Texto para Discussão

2420

Rio de Janeiro, outubro de 2018

ipea
SUMÁRIO EXECUTIVO

O DESAFIO DO PAREAMENTO DE GRANDES BASES DE DADOS: MAPEAMENTO DE MÉTODOS DE *RECORD LINKAGE* PROBABILÍSTICO E DIAGNÓSTICO DE SUA VIABILIDADE EMPÍRICA

Peng Yaohao
Pesquisador do Programa de Pesquisa para o Desenvolvimento Nacional (PNPD) na Assessoria Técnica (Astec) do Ipea. *E-mail*: <peng.yaohao@ipea.gov.br>.

Lucas Ferreira Mation
Técnico de planejamento e pesquisa na Astec/Ipea. *E-mail*: <lucas.mation@ipea.gov.br>.

O *record linkage* (RL) consiste no problema associado ao pareamento de registros de bases de dados distintas, combinando informações de cada uma dessas bases relativas a uma única unidade de observação. O desafio é realizar o pareamento de forma que dados disponíveis da mesma pessoa não deixem de ser pareados, ao mesmo tempo que informações de pessoas diferentes não sejam combinadas como se fossem provenientes do mesmo indivíduo.

Métodos de RL dividem-se basicamente em pareamento determinístico e probabilístico. No primeiro, os pares de registros precisam ser iguais em um determinado conjunto de indexadores para que dois deles sejam considerados como um par; já a abordagem probabilística permite que a similaridade entre os registros possa ser parcial, averiguada de acordo com alguma métrica de distanciamento. Esta segunda abordagem possibilita critérios de pareamento mais flexíveis e menos restritivos, viabilizando, por exemplo, parearem-se registros de nomes com erro de digitação em alguma das bases comparadas.

No contexto da avaliação de políticas públicas, o pareamento de registros admite compreender o perfil dos habitantes de uma nação ou região, bem como em que medida o indivíduo está conectado e inserido à estrutura socioeconômica macroscópica, propiciando um mapeamento mais acurado da realidade do país. Essa compreensão é fundamental para a tomada de decisão nas esferas federal, estadual e municipal.

Este trabalho realizou um mapeamento dos principais métodos de RL determinístico e probabilístico na literatura científica, descrevendo suas principais diferenças conceituais e metodológicas, discutindo tanto modelos clássicos bem estabelecidos na área quanto desenvolvimentos recentes envolvendo métodos de *machine learning* (aprendizado de máquinas). Foram também registrados os índices mais utilizados para a avaliação do desempenho preditivo dos pareamentos, das técnicas mais relevantes para a definição de chaves de pareamento (*blocking*) e das principais implementações computacionais existentes em linguagem R.[1]

Verificou-se empiricamente o desempenho preditivo de algoritmos de pareamento probabilístico com a utilização de subconjuntos das bases do Cadastro Único e do Registro Nacional de Condutores Habilitados (Renach), testando-se diferentes abordagens computacionais, combinações de chaves de pareamento, funções de distanciamento de *strings* (sequência de caracteres) e algoritmo de pareamento fonético, com a análise da qualidade desses pareamentos e de sua exigência computacional. Ademais, realizou-se uma avaliação acerca do pareamento determinístico que serviu como controle para os pares verdadeiros, identificando padrões de discordância de variáveis referentes ao mesmo indivíduo em bases de dados distintas.

Os resultados fornecem evidências de um *trade-off* (dilema de escolha) entre os erros tipo I (falsos positivos) e tipo II (falsos negativos) e em que medida esse *trade-off* é sensível a mudanças nos parâmetros (seleção de variáveis de *blocking*, função de distanciamento de *strings* etc.) definidos para o pareamento probabilístico. Dessa forma, os achados deste trabalho podem servir como consulta para estudos futuros de pesquisadores de variadas áreas do conhecimento que envolvam pareamento de registros, permitindo uma escolha de parâmetros bem ajustada aos custos relativos de erros tipo I ou tipo II

1. R é uma linguagem de programação estruturada de acesso gratuito com foco em análises estatísticas e manipulação de dados.

associados ao problema de interesse específico. Além disso, podem subsidiar a identificação de padrões e boas práticas para a aplicabilidade eficiente do RL probabilístico para grandes bases de dados.

SUMÁRIO EXECUTIVO