

Tratamento de dados, vieses e privacidade

Aline C. C. S. Azevedo
Instituto de Computação
UNICAMP
Campinas, Brasil
a189593@dac.unicamp.br

Beatriz C. Nascimento
Instituto de Computação
UNICAMP
Campinas, Brasil
b247403@dac.unicamp.br

Daniel G. Gratti
Instituto de Computação
UNICAMP
Campinas, Brasil
d214729@dac.unicamp.br

I. INTRODUÇÃO

No cenário atual, a avaliação e classificação de áreas com índices de criminalidade elevados têm emergido como um desafio crítico e complexo. A aplicação de técnicas de aprendizado de máquina permite auxiliar na análise desses índices através de uma vasta quantidade de dados. A capacidade de identificar e compreender essas regiões desempenha um papel fundamental para a formulação de estratégias de segurança pública. No entanto, é necessário também considerar fatores sociais para que não contribua ainda mais com vieses discriminatórios.

Apresentamos uma proposta que visa aprofundar a análise dos índices de criminalidade nas diferentes regiões de São Paulo, adotando uma abordagem ética e enfatizando a compreensão transparente do modelo. Nosso objetivo é identificar quais fatores socioeconômicos exercem maior influência na categorização do nível de segurança de uma área específica. Nesse sentido, nosso estudo se propõe a desenvolver um sistema para estimar o índice de criminalidade em regiões específicas e, concomitantemente, investigar o perfil socioeconômico das áreas mais impactadas pela violência urbana.

Para este trabalho utilizaremos duas bases de dados da cidade de São Paulo, uma de dados criminais providos pelo Núcleo de Estudos da Violência, da USP, de boletins de ocorrência coletados na cidade de São Paulo, entre 2006 e 2017, dos crimes de roubo e furto. A segunda base de dados consiste de 21 características socioeconômicas e de infraestrutura urbana, que caracterizam o perfil social de uma dada região, principalmente obtidos através do censo brasileiro realizado pelo Instituto Brasileiro de Geografia. Nossa proposta é verificar se os dados socioeconômicos de uma região são suficientes para a previsão do padrão de crime observado pela frequência de crimes, aproximada pelos boletins de ocorrência dados nas regiões observadas.

II. TRABALHOS RELACIONADOS

Fizemos uma revisão bibliográfica com estudos que utilizam aprendizado de máquina no contexto de criminalidade, com o foco em 4 artigos que discutimos a seguir, que possuem bases de dados similares à do nosso trabalho.

O artigo *CrimAnalyzer: Understanding Crime Patterns in São Paulo* [1] tem objetivo de apresentar uma ferramenta de análise visual, *CrimAnalyzer*, que permite aos usuários estudar o comportamento dos crimes em regiões específicas

da cidade de São Paulo - Brasil. O sistema ajuda a identificar áreas focais e os padrões de crimes associados, mostrando como esses padrões mudam ao longo do tempo. O *CrimAnalyzer* foi desenvolvido para auxiliar especialistas em criminologia e enfrenta três desafios principais: flexibilidade para explorar regiões locais e compreender os padrões de crimes, identificação de áreas focais espaciais de crimes importantes para investigação e compreensão da dinâmica dos padrões de crimes ao longo do tempo.

Outro artigo relacionado a estudo sobre análise de crimes na cidade de São Paulo é o *CriPAV: Street-Level Crime Patterns Analysis and Visualization* [2], onde é apresentada uma nova metodologia para lidar com análise de padrões de crime espaço-temporais em níveis detalhados de rua. A abordagem proposta consiste em dois componentes principais: um mecanismo estocástico para identificar concentrações prováveis de crimes e um mecanismo de aprendizado profundo para mapear séries temporais de crimes em um espaço cartesiano, permitindo a identificação de locais com comportamento semelhante. Esses componentes são integrados na ferramenta *CriPAV*, uma plataforma *web* de análise e visualização de padrões de crimes. *CriPAV* foi desenvolvida em colaboração com especialistas da área e validada por meio de estudos de caso com dados reais de crimes em São Paulo. Os experimentos demonstram que o *CriPAV* é eficaz em identificar padrões, incluindo locais de crimes pouco intensos, mas altamente prováveis, e locais distantes com padrões de crimes similares.

O estudo *Exploring counterfactual antecedents to reduce criminality in Rio de Janeiro* [3] concentrou-se na análise do impacto de variáveis socioeconômicas e urbanas nas taxas de criminalidade na cidade do Rio de Janeiro. Utilizando dados por bairro, foram consideradas as taxas per capita de três categorias de crimes (contra pedestres, estabelecimentos comerciais e veículos), além de variáveis socioeconômicas e urbanas. A análise identificou áreas com maiores índices de cada tipo de crime e realizou cenários contrafactuais para locais específicos. Os resultados mostraram que tipos distintos de crimes ocorrem em diferentes partes da cidade. As análises contrafactuais revelaram variações significativas dependendo do bairro.

O estudo *Crime Analysis Through Machine Learning* [4] aborda a previsão de crimes usando aprendizado de máquina para análise de dados criminais de Vancouver ao longo de 15 anos com duas abordagens de processamento de dados

distintas. Foram empregados modelos preditivos, incluindo o método dos k-vizinhos mais próximos e árvores de decisão impulsionadas. Os resultados demonstram que a previsão de crimes em Vancouver alcançou entre 39% e 44% de acurácia por meio desses modelos.

O artigo da ferramenta *CrimAnalyzer* [1] é o mais semelhante da proposta do nosso projeto, que também será focada na cidade de São Paulo. No entanto, trataremos de uma abordagem voltada para explicabilidade do modelo de aprendizado de máquina para possibilitar o entendimento e gerar possíveis soluções para problemáticas éticas que possam surgir.

III. DESCRIÇÃO DO TRATAMENTO DOS DADOS

Para transformar os dados no espaço encontrado até o espaço dos dados preparados, o *dataset* foi tratado anteriormente utilizando proporções e categorizações, principalmente para cobrir casos em que o número absoluto não é representativo ou é ambíguo. Os atributos Áreas de risco, Densidade Demográfica, Expansão Urbana e Homicídio são divididos em categorias ordinais de probabilidade de acidente, quartis, período de urbanização e taxas de homicídio, respectivamente. Outros atributos foram colocados em forma de porcentagem em relação ao todo, como é o caso dos atributos ligados a população e domicílios.

Uma última preparação está na criação de atributos de mobilidade urbana, número de viagens e tempo médio de deslocamento, que foram construídos a partir de dados de pesquisa de Origem - Destino, realizado pelo Metrô/SP em 2007 e dados censitários do IBGE. Estes indicadores foram incluídos como um *proxy* para medir as condições de mobilidade urbana dos setores.

É importante, também, explorarmos as estratégias para lidar com os diversos tipos de vieses presentes nesses dados. Embora existam várias formas de vieses, nossa análise se concentrará nos cinco que foram discutidos em sala de aula, e analisaremos de forma geral para a nossa base de dados e não atributo a atributo.

- *Viés social*: Como nossa base de dados abrange as diferentes regiões geográficas da cidade de São Paulo, é possível identificar a presença de vieses sociais decorrentes de um histórico de subinvestimento e da falta de oportunidades enfrentadas por grupos sociais marginalizados. Estes vieses inclusos nos dados são o objeto de estudo deste trabalho. Utilizaremos técnicas de *fairness* em aprendizado de máquina para identificar possíveis desigualdades refletidas nos atributos e, portanto, os vieses inclusos em cada um deles.
- *Viés de representação*: Em nosso *dataset* existe um viés de representação pois existem poucos dados sobre a criminalidade da maioria dos bairros analisados, enquanto possuímos muitas informações sobre a criminalidade em alguns bairros específicos. Por conta da grande disparidade de dados criminais entre setores, é possível que a correlação entre frequência de crimes e os dados socioeconômicos do respectivo setor não seja representativa

para bairros com pouca incidência de crimes, enquanto bairros com maior incidência, apesar de serem bem representativos, correspondem a minoria dos setores. Para balancear as amostras utilizaremos métricas balanceadas no número de amostras de setores enquanto tentamos melhor caracterizar regiões com menos incidência de crimes.

- *Viés temporal*: Nossos dados sofrem de um viés temporal quanto a diferentes formas de coleta. Alguns atributos foram coletados em frequências e intervalos diferentes, como é o caso dos indicadores sociais do IBGE, que são coletados de 10 em 10 anos, entre 1980 e 2010, exceto por alguns atributos, como chefe de família de alta renda, cujas coletas iniciaram-se em 1991 e dados de mobilidade urbana, que são apenas de 2007. Para contornar problemas relacionados a diferentes distribuições conforme os anos, utilizamos os atributos temporais como características históricas de nossas amostras regionais e tentamos correlacioná-las apenas para o período de 2006 a 2017 relacionado a crimes neste período. Deste modo, os atributos, ainda que apresentem alta correlação interna, são importantes para caracterizar a forma de crescimento socioeconômico e geográfico dos setores e, portanto, devem ser cuidadosamente tratados para manter sua função temporal nas amostras.
- *Viés de preparação de dados*: Um atributo incluso no *dataset* compõe um *proxy* para uma característica de deslocamento urbano que foi gerado através de amostragem de outros indicadores e de um estimador *Bayesiano* por setor. Este procedimento não pôde ser testado para verificação de possíveis vieses injetados em sua inclusão. Desta forma, o atributo de deslocamento médio pode conter vieses de preparação. As demais informações do *dataset* foram coletadas de diversas fontes que seguem um código de ética para coleta e utilização de dados e, portanto, consideramos que não há vieses de preparação presentes neles.
- *Envenenamento de dados*: Não foi encontrado nenhum envenenamento de dados, a princípio, na nossa base de dados.

IV. DESCRIÇÃO DO USO DOS DADOS

Utilizaremos os dados coletados para a confecção de um modelo de aprendizado de máquina que, através dos dados de censo, seja capaz de prever o perfil de criminalidade local. Este perfil será extraído por outro modelo matemático, por meio da base dados de crime. Uma abordagem possível de ser feita é através da extração de características dos relatos e boletins de ocorrência, por exemplo, pela determinação de parâmetros da distribuição de probabilidade do tempo entre ocorrências num setor, ou de forma automática, por outro modelo de aprendizado de máquina.

Com relação às questões de privacidade, nosso *dataset* segue a política de sigilo estatístico, que deve fornecer instruções e orientações sobre a preservação do sigilo estatístico e o risco de revelação de dados individuais nos processos de produção

e disseminação de informações estatísticas, estabelecendo, ainda, os procedimentos a serem adotados para garantia do sigilo estatístico e a não identificação de informantes, como, por exemplo, desidentificação de tabelas, anonimização de bases de dados, entre outros ¹. Portanto, não são necessários outros procedimentos para garantir a privacidade das pessoas cujos dados foram coletados para essa base.

REFERÊNCIAS

- [1] Germain Garcia, Jaqueline Silveira, Jorge Poco, Afonso Paiva, Marcelo Batista Nery, Claudio T Silva, Sergio Adorno, and Luis Gustavo Nonato. Crimanalyzer: Understanding crime patterns in sao paulo. *IEEE transactions on visualization and computer graphics*, 27(4):2313–2328, 2019.
- [2] Germain Garcia-Zanabria, Marcos M Raimundo, Jorge Poco, Marcelo Batista Nery, Cláudio T Silva, Sergio Adorno, and Luis Gustavo Nonato. Cripav: Street-level crime patterns analysis and visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4000–4015, 2021.
- [3] Vitória Aquino Guardieiro. Exploring counterfactual antecedents to reduce criminality in rio de janeiro. Technical report, 2021.
- [4] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime analysis through machine learning. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 415–420. IEEE, 2018.

¹https://ftp.ibge.gov.br/Informacoes_Gerais_e_Referencia/Codigo_de_Boas_Praticas_das_Estatisticas_do_IBGE_2a_edicao.pdf