International Journal of Population Data Science

Journal Website: www.ijpds.org

The Use of Density-Based Spatial Clustering of Application With Noise (DBSCAN) for Record Linkage in An Observational HIV Cohort

Olago, V[1], Bartels, L[2], Dhokotera, T[2], Bohlius, J[2], Egger, M[2,3], Singh, E[1,4], and Mazvita, S[1,4]

[1]National Health Laboratory Service (NHLS), National Cancer Registry (NCR), Johannesburg, South Africa

[2]Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland

[3]Centre for Infectious Disease Epidemiology and Research (CIDER), School of Public Health and Family Medicine, University of Cape Town, South Africa

[4]Division of Epidemiology and Biostatistics, School of Public Health, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, South Africa

Introduction

The South African HIV Cancer Match (SAM) study is a probabilistic record linkage study involving creation of an HIV cohort from laboratory records from the National Health Laboratory Service (NHLS). This cohort was linked to the pathology based South African National Cancer Registry to establish cancer incidences among HIV positive population in South Africa. As the number of HIV records increases, there is need for more efficient ways of de-duplicating this big-data. In this work, we used clustering to perform big-data deduplication.

Objectives and Approach

Our objective was to use DBSCAN as clustering algorithm together with bi-gram word analyser to perform big-data deduplication in resource- limited settings. We used HIV related laboratory records from entire South Africa collated in the NHLS Corporate Data Warehouse for period 2004-2014. This involved data pre-processing, deterministic deduplication, ngrams generation, features generation using Term Frequency Inverse Document Frequency vectorizer, clustering using DBSCAN and assigning cluster labels for records that potentially belonged to the same person. We used records with national identification numbers to assess quality of deduplication by calculating precision, recall and f-measure.

Results

We had 51,563,127 HIV related laboratory records. Deterministic deduplication resulted in 20,387,819 patient record deduplicates. With DBSCAN clustering we further reduced this to 14,849,524 patient record clusters. In this final dataset, 3,355,544 (22.60%) patients had negative HIV test, 11,316,937 (76.21%) had evidence for HIV infection, and for 177,043 (1.19%) the HIV status could not be determined. The precision, recall and f-measure based on 1,865,445 records with national identification numbers were 0.96, 0.94 and 0.95, respectively.

Conclusion / Implications

Our study demonstrated that DBSCAN clustering is an effective way of deduplicating big datasets in resource-limited settings. This enabled refining of an HIV observational database by accurately linking test records that potentially belonged to the same person. The methodology creates opportunities for easy data profiling to inform public health decision making.

*Corresponding Author:
Email Address: VictorO@nicd.ac.za (V Olago)

https://doi.org/10.23889/ijpds.v5i5.1422