

Privacy-Preserving Data Mining Techniques on an Open Source Toolkit

Eun-Joo Lee*

Department of Computer Science,
East Stroudsburg University of Pennsylvania,
327 Science and Technology Center,
East Stroudsburg, PA 18301-2999, USA

February 12, 2009

Abstract

A number of commercial tools that are available on the market are excellent in handling very large datasets. But they are limited in fixing and tuning things that suit our specific needs. Commercial vendors, naturally, need to be convinced of the usefulness of implementing new algorithms. On the other hand, a vast selection has been available for deployment in R¹ for a long time. In this tutorial I will show how we achieve privacy preserving data mining using open source Rattle package² for R. A privacy sensitive dataset is first recoded properly and distorted via the data distortion algorithm we developed. Multiple regression models are then fitted to the distorted dataset on Rattle GUI. Comparisons of the predictive power of the regression models fitted to the undistorted and distorted datasets will be presented.

*Email: elee@po-box.esu.edu. URL: <http://www.esu.edu/~elee>. This author's research work was supported by Keystone Innovation Starter Kit Grant Program - Pennsylvania Department of Community and Economic Development.

¹R is a free open source statistical language/software environment for statistical computing and graphics. It compiles and runs on a wide variety of Linux/UNIX platforms, Windows and MacOS.

²Rattle (the R Analytical Tool To Learn Easily) is an intuitive interfaced data mining toolkit used to analyse very large collections of data. It runs under GNU/Linux, Macintosh OS/X, and MS/Windows and is open source and freely available from Togaware.