Universidade de Coimbra

Mestrado em Engenharia e Ciência de Dados

Disciplina de Segurança e Privacidade

UNIVERSIDADE D
COIMBRA

Joana Simões, n.º 2019217013

Tomás Ferreira, n.º 2019224786

Introdução

Esta segunda parte do projeto teve como objetivo explorar diferentes formas de proteger a privacidade dos dados, onde a ideia principal de todas é "mascarar" os dados, do lado da ControlER, antes de os enviar para a Dellenture, assumindo que esta possa ser comprometida. Foram exploradas três técnicas para proteger os dados: anonimização e generalização deles; differential privacy; criação de dados sintéticos.

ARX — anonimização de dados

Classificação dos dados

O primeiro passo desta etapa do projeto foi classificar o tipo de dados em “Quai-Identifiers”, “Dados sensíveis”, “Dados insensíveis” e “Identificadores”.

Todos os dados relacionados com o valor monetário dos clientes, ou relacionado com os empréstimos dos clientes no passado foram considerados dados sensíveis. A razão deve-se ao facto de muitas culturas ser tabu falar sobre este tema. Ao mesmo tempo, considerar estes dados como sensíveis garante serem pouco destruídos pelo ARX, uma vez que são dos dados mais utilizados na análise.

Como Quasi-Identifiers temos os dados, são todos os dados relacionados com o perfil da pessoa: idade, educação, estado da família, trabalho da pessoa, etc., dos quais se consegue extrair informação que pode levar à identificação da pessoa.

Os dados identificadores são o identificador do empréstimo e / ou pessoa. Estes dados não devem de todo ser publicados, uma vez que identificam inequivocamente os participantes.

Os restantes parâmetros do dataset foram considerados insensíveis uma vez que nem identificam as pessoas e, em muitos casos, nem são importantes para a análise.

Type of data	Column in dataset
Quasi-identifiers	num_family_members, gender, income_type, education, family_status, age
Sensitive	infringed, annual_income, credit_amount, credit_annuity, goods_evaluation, past_{...}
Identifiers	Loan_id, SK_ID_CURR

Type of data	Column in dataset
Insensitive	All the others

Tendo os quasi-identifiers (QIDs) identificados, foi a analisada a separação e distinção dada por eles. Ao olhar para os resultados de utilizar apenas um Quasi-Identifiers, todos eles apresentam uma distinção muito baixa, e uma separação entre que varia entre os 43% e os 98%. Como também era de esperar, ao utilizar todos os QIDs, obtemos o máximo de distinção (3,47%) e de separação dos dados (99.89%). É necessário ter cuidado com estes valores, uma vez que com eles, conseguimos descobrir 3.47% dos indivíduos de uma classe.

Após aplicar a anomização e generalização dos dados, conseguimos reduzir estes valores máximos para apenas 0.26% de distinção e separação de 99.1%. Pelo que, agora, torna a probabilidade de identificação de indivíduos de uma classe menor. Mesmo que se utilize apenas um QID, tanto a distinção como a separação dos dados reduz-se bastante.

Configurações do ARX

Foram testadas diversas configurações até chegar à configuração descrita abaixo. Para avaliar os dados, foi analisada a utilizadde dos dados transformados e os riscos obtidos.

Os modelos finais utilizados foram um K-anonymity com k=8 e um L-diversity com L=2. Depois foi adicionado um limite à supressão de 25%, tendo antes experimentado com um limite de 40%, mas que se achou muito alto e que destruía muito os dados.

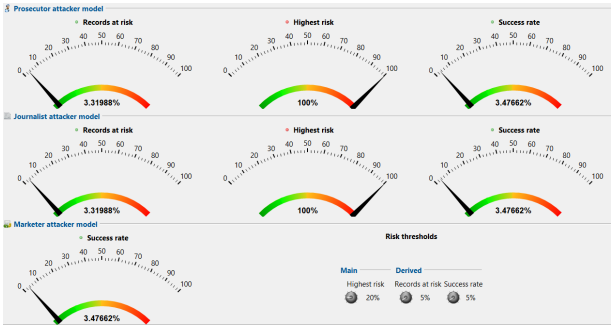
Foram, ainda, adicionados pesos aos diferentes QIDs, dando mais importância aos QIDs de género, educação, idade e estado da família, se seguida. Os pesos dados aos diferentes QID não parecem alterar os resultados obtidos, no entanto, deixaram-se estes pesos por se considerar que alguns QIDs deveriam ter maior peso que outros.

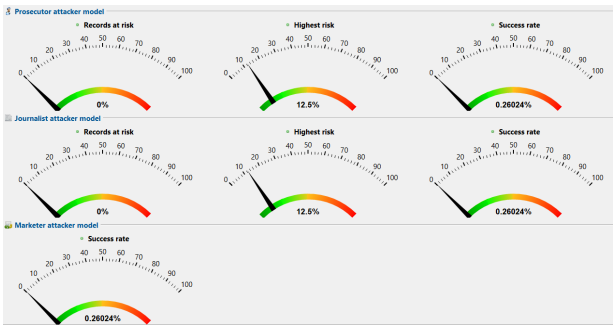
Por último, adicionaram-se generalizações aos Quasi-Identifiers, consoante o tipo de dados em causa. Para generalizar os parâmetros com valores numéricos, foram criados intervalos de dados. Por outro lado, para os valores não numéricos, foi adicionada uma hierarquia, na qual os valores poderiam pertencer a uma classe que continham um ou mais dos valores iniciais. Foi ainda testado mascarar os valores não numéricos, pelo que se considerou uma má opção, uma vez que todos tinham valores diferentes e era possível distingui-los apesar de um elevado nível de generalização.

Resultados do ARX

Ao analisar os resultados do ARX, é possível constatar que o risco de identificar

pessoas diminui bastante, pelos resultados dos modelos de ataque na figura **X**.





Pode-se observar que todos os riscos descem consideravelmente após generalizar hierarquicamente os Quasi-identifiers. Ao mesmo tempo, olhando para os resultados do "Highest Risk" para os modelos de ataque Prosecutor e Journalist, continuam na ordem dos 12.5% e existe um risco médio de 0.26% de identificar uma pessoa numa determinada classe. Estes valores são aceitáveis porque um risco menor poderia significar que destruímos muitos os dados ao anonimizá-los, pelo que deixariam de ter utilidade. Também o contexto do problema, ou seja, a identificação de pessoas segundo os empréstimos que elas pedem, identificar uma pessoa não é tão crítico como noutros contextos.

Comparando as análises quer com o dataset original, quer com o dataset anonimizado, verifica-se que os resultados pouco alteram, pelo que se pode considerar que a utilidade dos dados transformados se mantém relativamente intacta.

Análise do dataset transformado

A análise do dataset efetuada na meta 1 do trabalho não previa que agora na segunda meta muitos valores deixa-sem de ser numéricos, referindo mais concretamente aos QIDs. Assim, foi necessário criar uma análise dos dados finais, para que pudesse analisar tanto os dados antes como após a anonimização.

Pelos dados obtidos, observa-se que existe pouca diferença nos resultados obtidos com ambos os datasets. Assim, temos a vantagem de ser difícil para um atacante identificar os dados de uma pessoa, e, ao mesmo tempo, continuamos não perdemos a utilidade dos dados. A desvantagem desta técnica são que muitas vezes torna-se difícil obter informações relevantes dos dados transformados.

Differential Privacy

Outra abordagem explorada no projeto para anonimizar os dados foi a *differential privacy*. Esta técnica consiste em apenas enviar, neste caso da ControlER para a

Dellenture, apenas os dados requeridos pela segunda empresa, que analisará os dados. Os dados enviados vão com um erro adicionado, para não ser possível à Dellenture saber o valor real, apenas uma aproximação.

Para “mascarar” os dados, é-lhes adicionado um ruído que segue a distribuição de Laplace. O erro adicionado aos valores reais tem em conta a sensibilidade dos dados em causa, e um certo *epsilon.* Neste projeto, foi utilizado um *epsilon* de 0.01, e a sensibilidade é calculada em cada vez que a função de “mascarar” os dados é chamada.

Neste projeto apenas foram utilizadas somas e contagem dos dados de uma determinada categoria, pelo que a sensibilidade é fácil de calcular. Nestes casos, uma vez que a sensibilidade é o máximo absoluto entre a função aplicada aos dados menos a função aplicada aos dados, mas que não contém uma certo ponto (linha do dataset), o valor é fácil de calcular. Quando estamos perante uma soma, a sensibilidade é dada pelo maior valor presente no dataset em causa. Por outro lado, quando estamos perante uma contagem, a sensibilidade será sempre 1, já que é dada pela diferença entre o tamanho (linhas) do dataset e pelo n.º de linhas do dataset menos uma linha.

Os resultados do *differential privacy* mostram que a percentagem de erro varia bastante. É, ainda, possível observar que quando são utilizados muitos dados, o erro entre o valor original e o valor com ruído é muito pequeno. Isto acontece, por exemplo, quando queremos analisar os dados onde não existem infrações (infringed=0). Em oposição, ao utilizar menos dados, por exemplo, quando queremos analisar dados onde houve infrações, (infringed=1), o erro entre o valor real e transformado é relativamente grande em muitos casos. Apesar de o erro variar muito nalguns casos, ao comparar os resultados com a análise feita ao dataset original, verifica-se que as diferenças são pequenas.

A vantagem de utilizar esta abordagem é semelhante à técnica interior, uma vez que ao adicionar ruídos aos dados que são enviados à Dellenture, evita-se o risco de os dados serem utilizados para atos indevidos. No entanto, é necessário ter cuidado em quantas vezes se autoriza à Dellenture requisitar os dados, uma vez que ao requisitar os dados diversas vezes, a média os valores resultantes converge para o valor real, ou seja, deve-se limitar o n.º de pedidos ou então guardar o valor e reenviá-lo quando requisitado. Esta técnica tem ainda a desvantagem de necessitar que a ControlER esteja constantemente a enviar dados para a Dellenture, conforme necessários na análise, não sendo como nas outras abordagens em que se manda o dataset inteiro para ela analisar.

Criação de dados sintéticos

A última abordagem explorada no projeto foi a criação de um dataset sintético com o uso da livraria SVD. O Synthetic Data Vault (SDV) é um gerador de dados sintéticos, tal que os dados gerados têm o mesmo formato e propriedades estatísticas que os dados originais. Assim, apesar de apresentarem semelhanças, o dataset sintético não é igual aos dados originais. Consequentemente, com o uso desta técnica ocorre uma maior proteção na privacidade de cada indivíduo, uma vez que caso ocorra fuga de informação, o atacante não conseguirá reunir informações reais acerca dos indivíduos. No contexto do projeto, este passo da geração de dados sintéticos, deve ser elaborado por parte da ControlER antes de esta enviar os dados para a Dellenture analisar.

Durante a execução deste exercício foram testados três tipos de modelos diferentes, o CTGAN, o GaussianCopula e o CopulaGAN, sendo no fim comparados os desempenhos deles. Durante a resolução das experiências, verificou-se que os modelos de treino apresentavam diferentes tempos de execução e resultados. O modelo CopulaGAN, apesar de mais demorado, foi o que apresentou melhores resultados, com um "*score*" de 0,91. Por outro lado, o modelo mais rápido GaussianCopula apresenta um "*score*" mais baixo, de 0,90, mostrando que tem de existir um *trade off* entre velocidade de treino/geração e qualidade dos dados gerados.

Ao analisar os dados gerados pelo melhor modelo, acima descrito, verificou-se que os dados gerados perdem muitas das características dos dados originais. Observou-se que as relações entre diferentes variáveis se alteram completamente nos novos dados. Pelo que pode não ser uma boa opção quando queremos manter as distribuições das diferentes variáveis. Este problema encontrado também se pode dever ao facto de se ter reduzido o tamanho do dataset antes de treinar o modelo, ou seja, as amostras selecionadas podem ter afetado as distribuições, por exemplo, por se terem selecionados muitos outliers comparando com a sua densidade no dataset original. Uma solução para tal passaria por remover, primeiramente, todos os outliers do dataset original. Esta solução, acaba por ter benefícios em termos de privacidade e segurança dos dados, uma vez que, os outliers não passam, neste contexto, de pessoas com características únicas no dataset, e por isso, fáceis de reconhecer.

Após a análise dos resultados, observa-se que esta abordagem tem algumas vantagens e desvantagens em relação às restantes técnicas exploradas. A vantagem desta técnica é garantir que os dados dos clientes não são explorados indevidamente por terceiros. No entanto, esta técnica também acarreta algumas desvantagens, nomeadamente ser necessário reduzir o tamanho do dataset original para os modelos

serem treinados. Ao mesmo tempo, os modelos utilizados na geração de dados sintéticos não têm em conta alguns padrões que os dados contêm, este facto é ainda mais evidente quando apenas utilizamos uma parte do dataset para treinar os modelos.

Conclusão

Para o problema em questão, talvez a melhor opção seja a primeira, se anonimizar os dados através do ARX, por exclusão de partes. Por um lado, como temos um dataset muito grande, para gerar dados sintéticos é necessário reduzir o tamanho do dataset, e com isso, perdemos alguns padrões dos dados. Por outro lado, a *differential privacy* necessita de mais trabalho por parte da ControlER que tem de estar constantemente a enviar os dados requisitados à Dellenture.

Em suma, todas as abordagens exploradas apresentam vantagens e desvantagens, sendo necessário ver qual a melhor para o tipo de problema. A grande vantagem, comum a todas elas, é o facto de permitir “mascarar” os dados reais para que, caso sejam atacados, o atacante não pode tirar muitas conclusões deles, e as que retira têm sempre com elevada incerteza. Ao mesmo tempo, garantem a privacidade dos dados dos utilizadores, tão importante atualmente.