SyntImbNoisyDataForClassification

A synthetic imbalanced data set collections for binary classification task. Each data set consists of 4 parts. The parts of the datasets in the train folder contain different level of label noise. The parts of the datasets in the test folder are noise-free.

Content of the data files in the train folder

Each file consists of two parts, a header and a data part.

Structure of the headers:

@relation Name of the data set
@attribute X0 [min_value, max_value]
@attribute X1 [min_value, max_value]
...
@attribute X{D-1} [min_value, max_value]
@attribute cluster_idx {0,1,2}
@attribute is_noise {0,1}[^1]
@attribute Class {0,1}
@inputs X0, X1, X2, ..., X{D-1},cluster_idx,is_noise
@output Class

[^1]: If the data set is noise free, the set is {0}.

Warning

Make sure not using the last two inputs (extra information about the samples) as a part of the feature vectors.

Structure of the name

SyntheticData-N_v0-D_v1-Nmin_v2-Zmin_v3-CL_v4_v5_R0_extra_info.dat

N_v0: v0 is the number of samples; {600, 1200}
D_v1: v1 is the number of features; {4, 8, 16}
Nmin_v2: the number of the samples in the minority class; {60, 100, 140}
Z_v3: the noise level, expressed as a percentage of the minority class; {0, 4, 8, 16, 32}
CL_v4_v5: v4 and v5 are the number of the clusters in the minority and the majority classes {1, 2, 3}

Data part

The data part starts with the line after the "@data" string. Each line contains D values (X0, ..., X{D-1}) belonging to one sample, followed by three other values (cluster_idx, is_noisy, Class) that provide additional information about the sample:

- cluster_idx: The cluster index of a sample. Samples in the same cluster have the same index, samples in different clusters have different indices.
- is_noise: The is_noise is 0 if the class label of the sample is correct, otherwise 1.
- Class: The sample's class label. The label of the majority class is 0, the label of the minority class is 1. Remember, this label may be corrupted (see is_noise).

@data Comma separated values:
x0,x1,x2,...,x{D-1},cluster_idx,is_noise,class_label

Content of the data files in the test folder

Similar to the training data, but without the additional information (cluster index, noise level). The training sets do not contain label noise.

The related files in the train and test folders

A data set in a train folder named SyntheticData-N_v0-D_v1-Nmin_v2-Zmin_v3-CL_v4_v5_R0_extra_info.dat is associated with three data sets in a test folder named:

- SyntheticData-N_v0-D_v1-Nmin_v2-Zmin_0-CL_v4_v5_R0_test_1.dat
- SyntheticData-N_v0-D_v1-Nmin_v2-Zmin_0-CL_v4_v5_R0_test_2.dat
- SyntheticData-N_v0-D_v1-Nmin_v2-Zmin_0-CL_v4_v5_R0_test_3.dat

The possible values of v0, v1, ... v5 can be see under the "Structure of the name" section.

The data sets were generated by

Attila Fazekas, University of Debrecen, and
Szilvia Szeghalmy, University of Debrecen for the study entitled "A Comparative Study on Noise Filtering of Imbalanced Data Sets".