Disciplinas: Algoritmos de Reconhecimento de Padrões

Predição de crimes

Jean Lucas Monte Carvalho

INSTITUTO DE INFORMÁTICA
UFG

Introdução

O propósito do estudo em questão é analisar os padrões em ocorrências policiais na cidade de Chicago de forma a determinar a importância da localidade na ocorrência de um crime, e também se essa váriavel pode nos levar a predição de um tipo de crime dado o seu acontecimento. Podendo então ajudar a cidade numa questão tão [1]preocupante que é a criminalidade crescente na região.

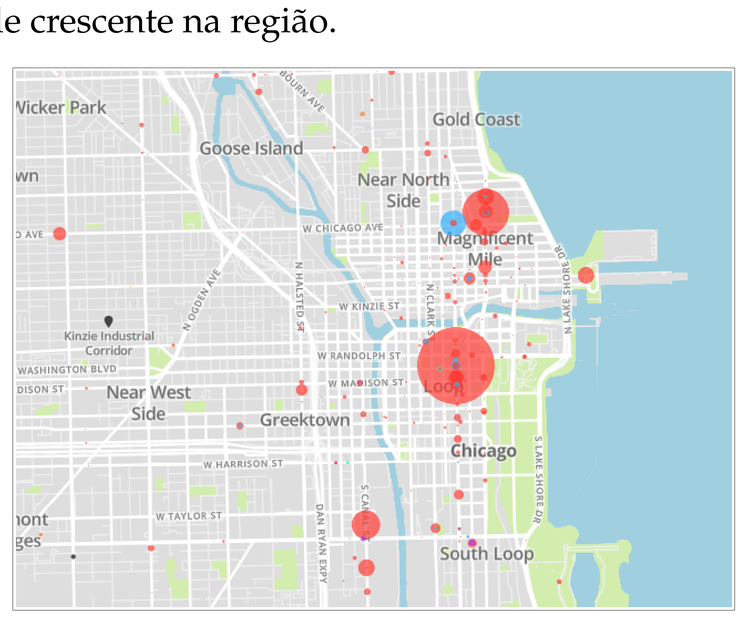


Figura 1: Distribuição geográfica da ocorrência de crimes.

Atualmente a conabilidade nos dados registrados e gerados diariamente é um grande problema para aqueles que dependem destes diretamente. O desenvolvimento do trabalho poderia auxiliar na conrmação dos dados obtidos, garantindo uma melhor cobertura nos serviços de segurança da cidade.

Materiais e métodos

Os métodos utilizados para o desenvolvimento foram:

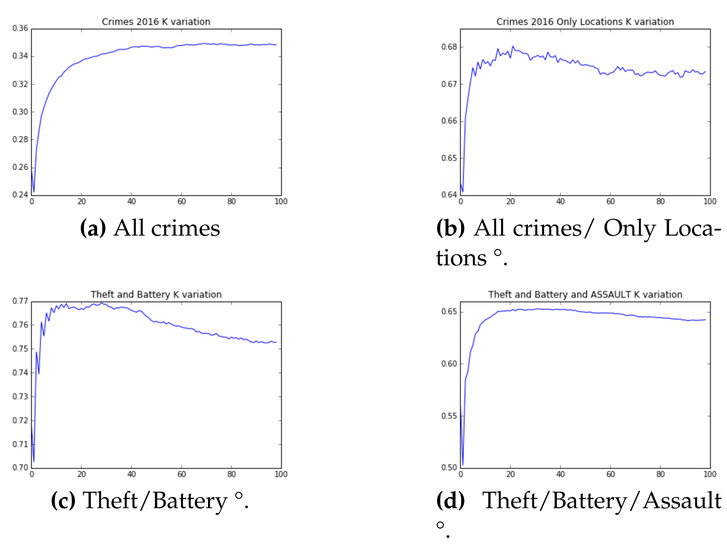
- Normalização
- Feature Selection (Kernel PCA)
- Naive Bayes
- Árvore de decisão(Random Forest)
- KNN
- Cross Validation w/ K Folds
- Matriz de confusão

Durante o desenvolvimento do trabalho, foi necessário uma etapa de engenharia de dados, que tem como objetivo limpar os dados para o projeto, descartando váriavéis inutilizáveis para o processo de aprendizagem e também a restruturando-os para melhor manuseio dos algoritmos. Os dados utilizados no projeto são fornecidos pelo portal de disponibilização de dados da cidade de Chicago. Aplicamos então as etapas de normalização, logo em sequência utilizamos da técnica Kernel PCA para extração das variáveis mais importantes para nossa classificação.
O primeiro algoritmo escolhido foi Naive Bayes tendo em vista sua facilidade de implementação e a distribuição dos dados, nos permitindo então afirmar de forma probabilística o tipo de crime.
Também foi utilizado KNN, este algoritmo tende a agrupar conjuntos de dados por semelhanças com seus k-vizinhos mais próximos, e portanto assumimos que devido ao agrupamento de crimes por localidade, como mostrado na Figura 1, o algoritmo seria uma boa escolha.
Por último foi utilizado de Random Forest, um algoritmo que constrói várias árvores de decisão de forma aleatória e as utiliza para classificar os dados de acordo com grupos. Este seria um bom candidato tendo em vista a capacidade de evitar o sobre-ajuste.
A técnica de cross validation foi aplicada para nos permitir determinar se a capacidade de generalização do nosso modelo sobre toda a base de dados era boa. Para avaliar a performance dos algoritmos foram utilizadas Matrizes de Confusão, sendo possivel aferir a acurácia observando os verdadeiros positivos, bem como os falsos positivos.

Resultados

Durante a implementação dos passos descritos, foi necessário monitorar a execução de alguns algoritmos para melhor ajuste de seus parâmetros de forma a garantir uma melhor acurácia.

Exploração do valor de K para o KNN



(a) All crimes

(b) All crimes/ Only Locations °.

(c) Theft/Battery °.

(d) Theft/Battery/Assault °.

Figura 2: Variação no valor de K para o KNN.

É possível observar de acordo com os gráficos que o valor de K é fundamental e que se comporta de maneira semelhante a cada variação dos datasets utilizados, o valor de K atinge em certo ponto, o máximo de performance, passando a ser menos efetivo após tal ponto. Com a utilização dos gráficos é possível escolher então o valor que melhor maximiza a performance e tempo de execução do algoritmo tendo em vista que a técnica KNN tem um custo de execução alto.
Para o algoritmo de Random Forest foi observado uma demora na execução pois o dataset utilizado é de grande capacidade e portanto a etapa de treinamento e geração das árvores aleatórias leva certo tempo, porém após o modelo ser gerado nota-se a rápida capacidade de classificação desta técnica. Após a execução dos algoritmos foi possível observar os seguintes resultados.

Algoritmos	Naive Bayes	KNN	Random Forest
Theft And Battery	35.6%	47.0%	26.0%
Theft, Battery and Assault	12.4%	18.0%	27.1%
Only Locations	85.0%	41.2%	68.0%
All Crimes	2.8%	0.0%	05.7%

Tabela 1: Eficiência dos algoritmos de machine learning.

Conclusão

Com o que foi apresentado, podemos determinar a confiabilidade dos dados já inseridos no banco, como também aqueles que serão inseridos diariamente, além de determinar a confiabilidade das ligações feitas à polícia, minimizando as chances de falsas ocorrências e garantindo que as ligações de maior prioridade sejam de fato as de maior confiabilidade.

Para mais informações

- Email: `jmontecarvalho@gmail.com`.
- Repositório de código fonte: `https://github.com/jeanlks/ARP/tree/master/finalProject`.
- Linkedin: `https://www.linkedin.com/in/jeanlucasmontecarvalho`
- Lattes: `http://lattes.cnpq.br/3167384947767536`

Referências

[1] Rosa Flores e Mallory Simon. *762 murders. 12 months. 1 American city.* [Online; acessado 9-Agosto-2017]. URL: `http://edition.cnn.com/2017/01/02/us/chicago-murder-rate-2016-visual-guide/`.

Professores orientadores: Gustavo Teodoro Laureano , Anderson da Silva Soares