Mineração de *Itemsets* Frequentes e Descoberta de Subgrupos em Análise de Dados Carcerários

Gabriel Bastos[1], **Fernanda G. Araújo**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{bastos.gabriel,fernandaguim}@dcc.ufmg.br

Resumo. *Neste trabalho, apresentamos uma metodologia e um arcabouço de análise para bases de dados carcerários, utilizando técnicas de aprendizado de máquina descritivo, em particular a mineração de itemsets frequentes e descoberta de subgrupos. Ao analisar uma base de dados estadunidense de registros prisionais, nossos resultados revelam padrões peculiares, que podem ser de interesse para especialistas das ciências sociais. Nosso arcabouço permite a fácil reprodução dos resultados apresentados, bem como análises mais amplas da base de dados.*

1. Introdução

O encarceramento pode trazer prejuízos que vão além da privação da liberdade, como o comprometimento de relacionamentos pessoais [Apel et al. 2010], a tendência à reincidência criminal [Bales e Piquero 2012], a dificuldade de ressocialização [Thomas 1973], e até mesmo problemas de saúde [Plugge et al. 2009]. O amplo impacto do sistema prisional na vida das pessoas, bem como na sociedade como um todo, motiva o estudo dos seus efeitos e dinâmicas.

A literatura atual apresenta diversas metodologias para análise de dados carcerários, em duas principais vertentes. A primeira utiliza técnicas demográficas tradicionais, que exploram análises de estatística descritiva para caracterizar os dados carcerários. Já a segunda, utilza métodos de ciência das redes para análise de núcleos familiares e sociais onde indivíduos experimentam passagem pela prisão. [Chung e Peter 2018]

Neste trabalho, propomos a abordagem de aprendizado de máquina descritivo para análise de dados carcerários. Apresentamos uma técnica de modelagem dos dados para a utilização de algoritmos bem estabelecidos na área, bem como sua aplicação em uma base de dados de registros prisionais estadunidense. Consolidamos nossa metodologia em um arcabouço de análise, que permite a fácil reprodução dos resultados, bem como análises mais profundas com novos parâmetros.

Nossos resultados demonstram duas categorias de padrões e subgrupos. A primeira constitui implicações diretas das leis, e não aprentam grande utilidade além da confimação de que o sistema judiciário pratica de fato a legislação. Já na segunda, observamos características peculiares que indicam dinâmicas não projetadas pelas leis. Como trabalhos futuros, sugerimos a análise destes resultados por especialistas das ciências sociais, bem como a confirmação da relevância destes padrões através de testes de significância estatística.

2. Base de Dados

A base de dados selecionada para estudo fornece informações colhidas de prisões estaduais e federais pelo Programa Nacional de Relatórios de Correções dos Estados Unidos da América. [United States Department of Justice 2016] Os dados constituem registros individuais de cada estadia na prisão, incluindo um identificador único para cada pessoa. Consideramos como reincidentes os registros de nova admissão de um mesmo detento, desconsiderando o mais antigo. As demais características são descritas na tebela 1.

Tabela 1. Características apresentadas na base de dados

Característica	Tipo	Instâncias
Identificador	I	um valor distinto para cada indivíduo
Gênero	C	homem, mulher
Tipo de admissão	C	retorno de liberdade condicional, novo, outro
Categoria do crime	C	ordem pública, drogas, violência, propriedade, outras
Etnia	C	branco, hispânico, negro, outra
Idade na admissão	C	18–24, 25–34, 35–44, 45–54, 55 anos ou mais
Tempo servido	C	0–1, 1–2, 2–5, 5–10, 10 anos ou mais
Tipo de soltura	C	condicional, incondicional, outras
Sentença máxima	C	0–1, 1–2, 2–5, 5–10, 10–25, 25 anos ou mais, perpétua
Categoria detalhada	C	fraude, propriedade, abuso sexual, homicídio doloso, roubo, homicídio culposo, furto veicular, violência, ordem pública, assalto, furto, outras
Idade na soltura	C	18–25, 25–34, 35–44, 45–54, 55 ou mais
Admissão	A	1950–2014
Data de soltura	A	1971–2014
Soltura mandatória	A	1927–9997 [1]
Soltura projetada	A	1900–9997 [1]
Liberdade condicional[3]	A	1966–9997 [1]
Estado	C	41 estados distintos
Nível de escolaridade	C	$\varnothing$ [2]

Tipo I: Identificador Tipo C: Categórico Tipo A: Ano, numérico

Considerando as características selecionadas em nossa metodologia, a base de dados totaliza cerca de 8 milhões de instâncias completas. Nestas, temos recortes notáveis de 89% homens e 11% mulheres, 39,7% brancos e 39,4% negros, e 17,9% registros de reincidência. É importante notar que, apesar da proporção entre brancos e negros estar balanceada, ela não corresponde à totalidade da sociedade estadunidense na época, que constituia cerca de 80% brancos e 13% negros. [United States Department of Commerce 2012]

[1] Este campo aprenseta valores inconcebíveis, e portanto foi desconsiderado em nossas análises.
[2] Em todas instâncias da base de dados, este campo é aprensentado como faltante.
[3] Data de elegibilidade. A data de soltura corresponde a qualquer saída, condicional ou não.

3. Metodologia

Apresentamos um arcabouço de análise do conjunto de dados, [Bastos e Araújo 2020] incluindo uma nova implementação genérica e paralela do algoritmo *DCI-Closed* [Lucchese e Orlando 2004] para a mineração de *itemsets* fechados frequentes, bem como ferramental e métricas para descoberta de subgrupos. Nosso arcabouço é capaz de realizar recortes baseados em variantes de características relevantes na base de dados, e a mineração de padrões e subgrupos nestes recortes.

3.1. Mineração de *Itemsets* Frequentes

Com o objetivo de descobrir e investigar padrões notáveis no conjunto de dados, optamos por investir em uma metodologia de mineração de *itemsets* frequentes. A partir de um conjunto arbitrário de itens $\mathcal{I}$, define-se *itemsets* como subconjuntos de $\mathcal{I}$, e a base de dados $\mathcal{D}$ como uma lista de transações, onde cada transação constitui um *itemset*. Em seguida, define-se o suporte de um *itemset* como o número de transações que contém tal *itemset*:

$$\sup(i) = \left|\left[\, t \mid \forall\, t \in \mathcal{D} : i \subseteq t \,\right]\right|$$

De forma a reduzir o espaço de busca por *itemsets* e promover resultados mais palatáveis, o conceito de *itemsets* fechados é estabelecido. Um *itemset* fechado é um representante da sua classe de equivalência, não apresentando perda de informação em relação às definições anteriores, o que satisfaz nosso caso de uso de forma equivalente. Um *itemset* é fechado se, e somente se não há nenhum superconjunto deste *itemset* com o mesmo suporte: [Lucchese e Orlando 2004]

$$i \text{ é fechado} \iff \nexists\, j \supset i : \sup(j) = \sup(i)$$

Finalmente, define-se um valor arbitrário como suporte mínimo, o que permite a definição dos *itemsets* frequentes: aqueles cujo suporte é maior ou igual ao suporte mínimo. Os *itemsets* frequentes constituem o resultado do algoritmo de mineração, e podem fornecer uma melhor compreensão dos padrões e dinâmicas envolvidas na base de dados.

Para aplicar tal metodologia em nosso estudo, adaptamos nossa base de dados ao formato adequado, e escolhemos um algoritmo capaz de lidar com a dimensão da nossa aplicação. Apresentamos o seguinte método para mapear a base de dados para um conjunto de transações constituídas de itens:

1. Interpretamos cada registro como uma transação, o que implica na definição dos resultados como os *itemsets* frequentes nos registros prisionais.
2. Realizamos a codificação *one-hot* das características apresentadas na primeira seção da tabela 1, o que corresponde à interpretação de cada variante de cada característica como um item distinto, e um *itemset* como um conjunto destas variantes. As demais características apresentam dados redundantes, desbalanceados ou intratáveis, e portanto foram desconsideradas.

Optando pela mineração de *itemsets* fechados frequentes, o resultado desta metodologia são os conjuntos fechados frequentes de variantes nos registros prisionais. Tal informação é útil para análise das dinâmicas envolvidas no sistema carcerário estadunidense, e fornece uma base para estudos conseguintes na área das ciências sociais.

3.2. Descoberta de Subgrupos

A descoberta de subgrupos é uma técnica de mineração de dados que extrai regras interessantes em relação a uma variável de destino (*target*), combinando a indução preditiva e descritiva. Os padrões extraídos são normalmente representados na forma de regras, sendo chamados de subgrupos.

Desta forma, a descoberta de subgrupos se enquadra entre o aprendizado descritivo supervisionado e o não supervisionado, sendo considerada um ponto médio entre a extração de regras de associação e a obtenção de regras de classificação. Wrobel define esta metodologia da seguinte forma:

"Na descoberta de subgrupos, assumimos que recebemos uma chamada população de indivíduos (objetos, clientes, ...) e uma propriedade daqueles indivíduos nos quais estamos interessados. A tarefa de descoberta de subgrupos é, então, descobrir os subgrupos da população que são estatisticamente "mais interessantes", ou seja, são os maiores possíveis e têm as mais incomuns características estatísticas (distribucionais) com respeito à propriedade de interesse". [Wrobel 2001]

Uma regra R, que consiste em uma descrição de subgrupo induzida, pode ser formalmente definida como: [Gamberger e Lavrac 2011]

$$R : \text{Cond} \rightarrow \text{Target}_{\text{Value}}$$

onde *Target* é uma variável de interesse, e *Cond* é um conjunto de características que é capaz de descrever uma estatística incomum da distribuição em relação ao valor *Value*.

Considerando nossa base de dados, um exemplo de uma possível regra seria:

$$R_x : \text{liberdade condicional, revogação da condicional} \rightarrow \text{reincidente}$$

Aqui, caso o individuo esteja em liberade condicional, e esta seja revogada, ele é reincidente criminal.

Optamos por utilizar a biblioteca *pysubgroup* [Lemmerich e Becker 2018] como implementação para descoberta de subgrupos. Além de reportar os padrões descobertos, a mesma reporta ainda um *score* de qualidade, baseado na acurácia ponderada:

$$AP = \frac{|subgroup|}{|dataset|} \cdot (p_{subgroup} - p_{dataset})$$

onde $p_{subgroup}$ é o número de positivos no subgrupo, e $p_{dataset}$ é o número de positivos na base de dados.

Em nossa metodologia, optamos por utilizar o algoritmo *Beam Search* [Greenberg et al. 2018], que consiste numa busca heurística que explora os *k* estados de um grafo, ao invés de somente um. Em particular, difere da abordagem gulosa tradicional, que seleciona o melhor estado atual e descarta o restante. O *Beam Search* mantém o controle de *k* estados, selecionando o melhor resultado de dada largura. Não obstante, outros algoritmos fornecidos pela biblioteca também podem ser explorados.

4. Resultados

4.1. Mineração de *Itemsets* Frequentes

Considerando a grande quantidade de registros na base de dados, arbitramos um suporte mínimo de 5% para a mineração de *itemsets* frequentes, o que representa cerca de 400.000 registros. A mineração utilizando tal parâmetro produz diversos padrões frequentes, dos quais observamos duas principais categorias.

A primeira e mais frequente categoria, constitui padrões decorrentes da dinâmica legal dos sistemas jurídico e prisional estadunidense. Padrões como por exemplo $\{\text{sentença de 2 a 5 anos}, 0 \text{ a } 1 \text{ anos servidos}, \text{saída condicional}\}$ com suporte 21.9%, não constituem informação notável, pois são uma mera consequência das leis. Tal categoria de padrões é recorrente em nossos resultados, e deve ser identificada e desconsiderada por especialistas durante análise.

A segunda categoria é a que constitui o extrato de valor de nosso estudo. Padrões cuja ocorrência não aparenta decorrer de algum aparato legal foram observados, mas com a devida ressalva. O real significado destes padrões deve ser apontado por especialistas das ciências sociais em trabalhos futuros, e nossas observações são meros palpites e sugestões de padrões inusitados. Apresentamos algumas ocorrências destes padrões nas tabelas 2 e 3.

Tabela 2. Padrões com enfoque étnico

Variantes de características	Suporte	Variantes de características	Suporte
propriedade, negro	10.5%	negro, condicional	27.5%
propriedade, branco	14.6%	branco, condicional	28.5%
drogas, negro	14.0%	negro, incondicional	11.5%
drogas, branco	9.5%	branco, incondicional	10.3%
violência, negro	10.0%	nova admissão, negro, condicional	16.5%
violência, branco	8.5%	nova admissão, branco, condicional	18.4%
drogas, negro, 0 a 1 a. s.	8.5%	negro, 0 a 1 a. s., condicional	15.4%
drogas, branco, 0 a 1 a. s.	6.3%	branco, 0 a 1 a. s., condicional	17.1%
negro, 0 a 1 a. s.	22.5%	negro, 0 a 1 a. s., incondicional	6.8%
branco, 0 a 1 a. s.	23.8%	branco, 0 a 1 a. s., incondicional	6.2%
negro, 1 a 2 a. s.	7.7%	negro, entre 18 e 24 anos	10.3%
branco, 1 a 2 a. s.	8.0%	branco, entre 18 e 24 anos	8.2%
negro, 2 a 5 a. s.	6.5%	branco, entre 25 e 34 anos	13.9%
branco, 2 a 5 a. s.	5.7%	negro, entre 25 e 34 anos	13.4%
		branco, entre 35 e 44 anos	11.1%
a. s.: anos servidos		negro, entre 35 e 44 anos	10.0%

Na tabela 2, apresentamos alguns padrões que demonstram diferenças entre as etnias branca e negra, apesar da base de dados ser balanceada com 39,7% brancos e 39,4% negros. Primeiro, notamos uma distinção entre as tipificações dos crimes cometidos por

ambas etnias. Adiante, observamos padrões cuja ocorrência levanta a suspeita do racismo. Notavelmente, os negros apresentam menor participação relativa em benefícios como liberdade condicional e penas brandas. Além disso, na tabela 3 mostramos que negros são consideravelmente mais reincidentes que brancos.

Tabela 3. Padrões em reincidentes

Variantes de características	Suporte
negro	48.1%
branco	38.6%
entre 18 e 24 anos	14.3%
entre 25 e 34 anos	38.6%
entre 35 e 44 anos	29.9%
entre 45 e 54 anos	14.5%
sentença de 2 a 5 anos	39.3%
sentença de 2 a 5 anos, condicional	28.2%
sentença de 2 a 5 anos, 0 a 1 anos servidos	17.6%
sentença de 2 a 5 anos, 1 a 2 anos servidos	14.4%

Na tabela 3, observamos que a grande massa de reincidentes são pessoas com mais de 24 anos, ao contrário do que se observa no conjunto de dados completo, onde 24.1% possuem 24 anos ou menos. Em seguida, notamos que apesar das sentenças mais severas de 2 a 5 anos, grande parte dos reincidentes são bem sucedidos em obter liberdade condicional, servindo menos que 2 anos.

4.2. Descoberta de subgrupos

Da mesma forma que na mineração de *itemsets* frequentes, a descoberta de subgrupos produz diversos resultados, dos quais grande parte são uma consequência direta das leis. Novamente, cabe aos especialistas a detecção e filtragem destes subgrupos.

Tabela 4. Amostra dos subgrupos descobertos

Subgrupo	Tamanho
liberdade condicional, tempo servido $<$ 10 anos	2
liberdade condicional, retorno da condicional	2
liberdade condicional, categoria da ofensa $=$ outra	2
idade $<$ 55, liberdade condicional, etnia $\neq$ outra	3
idade $<$ 55, liberdade condicional, tempo servido $<$ 5	3
idade $<$ 55, liberdade condicional, tempo servido $<$ 10, homem	4
idade $<$ 55, liberdade condicional, tempo servido $<$ 10, sentença $\neq$ perpétua	4

Na tabela 4, apresentamos uma amostra dos subgrupos observados. O *score* de todas instâncias são próximos de 0,061. Notavelmente, os subgrupos compartilham a característica da liberadade condicional, acompanhadas de outras características distintas, como raça, tempo servido e sentença.

5. Conclusão

Apresentamos um arcabouço robusto de análise da base de dados, implementando metodologias de mineração de *itemsets* frequentes e descoberta de subgrupos. Através deste, fornecemos a facilidade de estudo sobre a base e os resultados obtidos. Os resultados são facilmente reprodutíveis, e a variação dos parâmetros também pode ser realizada, promovendo análises mais completas dos dados. Convidamos os leitores a explorar a base de dados utilizando nosso arcabouço.

Nossos resultados apresentam interessantes perspectivas, propiciando novas análises por especialistas. Dentre os pontos chave, observamos padrões que levantam a suspeita de racismo, além de padrões e subgrupos que indicam dinâmicas inusitadas para reincidentes, bem como liberdade condicional.

Para trabalhos futuros, sugerimos a análise de nossos resultados por especialistas das ciências sociais, de forma a agregar conclusões com um aspecto crítico bem fundamentado. Além disso, sugerimos a confirmação da validez dos padrões apresentados por meio de testes de significância estatística, confirmando sua representatividade mediante ao universo dos dados.

6. Trabalhos Relacionados

Compton explora como o crime contra a propriedade pode afetar o comportamento de equilíbrio geral estático e dinâmico de famílias e empresas. [Compton 2019] Em contraste com as literaturas a seguir, trata-se o crime como uma transferência, e não como uma produção doméstica.

Seguindo na linha de raciocínio sobre o sistema de Justiça, temos ainda o tema de justiça para adultos emergentes e abordagens adequadas à idade. [Perker et al. 2019] Este relatório examina as implicações da acusação automática no sistema de justiça criminal do estado de Illinois, de todos os jovens com 18 anos ou mais, da mesma maneira que acusa e condena pessoas de 40 ou 50 anos.

Em outra perspectiva, Chung apresenta um estudo sobre raça nos EUA no século *XX* [Chung 2019], onde examina o risco e a prevalência de prisão dentro das redes familiares de americanos negros e brancos durante o "boom da prisão" (1985-1995). Notavelmente, estima-se que o americano negro médio, nascido no auge do boom da prisão, experimentou a prisão de um parente pela primeira vez aos 7 anos, e aos 65 anos pertencerá a uma família em que pelo menos 1 em 7 parentes em idade produtiva já foram presos.

Adiante, Chung e Peter apresentam um estudo sobre prisão em massa e família extensa, [Chung e Peter 2018]. Em constraste ao estudo apresentado anteriormente, conclui que o americano branco médio experimenta a prisão de um parente a partir dos 39 anos, e aos 65 pertence a uma família na qual 1 em cada 20 parentes em idade produtiva já foi preso.

Tais trabalhos apresentam diversas perspectivas sobre a mesma base de dados carcerários utilizada em nosso estudo, decorrentes de análises por especialistas das ciências sociais. De forma alternativa, apresentamos novas possibilidades de estudo e compreensão para os especialistas.

Referências

[Apel et al. 2010] Apel, R., Blokland, A. A. J., Nieuwbeerta, P., e van Schellen, M. (2010). The impact of imprisonment on marriage and divorce: A risk set matching approach.

[Bales e Piquero 2012] Bales, W. D. e Piquero, A. R. (2012). Assessing the impact of imprisonment on recidivism. *Journal of Experimental Criminology*, 8(1):71–101.

[Bastos e Araújo 2020] Bastos, G. e Araújo, F. G. (2020). Arcabouço de análise descritiva da base de dados ICPSR–36404. https://github.com/gahag-ml/icpsr-36404-analysis.

[Chung 2019] Chung, P. H. (2019). Race and Family in 20th Century United States.

[Chung e Peter 2018] Chung, P. H. e Peter, H. (2018). Mass imprisonment and the extended family. *Sociological Science*, 5(15):335–360.

[Compton 2019] Compton, A. (2019). Decomposing the Societal Opportunity Costs of Property Crime. MPRA Paper 97002, University Library of Munich, Germany.

[Gamberger e Lavrac 2011] Gamberger, D. e Lavrac, N. (2011). Expert-guided subgroup discovery: Methodology and application. *CoRR*, abs/1106.4576.

[Greenberg et al. 2018] Greenberg, C. S., Monath, N., Kobren, A., Flaherty, P., McGregor, A., e McCallum, A. (2018). Compact representation of uncertainty in clustering. In *Proceedings of the 32nd International Conference on Neural Information Processing*.

[Lemmerich e Becker 2018] Lemmerich, F. e Becker, M. (2018). pysubgroup: Easy-to-use subgroup discovery in python. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 658–662.

[Lucchese e Orlando 2004] Lucchese, C. e Orlando, S. (2004). Dci closed: A fast and memory efficient algorithm to mine frequent closed itemsets. In *Proc. of the IEEE ICDM 2004 Workshop on Frequent Itemset Mining Implementations (FIMI'04)*.

[Perker et al. 2019] Perker, S. S., Chester, L. E. H., e Schiraldi, V. N. (2019). Emerging adult justice in illinois: Towards an age-appropriate approach.

[Plugge et al. 2009] Plugge, E. H., Foster, C. E., Yudkin, P. L., e Douglas, N. (2009). Cardiovascular disease risk factors and women prisoners in the UK: the impact of imprisonment. *Health Promotion International*, 24(4):334–343.

[Thomas 1973] Thomas, C. W. (1973). Prisonization or resocialization?: A study of external factors associated with the impact of imprisonment. *Journal of Research in Crime*.

[United States Department of Commerce 2012] United States Department of Commerce, Census Bureau. (2012). Population and housing unit estimates, national intercensal tables: 2000-2010. Acesso em 15 de outubro de 2020: https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2000-2010-national.html.

[United States Department of Justice 2016] United States Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. (2016). National corrections reporting program, 1991-2014: Selected variables. ICPSR–36404.

[Wrobel 2001] Wrobel, S. (2001). *Inductive Logic Programming for Knowledge Discovery in Databases*, pages 74–101. Springer Berlin Heidelberg, Berlin, Heidelberg.