Datasets

The **annthyroid** dataset is derived from the "Thyroid Disease" dataset from the UCIMLR. The original data has 7200 instances with 15 categorical attributes and 6 real-valued attributes. The class labels are "normal", "hypothyroid", and "subnormal". For anomaly detection, the "hypothyroid" and "subnormal" classes are combined into 534 outlier instances, and only the 6 real-valued attributes are used.

The **arrhythmia** dataset is derived from "Arrhythmia" dataset from the UCIMLR. The original dataset contains 452 instances with 279 attributes. There are five categorical attributes which are discarded, leaving this as a 274-dimensional dataset. The instances are divided into 16 classes. The eight smallest classes collectively contain 66 instances and are combined into the outlier class.

The **breastw** dataset is also derived from the "Breast Cancer Wisconsin (Original)" dataset. This is a 9-dimensional dataset containing 683 instances of which 239 represent malignant tumors and are treated as the outlier class.

The **cardio** dataset is derived from the "Cardiotocography" dataset. The dataset is composed of measurements of fetal heart rate and uterine contraction features on cardiotocograms. The are each labelled "normal", "suspect", and "pathologic" by expert obstetricians. For anomaly detection, the "normal" class forms the inliers, the "suspect" class is discarded, and the "pathologic" class is downsampled to 176 instances forming the outliers. This leaves us with 1831 instances with 21 attributes in the dataset.

The **cover** dataset is derived from the "Covertype" dataset. The original dataset contains 581,012 instances with 54 attributes. The dataset is used to predict the type of forest cover solely from cartographic variables. The instances are labelled into seven different classes. For outlier detection, we use only the 10 quantitative attributes, type 2 (lodgepole pine) as the inliers, and type 4 (conttonwood/willow) as the outliers. The remaining classes are discarded. This leaves us with a 10-dimensional dataset with 286,048 instances of which 2,747 are outliers.

The **glass** dataset is derived from the "Glass Identification" dataset. The study of classification of types of glass was motivated by criminological investigations where glass fragments left at crime scenes were used as evidence. This dataset contains 214 instances with nine attributes. While there are several different types of glass in this dataset, class 6 is a clear minority with only nine instances and, as such, points in class 6 are treated as the outliers while all other classes are treated as inliers.

The **http** dataset is derived from the original "KDD Cup 1999" dataset. It contains 41 attributes (34 continuous and 7 categorical) which are reduced to 4 attributes (service, duration, src_bytes, dst_bytes). Only the "service" attribute is categorical, dividing the data into http, smtp, ftp, ftp_data, others subsets. Here, only the "http" data is used. The values of the continuous attributes are centered around 0, so they have been log-transformed far away from 0. The original data contains 3,925,651 attacks in 4,898,431 records. This smaller dataset is created with only 2,211 attacks in 567,479 records.

The **ionosphere** dataset is derived from the "Ionosphere" dataset. It consists 351 instances with 34 attributes. One of the attributes is always 0 and, so, is discarded, leaving us with a 33-dimensional dataset. The data comes from radar measurements of the ionosphere from a system located in Goose Bay, Labrador. The data are classified into "good" if the radar returns evidence some type of structure in the ionosphere, and "bad" if not. The "good" class serves as the inliers and the "bad" class serves as the outliers.

The **lympho** dataset is derived from the "Lymphography" dataset. The data contain 148 instances with 18 attributes. The instances are labelled "normal find", "metastases", "malign lymph", and "fibrosis". The two minority classes only contain a total of six instances, and are combined to form the outliers. The remaining 142 instances form the inliers.

The **mammography** dataset is derived from the original "Mammography" dataset provided by Aleksandar Lazarevic. Its goal is to use x-ray images of human breasts to find calcified tissue as an early sign of breast cancer. As such, the "calcification" class is considered as the outlier class while the "non-clacification" class is the inliers. We have 11,183 instances with 6 attributes, of which 260 are "calcifications."

The **mnist** dataset is derived from the classic "MNIST" dataset of handwritten digits. Digit-zero is considered the inlier class while 700 images of digit-six are the outliers. Furthermore, 100 pixels are randomly selected as features from the original 784 pixels.

The **musk** dataset is derived from its namesake in the UCIMLR. It is created from molecules that have been classified by experts as "musk" or "non-musk". The data are downsampled to 3,062 instances with 166 attributes. The "musk" class forms the outliers while the "non-musk" class forms the inliers.

The **optdigits** dataset is derived from the "Optical Recognition of Handwritten Digits" dataset. Digits 1–9 form the inliers while 150 samples of digit-zero form the outliers. This gives us a dataset of 5,216 instances with 64 attributes.

The **pendigits** dataset is derived from the "Pen-Based Recognition of Handwritten Digits" dataset from the UCI Machine Learning Repository. The original collection of handwritten samples is reduced to 6,870 points, of which 156 are outliers.

The **pima** dataset is derived from the "Pima Indians Diabetes" dataset. The original dataset presents a binary classification problem to detect diabetes. This subset was restricted to female patients at least 21 years old of Pima Indian heritage.

The **satellite** dataset is derived from the "Statlog (Landsat Satellite)" dataset. The smallest three classes (2, 4, and 5) are combined to form the outlier class while the other classes are combined to form the inlier class. The train and test subsets are combined to produce a of 6,435 instances with 36 attributes.

The **satimage-2** dataset is also derived from the "Satlog (Landsat Satellite)" dataset. Class 2 is downsampled to 71 instances that are treated as outliers, while all other classes are combined to form an inlier class. This gives us 5,803 instances with 36 attributes.

The **shuttle** dataset is derived from the "Statlog (Shuttle)" dataset. This

are seven classes in the original dataset. Here, class 4 is discarded, class 1 is treated as the inliers and the remaining classes, which are comparatively small, form an outlier class. This gives us 49,097 instances with 9 attributes, of which 3,511 are outliers.

The **smtp** is also derived from the "KDD Cup 1999" dataset. It is preprocessed in the same way as the **http** dataset, except that the "smtp" service subset is used. This version of the dataset only contains 95,156 instances with 3 attributes, of which 30 instances are outliers.

The **thyroid** dataset is also derived from the "Thyroid Disease" dataset. The attribute selection is the same as for the **annthyroid** dataset but only the 3,772 training instances are used in this version. The "hyperfunction" class, containing 93 instances, is treated as the outlier class, while the other two classes are combined to form an inlier class.

The **vertebral** dataset is derived from the "Vertebral Column" dataset. 6 attributes are derived to represent the shape and orientation of the pelvis and lumbar spine. Each instance comes from a different patient. The "Abnormal (AB)" class of 210 instances are used as inliers while the "Normal (NO)" class is downsampled to 30 instances to be used as outliers.

The **vowels** dataset is derived from the "Japanese Vowels" dataset. THE UCIMLR presents this data as a multivariate time series of nine speakers uttering two Japanese vowels. For outlier detection, each frame of each time-series is treated as a separate point. There are 12 features associated with each time series, and these translate as the attributes for each point. Data from speaker 1, downsampled to 50 points, form the outlier class/ Speakers 6, 7, and 8 form the outlier class. The rest of the points are discarded. This leaves is with 1,456 points in 12 dimensions, of which 50 are outliers.

The **wbc** dataset is derived from the "Wisconsin-Breast Cancer (Diagnostics)" dataset. The dataset records measurements for breast cancer cases. The benign class is treated as the inlier class, while the malignant class is downsampled to 21 points and serves as the outlier class. This leaves us with 278 points in 30 dimensions.

The **wine** dataset is a collection of results of a chemical analysis of several wines from a region in Italy. The data contain 129 samples having 13 attributes, and divided into 3 classes. Classes 2 and 3 form the inliers while class 1, downsampled to 10 instances, is the outlier class.