Previsão e análise de homicídios nos Estados Unidos por meio de algoritmos de classificação

Alice Cabral
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
alicecamarques@gmail.com

Ana Carolina Manso
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
acmsilverio@sga.pucminas.br

Anna Puga
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
annapugac@gmail.com

Arthur Leandro
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
aleandro@sga.pucminas.br

Danniel Henrique Vieira
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
danniel.hevieira@gmail.com

João Victor Amorim
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
amorimvictorjoao@gmail.com

Juliana Silvestre
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
juliana.silvestresilva@hotmail.com

Larissa Kaweski
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
larissakaweski15@gmail.com

Cristiane Neri Nobre
Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, MG, Brasil
nobre@pucminas.br

RESUMO

O presente artigo apresenta um trabalho prático na área de Inteligência Artificial realizado com o intuito de prever, por meio de algoritmos de classificação, os tipos dos crimes de homicídio nos Estados Unidos. Para esse objetivo, foi escolhida uma base de dados que contém dados de homicídios nos Estados Unidos de 1976 a 2014. A metodologia do trabalho consiste em um pré-processamento da base de dados e na execução de três algoritmos de classificação: Random Forest, J48 e Rede Neural. Para a análise e discussão dos resultados, algumas métricas foram avaliadas, as classes foram analisadas individualmente, e foi mostrado o modelo com melhor aproveitamento.

PALAVRAS-CHAVE

Inteligência artificial, Ciência de dados, Algoritmos de classificação, Homicídio, Estados Unidos.

1 INTRODUÇÃO

De acordo com o FBI, um homicídio é cometido nos Estados Unidos a cada 26 minutos. Os homicídios podem ser classificados em homicídio doloso (*murder)* ou em homicídio culposo (*manslaughter*), de acordo com o direito penal norte-americano. Nessa perspectiva, este estudo utilizou a base de dados "Homicide Reports 1980-2014", disponível na plataforma Kaggle, para prever os tipos dos crimes de homicídio nos Estados Unidos e analisar o perfil destes, constituindo o problema a ser atacado pelo presente estudo.

2 A BASE DE DADOS

A base de dados deste estudo é a "Homicide Reports, 1980-2014" (Registros de Homicídios, 1980-2014), com 638454 instâncias e 24 atributos, sendo 16 deles Strings, 6 inteiros, 1 booleano, e 1 do tipo ID. É uma base de dados de homicídios nos Estados Unidos, e inclui dados do Relatório de Homicídios do FBI de 1976 a 2014, além de incluir informações de mais de 22.000 homicídios que não foram reportados para o Departamento de Justiça. Os dados foram compilados e disponibilizados pelo Murder Accountability Project, fundado por Thomas Hargrove.

Descrição de atributos

A tabela 1 indica quais são os atributos da base de dados "Homicide Reports, 1980-2014" e os descreve por meio do seu nome, tipo e significado.

Tabela 1 - Atributos da base de dados "Homicide Reports 1980-2014"

NOME	TIPO	SIGNIFICADO
Record ID	Numérico	Identificador
Agency Code	Nominal	Código da agência da polícia
Agency Name	Nominal	Nome da agência da polícia
Agency Type	Nominal	Tipo da agência da polícia
City	Nominal	Cidade onde ocorreu o crime
State	Nominal	Estado onde ocorreu o crime
Year	Numérico	Ano em que ocorreu o crime
Month	Nominal	Mês em que ocorreu o crime
Incident	Numérico	Nº de incidentes de uma agência até a data do crime
Crime Type	Nominal	Tipo do crime
Crime Solved	Nominal	Se o crime já foi solucionado ou não
Victim Sex	Nominal	Sexo da vítima
Victim Age	Numérico	Idade da vítima
Victim Race	Nominal	Raça da vítima
Victim Ethnicity	Nominal	Etnia da vítima
Perpetrator Sex	Nominal	Sexo do criminoso
Perpetrator Age	Numérico	Idade do criminoso
Perpetrator Race	Nominal	Raça do criminoso
Perpetrator Ethnicity	Nominal	Etnia do criminoso
Relationship	Nominal	Relação do criminoso com a vítima
Weapon	Nominal	Arma utilizada no crime
Victim Count	Numérico	Quantidade de vítimas
Perpetrator Count	Numérico	Quantidade de criminosos envolvidos
Record Source	Nominal	Fonte do registro

3 REFERENCIAL TEÓRICO

A fim de melhor compreender os crimes de homicídio no contexto do direito penal norte-americano, bem como prever os tipos desses crimes por meio de algoritmos de classificação, faz-se necessário o entendimento de alguns conceitos, apresentados a seguir.

Crimes de Homicídio nos Estados Unidos

Nos Estados Unidos, cada estado possui seu próprio sistema processual e leis penais, ou seja, a maior parte do poder reside nos estados e não no governo central, ao contrário do Brasil, em que a maior parte do poder reside no governo central. Desse modo, nos EUA, cada estado é livre para instituir sua legislação criminal. Dito isso, a classificação a seguir diz respeito à maioria dos estados dos EUA.

Um homicídio pode ser classificado como *murder* ou *manslaughter*. *Murder*, ou assassinato, na linguagem popular, é homicídio doloso, voluntário, em que há a intenção de matar. Um crime desse tipo pode ser dividido em primeiro grau, que seria um homicídio doloso qualificado, cometido mediante premeditação, envenenamento, emboscada, fogo, etc, e em segundo grau, que seria um homicídio doloso simples, em que há a intenção de matar, mas não apresenta qualificadores. Já *manslaughter* caracteriza-se como homicídio culposo. Ele pode ser voluntário, quando ocorrido no calor da paixão ou quando reflete reação a súbita provocação, ou involuntário, quando se desdobra da negligência do réu. Negligência é a displicência no agir, a falta de precaução, a indiferença do agente que, podendo adotar as cautelas necessárias, não o faz.

Na base de dados “Homicide Reports 1980-2014”, o atributo *Crime Type* classifica os crimes de homicídio em dois valores: *Murder or Manslaughter*, que diz respeito a todos os graus de homicídio doloso e ao homicídio culposo voluntário, e *Manslaughter by Negligence*, que diz respeito ao homicídio culposo involuntário por negligência. Essas serão as classes utilizadas para o problema de classificação.

Problemas de Classificação

A Classificação é uma das categorias de problemas de Aprendizado de Máquina mais utilizadas e consiste em extrair modelos que descrevem classes de dados e são capazes de prever tendências, ou seja, os algoritmos de classificação são usados para prever atributos nominais a partir de um conjunto de dados. Entre os métodos mais utilizados em problemas de classificação estão as arvores de decisão, como o algoritmo Random Forest.

Nesses tipos de problema, a partir de uma base de dados rotulada, são formados dois subconjuntos diferentes: a base de treino, que geralmente contém 70% dos dados originais, e base de teste, com o restante dos dados. Posteriormente, a base de treino é submetida ao modelo para que os dados possam ser utilizados na busca de padrões, em seguida, ocorre a predição de classes, etapa na qual as instâncias pertencentes à base de teste são expostas ao modelo gerado para que este realize a predição de suas classes. Para avaliar a qualidade do modelo gerado é feita a comparação entre as classes preditas e as classes verdadeiras da base de teste, o que permite medir a habilidade de classificar corretamente exemplos não vistos durante o treinamento.

Para estimar o desempenho de um classificador são utilizadas métricas como a acurácia e a matriz de confusão. A acurácia de modelo representa a taxa total de instâncias que foram classificadas corretamente. Já a matriz de confusão fornece detalhes acerca do desempenho do modelo, mostrando, para cada classe, o número de classificações corretas e o número de classificações preditas. Com base nesses valores apresentados, é possível ainda calcular outras métricas, como as taxas de verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo.

Trabalhos Relacionados

Atualmente, alguns estudos têm se desenvolvido a fim de analisar bases de dados de violência. Tendo isso em vista, foram encontrados três artigos com bases de dados similares e desenvolvimento coincidente ao deste trabalho.

O artigo “Homicide Profiles Based on Crime Scene and Victim Characteristics”[7] foi desenvolvido com o objetivo de analisar as características de um homicídio a partir de informações sobre a vítima, o criminoso e a cena do crime. Para isso, foi selecionada uma base de dados proveniente do Projeto de Revisão e Homicídios liderado pela Secretaria de Estado e Segurança da Espanha que inicialmente possuía 684 instâncias e 24 atributos. Uma vez definida a metodologia a ser usada, Pecino-Latorre, Pérez-Fuentes e Patró-Hernández realizaram etapas de pré-processamento que envolveram a remoção de instâncias com dados ausentes e a redução do número de atributos apresentados. Em seguida, foi utilizado o algoritmo CART para gerar uma árvore de classificação para cada um dos atributos relacionados ao autor do crime, em que cada caminho da árvore representa uma regra capaz de caracterizar o homicídio de acordo com a variável em questão. Após análise dos resultados, foi possível concluir que características como idade e sexo do criminoso e sua relação com a vítima influenciam notavelmente as características do homicídio, tais como arma utilizada e local do crime. Dessa maneira, o estudo em questão demonstra a relevância da análise de dados durante o processo de investigação criminal, dando destaque para métodos de classificação, os quais permitiram obter potenciais informações a respeito dos autores dos crimes.

Kim, Dunham, Muljono, Lee e Wang realizaram um estudo a fim de fornecer uma ferramenta para as agências governamentais norte-americanas que seja capaz de ajudar a solucionar casos que envolvam mortes violentas. Para isso, os autores do artigo “Discovery of Association Rules in National Violent Death Data Using Optimization of Number of Attributes”[6] aplicaram uma série de técnicas de mineração de dados na base National Violent Death Reporting Database, a qual possui 37 atributos que fornecem informações a respeito da vítima, do crime e da relação entre vítima e suspeito. A partir da análise previa dos dados, Kim, Dunham, Muljono, Lee e Wang definiram como objetivo principal obter regras de associação que indicam uma combinação de atributos capaz de estabelecer se a morte foi um homicídio ou suicídio. Dessa maneira, inicialmente os dados foram submetidos a uma etapa de pré-processamento e foi utilizado algoritmo Apriori para obter associações entre os atributos. Os resultados obtidos pelos autores indicaram que as regras geradas com pelo menos 17 atributos possuem 0,8 de confiança, valor que foi considerado ideal para o proposito apresentado. Além disso, a análise dos resultados demonstra a capacidade do algoritmo Apriori na compreensão padrões ocultos em um grande conjunto de dados.

O artigo "Time Series Analysis and Crime Pattern Forecasting of City Crime Data"[5] trata de uma análise feita em uma base de dados coletada diretamente das fontes históricas da Distrito de Polícia de Manila (MPD), com informações referentes a 16 distritos e dados de 2012 a 2016. A base passa pela etapa de pré-processamento, onde tem-se a limpeza dos dados, consistindo na remoção de dados faltantes e inconsistentes, dados duplicados, ou formatações impróprias. Os dados foram filtrados e alguns atributos foram criados derivados de outros. A respeito dos modelos, o primeiro aplicado foi para geração de regras de associação relacionando um crime com seus atributos, utilizando o algoritmo Apriori. Analisando as regras geradas foi possível entender quais tipos de crime eram mais aleatórios, e quais seguiam certas características e, a partir delas, ações podem ser tomadas a fim de evitá-las. O segundo modelo aplicado foi um modelo classificador, que utilizou a métrica de avaliação Mean Absolute Percentage Error (MAPE), com o objetivo de conseguir inferir o número de crimes em uma data futura. Os algoritmos avaliados para para predizer o valor diário de crimes ou semanal foram o Linear Regression, o Gaussian Processes, o MultiLayer Perceptron, e o Sequential Minimum Optimization Regression, sendo que o melhor resultado foi para o MultiLayer Perceptron para dados semanais. Através dos dados encontrados e das análises geradas espera-se que a eficiência ao se lidar com crimes seja elevada.

4 PRÉ-PROCESSAMENTO

Sabe-se que o pré-processamento dos dados é essencial para obter uma boa análise que gere resultados satisfatórios, portanto, algumas técnicas são utilizadas sob a base de dados para este fim. Neste trabalho, foram utilizadas as plataformas Weka e R para realizar o pré-processamento e gerar os resultados.

Remoção de Atributos

A redução do número de atributos pode tanto melhorar o desempenho do modelo induzido, reduzindo seu custo computacional, quanto tornar os resultados obtidos mais compreensíveis. Em uma melhor análise e desenvolvimento dos dados obtidos, foi necessária a redução do número de atributos disponibilizados pela base. Foram removidos os seguintes atributos:

- *Record ID* → Atributo numérico de identificação de instâncias, considerado desnecessário para a classificação.
- *Agency Code* e *Agency Name* → Estes atributos não são relevantes para o objetivo do projeto por terem uma densidade de valores muito baixa, ou seja, poucos valores repetidos, além de fazerem referência ao atributo *Agency Type* que permaneceu na base. Portanto, também foram considerados como atributos redundantes.
- *City* → Removido a fim de simplificação, para permitir somente um atributo de localização que carrega mais informação: State.
- *Victim Count*, *Perpetrator Count* e *Incident* → Foram descartados por não trazerem informações relevantes para análise do problema.
- *Perpetrator Ethnicity* e *Victim Ethnicity* → Apresentam pouca variedade de valores (*Unknown*, *Hispanic* e *Not-Hispanic*), além de que um dos valores é desconhecido e se refere à maioria das instâncias.
- *Crime Solved* → Atributo nominal removido uma vez que apresentava apenas valor igual a sim e, portanto, não seria interessante para a classificação.

Remoção de Duplicatas

Não foram encontradas instâncias iguais na base de dados, portanto, não houve necessidade de remover duplicatas.

Substituição de Dados Ausentes

Base de dados podem conter dados ausentes por diversos motivos como, por exemplo, problemas na transmissão e no armazenamento dos dados, problemas nos equipamentos que realizam a coleta, entre outros. Devido ao fato de isso apresentar dificuldades relacionadas à qualidade dos dados, os mesmos devem ser tratados por meio de alguma técnica. Neste trabalho, optou-se simplesmente pela remoção das instâncias com dados faltantes, devido a, como dito anteriormente, o alto volume de dados de nossa base.

Remoção de Instâncias com Valores Inválidos

A existência de instâncias inválidas pode prejudicar a análise final de uma base de dados e levar a resultados incorretos, visto que aquele valor não indica um valor real. Nos atributos *Victim Age* e *Perpetrator Age* havia os valores 0, 998 e "?", considerados valores inválidos. A partir disso, todas as instâncias que estavam com esses valores foram removidas, principalmente porque os valores dos outros atributos da maioria dessas instâncias eram desconhecidos (*Unknown*).

Remoção de Instâncias com Valores Desconhecidos

Uma alta quantidade de dados desconhecidos impacta diretamente na qualidade do atributo, o que também ocorre para dados ausentes. Deixar esses valores no conjunto pode comprometer a qualidade dos resultados, levando a uma análise menos eficiente. Sendo assim, optou-se por remover todas as instâncias de todos os atributos que apresentavam o valor *Unknown* (desconhecido). Por existir uma grande quantidade de instâncias na base de dados original, percebeu-se como mais vantajosa a exclusão dessas instâncias, em contraste com a substituição pela moda ou média.

Balanceamento

Bases de dados desbalanceadas são aquelas em que dados de um subconjunto de classes aparecem com uma frequência maior do que os dados das demais classes, o que é o caso da base escolhida para este trabalho. Para evitar o favorecimento da classificação de novos dados na classe majoritária e evitar resultados errôneos, foi realizado o balanceamento da base de dados, após a realização de todas as etapas anteriores. Para o balanceamento, foi realizado um método de *undersampling*, que consiste em remover instâncias da classe majoritária. Com isso, permaneceram 5780 instâncias para cada classe (*Murder or Manslaughter* e *Manslaughter by Negligence*).

5 RESULTADOS

Após a realização das etapas de pré-processamento, os dados foram submetidos a três algoritmos de classificação na plataforma Weka, sendo eles o Random Forest, J48 e Rede Neural, garantindo assim a possibilidade de comparar a qualidade dos modelos gerados. Dessa maneira, os algoritmos foram escolhidos com base no seu potencial de ajuste de parâmetros, e consequentemente, por oferecerem melhores resultados.

Inicialmente, a fim de garantir que o sistema seja capaz de classificar corretamente a base de dados, foi realizada uma divisão de modo que 20% dos dados sejam destinados a testes e o restante utilizado no treinamento dos algoritmos selecionados. Além disso, com o objetivo de obter o melhor desempenho do modelo, foi utilizado o método *Random Search*, responsável por definir os

melhores valores que podem ser aplicados aos hiperparâmetros dos algoritmos.

Os resultados obtidos permitem a comparação dos três algoritmos utilizados. Para isso, a tabela 2 apresenta a porcentagem de instâncias corretamente classificadas para cada algoritmo e a tabela 3 expõe as métricas utilizadas para avaliar a qualidade dos modelos gerados, sendo elas a precisão, *recall* e *f-measure*.

Tabela 2 - Classificação das instâncias

Algoritmo	Instâncias corretamente classificadas (%)	Instâncias incorretamente classificadas (%)
Random Forest	81,1851	18,8149
J48	79,0225	20,9775
Rede Neural	79,1090	20,8910

Tabela 3 - Resultado das métricas utilizadas na avaliação dos modelos gerados

Algoritmos	Classe	Precisão	Recall	F-measure
Random Forest	Murder or Manslaughter	0,798	0,821	0,809
	Manslaughter by Negligence	0,826	0,803	0,814
J48	Murder or Manslaughter	0,766	0,819	0,792
	Manslaughter by Negligence	0,817	0,763	0,789
Rede Neural	Murder or Manslaughter	0,775	0,804	0,789
	Manslaughter by Negligence	0,808	0,779	0,791

A precisão de um modelo corresponde a taxa de instâncias corretamente classificadas como pertencentes a classe em questão dentre todos os que foram classificados nesta classe. Já o *recall* corresponde a taxa de instâncias corretamente classificadas como pertencentes a classe em questão dentre todos os que realmente são da classe em questão. O *f-measure,* por sua vez, é uma média harmônica entre a precisão e o *recall* do modelo.

A partir das tabelas, é possível gerar certas análises gerais dos resultados: observa-se que o algoritmo Random Forest obteve os melhores resultados baseando-se na porcentagem de instâncias corretamente classificadas, com aproximadamente 81% de acertos - em comparação com os 79% dos algoritmos J48 e Rede Neural. Porém, ao separar a análise para as classes obtidas na classificação, alguns padrões são verificados.

A precisão de todos os algoritmos utilizados foi melhor para a classe *Manslaughter by Negligence*, com valores acima de 0.8. Para a classe *Murder or Manslaughter* tal métrica se manteve sempre abaixo de 0.8. Em contraste, o *recall* obteve valores maiores para a segunda classe mencionada.

Levando em consideração a classe *Manslaughter by Negligence*, isso significa que a taxa de instâncias corretamente classificadas como pertencentes à classe dentre todas as instâncias admitidas pelo algoritmo como desta classe foi maior. Mas, entre todas as que realmente fazem parte da classe em questão, o desempenho foi menor. Para a classe *Murder or Manslaughter*, portanto, o contrário.

Ademais, para o *f-measure* todos os algoritmos utilizados resultaram em valores pouco variados, sempre dentro do intervalo de 0.78 e 0.81 para ambas as classes. O algoritmo com melhor resultado nesta métrica foi o Random Forest, que alcançou valores acima de 0.8.

6 CONCLUSÃO

De acordo com a análise de resultados, foi possível observar, portanto, que o melhor algoritmo de classificação para a base de dados selecionada foi o Random Forest. Além disso, o presente trabalho permite a análise dos tipos de homicídio cometidos no Estados Unidos, por meio da previsão do tipo de crime, ou seja, se determinado crime corresponde a um homicídio doloso/culposo voluntário (*Muder or Manslaughter)* ou a um homicídio culposo involuntário (*Manslaughter by Negligence*). Para isso, o modelo gerado correlaciona fatores como idade, sexo, raça e relacionamento da vítima e do criminoso e arma utilizada.

Dessa maneira, o estudo em questão além de demonstrar o potencial dos algoritmos de classificação na previsão de tendências, evidencia o potencial do modelo gerado no auxílio da resolução de futuros crimes que contenham padrões e características semelhantes aos encontrados na base de dados “Homicide Reports 1980-2014”.

REFERÊNCIAS

[1] Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(3), 139–154. doi:10.1002/sam.11312

[2] He, Z., Tao, L., Xie, Z., & Xu, C. (2020). Discovering spatial interaction patterns of near repeat crime by spatial association rules mining. Scientific Reports, 10(1). doi:10.1038/s41598-020-74248-w

[3] Dao, T. H. D., & Thill, J.-C. (2016). The SpatialARMED Framework: Handling Complex Spatial Components in Spatial Association Rule Mining. Geographical Analysis, 48(3), 248–274.7

[4] Chen, P., & Kurland, J. (2018). Time, Place, and Modus Operandi: A Simple Apriori Algorithm Experiment for Crime Pattern Detection. 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). doi:10.1109/iisa.2018.8633657

[5] Charlie S. Marzan, Maria Jeseca C. Baculo, Remedios de Dios Bulos, and Conrado Ruiz. 2017. Time Series Analysis and Crime Pattern Forecasting of City Crime Data. In Proceedings of the International Conference on Algorithms, Computing and Systems (ICACS '17). Association for Computing Machinery, New York, NY,USA,113–118. DOI:https://doi.org/10.1145/3127942.3127959

[6] S. Kim, C. Dunham, S. Muljono, A. Lee and T. Wang, "Discovery of Association Rules in National Violent Death Data Using Optimization of Number of Attributes," 2009 WRI World Congress on Computer Science and Information Engineering, 2009, pp. 616-621, doi: 10.1109/CSIE.2009.721.

[7] Pecino-Latorre, M.d.M.; Pérez-Fuentes, M.d.C.; Patró-Hernández, R.M. Homicide Profiles Based on Crime Scene and Victim Characteristics. Int. J. Environ. Res. Public Health 2019, 16, 3629. https://doi.org/10.3390/ijerph16193629