Previsão de crimes em Valência: Uma abordagem multi-rótulo

Crime prediction in Valencia: A multi-label approach

Luís Eduardo Freire da Câmara[1], Alexandre Rodrigues Loureiros[2], Jorge Mateu Mahiques[3], Flávio Miguel Varejão[4]

(1) UFES, Brasil, Espírito Santo, luis.camara@edu.ufes.br.
(2) UFES, Brasil, Espirito Santo, arodrigues.ufes@gmail.com.
(3) Universitat Jaume I, Espanha, Castellón, mateu@mat.uji.es.
(4) UFES, Brasil, Espírito Santo, fmvarejao@gmail.com.

Resumo:
O aumento da criminalidade em áreas com alta concentração populacional torna as cidades uma das principais fontes de violência. Compreender as características de ocorrências criminosas e sua relação com a escala urbana transcende as questões acadêmicas e se torna central para a sociedade atual. Destarte, propomos uma abordagem de classificação multirrótulo para prever a ocorrência de crimes nas ruas da cidade de Valência, Espanha. Para tal proposta, utilizamos a base de dados criminais de Valência entre os anos de 2010 a 2020. Transformamos o problema de previsão de crimes nas ruas em um problema de classificação multirrótulo. Adicionalmente, aspectos meteorológicos e espaciais foram considerados para auxiliar na previsão dos crimes. Devido ao alto número de rótulos, utilizamos métodos de classificação extrema multirrótulo, em especial, o método Bonsai. Aplicando a métrica precisão@k, foi possível notar que o método Bonsai foi superior aos métodos multirrótulo tradicionais. A metodologia desenvolvida para o problema proporciona diversos avanços na área de previsão de crimes.
Palavras-chave: Aprendizado de máquina, previsão de crimes, multirrótulo.

Abstract:
The increase in crime in areas with high population concentration makes cities one of the main sources of violence. Understanding the characteristics of criminal occurrences and their relationship with the urban scale transcends academic issues and becomes central to today's society. Thus, we propose a multi-label classification approach to predict crimes in the streets of Valencia, Spain. For this proposal, we used a crime database between the years 2010 to 2020. We transform a crime prediction into a multi-label problem. In addition, meteorological and spatial aspects were considered to assist in crime prediction. Due to the high number of labels, extreme multi-label methods were considered, especially the Bonsai method. Applying the precision@k metric to evaluate the models, it was possible to notice that the Bonsai method outperformed traditional approaches. The proposed methodology provides several advances in the area of crime prediction.
Keywords: Machine learning, crime prediction, multi-label.

1 Introdução

De acordo com a ONU (Organização das Nações Unidas), a população vem diminuindo sua taxa de crescimento populacional mundial desde 1950 (NAÇÕES UNIDAS BRASIL, 2022), no entanto, a elevação da população durante os anos gera diversos desafios de gestão de recursos, um deles sendo a segurança pública. Estudos mostram que existem diversos fatores que podem influenciar a taxa de insegurança da população. Em uma análise de agentes socioeconômicos e espaciais para taxa de assaltos, CHANG (2011) verificaram em seu estudo que houve uma taxa maior de crimes em locais onde apresentam mais casas e estabelecimentos comerciais, como também que o tipo de estrada, como um fator estrutural espacial, mostrou-se mais correlacionado com as taxas de invasão do que qualquer outro fator. Outros estudos sobre fatores espaciais foram tratados em ANSELIN et al. (2000), no que tange, a taxa de crimes em uma determinada localidade.

A ampla gama de fatores que influenciam as taxas de crimes, evidencia a importância do uso de métodos preditivos para auxiliar na gestão da segurança pública. De fato, recentes pesquisas focaram no uso de aprendizado de máquinas/estatística para criar ferramentas para auxiliar o combate ao crime. HOSSAIN (2020) investigaram o uso de uma metodologia utilizando técnicas de aprendizado supervisionado clássicos para prever crimes. BOGOMOLOV (2014) propuseram um novo método de previsão de

crimes geoespaciais a partir de múltiplas fontes de dados, em especial telefone móveis e dados demográficos. RODRIGUES (2010) utilizaram métodos de estatística espacial, como geoestatística e processos pontuais, para propor um sistema de vigilância para detectar aglomerados espaço-temporais emergentes na distribuição de homicídios na cidade brasileira de Belo Horizonte.

Este artigo propõe uma metodologia baseada em classificação multirrótulo utilizado dados temporais e espaciais dos crimes para prever tais delitos na cidade de Valência (Espanha), que passa por uma onda de crimes desde o início de 2021 (EPDATA, 2022). Portanto, há uma clara necessidade de pesquisa e desenvolvimento de tecnologias para encontrar soluções que ajudem a prevenir o aumento do número de casos de violência.

O restante deste artigo continua da seguinte forma: a Seção 2 descreve o objetivo proposto pelo artigo. A Seção 3 trata da seleção de técnicas, recursos e métodos de pesquisa. Nesta seção são apresentados a base de dados utilizada, os métodos multirrótulo, a descrição dos experimentos realizados e a métrica de avaliação. A Seção 4 apresenta os resultados obtidos. Por fim, a Seção 5 apresenta as considerações finais do artigo.

2 Objetivos

Este artigo investiga técnicas de classificação multirrótulo para propor uma metodologia para prever a ocorrência de crimes nas ruas da cidade de Valência, Espanha.

3 Procedimentos Metodológicos

Esta seção descreve a metodologia utilizada neste artigo. Todos os principais passos do método proposto estão apresentados no fluxograma da Figura 1. O método tem como entrada o conjunto de dados criminais, onde é atribuída uma abordagem multirrótulo aos dados. Com esta entrada, é feita uma divisão do conjunto para o treino do método preditivo. Antes de utilizar o modelo, o método é avaliado na base de teste. As seções a seguir descrevem os principais pontos do método proposto.

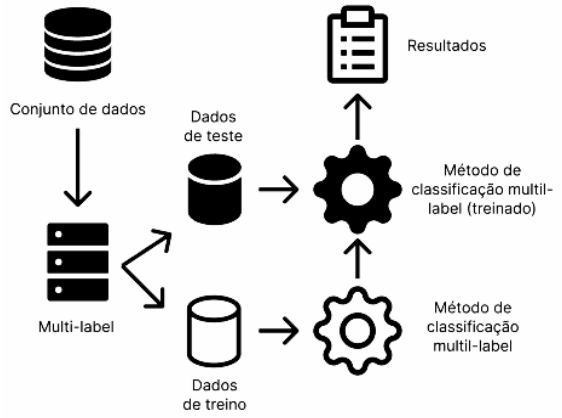

Figura 1 - Fluxograma com visão global da metodologia proposta.

3.1 Conjunto de dados

Originalmente, o conjunto de dados continha atividades criminosas em toda a cidade de Valência de 2010 a 2020. Neste conjunto de dados, haviam dados históricos dos crimes como latitude e longitude, a data, as distâncias de locais importantes envolvidos no crime, como bar, boate, caixa eletrônico, entre outras informações.

Para adicionar mais informações sobre o fato, foram adicionados dados meteorológicos extraídos da estação meteorológica da região. As características coletadas foram temperatura mínima, máxima e média, pressão mínima, máxima e média referentes a data do acontecido.

Todas as informações obtidas foram cedidas pelos serviços de dados abertos do governo.

3.2 Multirrótulo

O paradigma de classificação multirrótulo (*multi-label*) atribui a cada amostra um conjunto de rótulos alvo. Formalmente, para um χ denominando o espaço de características e $\omega = \{\square_1, \square_2, \square_3, \ldots, \square_\square\}$ sendo um conjunto de rótulos. Então, uma amostra é definida como um par de vetores $(\square, \square)$ onde $x \in \chi$ e $\square$ sendo uma combinação de rótulos, representado por um vetor binário em que $\square = (\square_1, \square_2, \ldots, \square_\square)$, onde $\square_\square = 1$, apenas em amostras que sejam associadas a um rótulo $\square_\square$ (MELLO; VAREJÃO; RODRIGUES, 2022).

No nosso contexto, foi utilizada a estratégia multirrótulo de forma que os dados foram organizados como uma matriz de dias por rua, onde cada dia representa uma amostra e as ruas da cidade representam os rótulos. Cada amostra (dia) é associada a um ou mais rótulos (ruas) considerando a ocorrência de crimes naquele dia, isto é, cada amostra é associada às ruas que tiveram a ocorrência de 1 ou mais crimes naquele dia. Para conhecer a rua representativa do crime, foi utilizada a rua mais próxima da sua localização. Considerando o período de estudo (2010 a 2020), obtivemos ao todo 3.652 amostras (dias) e 2078 rótulos (ruas). Para compor o conjunto de características χ foram consideradas informações de 3 fontes: distância, meteorológicas e históricas. Maiores detalhes sobre as características são apresentados na Seção 3.3. Após a fase de treinamento, podemos usar o modelo para predizer as ruas com maiores probabilidades de ocorrência de crimes no dia seguinte (usar a informação de hoje para ver crimes de amanhã).

Devido ao alto número de rótulos (2078 ruas), optamos por utilizar métodos XMC (*Extreme Multi Label Classification*), onde são melhor explicados em SHEN et al. (2020), onde é desenvolvido um classificador que aponta para um subconjunto de rótulos mais relevantes de um conjunto muito grande de rótulos.

A principal qualidade da abordagem de algoritmos multirrótulo é sua capacidade de tratar dados esparsos. Como abordagem principal, utilizamos o algoritmo Bonsai, que generaliza o conceito de representação de rótulos em XMC através da representação do espaço de rótulos em árvores rasas (KHANDAGALE; XIAO; BABBAR, 2020). O Bonsai pode ser visto com uma extensão do método Parabel (PRABHU et al., 2018), que é muito usado em aplicações de XMC.

Para investigar esses algoritmos, foi utilizada a biblioteca Omikuji, que é uma implementação eficiente de "árvores de rótulos particionadas" e sua variante para XMC. Esta biblioteca oferece a possibilidade de treinar/testar algoritmos de XMC alterando hiperparâmetros. (DONG; SUOMINEN, 2022)

3.3 Organização de experimentos

Inicialmente, todas as características utilizadas neste problema multirrótulo foram divididas em 3 conjuntos. O primeiro é composto por características históricas, representando a quantidade de acidentes registrados anteriormente em cada rua. O segundo conjunto de características é composto por medidas meteorológicas diárias. Por fim, o terceiro conjunto usa informações espaciais, representando as distâncias médias, mínimas e máximas entre pontos de interesse da localização (boates, bares, etc) e os crimes.

As amostras foram distribuídas em 70% para treino e 30% para teste. A partição é sequencial para garantir a propriedade de continuidade dos dados históricos. Ou seja, os 70% primeiros dias são considerados amostra de treinamento e os 30% demais dias são utilizados para avaliar o modelo.

O método Bonsai foi treinado e avaliado para cada possível combinação de conjunto de características. Como forma de comparação, outros métodos multirrótulo foram testados considerando a combinação de conjuntos de características que apresentou melhor resultado no método Bonsai. Um dos métodos utilizados foi o *Binary Relevance* (BR), sendo uma das abordagens mais básicas para classificação multirrótulo, (ZHANG et al., 2018). A principal característica deste método é que ele ignora os relacionamentos entre rótulos. Além deste, o método XMC Parabel também foi treinado e avaliado para permitir comparação com o método Bonsai.

3.5 Métrica de avaliação

Para avaliar os resultados obtidos pelos algoritmos, a métrica precisão@k (P@k) foi utilizada. Esta métrica é amplamente utilizada em diversas aplicações de sistemas de recomendações e XMC, veja por exemplo JASINSKA-KOBUS et al. (2020), PRABHU et al. (2018), MITTAL et al. (2022). A □ □ □ □ □ã□@□ é definida na Equação (1)

$$\square\square\square\square\square\tilde{a}\square@\square(\hat{y},\square) : = \frac{1}{\square} \sum_{\square \in \square\square\square\square(\hat{y})} \square_{\square} \quad (1)$$

A métrica Precisão@k é definida para um vetor de resultados previstos $\hat{y} \in \square^{\square}$ e um vetor contendo os resultados reais $\square \in \{0,1\}^{\square}$. A função $\square\square\square_{\square}(\square)$ retorna os k maiores índices de $\hat{y}$ (KHANDAGALE; XIAO; BABBAR, 2020).

P@k é uma medida que corresponde ao número de resultados relevantes (rótulos positivos) que pertencem ao conjunto dos k rótulos com maiores probabilidades estimadas de ocorrência. Por exemplo, P@10 ou "precisão em 10" corresponde ao percentual de resultados relevantes (rótulos positivos) nos 10 rótulos com maiores probabilidades estimadas de ocorrência. Quanto maior for esta medida, mais assertivo é o modelo de previsão. Para este trabalho, consideramos $k \in \{1,3,5\}$.

4 Resultados

Esta seção apresenta os resultados obtidos nos experimentos descritos na seção 3.6.

A melhor combinação encontrada para o método Bonsai foi a com dados históricos e meteorológicos, sendo os resultados de P@k de 29.51, 23.59 e 21.37 para k sendo 1, 3 e 5 respectivamente. Como forma de comparação, os métodos Parabel e *Binary Relevance* também foram avaliados usando este mesmo conjunto de características. Todos os resultados do método Bonsai e os resultados dos métodos alternativos são exibidos na tabela do apêndice A. Como podemos observar na tabela, o subconjunto de variáveis espaciais apresentam um menor grau de explicação do problema.

Para melhor visualizar as diferenças entres os métodos nos diferentes valores de k, foi criado um gráfico de linha (Figura 2). Neste gráfico, é possível notar que os resultados do método Bonsai são superiores ao Parabel nas métricas P@1 e P@5. Ao passo que em P@3, o Parabel obteve desempenho ligeiramente superior ao Bonsai.

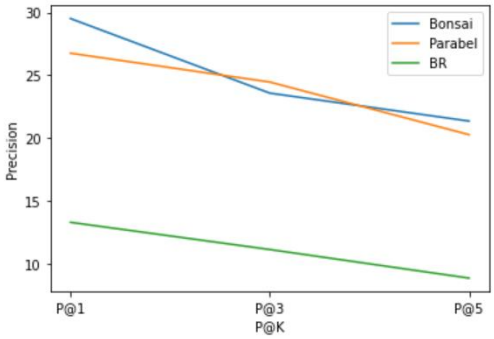

Figura 2 - Gráfico de comparação entre métodos (dados da pesquisa, 2022)

Adicionalmente, podemos perceber que o método BR apresentou resultados insatisfatórios e muito inferiores ao algoritmos de XMC. Uma explicação para isto é que o BR não considera correlação entre os rótulos, o que desfavorece seu resultado comparado com métodos XMC. Vale ressaltar que o esforço computacional necessário para obter os resultados do BR foram em média ≃ 80% maiores que para os métodos de XMC.

5 Conclusão ou Considerações Finais

Este artigo transforma o problema de previsão de crimes em Valência, Espanha, em problema de classificação multirrótulo. Devido ao alto número de rótulos e a relação entre os mesmos, os métodos de XMC se mostraram superiores em comparação aos métodos tradicionais de multirrótulo.

Para facilitar a visualização dos resultados, criamos um gráfico onde as ruas são coloridas baseadas nas suas respectivas probabilidades de ocorrências de crimes. A Figura 3 apresenta uma região com ruas com alta probabilidade de ocorrência de crimes. O mapa completo da cidade pode ser visto no apêndice B.

Figura 3 - Região com alta probabilidade de ocorrências criminosas

Metodologias utilizando a ideia de agrupamento entre ruas mais similares, o uso da análise de Componentes Principais (PCA) para redução dimensional, entre outras abordagens foram testadas. No entanto, todos resultados foram insatisfatórios e não reportados neste artigo.

Como trabalhos futuros destacamos o incremento de características no modelo para que as previsões tenham uma maior assertividade.

Referências

NAÇÕES UNIDAS BRASIL. População mundial chegará a 8 bilhões em novembro de 2022. [S.I] 2022. Disponível em <https://brasil.un.org/pt-br/189756-populacao-mundial-chegara-8-bilhoes-em-novembro-de-2022>. Acesso em: 05 set. 2022.

CHANG, Dongkuk. Social crime or spatial crime? Exploring the effects of social, economical, and spatial factors on burglary rates. Environment and behavior, v. 43, n. 1, p. 26-52, 2011

ANSELIN, Luc et al. Spatial analyses of crime. Criminal justice, v. 4, n. 2, p. 213-262, 2000.

EPDATA. Valencia - Crimen: asesinatos, robos, secuestros y otros delitos registrados en cada municipio. [S.I] [2022?]. Disponível em: <https://www.epdata.es/datos/crimen-asesinatos-robos-secuestros-otros-delitos-registrados-cada-municipio/6/valencia/7587> Acesso em: 8 set. 2022.

RODRIGUES, Alexandre Loureiros. Spatio-temporal models: low-rank approximation, inference and applications. 2010. Tese de Doutorado. Lancaster University.

HOSSAIN, Sohrab et al. Crime prediction using spatio-temporal data. In: International Conference on Computing Science, Communication and Security. Springer, Singapore, 2020. p. 277-289.

BOGOMOLOV, Andrey et al. Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th international conference on multimodal interaction. 2014. p. 427-434.

MELLO, Lucas Henrique Sousa; VAREJÃO, Flávio Miguel; RODRIGUES, Alexandre Loureiros. A Worst Case Analysis of Calibrated Label Ranking Multi-label Classification Method. Journal of Machine Learning Research, v. 23, n. 168, p. 1-30, 2022.

SHEN, Yanyao et al. Extreme multi-label classification from aggregated labels. In: International Conference on Machine Learning. PMLR, 2020. p. 8752-8762.

KHANDAGALE, Sujay; XIAO, Han; BABBAR, Rohit. Bonsai: diverse and shallow trees for extreme multi-label classification. Machine Learning, v. 109, n. 11, p. 2099-2119, 2020.

PRABHU, Yashoteja et al. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In: Proceedings of the 2018 World Wide Web Conference. 2018. p. 993-1002.

ZHANG, Min-Ling et al. Binary relevance for multi-label learning: an overview. Frontiers of Computer Science, v. 12, n. 2, p. 191-202, 2018.

JASINSKA-KOBUS, Kalina et al. Probabilistic label trees for extreme multi-label classification. arXiv preprint arXiv:2009.11218, 2020.

DONG, Tom; SUOMINEN, Osma. Omikuji. Disponível em https://github.com/tomtung/omikuji. Acesso em: 15 de set. 2022.

MITTAL, Anshul et al. Multi-modal Extreme Classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 12393-12402.

Apêndice A - Tabela com resultados dos experimentos

Método	Dados Históricos	Dados Meteorológicos	Dados Espaciais	P@1	P@3	P@5
Bonsai	X	X	X	22.35	23.20	21.96
Bonsai	**X**	**X**		**29.51**	**23.59**	**21.37**
Bonsai	X		X	25.14	23.11	19.85
Bonsai		X	X	22.55	22.75	21.65
Bonsai	X			27.87	24.59	18.96
Bonsai		X		27.47	29.49	25.6
Bonsai			X	21.48	22.58	21.03
Parabel	X	X		26.76	24.47	20.29
BR	X	X		13.33	11.17	8.90

Apêndice B - Frequência de crimes por rua, (2010 - 2020)