MDCC - Privacidade de Dados - 2022

Trabalho 2 - k-Anonimato e l-Diversidade

Javam Machado

1 Objetivo:

O trabalho consiste em implementar um algoritmo que anonimize um conjunto de dados contra ataques de ligação ao registro atendendo o k-anonimato, e contra ataques de ligação ao atributo sensível atendendo o l-diversidade. Deverá ser implementação o modelo k-anonimato por meio da **generalização** de valores de atributos como descrito no artigo *L. Sweeney. k-anonymity: a model for protecting privacy. Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.* Para garantir a diversidade dos atributos sensíveis dentro da classe de equivalência será necessário aplicar a **perturbação** do atributo sensível, quando for necessário.

2 Especificação:

Carregue o conjunto de dados "alunos.csv". O conjunto de dados contém a relação de alunos entre os meses de janeiro e outubro que estiveram doentes, possuindo 3 atributos:

- *Semi-identificadores:* Data, Idade;
- *Sensíveis:* Doença.

Primeira Parte: O valor de k deve variar no conjunto $k = \{2, 4, 8\}$. Para cada valor de k, o conjunto de dados deve ser anonimizado de forma a atender o modelo k-anonimato. Essa anonimização deve ser feita por generalização – hierarquia de quatro níveis no atributo *Data* (data, mes_ano, ano, década) e hierarquia de três níveis no atributo $Idade$, você deve escolher os intervalos nos níveis 2 e 3. Você deve gerar três datasets anonimizados de nome "kAnon-Alunos.csv", onde $k = \{2, 4, 8\}$. Mostre na tela as classes de equivalência atendendo o k-anonimato. Procure também plotar um histograma das classes de equivalência.

Para medir a utilidade do processo de anonimização, calcule a precisão e o tamanho médio das classes de equivalência, ambas as métricas para cada um dos datasets gerados.

Segunda Parte: Para os conjuntos de dados anonimizado com $k = 8$, verificar se atende o l-diversidade dos dados sensíveis com $l = \{2, 3, 4\}$. Caso não atenda, aplicar a perturbação do dado sensível, seja modificando o valor atributo sensível por um valor já conhecido, ou gerando um novo valor. Você deve gerar três datasets anonimizados de nome "lAnonAlunos.csv", onde $l = \{2, 3, 4\}$. Mostre na tela as classes de equivalência com os registros ordenados pelo valor do atributo sensível.

Para medir a utilidade do processo de anonimização para a diversidade, calcule o erro absoluto da frequência dos valores do atributo sensível entre os datasets original e anonimizados por l-diversidade com $l = \{2, 3, 4\}$.

3 Requisitos

- Linguagens: C++ ou Python
- Duplas: as mesmas do Trabalho I
- Preparar uma Demo para explicar, mostrar o seu programa e os resultados durante a aula de entrega. Escreva um Readme.txt descrevendo o projeto.
- Zipar o seu projeto (código fonte e executável), os datasets anonimizados, os gráficos e o Readme.txt em um único pacote e submeter via **Classroom**.
- O trabalho deverá ser entregue até as 14h da segunda-feira, dia 30/05/2022 e explicado durante as aulas dos dias 30/05 e 01/06, seguindo a mesma sequência de apresentação das duplas do Trabalho I.

4 Avaliação

Na avaliação serão considerados os seguintes indicadores:

- **Corretude** do programa;
- **Precisão** pela comparação do dataset original com o dataset anonimizado;
- Clareza na **explicação** do programa durante a Demo;
- **Pontualidade** e **documentação/qualidade** do código-fonte.