Code & Data Appendix

Background

This appendix describes how to reproduce our experiments and include the source code, datasets, software & hardware setting, dependencies, and instructions for running.

Repository Files

- └───data/
  - └───[dataset]/
    - ├───[dataset]_features.npy: The preprocessed features of [dataset]
    - ├───[dataset]_labels.npy: The labels of [dataset]
    - ├───[dataset]_pairs_X.npy: The features of the pairs used for training the Siamese network of [dataset]
    - └───[dataset]_pairs_y.npy: The labels of the pairs used for training the Siamese network of [dataset]
- ├───src/: The source code implementation of our approach

Datasets

These are the datasets we used in our experiments and described in the paper.

As noted, the datasets were taken from ODDS.

Dataset	#Samples	#Dim	Outliers
Annthyroid	7200	6	7.42 (%)
Cardio	1831	21	9.6 (%)
Mammo	11183	6	2.32 (%)
Satellite	6435	36	32 (%)

Dataset	#Samples	#Dim	Outliers
Seismic	2584	11	6.5 (%)
Thyroid	3772	6	2.5 (%)
Vowels	1456	12	3.4 (%)
Yeast	1364	8	4.7 (%)

Dependencies

The required dependencies are specified in `environment.yml`.
For setting up the environment, use Anaconda:

```
$ conda env create -f environment.yml
$ conda activate adtta
```

Running the Code

NOTES:

- A valid dataset name (instead of the mammo dataset) could be from the one described above.
- It is important to run the scripts from being inside `src/` (i.e., `$ cd src/`)

- **Preprocessing**

  A preprocessing phase on the desired dataset.

```
$ cd src/
$ python preprocess.py --dataset mammo
```

From this script, 4 files are going to be generated in the desired dataset `data/` folder:

```
- mammo_features.npy
- mammo_labels.npy
```

```
- mammo_pairs_X.npy
- mammo_pairs_y.npy
```
With these files, you can run the train and test.

- ## **Train & Test**

To train the estimator, as well as the Nearest Neighbors models (both Euclidean and Siamese), run:

```
$ cd src/
$ python main.py --dataset mammo --neighbors 10 --augmentations 7
```