Uma Abordagem Dinâmica para Anonimização de Dados de Saúde por Separatrizes

Kristtopher K. Coelho[1], Maurício M. Okuyama[1], Michele Nogueira[2], Alex Borges Vieira[3], Edelberto Franco Silva[3], José Augusto M. Nacif[1]

[1]Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa (UFV - Campus Florestal)

[2]Departamento de Ciência da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)

[3]Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Federal de Juiz de Fora (UFJF)

{kristtopher.coelho,mauricio.okuyama,jnacif}@ufv.br

michele@dcc.ufmg.br,alex.borges@ufjf.edu.br,edelberto@ice.ufjf.br

Abstract. *Technological advances enable the integration of Internet of Things (IoT) devices to perform continuous and proactive patient monitoring. These devices collect a large volume of sensitive data that requires privacy. Anonymization provides privacy by removing or modifying information that identifies an individual. However, traditional anonymization techniques, such as k-anonymity, depend on a fixed and pre-defined k value, susceptible to attribute identification attacks. This article presents Dynamic Anonymization by Separatrices (DAS), an approach for defining the ideal value k, and for dynamic grouping of data to be anonymized using separatrices measurements. Results show that the proposed approach efficiently mitigates attribute identification attacks.*

Resumo. *Os avanços tecnológicos possibilitam a integração de dispositivos da Internet das Coisas (IoT) para realizar o monitoramento contínuo e proativo de pacientes. Esses dispositivos coletam um grande volume de dados, sendo muitos desses dados sensíveis, exigindo privacidade. A anonimização oferece privacidade ao remover ou modificar informações que identifiquem um indivíduo. Entretanto, as técnicas de anonimização tradicionais, tais como o k-anonimato, são dependentes de um valor k fixo e pré-definido, sendo suscetíveis a ataques de identificação de atributos. Este artigo apresenta a Anonimização Dinâmica por Separatriz (Dynamic Anonymization by Separatrices – DAS), uma abordagem para definição do valor ideal k e para o agrupamento dinâmica dos dados a serem anonimizados usando medidas de separatrizes. Os resultados mostram que a abordagem proposta é eficiente para mitigar ataques de identificação de atributos.*

1. Introdução

Nos últimos anos houve um crescimento maciço na adoção de tecnologias da informação para cuidados da saúde (*Healthcare Information Technology* – HIT). O avanço tecnológico e a miniaturização dos dispositivos (vestíveis e implantáveis) da Internet das Coisas

(IoT), quando aplicados à saúde (*Internet of Health Things* – IoHT), proporcionaram um aumento expressivo na geração de dados de saúde [Ketu and Mishra 2021, Shahid et al. 2022,Olatunji et al. 2022]. As análises de *big data* de saúde permitem melhorar a precisão dos diagnósticos e as tomadas de decisões clínicas, além de promover o monitoramento remoto de eventos adversos, reduzir custos e otimizar o tratamento de doenças [Fernandes et al. 2012, Abouelmehdi et al. 2018, Batko and Ślęzak 2022]. No entanto, essas análises ocorrem sobre dados sensíveis que requerem privacidade por lei, ao passo que também precisam ser compartilhados entre instituições.

Os métodos de privacidade tradicionais, como a criptografia, autenticação e esquemas biométricos permitem compartilhar informações e dados de saúde fornecendo alto nível de segurança e privacidade [Coelho et al. 2023]. Entretanto, tais métodos são computacionalmente caros para integrar todos os dispositivos IoHT. Em contrapartida, as técnicas de anonimização transformam os registros de conjunto de dados em dados genéricos e indistinguíveis, utilizando operações simples e computacionalmente eficientes [Onesimu et al. 2022]. Destacam-se as abordagens baseadas em k-anonimato, as quais garantem que, no conjunto de dados sempre haverá pelo menos k indivíduos com a mesma combinação de atributos quase-identificadores. As abordagens complementares como l-diversidade e t-proximidade trazem maior diversidade aos grupos de dados semelhantes, coibindo a divulgação de identidade. No entanto, ainda são suscetíveis a ataques de divulgação de atributos. Além disso, a literatura ainda não convergiu para definir um valor ideal de k, o que inviabiliza a aplicação em dados privados de saúde [Torra and Navarro-Arribas 2023].

Na literatura, destacam-se as abordagens baseadas em agrupamento para solucionar o vazamento e a exposição de informações privadas de dados de saúde através do uso de técnicas de anonimização. Entre elas, a proposta θ–Sensitive [Khan et al. 2020] adiciona ruído às tuplas de dados anonimizados de modo a aumentar a variedade entre as classes de equivalência. [Liu and Li 2018] apresentam um trabalho no qual os grupos de quase-identificadores sofrem generalização e supressão. O k-anonimato adaptativo [Arava and Lingamgunta 2020] adota métodos de agrupamento de alta complexidade computacional (k-*member*, *C-means*, *One-Pass* k-*mean*–OKA e *Efficient Systematic Clustering*–ESC) para anonimizar os dados confidenciais. Em [Onesimu et al. 2022], os autores propõem anonimizar atributos numéricos com uma abordagem de intervalo fixo, em que os valores originais dos dados de saúde são substituídos por um valor equivalente calculado. Esta última necessita que o valor de k seja previamente conhecido, o que reduz a eficiência da privacidade no caso de escolhas equivocadas.

Este artigo apresenta a anonimização dinâmica por separatriz (*Dynamic Anonymization by Separatrices* – DAS), uma nova abordagem para anonimização de dados baseada em grupos ordenados em partes definidas por k-percentil. Ela é capaz de modificar informações que identifiquem uma pessoa (atributos quase-identificadores numéricos) para garantir a preservação de privacidade e maior fidelidade em relação aos dados puros. A abordagem é dinâmica, pois define o valor ideal para a quantidade de grupos (k) utilizando o método estatístico conhecido como "método cotovelo" ou método de Elbow. Em seguida, os grupos são delimitados pelos respectivos percentis. Todos os valores de um agrupamento pertencente a um k-percentil são generalizados, substituindo-os pela respectiva média de valores deste intervalo. Em resumo, a abordagem DAS possui duas

contribuições principais. Primeiro, ela define o valor ideal para a quantidade de grupos (k). Segundo, ela realiza a anonimização de atributos numéricos com baixo custo computacional e agrupamentos definidos por separatrizes. A definição dinâmica do valor ideal de k e a aplicação da anonimização por separatrizes em cada tipo de atributo numérico proporcionam maior diversidade/heterogeneidade entre as classes de equivalências dos atributos.

A anonimização realizada pelo DAS não apenas fornece dados mais realistas, úteis [Ghinita et al. 2007, Ayala-Rivera et al. 2014], mas também evita que os dados sejam divulgados. Portanto, a avaliação foca na proteção oferecida pela abordagem DAS contra a divulgação de atributos considerando duas bases de dados distintas, sendo que uma delas é composta exclusivamente por dados médicos. Além disso, a avaliação da perda de informação das técnicas é feita por meio da métrica *Normalized Certainty Penalty* (NCP). A comparação de desempenho é confrontada com outras propostas baseadas em k-anonimato e anonimização por intervalo fixo [Onesimu et al. 2022]. Ademais, é apresentada a análise de complexidade computacional da abordagem DAS. Os resultados indicam que a abordagem DAS é mais eficiente contra ataques de divulgação de atributos quando comparadas a propostas relevantes da literatura. Para contribuir com o desenvolvimento de pesquisas futuras e garantir a reprodutibilidade dos resultados, o código fonte e as bases de dados encontram-se disponíveis publicamente [1].

O artigo prossegue como segue. A Seção 2 explora os trabalhos relacionados. Na Seção 3, detalha-se a abordagem DAS. A Seção 4 descreve a metodologia de avaliação, incluindo a descrição e discussão dos resultados. Por fim, a Seção 5 conclui o artigo.

2. Trabalhos Relacionados

O amplo desenvolvimento e a implantação de dispositivos IoT na área da saúde ameaçam a proteção dos dados sensíveis. Diversas pesquisas tratam o vazamento de dados sensíveis aplicando técnicas de baixo custo computacional. Por exemplo, para garantir privacidade, Ouazzani e Bakkali [El Ouazzani and El Bakkali 2018] propuseram um algoritmo para k-anonimato sem conhecimento prévio do valor máximo k. O algoritmo agrupa iterativamente as combinações idênticas em relação aos atributos quase-identificadores escolhidos, até que todos os indivíduos do mesmo grupo possuam dados com as mesmas características. Apesar de focar em grandes volumes de dados, a proposta é avaliada para uma tabela de teste de atributos quase identificadores fictícios, contendo nove linhas com três quase-identificadores. Isto resultou em um $k = 3$. Apesar deste trabalho não aprofundar a discussão sobre os resultados, os autores acompanham a literatura inferindo que um valor de k baixo implica em menor privacidade.

O método de k-anonimato adaptativo (AKA) [Arava and Lingamgunta 2020] promete aplicação para serviços de saúde em nuvem. Como o serviço de nuvem é cobrado conforme o uso, refazer a anonimização para obter *clusters* ideais e aliados a perda mínima de informações é custoso. AKA adota métodos estabelecidos de agrupamento (k-*member*, *C-means*, *One-Pass* k*-mean* – OKA e *Efficient Systematic Clustering*–ESC) para anonimizar os dados confidenciais. Porém, esses grupos possuem alto custo computacional. Além disso, é um desafio encontrar um bom valor inicial de k escolhendo de forma

[1] https://github.com/mauriciokuyama/das

aleatória. Portanto, os autores se baseiam no *Enhanced Clustering Method* – KOC para definir o valor k no k-anonimato.

A proposta θ–Sensitive [Khan et al. 2020] mitiga o ataque de variância sensível e o ataque de similaridade categórica. A abordagem proposta obtém o valor de θ multiplicando a variância (σ^2) de uma classe de equivalência diversa por um valor observado (μ). Além disso, se a privacidade desejada não for alcançada, acrescenta pequenas quantidades de tupla(s) de ruído para aumentar a variabilidade em uma classe de equivalência. Esta solução impede o risco de divulgação de atributos, ampliando a privacidade do conjunto de dados considerado. Entretanto, os parâmetros, como μ e a quantidade de ruído, interferem significativamente nos níveis de privacidade e utilidade dos dados.

[Onesimu et al. 2022] apresentam um esquema de publicação de dados com preservação de privacidade com foco em atributos numéricos e categóricos. Para atributos numéricos, eles adotam a abordagem de intervalo fixo, na qual os valores originais dos dados de saúde são substituídos por um valor equivalente calculado. Os atributos categóricos são protegidos por fatiamento l-diverso dos dados, horizontal e verticalmente, generalizando-os para evitar vazamentos de privacidade. Esta abordagem exige que o valor de k seja fornecido previamente, o que pode interferir diretamente na eficiência da privacidade. Além disso, o fatiamento aumenta a perda de informações.

Um outro método de anonimização baseado em agrupamento atribui quase-identificadores semelhantes a um mesmo grupo para preservação da privacidade dos dados coletados por dispositivos vestíveis [Liu and Li 2018]. Os autores unificam os quase-identificadores nos mesmos grupos por meio de generalização e supressão. Todos os registros no mesmo conjunto equivalente são semelhantes entre si. Além dos modelos recentes de privacidade baseados em k-anonimato , é importante discutir suas respectivas fragilidades quanto a ataques de divulgação de atributos. Em [Torra and Navarro-Arribas 2023], os autores mostram que as técnicas MDAV [Domingo-Ferrer and Mateo-Sanz 2002] e Mondrian [LeFevre et al. 2005a] não apresentam diferenças fundamentais entre si e não são imunes ao risco de divulgação de atributos. MDAV [Templ 2008] é um método heurístico de micro-agregação multivariada para espaços n-dimensionais (>1). Já o Mondrian [LeFevre et al. 2005b] consiste na divisão recursiva de baixo para cima (*bottom-up*) da base de dados em duas até que cada *cluster* tenha entre k e $2k-1$ registros, também com valor de k definido pelo usuário.

Portanto, com o objetivo de avançar o estado da arte para anonimização de dados de saúde, em especial para dispositivos com poder computacional limitados, este artigo apresenta uma nova abordagem baseada em anonimização dinâmica por separatrizes. O DAS utiliza o método de Elbow [Bholowalia and Kumar 2014] como base para encontrar o melhor tamanho de agrupamento k de modo a reduzir o risco de ataques. Consequentemente, cada agrupamento é composto pelos quase-identificadores anônimos referentes ao respectivo k-percentil. Para assegurar a robustez e a eficiência da proposta acerca da privacidade dos dados, a proposta foi submetida à avaliação de risco de divulgação de atributos [Torra and Navarro-Arribas 2023] e quanto a avaliação de perda de informação técnica dos dados [Ghinita et al. 2007, Ayala-Rivera et al. 2014].

3. Anonimização Dinâmica por Separatriz

A abordagem DAS toma como base a construção por separatrizes. As separatrizes são valores que ocupam determinados lugares de uma distribuição de frequência [Correa 2003]. Podemos classificá-las de acordo com o número de partes iguais em que os dados são particionados. Dessa forma tem-se, por exemplo, quartis (4 partes), decis (10 partes) e percentis (100 partes) [Correa 2003]. Nesse sentido, as separatrizes permitem dividir uma distribuição em n partes iguais.

A abordagem DAS trata o problema de escolha do valor de k, uma vez que este é escolhido dinamicamente. Testar diversos valores de k é uma solução cara e escolher um valor arbitrário de k representa um risco significativo quanto à divulgação de atributos. Os exemplos na literatura indicam um valor de $k = 3$. Entretanto, alguns autores [Victor and Lopez 2020, Torra and Navarro-Arribas 2023] sugerem o valor de $k = 5$. Porém, quanto maior o k, menor o risco, uma consequência natural da dimensionalidade no agrupamento. Portanto, aliada à escolha dinâmica, a abordagem DAS contribui para garantir a eficiência quanto a privacidade dos dados, aumentando o valor associado a k. Além da eficiência inerente à privacidade, a abordagem DAS tem uma construção de agrupamento anônimo de dados com baixo custo computacional, o que lhe permite operar em *hardware* com recursos limitados.

No contexto de privacidade de dados, os atributos de um conjunto de dados são categorizados em identificadores, quase-identificadores e atributos sensíveis. Os *identificadores* são atributos que identificam unicamente um indivíduo, tais como nome e CPF. Nesse sentido, para proteger a privacidade dos pacientes, esses atributos devem ser removidos do conjunto de dados [Onesimu et al. 2022]. Os *quase-identificadores (QIs)* consistem em atributos que, isoladamente, não identificam um indivíduo, mas podem revelar sua identidade quando combinados [Olatunji et al. 2022]. Exemplos de quase-identificadores incluem características tais como gênero, idade e CEP. Por último, os *atributos sensíveis* referem-se a informações que se tornam sensíveis quando vinculadas a um indivíduo [Olatunji et al. 2022]. No contexto de aplicações em saúde, citam-se os medicamentos, as condições clínicas e os sinais fisiológicos.

Espera-se que os atributos QIs sejam fornecidos em uma estrutura de dados tabular potencialmente heterogênea com eixos (linhas e colunas) rotulados (índices e números). Além da categorização em termos de privacidade, os atributos também podem ser classificados em numéricos ou categóricos. Os *atributos numéricos* são representados por valores contínuos, enquanto os *atributos categóricos*, um caso especial do tipo discreto, têm apenas um número finito de valores [Dinh et al. 2021]. Em dados médicos, por exemplo, encontramos valores categóricos como nacionalidade, gênero e educação, assim como valores numéricos como idade, altura e peso [Dinh et al. 2021].

A abordagem DAS aplica o método de Elbow que é tradicionalmente utilizado para encontrar a quantidade ideal de *clusters* em técnicas de agrupamento, e este pode ser estendido à anonimização. Portanto, o método de Elbow é responsável por definir o valor ideal de k dinamicamente, o qual define o respectivo k-percentil conforme a distribuição dos atributos. O método de Elbow [Bholowalia and Kumar 2014], é um método gráfico utilizado para determinar um valor adequado para o número de *clusters* k, de forma que a adição de outro *cluster* não reduza significativamente a função de custo a ser minimizada. A ideia é começar com $k = 2$ e incrementá-lo em cada etapa, calculando o custo para

cada valor. O custo é baseado no cálculo da distorção, ou seja, a média das distâncias quadradas dos centros dos respectivos *clusters* até cada ponto de dados. Normalmente, a métrica de distância euclidiana é usada. Em algum ponto, o custo diminui drasticamente e depois se estabiliza para valores maiores [Kodinariya et al. 2013]. Este ponto representa o valor ótimo de k, que pode ser detectado automaticamente utilizando algoritmos como o *Kneedle* [Satopaa et al. 2011]. Embora existam outras abordagens na literatura para determinar o número ideal de *clusters*, o método de Elbow se destaca por seu tempo de execução menor, comparado com os outros métodos da literatura (*Gap Statistic*, *Silhouette Coefficient* e *Canopy*) [Yuan and Yang 2019]. O valor ótimo de k define as medidas de separatrizes, as quais ocupam determinados lugares de uma distribuição de frequência, delimitando assim os k grupos de atributos com características a serem generalizadas. A Figura 1 ilustra a distribuição em quartis.

Figura 1. Representação da divisão dos dados em medidas de separatrizes (quartil).

O Algoritmo 1 descreve o conjunto de instruções necessárias para realizar a anonimização dinâmica dos atributos numéricos usando separatrizes. O valor de k o qual será utilizado como limiar para definir os percentis, é definido individualmente para cada QI. Em seguida, o algoritmo processa cada QI de forma ordenada. Para obter cada valor separatriz, calcula-se o q-ésimo percentil dos dados ao longo do eixo (QI) especificado. A abordagem DAS utiliza o parâmetro "*closest_observation*", o qual estima a observação mais próxima de um valor ideal para quantidade de percentis [Hyndman and Fan 1996, Developers 2024].

O algoritmo processa cada quase-identificador individualmente. Na linha 2, o método de Elbow define o valor ideal de k (k-percentil) para o QI. Na Linha 3, o algoritmo ordena o QI de de forma crescente. Para todos os registros pertencentes ao QI, na Linha 6, calcula-se o valor do q-ésimo percentil. Em seguida na Linha 7, encontra-se o maior índice que corresponde ao respectivo q-ésimo percentil, delimitando assim a faixa de amostras as quais serão anonimizadas por generalização. Na Linha 8, calcula-se o valor médio das amostras pertencente ao respectivo q-ésimo percentil. Posteriormente, na Linha 9, todas as amostras pertencentes ao intervalo são substituídas pelo valor médio calculado no passo anterior. Na Linha 12, os valores anonimizados dinamicamente por agrupamento baseado em q-ésimo percentil são atribuídos à tabela de $QIs\_anonimos$. Este fluxo se repete até processar todos os QIs. Por fim, o algoritmo retorna a tabela anonimizada $QIs\_anonimos$, na linha 14.

A Tabela 1 ilustra como são representados os dados brutos, os quais serão tratados pela abordagem DAS. Já a Tabela 2, apresenta os respectivos dados anonimizados pela abordagem proposta. Neste exemplo, considerando $k = 3$, os QIs são anonimizados na ordem em que aparecem (idade, altura em centímetros e peso em quilograma). Observa-se pelos índices (ID) 1 e 4 da Tabela 2 que o algoritmo não satisfaz as condições do k-anonimato para o valor $k = 3$, definido por Elbow. Neste caso, para cada um dos registros

Algoritmo 1 Anonimização Dinâmica por Separatriz para atributos numéricos.

Entrada: QIs
Saída: QIs_anonimos
início
1 **para cada** QI ***em*** QIs **faça**
2 k = **Elbow** (**Unicos**(QI));
3 QI = **Ordena**(QI);
4 id_inicio = 0;
5 **para** $i \leftarrow 1$ ***até*** $k + 1$ **faça**
6 separatriz = **Percentil**(QI, ((100 / k) * i), metodo="closest_observation");
7 id_fim = **Maior**(onde(QI == separatriz));
8 QI_media = **Media**(QI, id_inicio, id_fim);
9 QI = **Substitui_media**(QI, QI_media, id_inicio, id_fim);
10 id_inicio = id_fim + 1;
11 **fim**
12 **Inserir**(QIs_anonimos, QI)
13 **fim**
14 **retorna** *QIs_anonimos*
fim

da tabela, não existem pelo menos $k - 1$ outros registros com valores idênticos de QIs. Entretanto, por não exigir a condição do k-anonimato, o DAS gera maior heterogeneidade entre os dados em respectivos agrupamentos por percentil e, portanto, menor perda de informação útil.

Tabela 1. Dados brutos

ID	Idade	Altura (cm)	Peso (Kg)
0	21	160	50,55
1	24	154	60,60
2	25	158	48,80
3	30	170	76,80
4	34	169	54,70
5	33	176	67,90
6	38	183	79,00
7	41	190	80,60
8	39	180	83,10

Tabela 2. Dados anonimizados

ID	Idade	Altura (cm)	Peso (Kg)
2	23	157	51,35
0	23	157	51,35
4	32	171	51,35
1	23	157	68,43
5	32	171	68,43
3	32	171	68,43
6	39	184	80,90
7	39	184	80,90
8	39	184	80,90

4. Avaliação

Esta seção descreve a metodologia de avaliação da abordagem DAS. Serão apresentadas as especificações das bases dados. Também, são apresentadas as métricas de avaliação e detalhadas as configurações da máquina utilizada para execução dos experimentos. Por fim, são divulgados e discutidos os resultados obtidos.

4.1. Base de Dados

Adult [Becker and Kohavi 1996] é uma base de dados construída para prever se o salário de um indivíduo ultrapassa $50 mil por ano, baseado em dados de senso, incluindo quase-identificadores como idade, sexo, educação, ocupação e raça; além do atributo sensível salário. Ela é utilizada para avaliação na maioria das propostas para métodos de k-anonimato [Bache and Lichman 2013, Khan et al. 2020, Onesimu et al. 2022, Torra and Navarro-Arribas 2023, Byun et al. 2007], inclusive em cenário de aplicação IoHT [Arava and Lingamgunta 2020]. A base de dados Adult é composta por 48.842 registros com 14 atributos, onde 7 são categóricos e 7 numéricos. Como o DAS destina-se a anonimizar dados numéricos, apenas estes foram considerados neste artigo. Os dados foram pré-processados para remover as entradas com atributos ausentes. Assim, o arquivo final contém 30.162 registros.

Também foi considerada a avaliação da técnica proposta em dados reais de saúde, comprovando assim a sua utilidade. Portanto, além da base Adult, a base de dados de saúde WEF – *wearable-exercise-frailty* [Sokas et al. 2022] também é utilizada neste trabalho. A base WEF contém dados reais de saúde, incluindo atributos quase-identificadores, tais como idade, sexo, altura e peso, além de atributos sensíveis, como eletrocardiograma (ECG) e aceleração triaxial (ACC). A base de dados é constituída por 80 registros e 45 atributos, sendo estes numéricos e categóricos. Para esta base de dados, foram considerados apenas os atributos numéricos disponíveis, os quais são, idade, altura e peso.

4.2. Métricas e Experimentos

A proposta de Anonimização Dinâmica por Separatriz foi implementada na linguagem de programação Python, com suporte das bibliotecas scikit-learn, yellowbrick, numpy, pandas e matplotlib. Os experimentos foram conduzidos utilizando uma máquina com 20 GB de memória RAM DDR4 3200MHz, processador AMD Ryzen 5 5500u. A identificação do valor ideal de k ocorreu através da média de 10 iterações do método de Elbow para cada atributo QI. Em seguida, os quase-identificadores foram anonimizados pelas abordagens comparadas (intervalo fixo [Onesimu et al. 2022], MDAV, Mondrian [Torra and Navarro-Arribas 2023] e DAS), empregando valores de k sugeridos pela literatura e o valor ideal obtido por Elbow. O desempenho da abordagem DAS foi comparado com o desempenho das demais técnicas, por meio de métricas de risco de divulgação de atributos e de perda de informação, descritas na sequência.

A divulgação de atributos ocorre quando um atacante tenta adquirir mais conhecimento sobre um indivíduo (por exemplo, diagnóstico). Quando o atacante consegue identificar registros QI de um indivíduo e correlacionar com algum conhecimento prévio, pode ocorrer a vinculação e consequentemente divulgação de identidade e exposição atributos sensíveis [Onesimu et al. 2022]. Uma métrica para avaliar o risco desse tipo de ataque é a vinculação de registros baseada em distância (*distance-based record linkage*) [Christen et al. 2020]. Este trabalho utiliza a implementação e descrição apresentada em [Jiang and Torra 2023]. O algoritmo realiza o ataque com objetivo de calcular a quantidade de registros os quais possam ser divulgados. Para cada registro $r1$ da base de dados anonimizada $DB1$, calcula-se a distância de $r1$ em relação a cada registro $r2$ da base de dados original $DB2$ (a função de distância utilizada foi a distância euclidiana). Em seguida, seleciona-se o registro $r'(r1)$ mais similar (próximo) de $r1$. Se o registro selecionado

$r'(r1)$ corresponder ao registro anonimizado $r1$, então os dois registros foram vinculados corretamente. Assim, a métrica descrita reflete o número de vínculos corretos obtido pelo algoritmo, em relação ao número total de registros. A correspondência entre os registros é conhecida pelo fato das bases de dados $DB1$ e $DB2$ terem sido geradas neste trabalho.

A avaliação da perda de informação das técnicas é feita por meio da métrica *Normalized Certainty Penalty* [Ghinita et al. 2007, Ayala-Rivera et al. 2014]. Sejam min_i e max_i o valor mínimo e máximo de um atributo numérico i, respectivamente. Um registro j pertence à uma classe de equivalência com valor máximo max_{ij} e mínimo min_{ij}, para o atributo i. O NCP de uma tabela anonimizada T^* é definido como:

$$NCP(T^*) = \frac{1}{|T| \times n} \times \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{max_{ij} - min_{ij}}{max_i - min_i} \quad (1)$$

onde $|T|$ é o número de registros e n o número de atributos. A métrica é baseada no conceito de que valores que representam um intervalo maior são menos precisos que valores que representam intervalos menores [Ayala-Rivera et al. 2014]. O valor do NCP varia entre 1 e 0, em que 0 representa nenhuma perda de informação (dados originais) e 1 representa perda total de informação. Assim, valores próximos a 0 são desejáveis [Ayala-Rivera et al. 2014].

4.3. Resultados e Discussões

O estado da arte busca soluções para ataques externos à privacidade de dados sensíveis, especialmente em IoHT. A divulgação de atributos é um dos principais responsáveis pelo vazamento de informações. A re-identificação das informações é originada a partir de atributos numéricos e categóricos. Exclusivamente, a abordagem DAS se destina a coibir os riscos à privacidade inerentes aos atributos numéricos. Ao considerar soluções baseadas em k-anonimato, um fator imprescindível passa pela escolha do valor de k. As soluções estáticas ([Onesimu et al. 2022], MDAV ou Mondrian) atribuem a responsabilidade da segurança ao usuário, o qual define os intervalos fixos de k. A literatura recomenda valores de $k = 3$ ou $k = 5$, porém, esta estimativa depende das características dos dados a serem anonimizados, o que implica diretamente no nível de proteção à privacidade e na utilidade dos dados.

Para o conjunto de dados WEF, ao ser aplicada a definição dinâmica para o valor de k, baseada no método de Elbow, encontra-se o valor ideal de $k = 6$ para peso e altura, como identificado pela intercessão da linha tracejada com a linha azul na Figura 2. A figura também exibe a quantidade de tempo de ajuste do modelo de *cluster* para cada k como uma linha verde tracejada. Em geral, o valor maior de k é comprovadamente mais seguro. Entretanto, inferir um valor arbitrariamente alto afeta o compromisso entre anonimização e utilidade dos dados, uma vez que a generalização total não representa características de todos os indivíduos. A Tabela 3 exibe informações sobre a divulgação de registros para os valores de k sugeridos na literatura e o valor obtido dinamicamente. Para a base de dados WEF, a abordagem de anonimização dinâmica baseada em separatriz obteve desempenho semelhante ao intervalo fixo com $k = 6$, porém, eliminando a responsabilidade de escolha do usuário e o respectivo custo. Em relação à perda de informação das técnicas, a Tabela 4, exibe informações que respaldam a aplicação do DAS mesmo

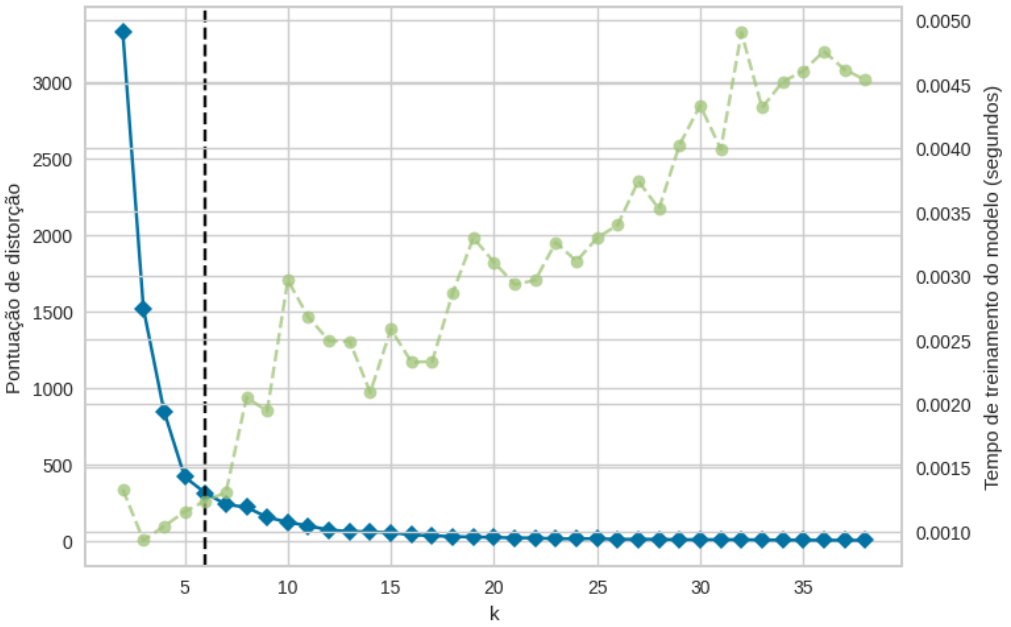

Figura 2. Definição do valor ideal de k usando Elbow para a base de dados WEF considerando o QI peso.

Tabela 3. Número de registros anonimizados vinculados corretamente aos 80 registros da base de dados original *wearable exercise frailty*.

WEF	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	17	21	13	-
$k = 5$	38	15	13	-
$k = 6$	48	12	8	-
$idade : k = 4$ $altura : k = 6$ $peso : k = 6$	-	-	-	51

Tabela 4. Perda de informação de dados originais para a base *WEF*.

WEF	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	0,184	0,094	0,160	-
$k = 5$	0,107	0,158	0,160	-
$k = 6$	0,086	0,185	0,240	
$idade : k = 4$ $altura : k = 6$ $peso : k = 6$	-	-	-	0,101

em conjuntos menores de dados, com um índice de perda de apenas 0,101, considerando os valores dinâmicos de k.

O conjunto de dados Adult é consideravelmente maior e com grande aderência junto às avaliações de técnicas de anonimização de dados. Para esta base de dados, o valor ideal atribuído pelo método de Elbow considerando o atributo horas por semana é $k = 9$, como ilustrado pela Figura 3. É importante ressaltar que, quanto maior a base de

dados, maior a possibilidade de divulgação de atributos [Torra and Navarro-Arribas 2023]. Na Tabela 5 é possível observar a descoberta dos atributos para MDAV e Mondrian para $k = 3$ e $k = 5$, a qual é significativamente reduzida ao aplicar o valor de $k = 9$. Além disso, o DAS obteve um desempenho superior aos outros métodos, com $k = 9$. Quanto à perda de informação das técnicas, a Tabela 6, ilustra o eficiente desempenho do método de anonimização DAS sob um vasto conjunto de dados, com um índice de perda de apenas 0,033, considerando os valores dinâmicos de k.

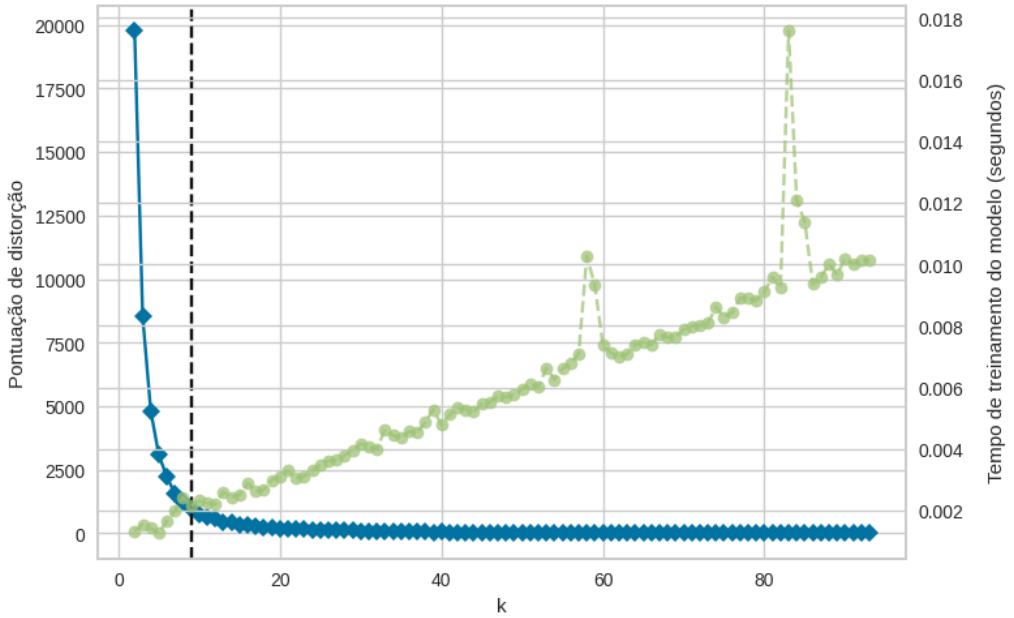

Figura 3. Definição do valor ideal de k usando Elbow para a base de dados Adult considerando o QI horas por semana.

Tabela 5. Número de registros anonimizados vinculados corretamente aos 30.162 registros da base de dados original *Adult*.

Adult	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	26	3367	2200	-
$k = 5$	85	2219	1186	-
$k = 9$	447	1471	683	-
$idade : k = 8$ $educacao\_num : k = 5$ $horas\_por\_semana : k = 9$	-	-	-	111

O desempenho computacional do DAS é intrinsecamente ligado à complexidade computacional do método de Elbow, etapa do algoritmo que possui maior custo. Entretanto, Elbow se destaca positivamente em relação aos demais métodos, (*Gap Statistic*, *Silhouette Coefficient* e *Canopy*). O DAS utiliza a implementação o método de Elbow *KElbowVisualizer* da biblioteca yellowbrick, que aplica o algoritmo Kneedle para encontrar o valor ideal de k, com complexidade conhecida de $O(n^2)$ em função da quantidade de registros [Satopaa et al. 2011]. Como os atributos possuem características unidimensionais, a entrada para o método de Elbow é limitada aos valores únicos dos atributos, reduzindo significativamente a quantidade de entradas e o tempo de processamento.

Tabela 6. Perda de informação de dados originais para a base *ADULT*.

Adult	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	0,318	0,008	0,021	-
$k = 5$	0,195	0,016	0,037	-
$k = 9$	0,092	0,028	0,056	-
$idade : k = 8$ $educacao\_num : k = 5$ $horas\_por\_semana : k = 9$	-	-	-	0,033

O desempenho da abordagem DAS na etapa de anonimização dos dados é obtida através da baixa complexidade computacional combinando o custo do algoritmo Kneedle com o custo da função de ordenação, a qual possui limite superior $NlogN$, multiplicado pela quantidade de atributos. Em contrapartida, as soluções baseadas em k-anonimato pertencem à classe de problemas NP-difícil [Onesimu et al. 2022]. Um fator determinante para alcançar o exímio desempenho quanto ao ataque de divulgação de atributos é o fato da proposta de anonimização dinâmica baseada em separatrizes gerar agrupamentos, ou classes de equivalência desbalanceadas. Deste modo, é possível coibir efetivamente ataques de divulgação de atributos. Além da segurança proporcionada, a anonimização dos dados permite que estes sejam divulgados publicamente para abastecer a comunidade científica com dados equivalentes aos dados puros. Consequentemente, possibilita assim extrair estatísticas e processamentos precisos e equivalentes aos obtidos com dados puros, ideal quando consideramos um cenário IoHT.

5. Conclusão

O compartilhamento dos dados é fundamental para prover avanços em pesquisas em diversas áreas, principalmente para cuidados com a saúde. Entretanto, as preocupações quanto à exposição e privacidade dos dados dificultam o compartilhamento entre as instituições. Portanto, nos últimos anos, diversas técnicas foram propostas com intuito de garantir a privacidade de dados sigilosos, e mitigar problemas de exposição pública. Entretanto, o compromisso entre segurança e inutilização dos dados é tênue. Neste sentido, a abordagem proposta neste artigo, a anonimização dinâmica por separatrizes, garante a privacidade dos atributos numéricos, mantendo-os o mais próximo das características puras. Tal compromisso com a privacidade dos dados é obtido por meio da identificação dinâmica da melhor configuração de agrupamento. Ademais, o DAS é aplicado individualmente a cada atributo, gerando maior segurança, utilidade/fidelidade aos dados devido à heterogeneidade entre os agrupamentos. Em trabalhos futuros, espera-se estender o DAS de modo a abranger os atributos categóricos, propondo uma abordagem dinâmica e eficiente para anonimizá-los.

Agradecimentos

Gostaríamos de agradecer o apoio financeiro da CAPES, Fapemig e CNPq.

Referências

Abouelmehdi, K., Beni-Hessane, A., and Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *Journal of big data*, 5(1):1–18.

Arava, K. and Lingamgunta, S. (2020). Adaptive k-anonymity approach for privacy preserving in cloud. *Arabian Journal for Science and Engineering*, 45(4):2425–2432.

Ayala-Rivera, V., McDonagh, P., Cerqueus, T., Murphy, L., et al. (2014). A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Priv.*, 7(3):337–370.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Batko, K. and Ślęzak, A. (2022). The use of big data analytics in healthcare. *Journal of big Data*, 9(1):3.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Byun, J.-W., Kamra, A., Bertino, E., and Li, N. (2007). Efficient k-anonymization using clustering techniques. In *International Conference on Database Systems for Advanced Applications*, pages 188–200. Springer.

Christen, P., Ranbaduge, T., and Schnell, R. (2020). Linking sensitive data. *Methods and techniques for practical privacy-preserving information sharing. Cham: Springer*.

Coelho, K. K., Tristão, E. T., Nogueira, M., Vieira, A. B., and Nacif, J. A. (2023). Multimodal biometric authentication method by federated learning. *Biomedical Signal Processing and Control*, 85:105022.

Correa, S. (2003). Probabilidade e estatística.

Developers, N. (2024). numpy.percentile.

Dinh, D.-T., Huynh, V.-N., and Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, 571:418–442.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, 14(1):189–201.

El Ouazzani, Z. and El Bakkali, H. (2018). A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k. *Procedia Computer Science*, 127:52–59. PROCEEDINGS OF THE FIRST INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING IN DATA SCIENCES, ICDS2017.

Fernandes, L. M., O'Connor, M., and Weaver, V. (2012). Big data, bigger outcomes. *Journal of AHIMA*, 83(10):38–43.

Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. (2007). Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

Jiang, L. and Torra, V. (2023). Data protection and multi-database data-driven models. *Future Internet*, 15(3).

Ketu, S. and Mishra, P. K. (2021). Internet of healthcare things: A contemporary survey. *Journal of Network and Computer Applications*, 192:103179.

Khan, R., Tao, X., Anjum, A., Kanwal, T., Malik, S. U. R., Khan, A., Rehman, W. U., and Maple, C. (2020). θ-sensitive k-anonymity: An anonymization model for iot based electronic health records. *Electronics*, 9(5):716.

Kodinariya, T. M., Makwana, P. R., et al. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005a). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005b). Multidimensional k-anonymity. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Liu, F. and Li, T. (2018). A clustering k-anonymity privacy-preserving method for wearable iot devices. *Security and Communication Networks*, 2018:1–8.

Olatunji, I. E., Rauch, J., Katzensteiner, M., and Khosla, M. (2022). A review of anonymization for healthcare data. *Big data*.

Onesimu, J. A., Karthikeyan, J., Eunice, J., Pomplun, M., and Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access*, 10:86979–86997.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a"kneedle"in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Shahid, J., Ahmad, R., Kiani, A. K., Ahmad, T., Saeed, S., and Almuhaideb, A. M. (2022). Data protection and privacy of the internet of healthcare things (iohts). *Applied Sciences*, 12(4).

Sokas, D., Butkuvienė, M., Tamulevičiūtė-Prascienė, E., Beigienė, A., Kubilius, R., Petrėnas, A., and Paliakaitė, B. (2022). Wearable-based signals during physical exercises from patients with frailty after open-heart surgery. *PhysioNet*.

Templ, M. (2008). Statistical disclosure control for microdata using the r-package sdcmicro. *Transactions on Data Privacy*, 1(2):67–85.

Torra, V. and Navarro-Arribas, G. (2023). Attribute disclosure risk for k-anonymity: the case of numerical data. *International Journal of Information Security*, 22(6):2015–2024.

Victor, N. and Lopez, D. (2020). Privacy preserving sensitive data publishing using (k, n, m) anonymity approach. *Journal of communications software and systems*, 16(1):46–56.

Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235.