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

WBC

This dataset consists of examples of different cancer types, benign or malignant. Examples of benign cancer are considered inliers, examples of malignant cancer are considered outliers. After downsampling the outliers, following Schubert et al. [1], 10 outliers remain. 234 instances are duplicates (231 inliers and 3 outliers), therefore 229 outliers were removed from the data set with duplicates and 226 outliers from the dataset without duplicates. Furthermore, we removed 16 instances with missing values, two of them being outliers and 14 inliers. The processed data set has 9 numeric attributes and 454 instances, namely 10 outliers (2.2%) and 444 inliers (97.8%). The same pre-processing has also been applied in [2] and [3].

References:

[1] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In Proc. SDM, pages 1047-1058, 2012.
[2] A. Zimek, M. Gaudet, R. J. G. B. Campello, and J. Sander. Subsampling for efficient and effective unsupervised outlier detection ensembles. In Proc. KDD, pages 428-436, 2013.
[3] H.-P. Kriegel, P. Kroeger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In Proc. SDM, pages 13-24, 2011.

Download all data set variants (57.1 kB). Access original data (breast-cancer-wisconsin.data)

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

File generated: 2016-07-05T21:48:14