrutadas por traficantes e morrem em confrontos entre grupos rivais ou com policiais, nas incursões bélicas a que, quase sempre, se resumem as chamadas “políticas de segurança” [2]. Além disso, civis em meio ao “fogo cruzado” acabam tendo suas vidas roubadas em virtude dos altos índices de criminalidade.

Os órgãos de gestão da segurança pública brasileira não possuem ainda, em sua maioria, os mesmos recursos utilizados em outros países. Este artigo propõe um ambiente de análises das informações para usuários especialistas, correlacionando os dados oriundos de ocorrências policiais por meio do uso de métodos e modelos computacionais (redes bayesianas e algoritmos genéticos) para a análise de dados e indicadores provenientes do setor de segurança pública, tendo como base as informações históricas disponíveis, avaliando os aspectos de causalidade e permitindo a plotagem em visualizações gráficas, resultantes da realização de inferências e análises de cenários, permitindo ao gestor possuir uma visão diferenciada do ambiente de estudo.

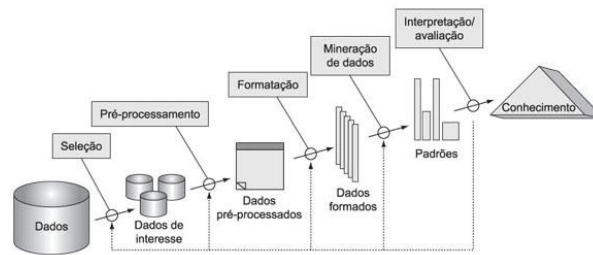
A integridade das soluções apresentadas pelo sistema depende, principalmente, das informações inseridas na ferramenta, portanto, os dados precisam ser trabalhados de forma a evitar incoerências que possam vir a gerar falhas no processo decisório. Para isso foram empregadas técnicas de descoberta de conhecimento em bases de dados (KDD – *Knowledge Discovery in Database*) que envolve fases de seleção, pré-processamento e mineração nos boletins de ocorrência provenientes do setor de segurança pública do Governo do Estado do Pará.

Este artigo encontra-se disposto da seguinte forma: A seção 2 trata da apresentação do domínio e o processo de extração do conhecimento. A seção 3 fala da otimização via banco de dados e da arquitetura da ferramenta. A seção 4 trata da análise de correlação. A seção 5 aborda a análise de cenários. A seção 6 fala sobre a apresentação de resultados em forma de visualização das informações e, por fim, a seção 7 apresenta as considerações finais.

## 2 Apresentação do Domínio e Etapas do Processo de Extração do Conhecimento

Os dados coletados e analisados neste estudo pertenciam ao órgão de gestão governamental do Estado do Pará, e foram obtidos a partir de boletins de ocorrência registrados nas delegacias da capital do Estado dos anos de 2002 a 2008. Essas informações estavam divididas em 27 variáveis, contabilizando 965.530 registros organizados em 73 planilhas, e apresentava grande quantidade de inconsistências como registros vazios ou irrelevantes para as análises, repetição de valores e notação incorreta das informações, como por exemplo, várias formas de escrita referenciando um mesmo logradouro.

O interesse em solucionar o problema de transformar dados em conhecimento, de forma que o processo não se utilize de métodos eminentemente manuais, tem fomentado várias pesquisas em um campo emergente chamado KDD [3], que desponta como uma tecnologia capaz de cooperar amplamente na busca do conhecimento embutido nos dados, sendo cada vez mais utilizado para auxiliar analistas no trabalho de investigação [4].



**Figura 1** – Etapas do processo KDD [3].

O processo KDD inicia com o entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados das quais se pretende extrair conhecimento. A partir dessas fontes é realizada uma seleção de dados, que serão pré-processados e submetidos aos métodos e ferramentas da etapa de *data mining* (DM) com o objetivo de encontrar padrões (conhecimento) a partir dos dados, envolvendo a criação de modelos apropriados de representação dos padrões e relações identificados a partir dos dados. O resultado desses modelos, depois de avaliados pelo analista, especialista e/ou usuário final, são empregados para prever os valores de atributos definidos pelo usuário final baseados em novos dados [3].

A qualidade do conjunto de dados está diretamente relacionada ao nível de ruído encontrado nos mesmos, que pode ser proveniente de dados desatualizados, erros de digitação ou transmissão de dados desprovidos dos atributos necessários para o reconhecimento de padrões ou contendo atributos irrelevantes à modelagem [5], o que faz com que seja necessário preparar os dados de modo que estejam adequados à aplicação da fase de DM, por meio de uma etapa conhecida como pré-processamento, e consiste no emprego de uma série de diferentes estratégias e técnicas, com a finalidade de transformar dados de entrada brutos em um formato adequado para posterior análise [6]. Esta fase consome cerca de 80% do esforço total de todo o processo de KDD [7].

As principais tarefas realizadas no pré-processamento, definidas neste estudo como sendo a junção das fases do KDD que vão da seleção dos atributos até a transformação dos dados são: (i) seleção dos atributos, que consiste em uma junção organizada de dados, buscando unir em apenas uma base, todas as informações que aparentam ser necessárias para a obtenção do conhecimento desejado [5], (ii) a limpeza da base, que se refere à garantia da qualidade (completude, veracidade e integridade) dos dados e pode ser obtida por meio de operações de padronização de dados, tratamento de valores ausentes, eliminação de dados errôneos e duplicatas [8], de forma que essas informações não comprometam a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD e (iii) a transformação de dados, a fim de evitar que alguns atributos, influenciem de forma tendenciosa adequando os dados à técnica de DM utilizada, podendo aplicar operações de normalização, discretização, conversão de valores simbólicos para valores numéricos e composição de atributos.

Tendo feito, as etapas descritas acima, iniciamos a próxima fase do processo de KDD, que é o processo de conversão de dados brutos em informações úteis (mineração de dados). Esta etapa consiste no processo de descoberta automática de informação útil em bases de dados extensas de modo a encontrar novos padrões úteis que poderiam permanecer desconhecidos [6], devido ao volume de dados. Existem cinco técnicas gerais de mineração de dados que englobam todas as outras formas de apresentação e permitem uma visão mais global e apropriada ao assunto. São elas a classificação, a estimativa, a previsão, a análise de afinidades e a análise de agrupamentos [9].

Vários algoritmos de aprendizado podem ser utilizados nas técnicas citadas anteriormente, dependendo da finalidade ao qual serão aplicados. Os algoritmos de interesse no contexto atual são as técnicas estatísticas por meio do uso de redes bayesianas e algoritmos genéticos.

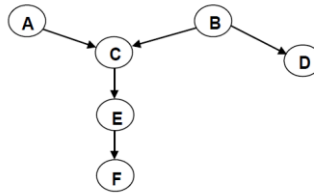
## 2.1 Redes Bayesianas

Uma rede Bayesiana, também chamada de rede de crença, rede probabilística ou rede causal, pode ser vista como um modelo que utiliza teoria dos grafos, condições de Markov e distribuição de probabilidades para representar uma situação, suas variáveis e estados e a partir disto realizar inferências [10].

O modelo bayesiano possui como componentes uma estrutura qualitativa, representando as dependências entre os nós, e quantitativa (tabelas de probabilidades condicionais - TPCs desses nós), avaliando, em termos probabilísticos, essas dependências [11]. Juntos, esses componentes propiciam uma representação eficiente da distribuição de probabilidade conjunta do conjunto de variáveis  $X_i$  de um determinado domínio [12].

A topologia da rede é formada por um conjunto de variáveis (nós) e arcos que ligam essas variáveis, formando um grafo dirigido acíclico em que cada nó possui uma distribuição condicional  $P(X_i|Pais(X_i))$  que quantifica o efeito dos pais sobre o nó. A figura 2 apresenta um esquema representativo em que os nós são definidos por  $A, B, C, D, E$  e  $F$ ; e os arcos direcionais representam a relação de causalidade entre as variáveis. Segundo [13] a grande vantagem no uso deste tipo de estrutura está em conseguir representar incerteza de forma gráfica através de nós e grafos.

Para a construção de uma rede bayesiana, são necessárias as dependências condicionais entre os atributos e suas TPCs. No caso das tabelas de probabilidade, são necessárias apenas as probabilidades a priori de ocorrência para os atributos que não possuem nenhum nó pai relacionado a ele (atributos  $A$  e  $B$ ), e para os que possuem, são necessárias as probabilidades do mesmo com relação aos seus respectivos pais (atributos  $C$ ,  $D$ ,  $E$  e  $F$ ), como pode ser visto no exemplo da figura a seguir.



**Figura 2** – Exemplo de uma rede Bayesiana [14].

A partir da RB gerada é possível realizar inferências a partir de cálculos probabilísticos por meio da utilização da fórmula de probabilidade condicional do teorema proposto por Bayes (1).

$$P(A = v_i | B) = \frac{P(B|A=v_i)P(A=v_i)}{\sum_{k=1}^n P(B|A=v_k)P(A=v_k)} \quad (1)$$

Uma vez que construímos a rede bayesiana, é possível realizar análises nas correlações entre os atributos da rede. Essa probabilidade *a posteriori* não é armazenada diretamente no modelo e, portanto, deve ser computada. De maneira geral, o cálculo de probabilidade de interesse, dado um modelo, é conhecido como inferência probabilística. É possível inferir sobre as dependências condicionais que se estabelecem entre as variáveis com base no gráfico da rede, bastando evidenciar a ocorrência de um determinado estado em uma ou mais variáveis da rede, propagando, dessa forma, o efeito das observações pela rede [14].

O processo de inferência bayesiana é a base do uso de RBs e consiste no processo de obtenção da probabilidade *a posteriori* a partir da probabilidade *a priori*, ou seja, extrair o conhecimento representado em uma rede já definida. O cálculo da probabilidade pode ser representado pelo produto das probabilidades dos nós ou, quando possuem pais, da sua probabilidade condicional. De maneira geral temos que (2):

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}, X_n = x_n) &= P(X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_2 = x_2, X_1 = x_1) = \\ P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_2 = x_2, X_1 = x_1) &P(X_{n-1} = x_{n-1}, \dots, X_2 = x_2, X_1 = x_1) = \\ \vdots & \\ \prod_{i=1}^n P(X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_1 = x_1) &= \prod_{i=1}^n P(X_i = x_i | Pais(X_i)) \end{aligned} \quad (2)$$

Segundo [15], a construção (ou aprendizado) das redes bayesianas pode ser feita usando apenas o conhecimento prévio do domínio, os dados ou a combinação de ambos. O aprendizado de estrutura é um importante problema a ser estudado pelo fato de o tamanho do espaço de busca de possíveis estruturas aumentar exponencialmente junto com o número de variáveis do modelo. Esse crescimento exponencial pode ser visualizado por (3) [16], que aponta o número de possíveis gráficos acíclicos dirigidos  $G$ , que podem ser gerados com um número  $n$  de variáveis:

$$G(n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} G(n-i) \quad (3)$$

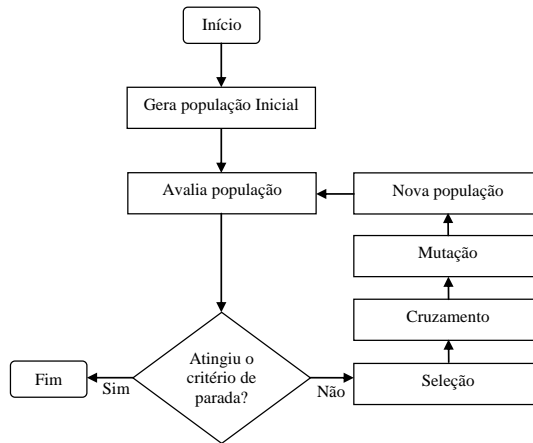
Para o aprendizado da estrutura da RB podemos utilizar o algoritmo de busca e pontuação K2, que procura entre as  $2^{n(n-1)/2}$  configurações possíveis a que maximiza a função de pontuação, sendo  $n$  o número de variáveis. Necessita, pois de um limite superior para o número de pais das variáveis e de uma ordenação prévia das mesmas para garantir a ausência de ciclos na estrutura da rede.

Três fatores têm motivado a utilização de RB no processo de *data mining* [17]: primeiro, a eficácia da manipulação de dados incompletos; em segundo lugar, a aprendizagem de relações causais entre as variáveis do domínio, o que facilita a sua análise; e, em terceiro lugar, as redes bayesianas permitem a combinação de conhecimento *a priori* do domínio com os dados disponíveis.

## 2.2 Algoritmos Genéticos

Algoritmos genéticos são algoritmos de busca baseados nos mecanismos de seleção natural e genética. Eles combinam a sobrevivência entre os melhores com uma forma estruturada de troca de informação genética entre dois indivíduos para formar uma estrutura heurística de busca [18].

As fases do processo evolutivo do algoritmo genético baseiam-se no mecanismo evolucionário proposto por Darwin, onde os indivíduos mais aptos apresentam maiores chances de sobrevivência e reprodução, perpetuando a espécie com características genéticas mais evoluídas.



**Figura 3** – Fluxograma clássico da execução de um algoritmo genético

De modo geral, podemos dizer que o AG inicia com a geração aleatória de uma população que passará por um processo de avaliação, caso a população gerada seja considerada a mais apta, por meio dos critérios estabelecidos para o algoritmo, finalizamos o processo, caso contrário a população é submetida às fases do processo evolutivo: seleção, cruzamento, mutação e geração de uma nova população a ser avaliada. Esse procedimento é executado até que o critério de parada seja atingido conforme pode ser observado no fluxograma apresentado na figura anterior.

### 3 Otimização via Banco de Dados e Arquitetura da Ferramenta

O Banco de Dados (BD) foi projetado de forma a modelar da maneira mais fiel possível um boletim de ocorrência generalizado, organizando os dados de maneira inteligente com o objetivo dar o melhor suporte para a aplicação. Deste modo, o BD apresenta as seguintes entidades (tabelas de dados): (i) Boletim de Ocorrência, (ii) Bairro, (iii) logradouro, (iv) local do crime (via pública, residência, bares etc.) e (v) crime.

Levando em consideração o prospecto do uso de RB, o modelo foi desenvolvido a fim de reduzir o custo computacional, decorrente da etapa de DM, de modo a contribuir significativamente para o melhor desempenho dos algoritmos de Inteligência Computacional no domínio estudado, que por sua vez apresenta variadas combinações entre os atributos, o grande número de registros e de espaço combinatório, garantindo um significativo aumento no desempenho dos mesmos.

A ferramenta proposta baseia-se no conceito de WEB 2.0 e foi desenvolvida utilizando modernas técnicas de programação e arquitetura de software, categorizando-se como um RIA (*Rich Internet Application*), que se destaca pela grande quantidade de recursos visuais permitindo visualizar informações de maneira diferenciada como a utilização de gráficos e mapas, que foram recursos explorados nas análises obtidas pelo sistema tendo em vista as informações disponíveis no banco de dados, como informações geográficas (local de ocorrência de um crime). Além disso, é importante frisar que todo o processamento do sistema é feito de maneira assíncrona, o que faz com que as requisições feitas pelo sistema sejam otimizadas.

### 4 Análise de Correlação

Para o aprendizado das relações existentes e aprendizado do modelo gráfico da rede bayesiana utilizou-se o algoritmo de busca e pontuação K2 [19], que permite encontrar a mais provável estrutura de rede de crença  $B$ 's a partir de um determinado conjunto de dados  $D$ . O algoritmo K2 aplica a pontuação bayesiana segundo (4).

$$P(B_s|D) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(r_i)}{\Gamma(r_i+N_{ij})} \prod_{k=1}^{r_i} \Gamma(N_{ijk} + 1) \quad (4)$$

onde,  $n$  é o número de nós;  $q_i$  é o número de configurações dos pais da variável  $X_i$ ;  $r_i$  é o número de possíveis valores do nó  $X_i$ ;  $N_{ijk}$  é o número de casos em  $D$  onde o atributo  $X_i$  é instanciado com o seu valor  $k$ , e a configuração dos pais de  $X_i$  é instanciada com o valor  $j$ ; e,  $N_{ij}$  denota o número de observações em que a configuração dos pais de  $X_i$  é instanciada com o valor  $j$ , sendo  $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ .

O algoritmo utilizado apresenta algumas particularidades, dentre elas, a exigência de que o especialista informe a ordem total das variáveis, ponto crucial para o aprendizado adequado da estrutura da RB, e faz com que o algoritmo evite circularidade na rede ao inferir a orientação dos arcos. Neste contexto, foram feitos os estudos nas variáveis do domínio a fim de obter a ordem ideal por índice de relevância (do maior para o menor), para que o algoritmo pudesse ser executado.

O diferencial proposto neste trabalho, diz respeito também à otimização do processo de aprendizagem da estrutura da RB, diminuindo o custo computacional decorrente da grande quantidade de cálculos matemáticos a serem realizados a partir de sua equação (4), que consiste na busca pelas frequências das correlações entre os estados dos atributos do domínio,

representado por  $N_{ij}$  e  $N_{ijk}$ , onde:  $i$  corresponde à variável estudada;  $j$  é a variável de correlação; e,  $k$  são os estados dos atributos que serão pontuados. Esse mecanismo percorre todas as combinações dos registros dos atributos na base de dados, a fim de obter as frequências correspondentes de cada relação para cada estado de cada atributo da RB, o que torna inviável a utilização da busca pontual para a base de dados extremamente grandes, como a utilizada para a realização deste trabalho, pois exigiriam um grande esforço computacional.

A solução encontrada foi aderir aos recursos de banco de dados, reduzindo assim o tempo de execução e processamento do algoritmo, de maneira que, ao invés de buscar registro por registro na base e dados obtemos apenas o resultado final de cada  $N_{ijk}$  através de comandos executados pela ferramenta no banco de dados, o que contribui para melhorias no processo de aprendizado e inferência da RB. Para exemplificar, o que acontece no processo de inferência bayesiana, utilizamos o modelo topológico apresentado na figura 2 para o cálculo de (5), onde a inferência é realizada por meio da aplicação do teorema de Bayes (1).

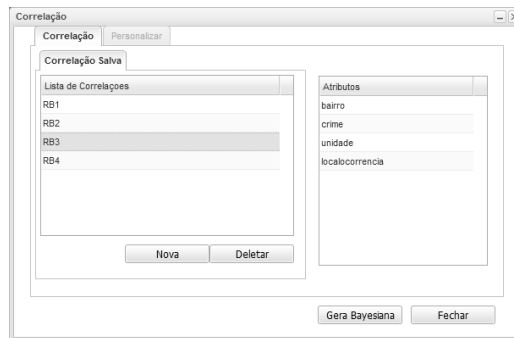
$$P(E|\bar{C}, D, F) = \frac{P(E\bar{C}\bar{D}\bar{F})}{P(E\bar{C}\bar{D}\bar{F}) + P(\bar{E}\bar{C}\bar{D}\bar{F})} \quad (5)$$

Para a execução da inferência supracitada, dado que existem incertezas quanto aos valores das variáveis  $A$  e  $B$ , faz-se necessário expandir o cálculo da probabilidade agregando todos os estados da RB, ressaltando que, quanto maior o número de variáveis não especificadas (incerteza), maior o número de cálculos envolvido no processamento (6).

$$\begin{aligned} P(F, \bar{E}, D, C, \bar{B}, A) &= P(F|\bar{E} \wedge D \wedge C \wedge \bar{B} \wedge A)P(\bar{E} \wedge D \wedge C \wedge \bar{B} \wedge A) \\ &\vdots \\ &= P(F|\bar{E})P(\bar{E}|C)P(D|\bar{B})P(C|\bar{B} \wedge A)P(\bar{B})P(A) \end{aligned} \quad (6)$$

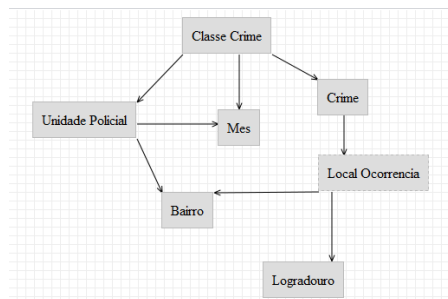
Com o uso de BD, é possível reduzir a complexidade e aumentar o desempenho com relação ao cálculo da inferência, pois não precisamos realizar todos os passos descritos em (6), bastando apenas buscarmos diretamente no banco o valor da interseção entre as variáveis de interesse  $P(F, \bar{E}, D, C, \bar{B}, A)$ .

No software de Segurança Pública, para realizar a correlação, os dados são extraídos da base e separados de acordo com sua categoria, dando ao usuário possibilidade de criar vários cenários de análise para a aplicação dos modelos computacionais de acordo com a análise desejada.



**Figura 4** – Janela de Seleção da Correlação

O resultado do estudo realizado apresenta a rede gerada a partir da base de dados, assim como uma tabela de probabilidades onde o especialista poderá realizar inferências sobre o comportamento dos atributos.



**Figura 5** – Rede Bayesiana gerada

Foram feitos testes comparativos da ferramenta proposta com as ferramentas WEKA e Genie 2.0 a fim de medir o desempenho e integridade das respostas obtidas, ou seja, o tempo para gerar uma rede bayesiana e se a rede gerada está coerente com o domínio estudado. Para a realização destes testes, foi utilizada uma amostra da base de dados contendo 158.361 registros e quatro variáveis (bairro, crime, local de ocorrência e unidade policial). Os resultados se apresentam conforme a Tabela 1:

**Tabela 1** – Análise de desempenho da ferramenta proposta.

Ferramenta	Resultado da Análise	Tempo de Execução (segundos)
WEKA	O formato de arquivo utilizado pela ferramenta não suportou a quantidade de dados utilizada.	–
Genie	A rede gerada não apresentou resultados satisfatórios, pois não estava de acordo com o conhecimento prévio tido do domínio, apresentando inconsistências na correlação das variáveis.	~9.56
Ferramenta Proposta	A rede apresentou resultados satisfatórios, de acordo com o conhecimento prévio a respeito do domínio.	~7.89

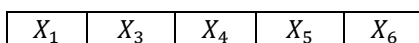
Não foram realizados testes comparativos a respeito do tempo de execução da inferência, pois cada uma das ferramentas, citadas na Tabela 1, utiliza um algoritmo de inferência diferente para a obtenção dos resultados probabilísticos.

## 5 Análise de Cenários

Neste trabalho, para a análise dos cenários, foi utilizado um modelo híbrido que consiste na união de AG e RB, proposto por [20], combinando as qualidades de algoritmos evolutivos para a pesquisa combinatória com o modelo probabilístico de inferência bayesiana, com o objetivo de identificar a melhor configuração, entre os possíveis valores (estados de nós em um RB) das variáveis do domínio, para uma ou mais variáveis no domínio em questão.

Diferentemente da forma como os AGs são usados na maioria dos sistemas híbridos propostos na literatura, onde são adotados para otimizar o processo de aprendizagem da estrutura da RB, aqui, a técnica é usada para a descoberta dos valores mais prováveis das variáveis de uma RB para a maximização do valor de um atributo de chave. A interação entre essas duas técnicas de inteligência computacional ocorre da seguinte maneira: o processo de descoberta cenário começa com o fornecimento da RB gerada a partir dos dados pré-processados e seus parâmetros, então, um AG é aplicado utilizando como função de *fitness* para os indivíduos (cenários) o valor da inferência gerada a partir da sua configuração, ao final do processo é apresentado o cenário ideal para alcançar uma meta particular.

Os cromossomos utilizados foram configurados na forma de um vetor de dados, onde a posição do vetor representava o atributo e o valor do vetor, o índice do valor candidato, por exemplo, para uma rede  $n$  atributos, o tamanho do cromossomo é da ordem de  $(n - 1)$ , ou seja, se na RB temos seis atributos, os cromossomos das populações terão cinco genes, pois retiramos do cromossomo a variável que foi inferenciada.



**Figura 6** – Exemplo de um cromossomo com a variável  $X_2$  inferenciada.

onde,  $X_1, X_3, X_4, X_5, X_6 \rightarrow$  são os índices dos valores candidatos para os atributos 1, 3, 4, 5, 6 respectivamente.

O algoritmo segue os procedimentos apresentados a seguir.

<ol style="list-style-type: none"> <li>1. DESCOBERTA_CENÁRIO (<i>rb</i>)</li> <li>2. /* retorna o cenário que melhor contribui para obtenção do valor meta para uma variável do domínio de aplicação */</li> <li>3. // argumentos: <i>rb</i> //rede bayesiana que codifica a distribuição conjunta <math>P(X_1, X_2, \dots, X_n)</math></li> <li>4. <i>população</i> <math>\leftarrow</math> POPULAÇÃO_INICIAL_ALEATÓRIA;</li> <li>5. repita</li> <li>6.   <i>nova_população</i> <math>\leftarrow</math> conjunto vazio</li> <li>7.   para <math>i \leftarrow 1</math> até TAMANHO(<i>população</i>) faça</li> <li>8.     <i>a</i> <math>\leftarrow</math> SELEÇÃO(<i>população</i>, FUNÇÃO_APTIDÃO(MÉTODO_INFERÊNCIA_I(<i>rb</i>)))</li> <li>9.     <i>b</i> <math>\leftarrow</math> SELEÇÃO(<i>população</i>, FUNÇÃO_APTIDÃO(MÉTODO_INFERÊNCIA_I(<i>rb</i>)))</li> <li>10.    se (TAXA_CROSSOVER é atendida)</li> <li>11.     <i>filho_ab</i> <math>\leftarrow</math> CROSSOVER(<i>a</i>,<i>b</i>)</li> <li>12.    se (TAXA_MUTAÇÃO é atendida)</li> <li>13.     <i>filho_ab</i> <math>\leftarrow</math> MUTAÇÃO(<i>filho_ab</i>)</li> <li>14.    incluir <i>filho_ab</i> em <i>nova_população</i></li> <li>15.    fim para</li> <li>16.   <i>população</i> <math>\leftarrow</math> <i>nova_população</i>;</li> <li>17. até algum indivíduo (cenário) está adaptado o suficiente ou até ser obtido um número <math>n</math> de gerações</li> <li>18. retornar o melhor cenário em <i>população</i>, de acordo com FUNÇÃO_APTIDÃO</li> </ol>
---

**Figura 7** – Algoritmo do processo de descoberta de cenários [20].

Na ferramenta desenvolvida neste trabalho, a primeira interação realizada para a análise de cenários é com a interface gráfica, onde o usuário irá escolher o atributo chave e seu valor, que será processado pelo algoritmo e retornará o cenário onde o valor do atributo escolhido estará maximizado, mudando também, a configuração das outras variáveis da rede bayesiana.

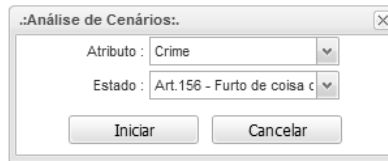


Figura 8 – Interface de seleção para a análise de cenários.

O resultado desse processo é a reestruturação das probabilidades *a posteriori* dos atributos da RB, para apresentar o melhor cenário à variável escolhida. Por exemplo, tendo uma RB composta de quatro variáveis (Bairro, Crime, Unidade Policial, Local Ocorrência) ao selecionarmos o atributo “Crime” em seu estado “Art. 156 – Furto de coisa comum”, conforme apresentado na interface de seleção para a análise de cenários, fazemos com que o algoritmo híbrido utilizado (AG+RB) retorne o melhor cenário probabilístico, considerando cada atributo da rede, de maneira que o atributo escolhido para análise de cenário seja maximizado no estado selecionado, ou seja, apresente o cenário ideal para este atributo. O resultado gerado por esse mecanismo apresenta os estados dos outros atributos da RB (Bairro, Unidade Policial, Local Ocorrência), que fazem com que o atributo escolhido tenha seu valor maximizado para o estado desejado.

## 6 Apresentação de Resultados

O processo de visualização e representação de informações serve como base de análise dos resultados obtidos, pois apresenta ao usuário especialista uma nova perspectiva a respeito dos mesmos. Nesse contexto, os dados contidos nos boletins de ocorrência são apresentados através de gráficos e referências geográficas como descritos a seguir.

A utilização do gráfico permite transmitir de forma simples e eficiente as informações contidas na base de dados, possibilitando ao usuário especialista visualizar de forma mais abrangente o cenário de estudo, por informar de forma satisfatória a contribuição percentual de cada valor em um total.

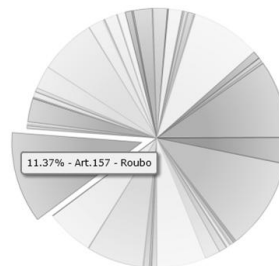


Figura 9 – Visualização de Informações no Gráfico (Relação Bairro/Crime).

Além disso, através da integração de um Sistema de Informações Geográficas (SIG) à ferramenta apresentada neste artigo, foi possível transformar os valores de endereço existentes em coordenadas conhecidas através da utilização de marcadores geográficos (*Place Marks*), os quais permitem georreferenciar, de forma aproximada, as localidades das ocorrências policiais. Segundo [21], o termo SIG é usado atualmente de forma genérica para sistemas computacionais que tenham capacidade de manipulação e geração de dados geográficos, através do cruzamento e análise de informações, com o objetivo de viabilizar a tomada de decisão por parte do usuário. A aplicação do SIG entra como uma expansão da aplicabilidade de redes bayesianas sobre o domínio, permitindo, dada a existência de uma variável que denote localização, a plotagem do resultado das inferências geograficamente, facilitando a análise e gestão das informações.

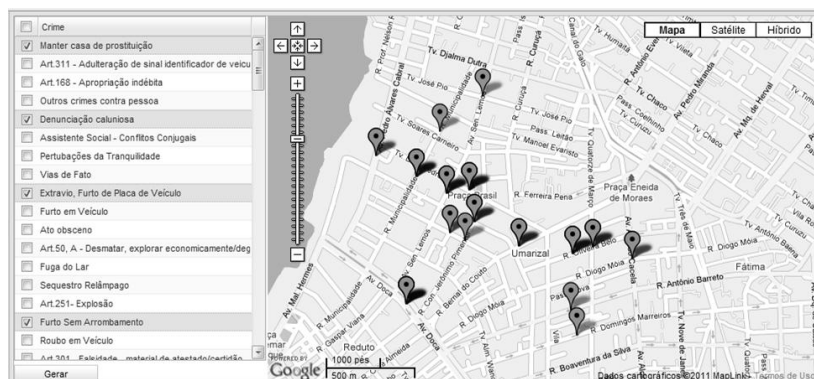


Figura 10 – Representação de Informações no Mapa

## 7 Conclusão

Sistemas de gestão capazes de processar grandes quantidades de dados em tempo real propiciam consultas eficientes e atualizadas sobre a situação das incidências estatísticas do domínio, constituindo-se ferramentas de grande utilidade no apoio ao planejamento e distribuição dos recursos públicos. Os órgãos de gestão da segurança pública brasileira não possuem ainda, em sua maioria, os mesmos recursos e instrumentos da Tecnologia do Conhecimento (métodos) e da Informação (técnicas e tecnologias) utilizados em outros países. Tais recursos, como os apresentados neste trabalho, vêm a contribuir possibilitando maior controle e compreensão, de forma eficiente, dos fenômenos políticos, econômicos e sociais, que têm influência no processo de tomada de decisão pelos órgãos gestores.

O trabalho apresentado, além de trazer a utilização de novas técnicas e recursos disponíveis, inova tecnicamente ao trazer uma nova abordagem de trabalho em análise de cenários com redes bayesianas e algoritmos genéticos em bases de dados extensas, explorando a estrutura de um banco de dados SQL, que contribui de maneira significativa para a melhoria do desempenho das análises combinatórias, considerando o enorme volume de informações e a complexidade dos algoritmos, além de permitir uma melhora na eficiência do processo decisório do gestor, pois conta com recursos que permitem elevar o nível de abstração dos dados outrora dispostos em forma de planilhas contendo um imenso volume de informações e, por isso, inviáveis de serem analisados de forma manual.

Ressalta-se ainda que, embora o sistema tenha sido desenvolvido inicialmente com o intuito de trabalhar com dados de segurança pública, pode ser utilizado nos outros setores da gestão governamental. Um dos objetos de trabalhos futuros é incluir módulos ao sistema que objetivam prover aos órgãos gestores, uma estrutura capaz de expandir sua utilização acrescentando informações de natureza social, econômica etc., de forma a enriquecer ainda mais as análises.

## 8 Referências

- [1] C. A. Ayres, Segurança Pública e Educação: Parceria Intersetorial e Ciclos de Formação Continuada, **Conferência internacional “Educação, Globalização e Cidadania: Novas Perspectivas da Sociologia da Educação”**, UFPI, (2008).
- [2] L. E. Soares, Novas Políticas de Segurança Pública: alguns exemplos recentes. **Jus Navigandi**, Teresina, ano 7, n. 65(2003).
- [3] Fayyad, U., Piatetsky-Shapiro, G. e Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communication of the ACM**, 11(1996), 27-34.
- [4] E. Turban, J. E. Aronson, Decision Support Systems and Intelligent Systems, 6ª Edição, **Prentice-Hall**, (2001).
- [5] Carvalho, D. R., Bueno, M., Neto, W. A. e Lopes, L. R. Ferramenta de Pré e Pós-processamento para Data Mining. **XII Seminco, PROEB**, Blumenau/SC, (2003).
- [6] P. Tan, M. Steinbach e V. Kumar. Introduction to Data Mining. **Boston, Pearson Education, Inc. /Addison Wesley**, (2006).
- [7] M. Klemettinen, H. Manilla, P. Ronkainen, H. Toivonen, A. I. Verkamo, Finding Interesting Rules from Large Sets of Discovered Association Rules, **Proc. of the Third Int’l Conf. on Information and Knowledge Management, Maryland**, (1994).
- [8] R. C. D. NEVES, Pré-processamento no Processo de Descoberta de Conhecimento em Banco de Dados. **PPGC-UFRGS**, (2003).
- [9] L. A. V. Carvalho, Data Mining: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. **Rio de Janeiro: Ciência Moderna**, (2005).
- [10] A. R. Gonçalves, André Ricardo. Fundamentos e aplicações de técnicas de aprendizagem de máquina, **TCC(Graduação) – UEL**, (2008).
- [11] Z. Chen, Data Mining and Uncertain Reasoning - an Integrated Approach, **John Wiley Professional**, (2001).
- [12] J. Pearl, Probabilistic Reasoning in Intelligent System, **Morgan Kaufmann Publishers**, (1988).
- [13] E. Charniak, Bayesian networks without tears: making Bayesian networks more accessible to the probabilistically unsophisticated, **AI Magazine**, 12 (1991) 50-60.
- [14] A. L. Santana, Estratégias para a melhoria da modelagem e interpretabilidade de Redes Bayesianas. **Tese (Doutorado) – UFPA**, (2008).
- [15] A. L. Santana, C. R. L. Francês, P. Tarso, C. W. A. Costa, P. T. Endo, A. B. R. KLAUTAU, Um Retrato da Aplicação de Recursos da Saúde e seu Impacto no IDH dos Municípios do Estado do Pará. **XXXI SEMISH**, (2004).
- [16] R. W. Robinson, Counting unlabeled acyclic digraphs, **Proceedings of the Fifth Australian Conference on Combinatorial Mathematics**, (1976), 28-4.
- [17] S. Russell, P. Norvig, Artificial Intelligence – A Modern Approach. 2ª Edição, **New Jersey, Prentice Hall**, (2003).
- [18] LINDEN, R. Algoritmos Genéticos: Uma importante ferramenta da Inteligência Computacional. 2ª Edição, **Brasport, Rio de Janeiro**, (2008).
- [19] D. Heckerman, Bayesian networks for Data Mining, **Journal of Data Mining and Knowledge Discovery, Kluwer Academic Publishers**, 1 (1997), 79-119.
- [20] C. A. J. Rocha, Estratégia de Otimização para a Melhoria da Interpretabilidade de Redes Bayesianas: Aplicações em Sistemas Elétricos de Potência. **Tese (Doutorado) – UFPA**, (2009).
- [21] T. Bernhardsen, Geographic Information Systems: an introduction. **John Wiley & Sons**, (1999) 372 p.