Um Estudo Comparativo entre Algoritmos de Proteção da Privacidade Aplicado à Bases de Dados na Área de Saúde

Francimaria Nascimento[1], **Karliane Vale**[2], **Flavius Gorgônio**[2]

[1]Departamento de Matemática e Informática Aplicada (DIMAP)
Universidade Federal do Rio Grande do Norte (UFRN)
Campus Universitário, Lagoa Nova, 59.078-970, Natal, RN, Brasil

francimariasantos@ufrn.edu.br

[2]Laboratório de Inteligência Computacional Aplicada a Negócios (LABICAN)
Universidade Federal do Rio Grande do Norte (UFRN)
Rua Joaquim Gregório, S/N, 59.300-000, Caicó, RN, Brasil

karliane@dct.ufrn.br, flavius@dct.ufrn.br

Abstract. *The growing increase in the volume of data which is collected, stored and shared by health institutions creates benefits for the process of decision making based on the knowledge obtained from applying data analysis and data mining techniques, aiming to achieve relevant information. Despite the obtained benefits, sharing this specific kind of data in its original raw format may compromise patients' privacy. In an attempt to validate solutions for this problem, this article considers and compares data anonymization and perturbation techniques, assessing their efficiency in providing privacy and safety of shared data, more specifically, when applied to databases of the health field.*

Resumo. *O crescente aumento no volume de dados coletados, armazenados e compartilhados por instituições da área de saúde gera benefícios para o processo de tomada de decisão com base no conhecimento adquirido a partir da aplicação de técnicas de análise e mineração de dados na extração de informações úteis. A despeito dos benefícios propiciados, o compartilhamento desses dados em seu formato original pode pôr em risco a privacidade dos pacientes. Na tentativa de validar soluções para este problema, este artigo compara algumas técnicas de anonimização e perturbação de dados, avaliando a eficácia dessas técnicas na garantia da privacidade e segurança de dados compartilhados, em particular, quando aplicadas a bases de dados na área de saúde.*

1. Introdução

Há uma crescente adoção de práticas de Tecnologia da Informação em pesquisas na área de saúde, o que resulta na coleta, armazenamento e compartilhamento de grandes volumes de dados, nos quais, a aplicação de técnicas de mineração de dados pode possibilitar a descoberta de informações potencialmente úteis. Entretanto, algumas instituições que realizam pesquisas na área de saúde e que necessitam fazer uso de várias fontes de dados, podem hesitar em compartilhar os dados entre si, pois seus registros normalmente possuem dados altamente sensíveis que não podem ser expostos, como por exemplo, informações pessoais dos pacientes [Kumari et al. 2012]. Por isso, nesses casos é importante compartilhar os dados, porém protegendo a privacidade dos registros.

O termo Preservação da Privacidade na Mineração de Dados (*Privacy Preserving Data Mining*) foi introduzido quase simultaneamente em [Agrawal and Srikant 2000] e [Lidell and Pinkas 2000]. Desde então, o assunto vem sendo amplamente estudado e é de grande relevância para a área de mineração de dados, dada a necessidade de algumas organizações extraírem o conhecimento existente em bases de dados compartilhadas, de forma colaborativa, porém com a garantia da proteção da privacidade.

Na literatura, estão descritas várias técnicas para garantir a proteção da privacidade dos dados, dentre elas, a Anonimização [Emam 2006, Byun et al. 2007] e a Perturbação [Agrawal and Srikant 2000], que são comparadas neste trabalho. Assim, faz-se necessário analisar a eficácia das técnicas utilizadas tanto com respeito à garantia da privacidade dos dados, quanto em relação ao nível de precisão dos resultados obtidos após a aplicação das técnicas. Pois, apesar das limitações e consequências, a utilização de técnicas para a manutenibilidade da privacidade dos dados é importante e deve ser aplicada quando se compartilham dados entre instituições a fim de se garantir a proteção dos registros. A não utilização dessas técnicas pode aumentar a vulnerabilidade dos dados, que podem vir a ser usados por terceiros mal-intencionados e causar diversos danos – problemas jurídicos, perda de privacidade, de patrimônio intelectual, entre outros – ao proprietário dos dados e/ou à instituição que os coletou.

2. Técnicas de Anonimização

Técnicas de anonimização realizam a remoção e/ou ofuscação de dados que possam auxiliar na identificação de um indivíduo em particular dentro de um conjunto de dados [Emam 2006]. Na literatura, estão descritos alguns modelos de anonimização, entre eles, estão o *k-anonymity* [Byun et al. 2007] e o *l-diversity* [Machanavajjhala et al. 2007]. O modelo *k-anonymity* tem como objetivo evitar que sejam feitas ligações entre atributos que identifiquem o proprietário do registro, desta forma, esse modelo exige que qualquer registro seja indistinguível de, pelo menos, $k - 1$ outros registros que possuam quasi-identificadores predeterminados (p.ex.: sexo, data de nascimento e CEP que, combinados, podem identificar o proprietário do registro).

Contudo, a homogeneidade de alguns valores sensíveis dentro de um conjunto de dados gera alguns problemas na proteção da privacidade com o modelo *k-anonymity*, pois a proteção dos *k*-indivíduos talvez não corresponda a todos os atributos sensíveis (ex.: diagnóstico de uma doença). Tendo em vista este problema, o modelo de *l-diversity* foi projetado para lidar com alguns problemas do modelo *k-anonymity* [Sinha and Kumar 2010].

O modelo *l-diversity* é baseado na premissa de que um conjunto de dados deve possuir pelo menos *l* atributos sensíveis "bem representados", para que a privacidade dos dados seja protegida [Machanavajjhala et al. 2007]. Algumas interpretações para o termo "bem representado" são demonstrados através dos seguintes princípios:

1 - Entropia - O cálculo da entropia *l-diversity* pela Equação 1, é usado para captar o número de grupos de atributos "bem representados", devido ao fato da entropia aumentar quando uma frequência se torna mais uniforme.

$$Entropia(E) = -\sum_{s \epsilon S} p(E, s) \log p_{(E,s)} \geq \log l \quad (1)$$

onde E representa uma classe de equivalência, s é o domínio de atributos sensíveis, e $p(E, s)$ é uma fração de registros em E que possuem atributos sensíveis s. Uma tabela é *l-diversity* se todas as classes de equivalência E, possuírem Entropia(E) $\geq \log l$.

Recursividade - Tem como finalidade certificar que o valor menos frequente não apareça raramente e que um valor muito frequente não apareça com muita frequência. Deste modo, dado um valor constante v, uma classe E satisfaz o princípio da recursividade se $r_i < v(r_l + r_{(l+1)} + \cdots + r_m)$, onde cada r representa o valor de um registro. Uma tabela possui recursividade *(v, l)-diversity*, se todas as classes também tiverem.

Para atingir o objetivo do *l-diversity*, um algoritmo denominado *Anatomize* foi proposto por [Xiao and Tao 2006]. Especificamente, este algoritmo separa os dados em duas tabelas, uma contendo os valores quasi-identificadores (QIT), que combinados podem identificar o usuário, e uma tabela sensível (ST), na qual são armazenados os atributos sensíveis.

3. Técnicas de Perturbação

Técnicas de perturbação adicionam ruídos aleatórios aos registros antes da etapa de mineração de dados, de forma que os resultados obtidos com os dados perturbados sejam aproximadamente os mesmos dos dados originais, possibilitando a extração de informações a partir da aplicação de técnicas de mineração de dados [Kedar et al. 2013, Sinha and Kumar 2010].

Com o objetivo de proteger dados de pesquisas médicas, [Liu et al. 2012] propuseram dois algoritmos para adicionar perturbações nas bases de dados em que são aplicadas técnicas de análise de agrupamento: i) Distância Aleatória no Domínio da Distância (*Random Distance in Distance Domain* - RDD), onde os registros primeiramente são agrupados com o algoritmo *k-means* e, em seguida, a cada registro é adicionado um ruído de uma distribuição Gaussiana, aplicando pequenos ajustes nas distâncias entre o dado e o centroide do grupo ao qual ele foi classificado com o *k-means*, com a finalidade de manter os registros no mesmo grupo antes e depois de serem perturbados; e ii) Rotação em Torno do Centro de Agrupamento (*Rotation Around the Center of Clustering* - RACC), que distintamente de alguns algoritmos utilizados para adicionar perturbação à bases de dados, não perturba o registro adicionando diretamente um ruído aos dados originais, mas por uma pequena medida de distância aleatório $d_j (0 < d_j \leq 1)$ e um ruído aleatório $\theta(\theta_1, \theta_2, ..., \theta_n)$, onde, $\theta \, \epsilon \, (0, 2\pi)$. O algoritmo RACC, calcula o valor do registro perturbado Q com base nas Equações 2 e 3:

$$r = d_j \times dis(R, C) \tag{2}$$

$$\begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{bmatrix} = \begin{bmatrix} C_{i1} \\ C_{i2} \\ \vdots \\ C_{in} \end{bmatrix} + r \begin{bmatrix} \cos(\theta_1) \\ \cos(\theta_2) \\ \vdots \\ \cos(\theta_n) \end{bmatrix}, 0 < \theta < 2\pi \tag{3}$$

onde, r é resultante do produto da proporção de distância aleatória d_j e $dis(R, C)$, que é a distância euclidiana do registro R ao centróide C do agrupamento ao qual R faz parte.

3.1. Metodologia dos Experimentos

Para realização dos experimentos foram selecionadas 5 bases de dados da área de saúde disponíveis no UCI *Repository of Machine Learning Databases*: *Breast Cancer Wisconsin*, com 699 instâncias e 10 atributos; *Fertility Data Set*, com 100 instâncias e 10 atributos; *Lung Cancer Data Set*, com 32 intâncias e 56 atributos; *Mammographic Mass Data Set*, com 961 instâncias e 6 atributos; e *SPECTF Heart Data*, com 267 instâncias e 44 atributos. Para fins de simplificação foram renomeadas, respectivamente, como: Base de Dados 1, Base de Dados 2, Base de Dados 3, Base de Dados 4 e Base de Dados 5.

As técnicas de anonimização e perturbação foram escolhidas por serem bastante utilizadas na proteção da privacidade de dados na área de saúde [Emam 2006, Liu et al. 2012]. Dentre os modelos citados de anonimização, foi utilizado o *l-diversity*, pois, segundo [Li et al. 2007], o modelo de *k-anonymity* não é suficiente para a proteção dos dados. O modelo *l-diversity* foi aplicado com $l = 2$ (diversidade igual a dois), porque o atributo sensível escolhido foi o atributo *classe* que em todas as bases de dados possui apenas dois possíveis valores.

Dentre os algoritmos citados que implementam a estratégia proposta pela técnica de perturbação, foi utilizado o algoritmo RACC, pois, segundo [Liu et al. 2012], com o aumento no nível de perturbação o algoritmo RDD diminui o nível de precisão do resultado da mineração de dados, entretanto, o algoritmo RACC mantém essa precisão estável. O algoritmo de análise de agrupamentos escolhido foi o *k-means*, com *k* igual a 3. A escolha do *k-means* para a etapa de análise de agrupamentos se deu por este ser um dos algoritmos de agrupamento mais conhecidos e ser amplamente usado, além de simples e de fácil implementação [Oliveira and Zaïane 2007].

3.2. Quantificação de Erros de Agrupamento

A Equação 4 foi utilizada para medir, em valores percentuais, a taxa de erros de agrupamento, denotada por E_c, que deveria ser a menor possível:

$$E_c = \frac{1}{N} \sum_{i=1}^{k} ((|Grupo_i(D)|) - (|Grupo_i(D')|)) \quad (4)$$

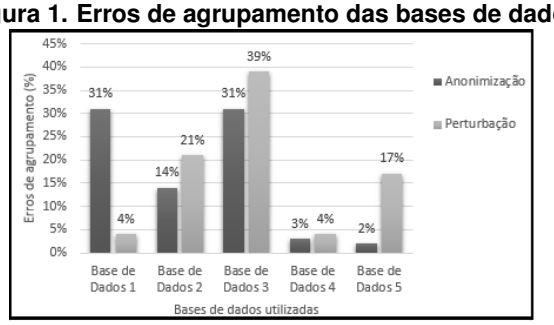
onde N representa o número de registros na base de dados original, k é o número de agrupamentos analisados, $|Grupo_i(D)|$ representa os registros do agrupamento dos dados originais e $|Grupo_i(D')|$ os registros do agrupamento dos dados distorcidos.

A Equação 4, proposta por [Oliveira and Zaïane 2010], foi escolhida por não considerar simplesmente a quantidade de pontos que cada agrupamento possui. Ao invés disso, é levado em consideração o agrupamento real de cada ponto, pois são comparados os rótulos dos agrupamentos de cada ponto antes e depois da distorção. Na Figura 1 são descritos os resultados obtidos no experimento que compara as técnicas de anonimização e perturbação, tendo sido utilizada a taxa de erro para medir a eficácia do algoritmo.

3.3. Quantificação da Privacidade

Além da medida de erros de agrupamento, [Oliveira and Zaïane 2010] propuseram quantificar a privacidade dos dados a partir de uma medida de segurança Sec, com base em uma medida de variância dada por $var(x - x')$, onde x representa um atributo original e x' o atributo distorcido, de modo que quanto maior o valor de $var(x - x')$ melhor o

Figura 1. Erros de agrupamento das bases de dados.



resultado. Esta medida pode ser expressa em uma escala invariante, no que diz respeito à variação de valores da variável original, descrita na Equação 5, onde quanto maior a variância melhor o resultado obtido.

$$Sec = \frac{var(x - x')}{var(x)} \tag{5}$$

Os resultados obtidos pela aplicação da técnica de perturbação nas bases de dados citadas são apresentados na Tabela 1, onde estão descritos o valor mínimo (V_{min}), o valor máximo (V_{max}), a média ($V_{méd}$) e o desvio padrão (σ), calculados a partir da Equação 5.

Tabela 1. Nível de privacidade do dados perturbados (%)

Bases de Dados	Vmin	Vmax	Vméd	Vσ
Base de Dados 1	0,97	1,82	1,15	0,22
Base de Dados 2	1,35	2,91	2,16	0,42
Base de Dados 3	4,38	11,41	7,44	1,51
Base de Dados 4	0,61	1,60	0,88	0,32
Base de Dados 5	5,05	8,13	6,67	0,65

A partir dos testes realizados, foi possível verificar que em todas as bases de dados em que foi aplicada a técnica de anonimização os resultados obtidos pela aplicação da Equação 5 foram iguais a zero. Isso acontece, como citado anteriormente, em função do algoritmo de anonimização utilizado não realizar distorções nos atributos dos registros, e sim, criar regras pelas quais mais de um registro possuam atributo sensível distinto e os atributos quasi-identificadores semelhantes, deste modo alguns registros que não se encaixam nas regras geradas são suprimidos da base de dados [Machanavajjhala et al. 2007].

Desta forma, a probabilidade de descobrir o valor do atributo sensível é $\frac{1}{l}$, onde l é a diversidade que se deseja alcançar. Assim, quanto maior o valor de l, menor a probabilidade de descoberta do atributo sensível. Por exemplo, uma base de dados que possui apenas dois possíveis valores para um atributo sensível, poderá ter no máximo l = 2, com isso o valor do atributo sensível pode ser reconstruído com 50% de probabilidade.

4. Conclusões e Trabalhos Futuros

A partir dos resultados obtidos, foi possível inferir que o algoritmo aplicado para perturbação dos dados mantém um certo *trade-off* entre privacidade dos dados e precisão dos resultados da análise de agrupamentos, ou seja, mais privacidade implica diretamente

em menos precisão. Também foi possível constatar que as bases de dados que possuem maiores dimensionalidades possuem melhores resultados quanto à privacidade dos dados.

Quanto à eficácia, verificou-se mais erros de agrupamentos nas bases de dados em que foi aplicada a técnica de perturbação, sendo que estes ocorrem devido à distribuição espacial dos dados, pois o algoritmo de perturbação utilizado rotaciona os dados em torno do centroide de cada agrupamento gerado pelo algoritmo. Desta forma, quando os agrupamentos são próximos, alguns registros ficam próximos do centroide do agrupamento vizinho, gerando assim erros de agrupamento.

Como os resultados obtidos nesta pesquisa são iniciais, é importante considerar como proposta de trabalho futuro uma análise mais profunda nas técnicas analisadas, averiguando diferentes configurações, como por exemplo modificando quantidade e tipos de atributos sensíveis, além da quantidade de agrupamentos usados.

Referências

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *ACM Sigmod Record*, volume 29, pages 439–450, California. ACM.

Byun, J., Kamara, A., Bertino, E., and Li, N. (2007). Efficient k-anonymization using clustering techniques. In *12th Int Con Database Syst Adv App*, pages 188–200, Berlin.

Emam, K. (2006). *Data anonymization practices in clinical research: a descriptive study*. CHEO Research Institute, Ottawa.

Kedar, S., Dhawale, S., Vaibhav, W., Kadam, P., Wani, S., and Ingale, P. (2013). Privacy preserving data mining. *Advanced Res. in Comp. and Com. Eng.*, 2(4):1677–1680.

Kumari, A., Rao, R., and Suman, M. (2012). Vector quantization for privacy preserving clustering in data mining. *Advance Computing*, 3(6):69–74.

Li, N., Li, T., and Venkatasubramaniam, S. (2007). t-Closeness. In *23rd International Conference on Data Engineering, 2007*, pages 106–115, Istambul. IEEE.

Lidell, Y. and Pinkas, B. (2000). Privacy-preserving data mining. In *International Cryptology Conference on Advances in Cryptology, 1880*, pages 36–54, California. Springer.

Liu, L., Yang, K., Hu, L., and Li, L. (2012). Using noise addition method based on pre-mining to protect healthcare privacy. *Journal Control Eng. App. Inf.*, 14(2):58–64.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on KDD*, 1(3):52.

Oliveira, S. and Zaïane, O. (2007). A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. *Com&Sec*, 26(1):81–93.

Oliveira, S. and Zaïane, O. (2010). Privacy Preserving Clustering by Data Transformation. *Journal of Information and Data Management*, 1(1):37–51.

Sinha, B. and Kumar, J. (2010). *Privacy Preserving Clustering In Data Mining*. PhD thesis, National Institute of Technology Rourkela, Rourkela.

Xiao, X. and Tao, Y. (2006). Anatomy: Simple and effective privacy preservation. In *Proc. of the 32nd Int. Conf. on Very Large Data Bases*, pages 139–150, Hong Kong.