Dataset	Number of features	Number of classes	Number of instances	Number of Points
Iris	4	3	150	16
Glass	10	7	214	1024
Vowel	10	11	990	1024
Wine	13	3	178	8192
Australian	14	2	690	16384
Letter	16	26	20,000	65536
Zoo	17	7	101	131072
Vehicle	18	4	846	262144
Segmentation	19	7	2310	524288
German	24	2	1000	16777216
WBCD	30	2	569	1073741824
Ionosphere	34	2	351	17179869184
Satellite	36	6	6435	
Lung Cancer	56	3	32	
Sonar	60	2	208	
LibrasMovement	90	15	360	
Hill-Valley	100	2	606	
Spectrometer	102	2	531	
Musk1	166	2	476	
Semeion	256	2	1593	
arrhythmia	279	16	452	
Madelon	500	2	4400	
Secom	591	2(-1,1)	1567	
Isolet5	617	2	1559	
Multiple Features	649	10	2000	
Gisette	5000	2	13500	
p53Mutants	5409	2	16772	
Amazon	10000	50	1500	
Arcene	10000	2(-1,1)	900	
Dexter	20000	2(-1,1)	2600	

Table 1: Datasets

Dataset	Number of features	Number of classes	Number of instances	Number of Points
Lymphography (Lymph)	18	4	148	
mushroom	22	2	5644	
Spect	22	2	267	
leddisplay	24	10	1000	
dermatology	34	6	366	
Connect4	42	3	44473	
soybeanlarge	35	19	307	
Chess	36	2	3196	
Splice	60	3	3190	

Table 2: Datasets2
mushroom: remove missing values (8124-2480=5644)

..

1 Discretise Dataset Sets

Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy

The four data sets we used are shown in Table 1. They have been extensively used in earlier studies [1], [13], [26], [27], [5]. The first two data sets, HDR-MultiFeature (HDR) and Arrhythmia (ARR), are also available on the UCI machine learning archive [28]. The latter two, NCI and Lymphoma (LYM), are available on the respective authors' Web sites. All the raw data are continuous. Each feature variable in the raw data was preprocessed to have zero mean-value and unit variance (i.e., transformed to their z-scores). To test our approaches on both discrete and continuous data, we discretized the first two data sets, HDR and ARR. The other two data sets, NCI and LYM, were directly used for continuous feature selection.

The data set HDR [6], [14], [13], [28] contains 649 features for 2,000 handwritten digits. The target class has 10 states, each of which has 200 samples. To discretize the data set, each feature variable was binarized at the mean value, i.e., it takes 1 if it is larger than the mean value and -1 otherwise. We selected and evaluated features using 10-fold Cross-Validation (CV). The data set ARR [28] contains 420 samples of 278 features. The target class has two states with 237 and 183 samples, respectively. Each feature variable was discretized into three states at the positions $Mean \pm Std$ ($Mean$ is the mean value and Std the standard deviation): it takes -1 if it is less than $Mean - mStd$, 1 if larger than $Mean + Std$, and 0 if otherwise. We used 10-fold CV for feature selection and testing.

The data set NCI [26], [27] contains 60 samples of 9,703 genes; each gene is regarded as a feature. The target class has nine states corresponding to different types of cancer; each type has two to nine samples. Since the sample number is small, we used the Leave-One-Out (LOO) CV method in testing. The data set LYM [1] has 96 samples of 4,026 gene features. The target class corresponds to nine subtypes of the lymphoma. Each subtype has two to 46 samples. The sample numbers for these subtypes are highly skewed, which makes it a hard classification problem. Note that the feature numbers of these data sets are large (e.g.,NCI has nearly 10,000 features).Thesedata sets represent some real applications where expensive feature selection methods (e.g., exhaustive search) cannot be used directly. They differ greatly in sample size, feature number, data type (discrete or continuous), data distribution, and target class type (multiclass or 2-class). In addition, we studied different mutual information calculation schemes for both discrete and continuous data and provided results using different classifiers and different wrapper selection schemes. We believe these data and methods provide a comprehensive testing suit for feature selection methods under different conditions.

2 Generate Datasets

- ADO.NET: Creating a DataSet with Code: ADO.NET http://www.knowdotnet.com/articles/cust

- R: mlbenck package, but only small number of features, continuous features
- Dataset Generator (datgen): http://www.datasetgenerator.com/