Predição de Localização de Crimes utilizando Random Forest

Ígor Arthur Lauxen[1]

[1]Instituto de Informática – Universidade Vale dos Sinos (Unisinos)
Caixa Postal 93.022 – 750 – São Leopoldo – RS – Brasil

igoral@edu.unisinos.br

Introdução

Este trabalho tem como principal objetivo ser uma continuação do trabalho apresentado na cadeira de Seminário de 2021/01, onde a proposta foi avaliar quais algoritmos poderiam ser utilizados para realizar a previsão de invasões domiciliares em uma dada cidade.

No trabalho anterior, foram vistos que algoritmos supervisionados são comumente utilizados para realizar previsões de onde as invasões ocorreriam e aparecem algoritmos como: Random Forest (RF), Regressão Logística e Risk Terrain Modeling (RTM).

A motivação inicial seria executar uma previsão com base nos dados coletados em São Leopoldo, porém, os dados da secretária de segurança do Rio Grande do Sul [1] apresentam somente informações totais de crimes, sem prover demais informações como localização, horário, número de vítimas, entre outros.

Portanto, para viabilizar a validação de pelo menos um dos modelos no artigo do seminário, foi coletado dados da plataforma Kaggle, Crimes in Chicago [2]. Esta base de dados não possuem todos os dados sugeridos pela análise feita anteriormente, pois não possui informações sobre o agressor e nem sobre a região, como econômica ou educacional. Entretanto, como possui o tipo de crime, data e posição geográfica, dados foram considerados.

Além disso, estes dados não restringem somente a um único tipo de crime, o que pode influência no resultado do modelo pois não foi validado qual tipo de análise ou dados seria necessário para prever cada um deles de forma apropriada. Neste artigo está sendo utilizado o mesmo tipo de processamento independente do crime, sem nenhum tipo de filtro pelas categorias levantadas na base.

Metodologia

Neste trabalho pretende-se executar o modelo Random Forest para previsão de locais onde crimes irão ocorrer. O modelo random forest é citado nos artigos de [Lin, Ying Lung et al. 2018], [Cichosz, Paweł 2020] e [Araújo, Adelson 2019].

Segundo [Araújo, 2019 *apude* VOMFELL 2018], Random Forest é um modelo que utiliza a metodologia "boosting", ou seja, ele treina uma árvore de decisões com diferentes funcionalidades de forma aleatória

Um dos desafios de utilizar a estrutura de dados do *Crimes in Chicago* é que os dados dão poucas informações da região onde os crimes ocorrerem e segundo [Cichosz, Paweł 2020] idealmente os dados não podem ser derivados de representações de coordenadas, mas sim de atributos derivados da geografia agregados por pontos de interesse, tornando as relações reusáveis e independentes da tipografia da cidade.

Segundo o artigo de [Lin, Ying Lung *et al*. 2018], o modelo do Random Decision Forest pode ter problemas de performance e poderia ser substituído pelo Deep Neural Network, entretanto, não é foco deste trabalho fazer comparações entre diferentes algoritmos para ver o mais performático, mas sim a validação se Random Forest consegue prever crimes de forma satisfatória com o data set encontrado.

Além disso, com o intuito de validar se haveriam *outliers* nos dados coletados, foi gerado um mapa com todos os pontos onde houve crimes. Essa estratégia é levantada por [Younes Charfaou 2019] em seu *blog* para que haja, além do benefício citado anteriormente, descobertas em cima dos dados.

Desenvolvimento do código pode ser visto no git hub, acessando https://github.com/igorlauxen/crime-prediction .

Afim de auxiliar na performance do algoritmo, foram reduzidos o numero de colunas utilizadas para a análise, removendo por exemplo colunas de Community Area, Location Description. Essas remoções podem ser vistas na classe *DataPreparator.py* no repositório.

Conclusão

O algoritmo levou 30 minutos, 40 minutos e 1 hora para processar 250k, 350k e todas as linhas subsequentemente e a variação da mediana dos erros não apresentou variações grande entre as quantidades de linhas, sendo de 0.60634, 0.0632, 0.0643. O que induz autor acreditar que os dados não auxiliam no uso de *random forest*, pois a taxa de acerto esta tão alta que se acredita haver alguma inconsistência na relação das variáveis empregadas.

Uma das grandes dificuldades é que existe uma variação muito grande de entradas nas colunas que são caracteres, portanto, quando o dado é traduzido para inteiros, de forma que o algoritmo consiga processar, o python acaba não conseguindo processar devido a uma falha na memória, por isso, o campo *Description* acabou sendo desconsiderado, embora autor considere-o uma entrada interessante a ser analisada.

[Nitta 2019] comprovou que locais onde houve crimes tem tendências a ocorrerem outros times novamente; dito isso, com a análise do mapa, percebeu-se que os pontos são bem próximos um dos outros e não existem *outliers* chamativos, portanto, não foi realizado uma mudança específica nos dados para cobrir essa necessidade, já que os pontos já estão aglomerados.

Referencias

Secretária da Segurança Pública “Indicadores Criminais”, Acessado em < https://ssp.rs.gov.br/indicadores-criminais> Data: 16 de out. de 2021.

Kaggle “Crime in Chicago”, Acessado em < https://www.kaggle.com/currie32/crimes-in-chicago> Data: 16 de out. de 2021.

Lin, Ying Lung et al. (2018) Grid-based crime prediction using geographical features. Journal ISPRS International Journal of Geo-Information. Volume 7. Issue 8.

Cichosz, Paweł (2020). Urban Crime Risk Prediction Using Point of Interest Data. ISPRS Int. J. Geo-Inf. Volume 9. Issue 7. Pagina 549.

Araújo, Adelson. (2019). Predspot: Predicting Crime Hotspots with Machine Learning. UFRN.

Younes Charfaou (2019) "Working with Geospatial Data in Machine Learning". Acessado em: < https://heartbeat.comet.ml/working-with-geospatial-data-in-machine-learning-ad4097c7228d > Data: 15 de out. de 2021.

NITTA, G.R., Rao, B.Y., Sravani, T. et al. LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type. SOCA 13, 187–197 (2019).