Supplementary Material for
On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study
by G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent and M. E. Houle
Data Mining and Knowledge Discovery 30(4): 891-927, 2016, DOI: 10.1007/s10618-015-0444-8

WDBC

This data set describes nuclear characteristics for breast cancer diagnosis. Again, we consider examples of benign cancer as inliers and malignant cancer as outliers. In the preprocessing, we follow Zhang et al. [1], downsampling the outliers to 10. The processed database has 30 numeric attributes and 367 instances, namely 10 outliers (2.72%) and 357 inliers (97.28%).

References:

[1] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In Proc. PAKDD, pages 813-822, 2009.

Download all data set variants (1.1 MB). Access original data (wdbc.data)

- WDBC (version#01)
- WDBC (version#02)
- WDBC (version#03)
- WDBC (version#04)
- WDBC (version#05)
- WDBC (version#06)
- WDBC (version#07)
- WDBC (version#08)
- WDBC (version#09)
- WDBC (version#10)

File generated: 2016-07-05T21:48:16