A List of Suggested Dataset Descriptors

By: Noureddin Sadawi

Mar 2014

1 By Noureddin

1. **ClassificationOrRegression** Whether it's a classification or regression dataset/problem
2. **Percentage of Nominal Attributes** (Number of Nominal attributes/Number of Attributes) * 100
3. **Percentage of Binary Attributes** (Number of Binary attributes/Number of Attributes) * 100
4. **Percentage of Numeric Attributes** (Number of Numeric attributes/Number of Attributes) * 100
5. **Has Missing Values** Whether the dataset has missing values (yes | No)
6. **Percentage of Present Values** (Number of non-missing values/Total Number of values) * 100
7. **Percentage of Missing Values** (Number of missing values/Total Number of values) * 100

2 From DMOP

1. **AverageAbsoluteFeatureCorrelation:** METAL characteristic: Average absolute correlation between continuous features.
2. **AverageCategoricalFeaturePairsMutualInformation:** METAL characteristic: Average mutual information between pairs of categorical features.
3. **AverageFeatureEntropy:** METAL characteristic: Average feature Entropy

4. **BetweenGroupsSumSquaresCrossProducts:** METAL characteristic: A matrix containing the difference between the matrix of total and the matrix of within-groups sums of squares and cross products.
5. **EigenValuesLinearDiscriminantFunctions:** METAL characteristic: A vector of eigen values of linear discriminant functions.
6. **NoiseSignalRatio:** METAL characteristic: Noise signal ratio
7. **NumberOfCategoricalFeatures:**
8. **NumberOfContinuousFeatures:**
9. **NumberOfFeatures:**
10. **NumberOfHOutliers:** METAL characteristic: Number of continuous features with outliers.
11. **NumberOfInstances:**
12. **NumberOfInstancesPerFeature:** From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.
13. **ProportionOfCategoricalFeatures:**
14. **ProportionOfHOutliers:** METAL characteristic: Proportion of continuous features with outliers.
15. **TotalSumSquaresCrossProducts:** METAL characteristic: Matrix of total sums of squares and cross products of features.
16. **WithinGroupsSumSquaresCrossProducts:** METAL characteristic: matrix of within-groups sums of squares and cross products of features.
17. **AverageSVMFeatureWeight:**

2.1 CategoricalLabeledDataSetCharacteristic

1. AverageMutualInformation: METAL characteristic: Average mutual information
2. AverageReliefFeatureWeight:
3. CanonicalCorrelationBestLinearCombination: METAL characteristic: Canonical correlation of the best linear combination of features to distinguish between classes.
4. ClassAbsoluteFrequencies: METAL characteristic: Absolute class frequencies. Stored in a vector indexed by each class value.

5. ClassCovarianceMatrices: METAL characteristic: Class covariance matrices. Stored in a vector indexed by class and each containing a matrix of (features x features)

6. ClassEntropy: METAL characteristic: Class entropy.

7. ClassRelativeFrequencies: METAL characteristic: Relative class frequencies. Stored in a vector indexed by each class value.

8. ErrorRateOf1NNClassifier: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

9. ErrorRateOfLinearClassifierLP: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

10. FeatureMutualInformationPerClass: METAL characteristic: For each categorical feature, the mutual information between the feature and the class. It is stored in a vector indexed by each categorical feature.

11. FeatureValueFrequenciesPerClass: METAL characteristic: For each k value of each j categorical feature and each i class, the proportion of cases that have the k value in the j feature and belong to the i class. It is stored in a vector indexed by each categorical feature and containing a flat contingency tables that combine the values of the categorical feature with the class values.

12. MaximumFeatureEfficiency: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

13. MaximumFishersDiscriminantRatio: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

14. MinimumSumOfErrorDistanceByLP: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

15. NonLinearityOf1NNClassifier: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

16. NonLinearityOfLinearClassifierLP: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

17. NumberOfClasses:

18. ProportionOfBoundaryPoints: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

19. ProportionPointsWithRetainedAdherence: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

20. RatioOfAverageIntraInterDistances: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

21. VolumeOfOverlapRegion: From Mitra Basu and Tin Kam Ho. Data Complexity in Pattern Recognition. Springer, 2006.

3 Questions

1. Each performance measure will convey some information and hide other
2. Therefore, there is an information tradeoff carried through the different metrics
3. Would it be a good idea to ask experts to vote for their favourite metrics and choose the best 15/20?
4. Does the sparsity of dataset affect the performance of [binary] classifiers?
5. For dataset descriptors, we have to decide whether datasets are always of the same nature
6. QSAR is usually a binary classification problem, OR, a regression problem with a threshold which makes it a binary classification problem
7. QSAR experts use metrics such as RIE, BEDROC, and pROC which emphasize, what they call, the *early recognition* problem specific to VS. These are not in WKEA!

4 Columns

1. TID 1
2. PREF_NAME Maltase-glucoamylase
3. ORGANISM Homo sapiens
4. TARGET_CHEMBL_ID CHEMBL2074
5. MOLREGNO 123534
6. MOLECULE_CHEMBL_ID CHEMBL307429
7. MEDIAN_PXC50

8. ACTIVITY_FLAG
9. std_smiles
10. FCFP4_1024
11. FCFP6_1024
12. ECFP4_1024
13. ECFP6_1024
14. ALogP: Molecular hydrophobicity (lipophilicity), usually quantified as log P (the logarithm of 1-octanol/water partition coefficient), is an important molecular characteristic in drug discovery.
15. Molecular_Weight
16. Num_H_Acceptors
17. Num_H_Donors
18. Num_Atoms
19. Num_RotatableBonds
20. Num_AromaticRings
21. TPSA
22. MDL2DKeys166
23. MDL2DKeys960
24. FCFP6_2048
25. ECFP6_2048
26. FCFP6_4096
27. ECFP6_4096