RELEVANCE AND MUTUAL INFORMATION-BASED FEATURE DISCRETIZATION[1]

Artur J. Ferreira[1,3]
arturj@isel.pt

Mário A. T. Figueiredo[2,3]
mtf@lx.it.pt

[1]Instituto Superior de Engenharia de Lisboa
[2]Instituto Superior Técnico
[3]Instituto de Telecomunicações, Lisboa, PORTUGAL

Saturday, 16 February 2013

[1]ICPRAM2013, Barcelona, Spain, 15-18 February, 2013

Outline

Introduction and Motivation

The use of feature discretization (FD) techniques for machine learning problems:

1. mandatory in some cases, optional in other cases
2. finds a convenient representation, ignoring irrelevant fluctuations on the data
3. yields compact data representations, reducing the memory requirements to store the data
4. reduces training time and improves classification accuracy
5. with or without a coupled *feature selection* (FS) technique, may improve the results of many learning methods

Feature Discretization: taxonomy

FD techniques are usually categorized among five axes:

1. **unsupervised** or **supervised**; the latter uses class labels to compute the discretization intervals
2. **static** (single pass assuming independent features) or **dynamic** (take dependencies into account)
3. **global** (discretizes the entire feature space) or **local** (discretizes some features, as needed)
4. **top-down** (splitting) or **bottom-up** (merging)
5. **direct** (sets a priori the number of bits per feature) or **incremental** (starts with a coarse discretization pass for all features and subsequently allocates more bits to each feature)

Feature Discretization: quality indicators

The quality of discretization is usually assessed by two indicators:

- the **generalization error**
- the **complexity** (number of intervals or equivalently the number of bits used to represent each instance)

Some facts from the FD literature:

- supervised FD may, in principle, lead to better classifiers
- however, it has been found that unsupervised FD methods perform well on different types of data
- no technique is uniformly better than all the others
- the performance of a FD method strongly depends on the data

Feature Discretization: unsupervised methods

Some commonly used **unsupervised** FD methods:

- *equal-interval binning* (EIB) - uniform quantization
- *equal-frequency binning* (EFB) - non-uniform quantization in which the number of occurrences in each interval is the same
- *proportional k-interval discretization* (PkID) - the number/size of intervals depend on the number of training instances
- *unsupervised Linde-Buzo-Gray* (U-LBG) - discrete features with minimum *mean square error* (MSE) with the original ones

U-LBG1/2 Algorithms (key ideas)

U-LBG1 and U-LBG2 are two versions of the U-LBG approach:

- rationale - low MSE between discrete and original features is adequate for learning
- the LBG algorithm is applied individually to each feature
- U-LBG1 uses a variable number of bits per feature, being stopped when:
  - the MSE distortion falls below some threshold Δ
  - or the maximum number of bits per feature q is reached
- U-LBG2 uses a fixed number of bits per feature, q

Feature Discretization: supervised methods

Some commonly used **supervised** FD methods:

- *information entropy minimization* (IEM) - uses the *minimum description length* (MDL) principle with a top-down approach
- *IEM variant* (IEMV) - uses the MDL principle to control the number of different values for a feature
- *class-attribute interdependence maximization* (CAIM) - maximizes the class-attribute interdependence
- the *class-attribute contingency coefficient* (CACC) - maximizes the class-attribute contingency coefficient

The first two methods are based on information theory whereas the other two are based on statistical measures

Our Proposals for FD

In this paper, we propose two FD methods:

1. a *static*, *global*, *top-down*, *incremental*, relevance-based method for unsupervised or supervised learning
   $\longrightarrow$ **Relevance-based LBG (R-LBG)**

2. a *static*, *global*, *top-down*, *incremental*, and *supervised* method based on the maximization of the *mutual information* (MI) between each feature and the class label
   $\longrightarrow$ **Mutual Information Discretization (MID)**

Proposal 1: R-LBG Algorithm (key ideas)

The main characteristics of the R-LBG algorithm are as follows:

- applies the (unsupervised) LBG algorithm, with an incremental number of bits per feature
- it uses a (supervised or unsupervised) relevance function, @*rel*, and a (nonnegative) stopping factor ϵ
- @*rel*, producing non-negative values, is applied after each discretization
- for each feature, $\widetilde{X}_i$, discretization is halted at b ($< q$) bits, whenever $@rel(\widetilde{X}_i^{(b)}) - @rel(\widetilde{X}_i^{(b-1)}) < \epsilon$
- setting $\epsilon = 0$, leads to the minimum number of bits that ensures maximum relevance
- it is a generalization of the U-LBG1 and U-LBG2 techniques

R-LBG Algorithm (relevance function)

Some choices for the unsupervised relevance function @*rel*:

- @*rel* = *MSE* between original and discrete features, we have the unsupervised U-LBG1/2 approaches
- the quotient between the variance of the discrete feature and the number of discretization intervals

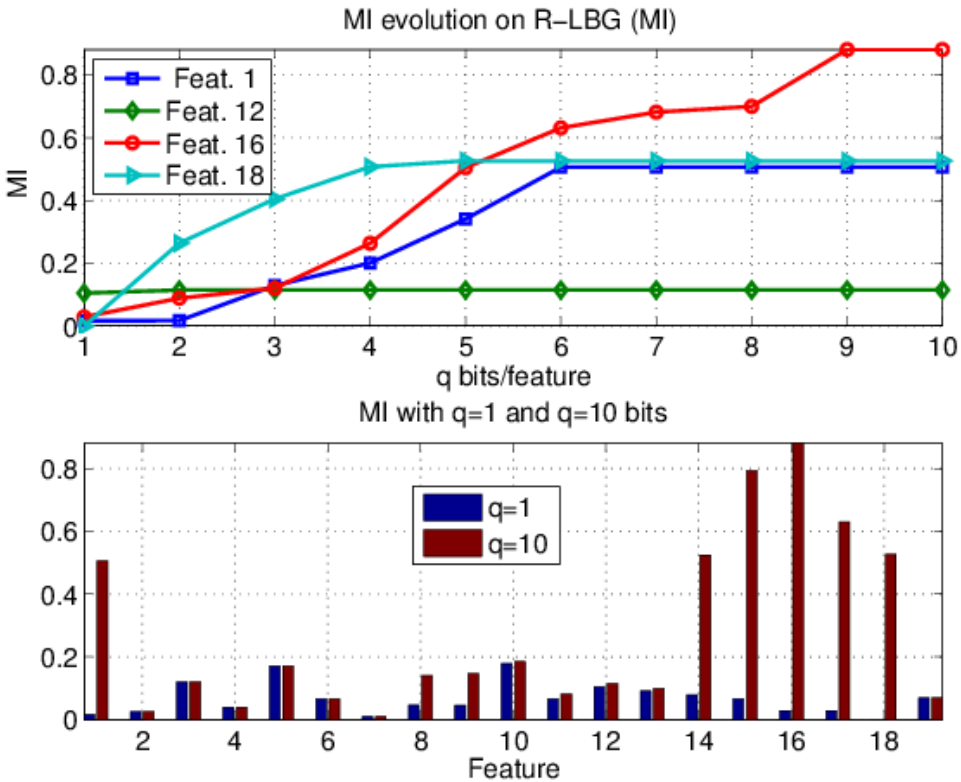
$$@rel(\widetilde{X}_i^{(b)}) = NVAR(\widetilde{X}_i) = \text{var}(\widetilde{X}_i) \ / \ 2^b$$

For the supervised case, @*rel* can be computed as:

- the MI between discretized features $\widetilde{X}_i$ and the class label **y**
- the well-known Fisher's ratio (using the same operands)
- many rank criteria used in feature selection methods

R-LBG: some insight on the relevance

R-LBG ($@rel = MI$) on the Hepatitis dataset ($d = 19$ features)



Top: MI as a function of the number of bits $q \in \{1, \ldots, 10\}$, for features 1, 12, 16, and 18. Bottom: MI with $q = 1$ and $q = 10$ bits.

$\longrightarrow$ **MI curves level-off at some point; categorical (feat. 12) and real features (feat. 1, 16, and 18) exhibit different behavior.**

Proposal 2: Mutual Information Discretization

The key motivations for the MID algorithm are as follows:

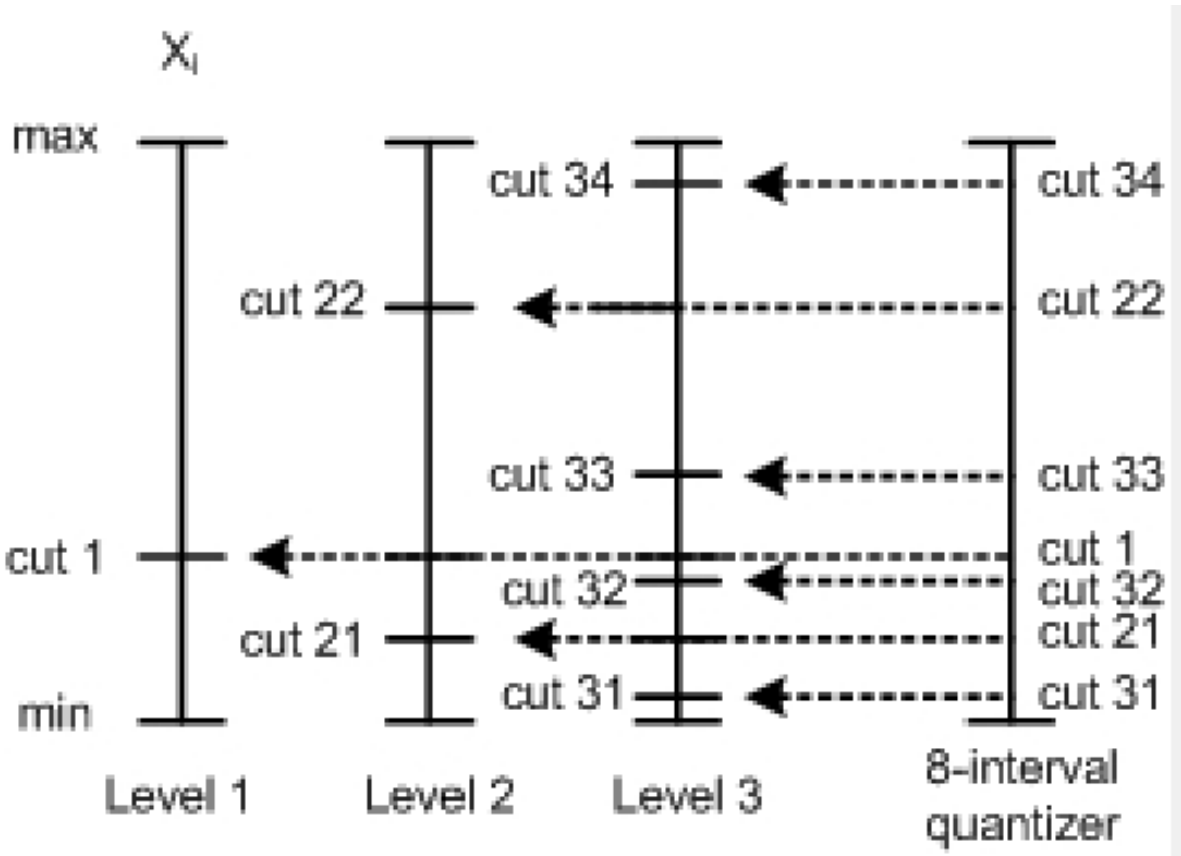
- good FS criteria will also be adequate for FD
- MI between features and class labels $\mathbf{y}$ is adequate for FS
- the Hellman-Raviv and Santhi-Vardi bounds relate the Bayes error with the MI

$$err_{Bayes}(\widetilde{X}_i) \leq \frac{1}{2} H(\widetilde{X}_i|\mathbf{y}) \qquad err_{Bayes}(\widetilde{X}_i) \leq 1 - 2^{-H(\widetilde{X}_i|\mathbf{y})}.$$

- since $MI(\widetilde{X}_i;\mathbf{y}) = H(\widetilde{X}_i) - H(\widetilde{X}_i|\mathbf{y})$, to maximize $MI(\widetilde{X}_i;\mathbf{y})$, will minimize the conditional entropies and the Bayes error!

MID: how does it work

- illustration for a given feature ($q = 3$ bit)
- it computes the cut-points that maximize the MI of the discrete feature with the class label



MID is an incremental and recursive partition algorithm

MID: fixed and variable versions

We have devised two versions of MID:

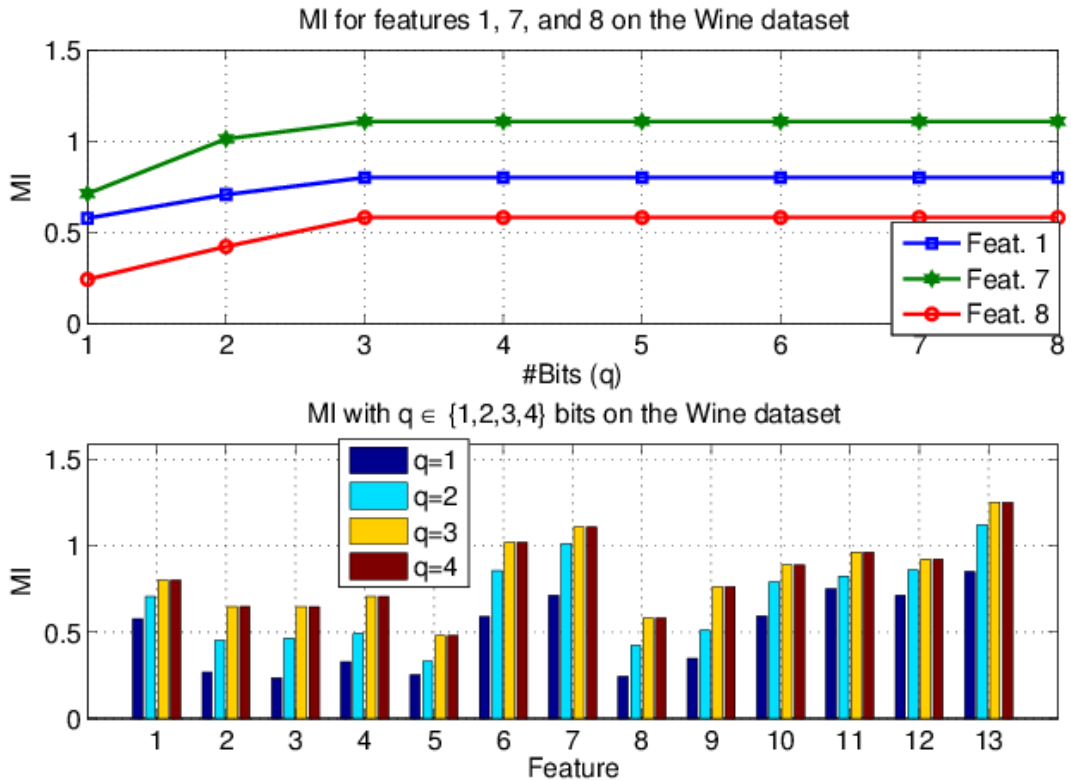
- MID-fixed, applies MID with q bits per feature
- MID-variable, allocates *up to* q bits per feature

MID-variable is controlled by a (nonnegative) stopping factor ϵ:

- $MI(\widetilde{X}_i^{(b)}; \mathbf{y})$ is computed
- for each feature, discretization is halted at b bits, whenever $MI(\widetilde{X}_i^{(b)}; \mathbf{y}) - MI(\widetilde{X}_i^{(b-1)}; \mathbf{y}) < \epsilon$
- setting $\epsilon = 0$ leads to the minimum number of bits that maximizes the MI
- for a given q, MID-variable will produce fewer discretization intervals than MID-fixed

MID: evolution of MI

MI evolution as a function of the number of bits on the Wine dataset ($d = 13$ features)



Top: MI for features 1, 7, and 8, with $q \in \{1, \ldots, 8\}$. Bottom: MI between discretized features and the class label, for $q \in \{1, 2, 3, 4\}$

Experimental Evaluation: Task and Datasets

- Supervised classification with linear *support vector machines* (SVM), naïve Bayes (NB), and k-nearest neighbors (KNN)
- 10-fold cross-validation strategy - learn a quantizer on training part and apply it to the test part
- UCI, microarray[$], and face image[#] datasets with d features, c classes, and n patterns (in some cases, $d \gg n$)

Dataset	d	c	n	**Dataset**	d	c	n
Wine	13	3	178	Leukemia1[$]	5327	3	72
Hepatitis	19	2	155	TOX-171[$]	5748	4	171
Ionosphere	34	2	351	Brain-Tumor1[$]	5920	5	90
Colon[$]	2000	2	62	ORL10P[#]	10304	10	100
SRBCT[$]	2309	4	83	Prostate-Tumor[$]	10509	2	102
AR10P[#]	2400	10	130	Leukemia2[$]	11225	3	72
PIE10P[#]	2420	10	210	GLI-85[$]	22283	2	85

Experimental Results: R-LBG and MID 1/3

R-LBG (@*rel* = *MI*) and *MID-variable* with $q = 4$ and linear SVM classifier
First row: complexity - total number of bits per instance
Second row: generalization error - test set error rate (%)

	R-LBG (MI)		**MID variable**	
Dataset / No FD	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0$	$\epsilon = 0.1$
Wine	52.0	30.6	38.3	**26.2**
3.9	2.8	**1.7**	3.4	2.8
Hepatitis	46.5	68.8	**28.5**	65.6
21.3	**15.5**	21.9	18.7	18.1
Ionosphere	129.0	102.4	**73.0**	85.0
12.8	14.0	12.5	9.4	**5.7**
Colon	7954.6	7564.0	**4682.0**	6151.9
17.7	19.4	**14.5**	19.4	**14.5**
SRBCT	9222.5	8827.7	**7144.2**	7180.3
0.0	**0.0**	**0.0**	**0.0**	**0.0**
AR10P	9599.8	9583.2	**8620.4**	8640.4
0.8	0.8	0.8	**0.0**	**0.0**
PIE10P	9679.9	9662.5	8550.7	**8543.4**
0.0	**0.0**	**0.0**	**0.0**	**0.0**

Experimental Results: R-LBG and MID 2/3

R-LBG (@$rel = MI$) and *MID-variable* with $q = 4$ and linear SVM classifier
First row: complexity - total number of bits per instance
Second row: generalization error - test set error rate (%)

	R-LBG (MI)		**MID variable**	
Dataset / No FD	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0$	$\epsilon = 0.1$
Leukemia1	21248.2	19818.7	**14636.9**	15555.9
8.3	**4.2**	5.6	8.3	6.9
TOX-171	22988.5	21439.2	**19012.4**	20070.8
14.6	**2.3**	2.9	4.1	4.1
Brain-Tumor1	23649.6	22174.4	17531.0	**17436.5**
11.1	**8.9**	10.0	10.0	10.0
ORL10P	41215.6	41195.1	37410.3	**37410.2**
1.0	**1.0**	**1.0**	2.0	2.0
Prostate-Tumor	41735.0	40431.3	**25493.1**	36598.8
10.8	**7.8**	**7.8**	**7.8**	**7.8**
Leukemia2	44300.1	40072.9	31124.0	**30255.4**
5.6	**1.4**	**1.4**	**1.4**	**1.4**
GLI-85	88561.7	84364.2	**54906.9**	72131.8
10.6	**8.2**	**8.2**	**8.2**	**8.2**

Experimental Results: R-LBG and MID 3/3

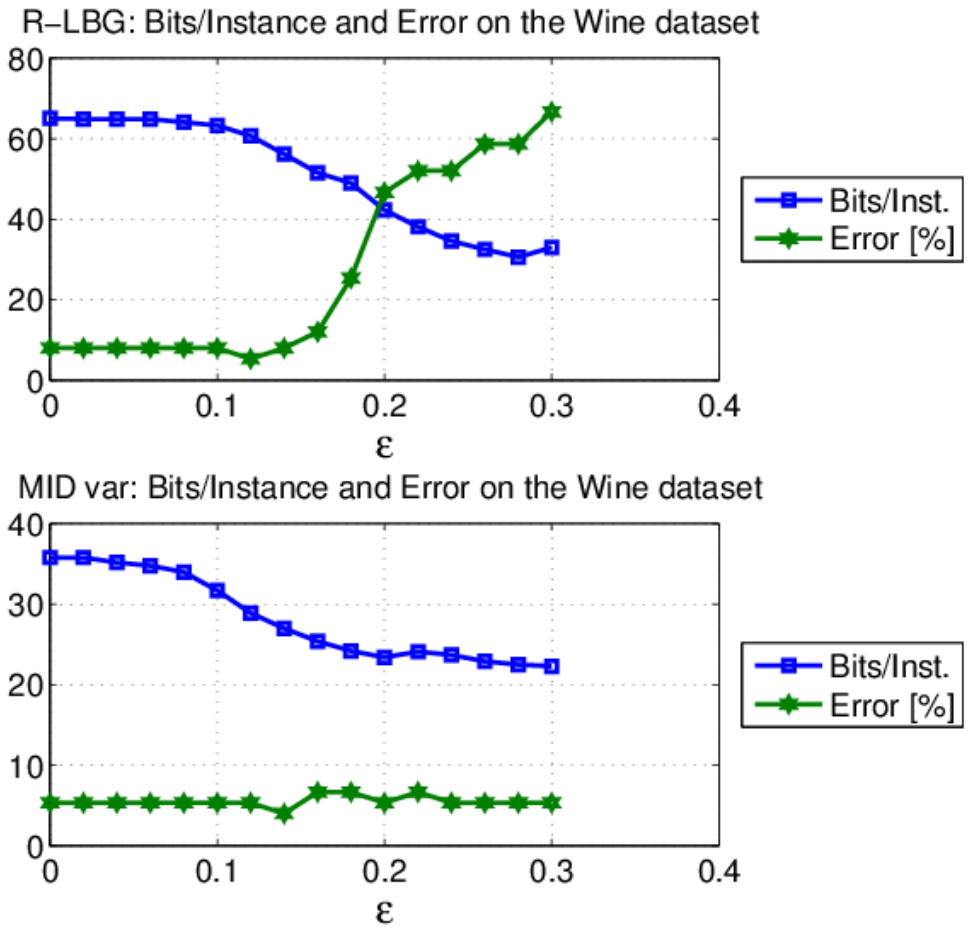
Some comments on these results:

- R-LBG with $\epsilon = 0$ usually leads to a larger number of bits per instance, as compared with $\epsilon = 0.1$
- R-LBG with $\epsilon = 0$ attains maximum relevance; with $\epsilon > 0$, the discretization process is halted earlier
- MID-variable with $\epsilon = 0$ yields the minimum bits that ensure the maximum MI
- MID-variable with $\epsilon = 0$ usually attains the best results with a few exceptions

The Friedman test reported a p-value of $0.04164 < 0.05$

R-LBG and MID: sensitivity on the ϵ parameter

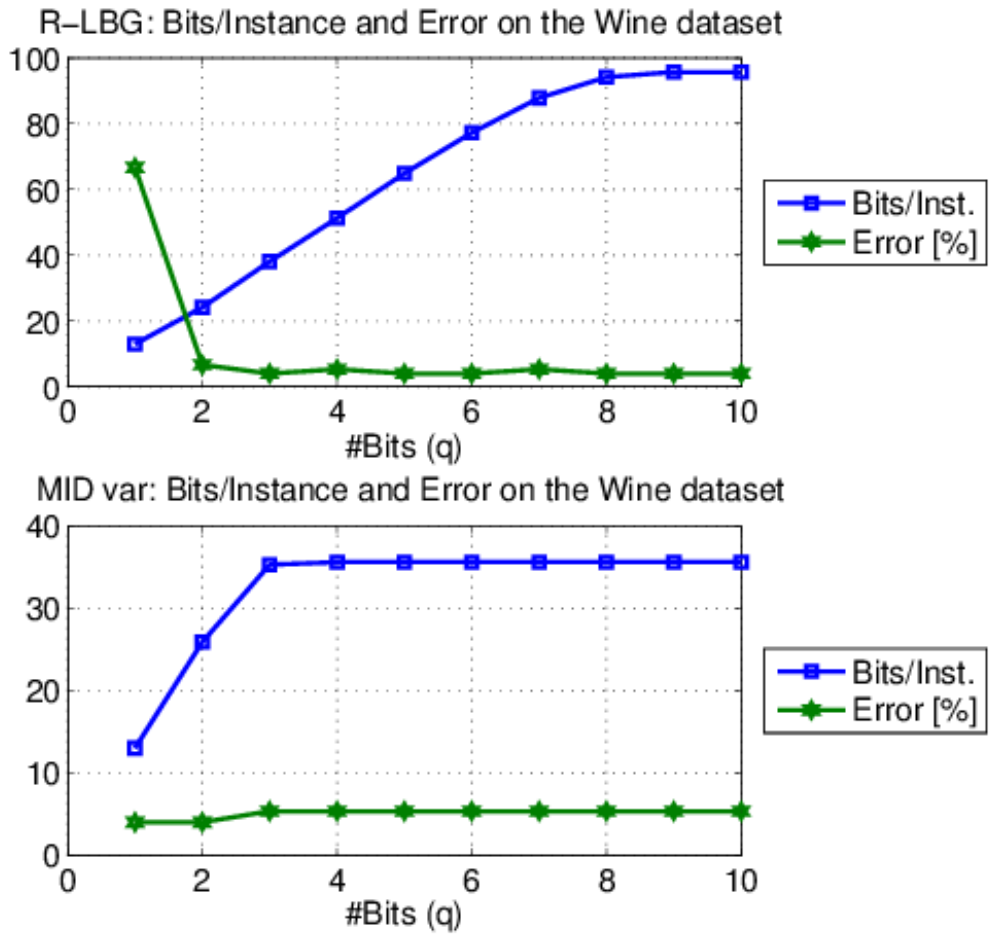
Average bits/instance and test set error rate (%, NB classifier) as function of the ϵ parameter



Wine dataset with $q = 5$ bits. Top: R-LBG (MI) Bottom: MID-variable

R-LBG and MID: sensitivity on the q parameter

Average bits/instance and test set error rate (%, NB classifier) as function of the q parameter



Wine dataset with $\epsilon = 0.05$. Top: R-LBG (MI) Bottom: MID-variable

Experimental Results: Unsupervised FD

Unsupervised FD with $q = 3$ bit/feature, @$rel = NVAR$, and $\epsilon = 0.25$
First row: complexity - total number of bits per instance
Second row: generalization error - test set error rate (%), linear SVM classifier

		Existing Methods					Proposed
Dataset	**No FD**	**EIB**	**EFB**	**PkID**	**U-LBG1**	**U-LBG2**	**R-LBG**
AR10P		7200.0	7200.0	9267.8	7200.0	7200.0	**6568.6**
	0.8	**0.0**	0.8	0.8	0.8	0.8	1.5
PIE10P		7260.0	7260.0	9680.0	7260.0	7260.0	**3774.4**
	0.0	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Leuk.1		15981.0	15981.0	15981.0	15981.0	15981.0	**5733.0**
	5.6	**2.8**	4.2	4.2	4.2	4.2	**2.8**
TOX-171		17244.0	17244.0	22992.0	17244.0	17244.0	**5847.6**
	9.9	**1.2**	1.8	**1.2**	1.8	1.8	8.2
B-Tumor1		17760.0	17760.0	23680.0	17760.0	17760.0	**6085.4**
	13.3	**8.9**	11.1	11.1	**8.9**	**8.9**	11.1
ORL10P		30912.0	30912.0	41216.0	30912.0	30912.0	**19385.4**
	1.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
P-Tumor		31527.0	31527.0	42035.4	31520.0	31527.0	**11394.0**
	10.8	**8.8**	**8.8**	**8.8**	**8.8**	**8.8**	**8.8**
Leuk.2		33675.0	33675.0	33675.0	33675.0	33675.0	**12431.4**
	4.2	**2.8**	**2.8**	**2.8**	**2.8**	**2.8**	4.2
GLI-85		66849.0	66849.0	66849.0	66849.0	66849.0	**25118.7**
	14.1	10.6	**8.2**	**8.2**	9.4	9.4	**8.2**

Experimental Results: Supervised FD

Supervised FD with @$rel = MI$ and $\epsilon = 0.1$
First row: complexity - total number of bits per instance
Second row: generalization error - test set error rate (%), linear SVM classifier

	Existing Methods				Proposed Methods		
Dataset	**IEM**	**IEMV**	**CAIM**	**CACC**	**R-LBG**	**MIDf**	**MIDv**
AR10P	12903.6	**7138.4**	7200.0	7200.0	7145.6	7200.0	7266.3
	2.3	20.0	0.8	**0.0**	**0.0**	**0.0**	**0.0**
PIE10P	9103.4	**5264.0**	7260.0	7260.0	7077.3	7260.0	7154.7
	0.0	1.9	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Leuk.1	28435.3	26034.7	*	*	14656.1	15981.0	**14278.0**
	40.3	56.9	*	*	4.2	**2.8**	**2.8**
TOX-171	36134.8	28253.7	*	*	**15725.6**	17244.0	15824.2
	5.8	**2.9**	*	*	**2.9**	3.5	4.7
B.-Tumor1	32808.3	27133.5	*	*	**15674.3**	17760.0	16343.6
	20.0	35.6	*	*	11.1	10.0	**8.9**
ORL10P	26475.7	**24176.8**	*	*	30863.0	30912.0	30786.9
	9.0	**1.0**	*	*	2.0	2.0	2.0
P.-Tumor	54395.6	51964.7	*	*	30695.0	31527.0	**28506.3**
	12.7	11.8	*	*	**5.9**	6.9	7.8
Leuk.2	48380.1	40447.3	*	*	28857.3	33675.0	**28670.8**
	8.3	6.9	*	*	**2.8**	**2.8**	**2.8**
GLI-85	135866.9	130689.1	*	*	64065.0	66849.0	**58633.4**
	11.8	12.9	*	*	**9.4**	**9.4**	10.6

Conclusions

Our FD methods (Relevance-LBG and MID):

- for both unsupervised and supervised FD, attain equal or better results than state-of-the-art techniques
- scale well for high-dimensional data, contrary to other approaches
- have shown the adequacy of MI between features and class labels for discretization
- attain adequate discretizations with a variable number of bits per feature

As future work, we will explore the embedding of feature selection in the discretization process