Classificação diferencialmente privada e não discriminatória utilizando árvore de decisão

Maria L. M. Silva[1] **Javam C. Machado**[1]

[1]LSBD – Departamento de Computação – Universidade Federal do Ceará

{malu.maia, javam.machado}@lsbd.ufc.br

Resumo. *Devido ao crescimento do uso de tecnologias para classificações de indivíduos e ao surgimento de leis de proteção aos dados, a preservação da privacidade e a preocupação quanto à discriminação propagada por algoritmos de classificação se tornaram temas relevantes para sociedade. Enquanto o processo de classificação pode associar indivíduos às suas características sensíveis, tais indivíduos podem ser discriminados uma vez associados a essas características. Para mitigar o problema de classificações discriminatórias e preservação da privacidade de indivíduos, propomos um algoritmo de árvore de decisão, onde aplicamos mecanismos de privacidade diferencial e propriedades de fairness para assegurar, ao mesmo tempo, a privacidade e a não-discriminação de indivíduos com dados pessoais envolvidos no processo. Nossa proposta introduz, portanto, equidade ou fairness em um algoritmo de classificação baseado em árvores de decisão.*

Nível: Mestrado

Programa: Programa de Pós-Graduação em Ciência da Computação

Ingresso: Março de 2020

Previsão de término: Março de 2022

Etapas concluídas: Revisão bibliográfica e Definição do problema

Defesa da qualificação: 10 de Agosto de 2021

Lista de publicações:

- Top-k reverso com privacidade sobre os dados públicos de COVID-19 no Ceará. Silva, M., Machado, J., Chaves, I. – *SBBD 2020 - Short and Vision and Industrial Papers*
- Detectando Doença de Parkinson - Uma Comparação de Modelos de Aprendizagem de Máquina com Redução de Dimensionalidade Diferencialmente Privada. Filho, M., Silva, M., Barros, P., Machado, J., Mattos, C. – *SBBD 2020 - Short and Vision and Industrial Papers*
- Private Reverse Top-k Algorithms Applied on Public Data of COVID-19 in the State of Ceará. Silva, M., Machado, J., Chaves, I. – *JIDM Vol. 12 (2021)*

1. Introdução

A utilização de modelos de aprendizado de máquina é cada vez mais comum para a classificação de indivíduos, seja para um empréstimo no banco, para a contratação em uma empresa ou para o recebimento de algum benefício. Como dados muitas vezes tendenciosos são utilizados para o treinamento desses modelos, a propagação da discriminação reflete em classificações injustas para indivíduos pertencentes a classes minoritárias. Além disso, é importante que pessoas que submetem seus dados para algum processo classificatório também tenham sua privacidade preservada para evitar a associação de um indivíduo a uma característica que possa influenciar na saída do algoritmo de classificação de forma discriminatória.

Quantidades massivas de dados são coletadas e publicadas por instituições com finalidade de pesquisa, transparência ou oferta de serviços. Como consequência, leis de regulamentação tornaram-se necessárias para garantir que a identidade e informações sensíveis a indivíduos, cujos dados foram publicados, não fossem descobertas. Um indivíduo pode ser re-identificado através de ataques de ligação [Newcombe et al. 1959], através da comparação entre dois ou mais bancos de dados, e informações sensíveis podem ser utilizadas contra ele de forma abusiva. Por exemplo, quando um plano de saúde aumenta a mensalidade de um sujeito após descobrir que ele é fumante. Estudos indicam que o tabagismo aumenta as chances do desenvolvimento do câncer de pulmão. Assim, ao relacionar um indivíduo à condição fumante através de ataques de ligação, o plano de saúde obtém a informação de que ele, possivelmente, terá câncer.

A não-discriminação em classificações é necessária para evitar que um indivíduo apto para o recebimento de um benefício seja rejeitado, devido a um atributo irrelevante para o problema. Por exemplo, uma mulher negra com renda fixa e sem dívidas sendo rejeitada para um crédito bancário, enquanto um homem branco com as mesmas características é aprovado. Além disso, indivíduos cujos dados foram submetidos podem sofrer ataques de ligações executados pela própria entidade responsável pela classificação, influenciando na tomada de decisão. Um exemplo disso é o empregador descobrir que um candidato a uma vaga tem AIDS e não aceitá-lo em decorrência disso.

O problema que procuramos atacar consiste em construir um algoritmo privado que classifique indivíduos em um conjunto de dados de forma não-discriminatória. Este trabalho visa acrescentar ao modelo de árvore de decisão um novo critério: o *private fair information gain*. Tal critério assegura que indivíduos sejam rotulados de forma que nenhum atributo passível de discriminação prejudique a classificação, ao mesmo tempo que garante a privacidade dos sujeitos a serem classificados.

2. Fundamentação Teórica

2.1. Privacidade Diferencial

Privacidade diferencial [Dwork 2006] é uma propriedade de alguns algoritmos aleatórios, chamados mecanismos. O objetivo é garantir que a presença ou ausência de um indivíduo em um conjunto de dados seja indistinguível após a análise da saída de uma consulta, ou seja, a inclusão ou exclusão de dados não terá grandes efeitos no resultado final. Além disso, várias saídas de uma mesma consulta podem ser obtidas com probabilidade similar, para impossibilitar que um adversário seja capaz de aprender algo

que ainda não sabia acerca de um indivíduo.

Definição 2.1 (ε-Privacidade Diferencial)**.** Um mecanismo M satisfaz ε-privacidade diferencial, onde $\varepsilon \geq 0$ se, e somente se, para todos os conjuntos de dados vizinhos D e D', isto é, conjuntos de dados que diferem em apenas um registro, e $T \subseteq Range(M)$, que denota uma possível saída contida no conjuntos de todas as saídas possíveis do mecanismo M, temos:

$$Pr[M(D) \in T] \leq \exp(\varepsilon) \times Pr[M(D') \in T], \tag{1}$$

em que ε, conhecido como *budget*, representa o limite para perda de privacidade permitida de uma consulta.

Os mecanismos funcionam adicionando uma quantidade apropriada de ruído aleatório à respostas de consultas. A magnitude do ruído é escolhida através da sensibilidade da consulta, ou seja, a maior diferença que um indivíduo pode causar na resposta de uma consulta. Quanto maior o conjunto de dados, melhor o funcionamento do mecanismo.

Definição 2.2 (L1-sensibilidade)**.** Seja f uma função sobre o conjunto de dados D e R^d, um vetor de números reais com dimensão d, para $f : D \rightarrow R^d$, a L1-sensibilidade de f é definida por

$$\Delta f = \max_{D \simeq D'} ||f(D) - f(D')||_1, \tag{2}$$

onde $D \simeq D'$ denota que o conjunto de dados D é vizinho do conjunto de dados D'.

Definição 2.3 (Mecanismo de Laplace)**.** Seja f uma consulta com retorno numérico, o mecanismo de Laplace [Dwork 2006] é definido como

$$M_L(D, f, \varepsilon) = f(D) + X, \tag{3}$$

onde X é uma variável aleatória independente e identicamente distribuída que segue $Laplace(\frac{\Delta f}{\varepsilon})$.

2.2. *Fairness* em aprendizado de máquina

Sistemas de suporte à tomada de decisão têm intensamente utilizado algoritmos de classificação em ambiente de alocação de empregos [Elbassuoni et al. 2020] e em aplicações de concessão de crédito [Lee and Floridi 2021]. O objetivo de algoritmos de *fairness* em aprendizado de máquina é garantir que um modelo não reproduza a discriminação de pessoas em suas classificações. *Fairness* pode ser alcançada por meio da alteração de dados tendenciosos utilizados para o treinamento do algoritmo, ou aplicando restrições que devem ser atendidas para que não haja mudança de tratamento na classificação de pessoas que possuam alguma característica protegida. Assim, valores de atributos que não devem influenciar na tomada de decisão, como por exemplo raça, religião, nacionalidade, gênero e cor, têm tratamento específico no processo de classificação.

De acordo com [Barocas et al. 2017], a maioria dos critérios de *fairness* envolve propriedades das distribuições conjuntas de um atributo sensível A, um *label* Y e de um classificador ou pontuação R a uma instância, como mostra a Tabela 1. Na propriedade

de independência, todos os grupos devem ter a mesma probabilidade de aceitação. Na separação, é permitida a correlação entre o classificador e o atributo sensível, quando o *label* Y é levado em consideração. Por fim, na suficiência também é utilizada probabilidade condicional e o classificador está relacionado a característica sensível para prever o atributo $label$.

Independência	Separação	Suficiência
$R \perp A$	$R \perp A\|Y$	$Y \perp A\|R$

Tabela 1. Critérios de não-discriminação.

Na literatura, existem vários outros termos equivalentes a independência, como *demographic parity, statistical parity* e *group fairness* [Dwork et al. 2012]. Para classificações binárias onde 1 representa "aceitação" e 0 "negação" de um benefício, e para todos os grupos a e b, as propriedades podem ser representadas pelas condições abaixo.

- **Independência.** Um classificador satisfaz independência se a característica sensível é estatisticamente independente da pontuação da instância.

$$P\{R = 1|A = a\} = P\{R = 1|A = b\}, \tag{4}$$

- **Separação.** Um classificador R satisfaz separação se a taxa de verdadeiros positivos for a mesma para todos os grupos e a taxa de falsos positivos também for igual para todos os grupos.

$$\begin{aligned} P\{R = 1|Y = 1, A = a\} &= P\{R = 1|Y = 1, A = b\}, \\ P\{R = 1|Y = 0, A = a\} &= P\{R = 1|Y = 0, A = b\}. \end{aligned} \tag{5}$$

- **Suficiência.** Uma variável aleatória R satisfaz suficiência para A, se e somente se, a característica *label* for independente do atributo sensível, dado um classificador.

$$P(Y = 1|R = r, A = a) = P(Y = 1|R = r, A = b), \tag{6}$$

$\forall r$, onde $r \in supp(R)$ e $supp(R)$ representa o suporte de R.

3. Trabalhos Relacionados

[Dwork et al. 2012] propõem um *framework* que trata o problema de classificações injustas através de um modelo de otimização linear, que utiliza uma métrica de distância para definir a similaridade entre indivíduos. Esta abordagem trata o que chamamos de *individual fairness*, em que indivíduos que possuem habilidades semelhantes para uma tarefa são classificados de forma semelhante. Além disso, a definição de *fairness* do artigo é uma generalização da noção de privacidade diferencial. Todos esses princípios têm como ligação a condição de *Lipschitz* sobre um classificador aleatório, que mapeia indivíduos a distribuições sobre as saídas. Por se tratar de um problema de otimização, sempre será encontrado o ótimo global, ou seja, o classificador que maximiza a não discriminação entre indivíduos. O problema de modelos desse tipo é o custo computacional que pode ser muito elevado considerando um grande conjunto de dados.

[Xu et al. 2019] ativam *fairness* e privacidade diferencial em regressão logística. Para *fairness* é utilizado o conceito de independência, visto na Seção 2.2. Para ativar

privacidade diferencial é utilizado um método adequado para modelos de otimização, o mecanismo funcional, que aplica ruído de Laplace aos coeficientes polinomiais da função objetivo. Os autores propõem duas contribuições, (i) PFLR que ativa *fairness* através de uma penalidade aplicada à função objetivo e (ii) PFLR* que ativa *fairness* no próprio mecanismo funcional, reduzindo a quantidade de ruído injetado e melhorando a acurácia do modelo. As soluções utilizam um mecanismo ideal para a forma polinomial da função logística, ao mesmo tempo que *fairness* é aplicada como um deslocamento da média de Laplace utilizada no mecanismo funcional, ativando assim privacidade e *fairness* simultaneamente. A acurácia do modelo PFLR* é superior ao PFLR, porém inferior ao modelo original de regressão logística.

[Pujol et al. 2020] apresentam os impactos dos algoritmos de privacidade diferencial nas tomadas de decisão de alocação. O artigo trata três abordagens e soluções para mitigar *unfairness*: (i) direito de voto para línguas minoritárias, ou seja, de pessoas que falam outras línguas e não estão habituadas à língua do país em que residem, onde a jurisdição local deve providenciar assistência linguística durante eleições, (ii) alocação de fundos para educação e (iii) distribuição de representantes legislativos. O objetivo das soluções é tratar populações semelhantes de forma semelhante. Para privacidade é adicionado ruído nas contagens referentes a cada abordagem. Todos os problemas descritos no artigo tratam de alocação de recursos, mas para cada situação é indicada uma solução. Apesar do trabalho atingir o objetivo de ativar privacidade e *fairness*, as soluções não são generalizadas para todas as abordagens, pois são focadas em situações políticas específicas.

A Tabela 2 mostra uma breve comparação entre este trabalho e os trabalhos relacionados. A ativação simultânea diz respeito a privacidade e *fairness* sendo ativadas na mesma etapa. Nessa característica, os dois primeiros trabalhos utilizam modelos de otimização, onde *fairness* é dada como critério, enquanto que nós propomos ativar ambas na etapa de treinamento do modelo. A literatura define dois principais tipos de *fairness*, (1) individual [Dwork et al. 2012] em que indivíduos semelhantes devem ser tratados de forma semelhante, e (2) grupal quando grupos diferentes têm a mesma chance de receber ou não um benefício, como explicado na Seção 2.2.

[Dwork et al. 2012] propôs um modelo generalizado, que serve para qualquer modelo de classificação, porém a ativação da privacidade é uma redução do modelo de *fairness*, [Xu et al. 2019] fez um modelo justo e privado de regressão logística, enquanto [Pujol et al. 2020] não utilizou nenhum modelo específico e sim conceitos estatísticos. Nosso trabalho visa construir um modelo justo e privado de árvore de decisão com altos níveis de acurácia, justiça e privacidade, equilibrando o *trade-off* existente entre eles.

Artigo	**Ativação Simultânea**	**Tipo de *Fairness***	**Modelo de classif.**
[Dwork et al. 2012]	✓	Individual	Generalizado
[Xu et al. 2019]	✓	Grupal	Regressão Logística
[Pujol et al. 2020]	x	Grupal	-
Este trabalho	✓	Individual	Árvore de Decisão

Tabela 2. Tabela comparativa entre trabalhos relacionados.

4. Metodologia

A metodologia empregada nessa pesquisa para atacar o problema científico passou pelo estudo conceitual de privacidade e de *fairness*, além da compreensão da estratégia de classificação por árvore de decisão. No passo seguinte, fizemos um levantamento do estado da arte em busca de outros trabalhos que tratam conjuntamente privacidade e *fairness* em algoritmos de classificação. Definimos então uma estratégia de ataque onde modificamos o processo de classificação com adição de ruído de Laplace e ativação de *fairness* como detalhado abaixo.

Antes das etapas do algoritmo responsáveis pela classificação, ativação de *fairness* e privacidade, é necessário o pré-processamento dos dados. Os campos numéricos devem ser normalizados, para que atributos com uma grande amplitude de valores não influenciem mais do que outros na predição. Além disso, transformamos os atributos categóricos através do *one-hot encoding*, para conseguirmos calcular a distância entre dois indivíduos, que define a similaridade entre eles. Indivíduos similares devem ser classificados da mesma forma, a fim de garantir a *individual fairness* [Dwork et al. 2012].

Após a etapa de pré-processamento, iniciamos a etapa de treinamento do modelo de árvore de decisão justa e privada. Este modelo possui algumas alterações em relação ao original [Quinlan 1986], onde o critério de *split* utilizado é o *information gain* com algumas modificações. Para a garantia de privacidade, aplicamos ruído de Laplace sobre a entropia, que é utilizada para o cálculo do *information gain*, e também sobre a contagem das classes retornadas pelos nós folhas, utilizadas para predição. Para ativar *fairness*, utilizamos o conceito de similaridade, em que um indivíduo x é similar a um indivíduo y se $l_\infty(x, y) \leq \tau$, sendo τ um limite de justiça.

- Dividimos o *budget* de privacidade em duas partes, utilizando o mecanismo de Laplace: ε_1 e ε_2, onde ε_1 é destinado a privacidade dos nós-não folhas, necessários para o cálculo da entropia, e ε_2 destinado às contagens das classificações dos nós folhas, responsáveis pela predição;
- Adicionamos ruído aos nós não-folhas, ε_1 é distribuído para cada nível da árvore;
- Adicionamos ruído às contagens referentes as classificações dos nós folhas, responsáveis pela tomada de decisão;
- Na entropia medimos a pureza de um conjunto medida pela proporção de valores a que um atributo A pode assumir em relação a todo o conjunto. Para ativar *fairness*, a proporção de valores sofre uma pequena alteração e considera a proporção de valores para indivíduos similares. Se um indivíduo 1 é similar a um indivíduo 2, os valores de atributos de 1 e 2 pertencerão à mesma proporção.

Para avaliarmos nosso modelo, utilizaremos conjuntos de dados conhecidos por serem tendenciosos: (i) o conjunto de dados *Adult* [Dua and Graff 2017], extraído do Censo de 1994 dos Estados Unidos, que associa uma renda a um indivíduo baseando-se em atributos pessoais; e (ii) o conjunto de dados COMPAS [Angwin et al. 2016], que contém dados coletados no uso da ferramenta de avaliação de risco COMPAS do Condado de Broward, na Flórida, onde cada amostra prevê o risco de reincidência de indivíduos com base em atributos pessoais e histórico criminal. As métricas utilizadas serão acurácia e a métrica de *fairness* apresentada em [Ranzato et al. 2021], que é a razão entre a quantidade de vezes que o classificador rotulou os indivíduos similares da mesma forma e o tamanho do conjunto de teste.

Em suma, visamos minimizar a discriminação das classificações enquanto garantimos a privacidade dos indivíduos. Por fim, avaliaremos os resultados e as possíveis melhorias que podem ser feitas através da mudança das métricas ou mecanismos de privacidade.

5. Conclusão

Revisamos definições básicas das áreas relacionadas e trabalhos que envolvem soluções para os problemas de privacidade de dados e *fairness* simultaneamente, além de propormos uma solução para o problema de classificações não privadas, tendenciosas e discriminatórias, utilizando árvore de decisão. Os próximos passos da pesquisa envolvem a análise dos resultados experimentais, avaliando a acurácia dos modelos de classificação quanto à privacidade e à não-discriminação.

A avaliação poderá gerar revisão na estratégia de ataque bem como no ajuste de parâmetros tanto na adição de ruído como no cálculo da entropia. Buscaremos uma solução com bons níveis de acurácia, preservando a identidade dos indivíduos a serem classificados, garantindo que atributos protegidos não influenciem na tomada de decisão.

Referências

Angwin, J., Larson, J., Mattu, S., , and Kirchner, L. (2016). Machine bias, *ProPublica*.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C. (2006). Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Elbassuoni, S., Amer-Yahia, S., and Ghizzawi, A. (2020). Fairness of scoring in online job marketplaces. *Trans. Data Sci.*, 1(4):29:1–29:30.

Lee, M. S. A. and Floridi, L. (2021). Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds Mach.*, 31(1):165–191.

Newcombe, H. B., Kennedy, J. M., Axford, S., and James, A. P. (1959). Automatic linkage of vital records. *Science*, 130(3381):954–959.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Ranzato, F., Urban, C., and Zanella, M. (2021). Fair training of decision tree classifiers. *arXiv preprint arXiv:2101.00909*.

Xu, D., Yuan, S., and Wu, X. (2019). Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 594–599.