Pycol - A Python package for dataset complexity measures

January 31, 2024

1 Validation

The validation of the implemented measures was divided into two groups. The first group was used for the measures already implemented in *pymfe*, whose results were compared to those given by the measures implemented in the *pycol* package. The remaining measures without an available implementation were tested using synthetic datasets.

The artificial datasets were created using the data generator introduced in git. With this generator, it is possible to create data clusters of variable size and topology. The generator divides the creation of samples into multiple regions, and the user can configure the number of regions, their size, shape, and location. For each type of shape available, an algorithm uniformly fills the area inside this region with safe and borderline samples. Afterward, the area around the region is populated with rare and outlier examples.

Results for the first group of measures were compared in four datasets of the KEEL repository (Derrac et al., 2015). The characteristics of these datasets are shown in Table 1. The datasets were chosen to have varying numbers of instances (from 205 to 2200), features (from 4 to 7), and binary and non-binary classification problems. For non-binary datasets, the OvO results are summarized using a mean.

Name	**#Samples**	**#Features**	**#Classes**
newthyroid	215	5	3
ecoli	335	7	8
balance	625	4	3
titanic	2200	4	2

Table 1: KEEL repository dataset characteristics.

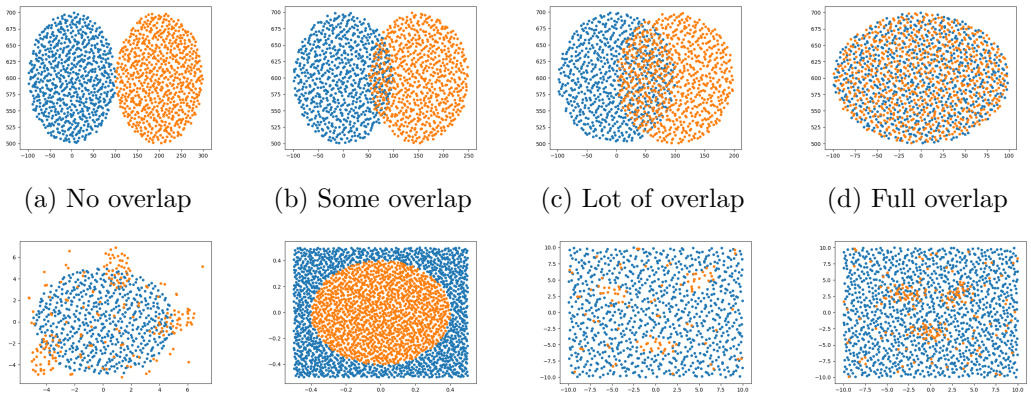
	newthyroid		ecoli		balance		titanic	
Measure	**pycol**	**pymfe**	**pycol**	**pymfe**	**pycol**	**pymfe**	**pycol**	**pymfe**
F1	**0.5429**	**0.5124**	**N.A**	**0.5677**	**0.8342**	**0.8306**	**0.8370**	**0.9030**
F1v	0.0498	0.0498	0.1240	0.1240	0.2292	0.2292	0.4356	0.4356
F2	0.0005	0.0005	0.000	0.0000	1.000	1.0000	1.000	1.000
F3	0.1349	0.1349	0.9569	0.9569	0.5980	0.5980	1.000	1.000
N1	0.1023	0.1023	0.3035	0.3035	0.2752	0.2752	0.3198	0.3198
N2	**0.2368**	**0.2478**	**0.4160**	**0.3966**	**0.4036**	**0.4231**	**0.0270**	**0**
N3	0.0279	0.0279	0.2083	0.2083	0.2128	0.2128	0.2221	0.2221
N4	0.0093	0.0093	0.1398	0.1398	0.1312	0.1312	0.4329	0.4329
LSC	0.7702	0.7702	0.9741	0.9741	0.9663	0.9663	0.9999	0.9999
T1	0.2279	0.2279	0.7529	0.7529	0.3648	0.3648	0.004	0.004

Table 2: Complexity measures results for the KEEL datasets

The results of the validation of the first group can be found in Table 2. All measures except F1 and N2 obtain the same result in both packages for every dataset, indicating the implementation is valid for at least those. As for F1, the difference in results is due to a slight change in the implementation where the means of each feature is not normalized, justifying variations between both approaches. Finally, for N2, the differences between the two packages are also very small, likely due to the default distance metrics used in each one, which are slightly different in terms of normalization.

For the second group of measures, two sets of tests were made. The first set of tests starts by creating two clusters or classes with 750 samples each. The overlapped region between these clusters is increased until the two clusters are completely overlapped (Figure 1a-1d). A second set of more complex artificial datasets was also used, which were taken from git. Table 3 presents the characteristics of the datasets and a 2D view of the datasets is also presented in Figure 1e-1h.

Results for the second group of metrics in each of the artificial datasets are presented in Table 4. Ideally, if the complexity measures are implemented correctly, their values will indicate a higher complexity for datasets with higher overlap. For example, an increase in overlap should be seen from 1a-1d.



(a) No overlap (b) Some overlap (c) Lot of overlap (d) Full overlap

(e) circles-2d (f) spheres-2d (g) paw3-2d (h) paw3-3d

Figure 1: Artificial datasets used. The first 4 figures represent the first set of datasets with increasing overlap. The next four figures represent the artificial datasets generated using git

Name	#Samples	#Features	#Classes	Class Ratio
cirles-2d	800	2	2	1:3
spheres-2d	3000	2	2	1:1
paw3-2d	1000	2	2	1:9
paw3-3d	1500	2	3	1:7

Table 3: Characteristics of the second set of artificial datasets

Overall, the results for the first set of tests show that all the metrics behave according to the expected, as when the overlapped region increases, their values increase, too. A notable exception to this rule are the SI, purity and neighbourhood separability measures,

	First set of tests				Second set of tests			
Measure	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**circles-2d**	**spheres-2d**	**paw3-2d**	**paw3-3d**
R value	0.003	0.1140	0.2953	0.7107	0.2050	0.01967	0.0819	0.0799
D3	[2,3]	[89,82]	[232,211]	[532,534]	[41,123]	[29,30]	[39,84]	[11,69]
CM	0.003	0.114	0.2953	0.7106	0.205	0.01967	0.082	0.08
kDN	0.0052	0.0957	0.2406	0.58413	0.2557	0.0334	0.1433	0.1366
DBC	0.0096	0.2181	0.5776	0.8535	0.3632	0.0538	0.2347	0.2095
SI	0.9966	0.7180	0.5776	0.6773	0.87625	0.9893	0.8393	0.927
input noise	0.4990	0.5927	0.7410	0.9983	0.9568	0.7958	0.9886	0.9760
borderline	0.8000	14.9300	40.7300	98.0670	33.625	5.433	16.7333	16.888
deg overlap	0.0147	0.1753	0.4346	0.9993	0.5787	0.0923	0.3800	0.3600
C1	0.1328	0.2003	0.3037	0.5011	0.1864	0.08047	0.0	1.9047
C2	0.1664	0.2267	0.3199	0.5031	0.3256	0.3682	0.0111	0.1815
Clst	0.004	0.1220	0.3366	0.7147	0.2987	0.02467	0.1653	0.1590
purity	0.0228	0.0247	0.0181	0.0001	0.024	0.008	0.003	0.037
neigh. sep.	0.2965	0.26976	0.2237	0.1228	0.2214	0.2881	0.0360	0.2568

Table 4: Complexity Measure Results in the Artificial Datasets. The first set of tests is represented in the first four rows, while the second is in the last four rows.

however these measures work differently from the rest where smaller values indicate higher complexity, so the values presented still indicate that the implementation is valid.

As for the second set of tests, as most of the experimented metrics take into account the local region around each sample, it is expected that the values for the measures will represent lower complexity, since as seen in Figure 1 these datasets present low local overlap, with very well-defined clusters. An exception is input noise, a feature-based metric that should have high values since none of the two features can separate any of the datasets linearly.

The results of these experiments are mostly within expectations, as most of the measures indicate a low complexity. The measures (between 0 and 1) are lower than 0.5 when 1 represents high complexity and higher than 0.5 when 1 represents low complexity. Also, as expected, input noise, being a feature-based metric, gives very high values, representing high complexity. The two measures that got results that defied expectations were purity and neighbourhood separability, which both have similar formulations. However, being multi-resolution metrics, this result is most likely due to the need for better parametrization, which is very dataset-dependent.

References

Artificial dataset generator repository, https://github.com/sysmon37/datagenerator. URL `https://github.com/sysmon37/datagenerator/tree/newsampling`. Accessed: 2022-01-22.

J Derrac, S Garcia, L Sanchez, and F Herrera. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Log. Soft Comput*, 17:255–287, 2015.