Exploiting the Bin-Class Histograms for Feature Selection on Discrete Data

Artur J. Ferreira[1,3] Mário A. T. Figueiredo[2,3]

[1] Instituto Superior de Engenharia de Lisboa
[2] Instituto Superior Técnico, Universidade de Lisboa
[3] Instituto de Telecomunicações, Lisboa,
PORTUGAL
arturj@isel.pt mario.figueiredo@lx.it.pt

Abstract. In machine learning and pattern recognition tasks, the use of feature discretization techniques may have several advantages. The discretized features may hold enough information for the learning task at hand, while ignoring minor fluctuations that are irrelevant or harmful for that task. The discretized features have more compact representations that may yield both better accuracy and lower training time, as compared to the use of the original features. However, in many cases, mainly with medium and high-dimensional data, the large number of features usually implies that there is some redundancy among them. Thus, we may further apply feature selection techniques on the discrete data, keeping the most relevant features, while discarding the irrelevant and redundant ones. In this paper, we propose relevance and redundancy criteria for supervised *feature selection* (FS) techniques on discrete data. These criteria are applied to the bin-class histograms of the discrete features. The experimental results, on public benchmark data, show that the proposed criteria can achieve better accuracy than widely used relevance and redundancy criteria, such as mutual information and the Fisher ratio.

Keywords: feature selection, feature discretization, discrete features, bin-class histogram, matrix norm, supervised learning, classification.

1 Introduction

High-dimensional (HD) datasets (*i.e.*, with a large number of features) are becoming increasingly common in many different application domains of machine learning and pattern recognition. For instance, we can find them in different areas, such as genomics, bioinformatics, computer vision, satellite image analysis, and multimodal audio-visual processing. When dealing with HD data, one often resorts to *feature discretization* (FD) [1] and *feature selection* (FS) [2] procedures. FS methods aim at finding an adequate subset of the original features, whereas FD looks for compact data representations, desirably ignoring irrelevant fluctuations on the data for the task at hand, and leading to more robust classifiers, and lower training time.

The literature on FD and FS is vast, with many unsupervised and supervised techniques. A comprehensive list of FS techniques can be found in [2]. Regarding FD, there are several comprehensive reviews, such as the recent survey in [3].

FS methods can be grouped into four classes [2]: wrappers, embedded methods, filters, and hybrid methods. A filter retains some of the features and discards others, based on a criterion that is independent of any subsequent learning algorithm. Although filters are the simplest and fastest approaches, thus expected to perform worse than the other types of methods, it is often the case that they are the only applicable option on HD datasets, where the other approaches can be computationally too expensive.

Regarding FD methods, the dynamic techniques that take into account feature interdependencies are usually preferable to their static counterparts, which discretize each feature individually. However, when dealing with HD data, dynamic FD methods have a prohibitive computational cost, and one has to resort to suboptimal static methods. Thus, when learning from HD data, it is useful to apply some FS filter after data discretization, in order to remove the remaining feature interdependencies. As a consequence, we can combine FD and FS techniques, yielding joint discretization and selection.

1.1 Our Contribution

In this paper, we propose four criteria for relevance and redundancy assessment for FS purposes, on discrete features. After running some FD technique, we apply one of our criteria in order to select and keep an adequate subset of features. After the FD process is carried out, the *bin-class histogram* (BCH) of each feature is computed. In a nutshell, the BCH for one feature holds the number of times that each discretization bin occurs among each class, considering all the available data patterns (see Section 3 for details).

The remainder of the paper is organized as follows. Section 2 briefly reviews some existing supervised FD and FS techniques. Section 3 details the proposed criteria for relevance and redundancy assessment. The experimental evaluation of our methods, compared against standard methods on public benchmark datasets is reported in Section 4. Finally, Section 5 ends the paper with some concluding remarks and directions for future work.

2 Short Review of Feature Discretization and Selection

2.1 Feature Discretization

Many datasets have continuous features (formally, real-valued, but in practice stored with a floating point representation). Some classification algorithms can only deal with discrete features; in this case, a discretization procedure is needed as a pre-processing stage. Regardless of the type of classifier used, discretized features may lead to better results, as compared to their original version [1].

The *information entropy maximization* (IEM) method [4] is a well-known supervised FD technique. It relies on the principle that the most informative

features to discretize are the most compressible ones. IEM adopts an entropy minimization heuristic for discretization into multiple intervals; it operates in a recursive, incremental, top-down fashion, computing the discretization bins that minimize the number of bits per feature.

The *class-attribute interdependence maximization* (CAIM) [5] algorithm aims at maximizing the class-attribute interdependence and to generate a (possibly) minimal number of discrete intervals. Similarly to IEM, CAIM does not require a predefined number of bins, being an incremental top-down approach. The experimental reported in [5], comparing CAIM with six other FD algorithms, show that the discrete features generated by CAIM almost always have the lowest number of bins and yield the highest classification accuracy.

2.2 Feature Selection

This section briefly reviews two well-known relevance criteria widely used as supervised FS filters (and in hybrid methods). Consider a supervised dataset, with d features and n instances (each known to belong to one of C classes, $\{1, 2, ..., C\}$), stored in $d \times n$ matrix X (*i.e.*, X_{ij} is the i-th feature of the j-th instance). The class labels are given in a $C \times n$ binary matrix Y, where $Y_{cj} = 1$, if and only if the j-th instance belongs to class c.

In the multi-class case ($C > 2$, assuming class labels in $\{1, 2, \ldots, C\}$), the Fisher ratio (FiR) for the i-th feature (see, *e.g.*, [6]) is given by

$$\mathrm{FiR}_i = \sum_{c=1}^{C} n_c \big(\mu_{ci} - \eta_i\big)^2 \Big(\sum_{j=1}^{c} n_c \sigma_{ci}^2\Big)^{-1}, \tag{1}$$

where $n_c = \sum_{j=1}^{n} Y_{cj}$ is the number of instances in class c,

- $\mu_{ci} = \frac{1}{n_c} \sum_{j=1}^{n} X_{ij} Y_{cj}$ is the sample mean of feature i in class c;
- $\eta_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$ is the global sample mean of the $i-$th feature;
- $\sigma_{ci}^2 = \frac{1}{n_c} \sum_{j=1}^{n} Y_{cj} (X_{ij} - \mu_{ci})^2$ is the sample variance of feature i in class c;

Another widely used measure of feature relevance is (the sample-based estimate of) the *mutual information* (MI) [7] between each feature and the class label. The MI is non-negative, being zero if and only if the two involved variables are statistically independent [7].

When using either the FiR or the MI for relevance-based FS on a d-dimensional dataset, we simply keep the $m \leq d$ top-ranked features, according to the adopted relevance measure. Both the FiR and the MI are *global* ranking measures, in the sense that they assign a number (a relevance value) to each feature. However, when dealing with discrete data, one can further analyze the distribution of the

discretization bins among patterns of all classes, thus having some *local* insight on the discriminative power of each feature. Despite its popularity [8], when dealing with HD data, MI may not be the best relevance criteria [9].

With HD data, it is often the case that we have redundant features, which convey the same information. Keeping all these features may have harmful consequences for the learning task, thus they should be discarded. In this context, some FS filters follow the *relevance-redundancy* (RR) approach, such as the *relevance-redundancy feature selection* (RRFS) method [10] , which finds the most relevant subset of features, and then efficiently searches for redundancy in this subset, selecting only highly relevant features with low redundancy.

3 Proposed Relevance and Redundancy Criteria

In this Section, we describe the proposed relevance and redundancy criteria for discrete data and its usage for FS filters. The relevance of the discrete feature is computed by checking the histogram of each discretization interval, across the different classes. The proposed method works as follows:

1. (independently) discretize all the d features in the dataset, with some FD method (e.g. one of the techniques mentioned in Section 2.1);
2. obtain the $b_i \times C$ *bin-class histogram* (BCH) matrix $\mathbf{B}^{(i)}$ for each feature i, where b_i is the number of discretization bins of the i-th feature (specifically, $B^{(i)}_{ac}$ is the number of times that feature i takes values in the a-th bin and the class label is c; Fig. 1 illustrates the BCH matrices for two discrete features with four bins ($b_1 = b_2 = 4$) in a three-class problem ($C = 3$), for a dataset with $n = 75$ patterns, 25 per class);
3. apply one of the criteria r_1, r_2, r_3, or r_4 (see below), to assess the relevance (and/or the redundancy) of each feature and keep the most relevant ones.

The key idea of this proposal is to use the local information provided by the BCH, in order to identify the most discriminative features. The rationale is that this local information may be more meaningful for this task, as compared to global indicators such as the FiR and the MI.

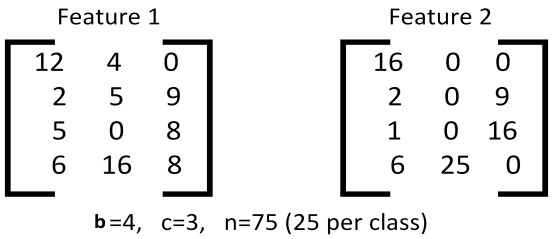

Fig. 1. The *bin-class histogram* (BCH) matrix for two discrete features with four bins in a three-class problem, with 25 instances per class.

The relevance of a discrete feature is proportional to the non-uniformity of its histogram across the discretization bins and classes. For instance, if a given row (discretization bin) of the BCH matrix has an (almost) uniform distribution, this shows that that discretization level dos not contribute to distinguishing among the classes. A special and interesting case is the occurrence of zeros in this matrix, which implies that a given discretization level never occured in the patterns of a given class.

We propose four criteria to assess the relevance of feature i based on its BCH matrix. The first two aim at assessing the non-uniformity of the BCH matrix:

- $r_1^{(i)}$: the number of zero entries in matrix $\mathbf{B}^{(i)}$;
- $r_2^{(i)}$: the sum of the absolute differences between all pairs of columns of $\mathbf{B}^{(i)}$:

$$r_2^{(i)} = \sum_{k=1}^{C-1} \sum_{m=k+1}^{C} \left\| \mathbf{B}_k^{(i)} - \mathbf{B}_m^{(i)} \right\|_1, \tag{2}$$

  where $\mathbf{B}_k^{(i)}$ and $\mathbf{B}_m^{(i)}$ denote the k-th and the m-th columns of $\mathbf{B}^{(i)}$.

The other two criteria are based on *matrix norms* and *matrix similarity* measures[11, 12]. The key idea is that an ideal BCH matrix (after normalizing each column to sum to one) is a "rectangular identity matrix" (*i.e.*, an identity with possibly several additional rows of zeros). In detail,

- $r_3^{(i)} = \text{trace}\big(\bar{\mathbf{B}}^{(i)}(\bar{\mathbf{B}}^{(i)})^T\big)$, where $\bar{\mathbf{B}}^{(i)}$ is the the normalized version of $\mathbf{B}^{(i)}$, such that its columns sum up to one. The maximum value of the trace of $\bar{\mathbf{B}}^{(i)}(\bar{\mathbf{B}}^{(i)})^T$ is C, which is achieved by the ideal matrix (as explained above).
- $r_4^{(i)} = \text{trace}\Big(\sqrt{\big(\bar{\mathbf{B}}^{(i)}\big)^T \bar{\mathbf{B}}^{(i)}}\Big)$, called the trace (or nuclear) norm, which is also equal to the sum of the singular values of $\bar{\mathbf{B}}^{(i)}$. The relevance of a feature is proportional to this value.

These criteria can be applied to both binary or multi-class problems. For the example in Fig. 1, these relevance values are: i) for feature 1, r_1=2, r_2=74, r_3=1.14, and r_4=1.67; ii) for feature 2, r_1=5, r_2=132, r_3=2.01, and r_4=2.39. Thus, feature 2 will be considered more relevant than feature 1, which is in accordance with the above considerations about the $\mathbf{B}$ matrix.

4 Experimental Evaluation

We report an experimental evaluation, carried out on public domain standard benchmark datasets, from the UCI [13] and the *gene selection model selector* (GEMS)[4] repositories. We perform a supervised classification task with the linear *support vector machines* (SVM) classifier from Weka[5], with its default parameters. The classification accuracy is assessed using 10-fold *cross validation*

[4] www.gems-system.org
[5] www.cs.waikato.ac.nz/ml/weka

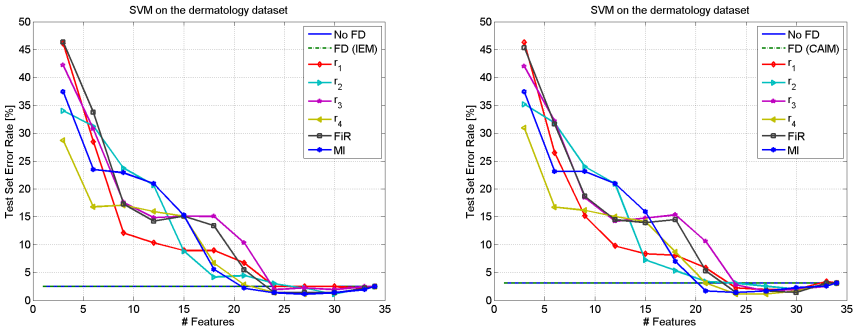

Fig. 2. SVM test set error rate (%), 10-fold CV, for the Dermatology dataset, as functions of the number of features for different FS relevance-only criteria: IEM discretization (left); CAIM discretization (right).

(CV). For each CV fold, the IEM and CAIM FD methods (see Section 2.1) are applied to the training partition to learn a quantizer, which is then applied to the test partition. We perform FS with relevance-only FS filters and relevance-redundancy filters, comparing our criteria against the FiR and the MI.

Fig. 2 shows the test set error rate as functions of the number of features for the relevance-based FS filters with our relevance measures: r_1, r_2, r_3, r_4, the FiR and the MI, after IEM and CAIM discretization. We use the Dermatology dataset, a skin disease diagnosis problem, with $d = 34$ features, $C = 6$ classes, and $n = 358$ instances. The horizontal lines refer to the baseline error without FS; one of these lines corresponds to the absence of FD whereas the other corresponds to FD by the corresponding discretization method (IEM or CAIM). The error rates reported by our four relevance criteria are competitive with those attained by FiR and MI. Regarding the impact of the FD method on the final results, we find no appreciable differences between these methods.

Fig. 3 shows the results of the test set error rate as functions of the number of features for the relevance-based FS filters, on the Hepatitis and Dexter datasets with IEM discretization. On the Hepatitis dataset, r_3 and r_4 achieve the best results, with error rates below the baseline values. On the Dexter dataset (with sparse data), r_2 is clearly more adequate than the other three criteria.

Table 1 reports the test set error rate attained by the same methods of Fig. 2, using IEM discretization, for several datasets, with quite different types of data. We rank the features according to the relevance criteria and we keep the $m < d$ top-ranked features. These results show that the proposed relevance measures attain results similar or better to those of FiR and MI, in different problems. Altough none of the relevance criteria outperforms all the others, we can observe that despite their simplicity r_1 and r_2 achieve good results; in fact, r_2 is the best for the sparse data of the Dexter dataset. The relevance criteria r_3 and r_4 also achieve results which are usually equal or better than those of FiR and MI.

We now assess the results of the filter RRFS method (see Section 2.2), using the same relevance criteria as in the previous experiments, for the same datasets

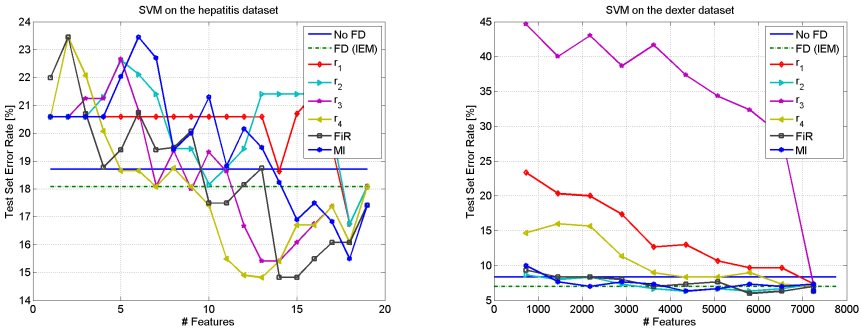

Fig. 3. SVM test set error rate (%), 10-fold CV, for the Hepatitis (left) and Dexter (right) datasets, as functions of the number of features, for different FS relevance-only criteria with IEM discretization.

Table 1. Test set error rate (%), for the SVM classifier, 10-fold CV, for datasets with c classes, n instances, with dimensionality d. We perform FS to select subsets with $m < d$ features. The best results (lower error) are in bold face.

			Relevance-Only Feature Selection					
Dataset $(d; c; n)$	m	***FD (IEM)***	r_1	r_2	r_3	r_4	***FiR***	***MI***
Wine (13;3;178)	8	2.25	**1.11**	2.22	3.48	2.22	2.75	2.22
Hepatitis (19;2;155)	12	18.08	20.58	19.45	16.66	**14.90**	18.15	20.15
Ionosphere (33;2;331)	30	11.42	11.71	**10.86**	12.29	12.00	12.26	11.97
Dermatology (34;6;358)	24	2.52	2.27	3.34	1.98	**1.11**	1.96	1.98
Sonar (60;2;208)	24	21.19	36.09	33.66	21.13	**20.18**	21.63	21.68
M-Libras (90;15;360)	63	**23.89**	26.11	27.44	28.00	29.44	29.67	30.11
Colon (2000;2;62)	800	18.81	19.05	19.05	**15.48**	17.14	17.14	17.38
Example1 (9947;2;50)	4370	**2.78**	3.11	**2.78**	3.11	**2.78**	**2.78**	**2.78**
Prost.-Tumor (10510;2;102)	2100	8.82	6.82	6.82	8.82	8.82	8.82	**5.73**
Leukemia1(5327;3;72)	2128	4.29	**2.86**	**2.86**	4.29	4.29	4.29	**2.86**
ORL10P (10304;10;100)	3090	2.0	**0.0**	**0.0**	3.0	1.0	1.0	**0.0**
Brain-Tumor2 (10367;4;90)	4144	20.00	24.00	24.00	**18.00**	**18.00**	**18.00**	24.00
Dexter (20000;2;2600)	2936	7.33	17.33	**6.67**	40.67	9.67	8.33	8.00

as in Table 1. The results reported in Table 2 suggest that the proposed relevance criteria are also useful for relevance-redundancy FS filters. In many datasets, the proposed criteria yield the lowest test set error rate.

5 Conclusions

We have proposed new criteria for supervised selection of discrete features, based on their bin-class histograms. Our experiments suggest that all the proposed criteria are useful for both relevance-only and relevance-redundancy FS filters. The classifiers learned on the features selected by our methods usually attain equal or better accuracy than those learned on the original discretized or non-discretized features. The proposed criteria attain equal or better results than two widely used FS criteria, on different types of data. We conclude that the proposed criteria deserve further study.

Table 2. Average number of features (m_*) and test set error rate (%), for the SVM classifier, 10-fold CV (the same datasets as in Table 1). The RRFS method uses the M_S values reported for each dataset. The best results (lower error) are in bold face.

	RRFS - Relevance-Redundancy Feature Selection					
Dataset; M_S	$m_1; r_1$	$m_2; r_2$	$m_3; r_3$	$m_4; r_4$	m_f***; FiR***	m_m***; MI***
Wine; 0.95	7.9; **2.2**	7.2; **2.2**	8.9; 3.9	8.9; 3.9	8.6; 3.3	9.6; 2.7
Hepatitis; 0.95	13.7; 21.2	15.3; 22.6	16.2; 23.1	16.2; 23.1	16.7; **17.9**	16.2; 20.5
Ionosphere; 0.95	32.0; 12.8	32.0; 12.8	32.0; **10.8**	32.0; **10.8**	32.0; 11.0	32.0; 11.9
Dermatology; 0.95	32.9; 2.2	33.0; **1.6**	33.0; 2.2	33.0; 2.2	32.9; 2.2	32.8; **1.6**
Sonar; 0.95	56.6; 22.1	56.3; 20.7	56.8; 20.2	56.8; 20.2	57.4; **19.7**	55.2; 21.2
M-Libras; 0.95	44.1; 30.2	52.6; **26.6**	62.1; **26.6**	62.1; **26.6**	33.3; 42.7	37.2; 35.0
Colon; 0.6	73.1; **14.7**	73.3; **14.7**	48.9; 17.8	48.9; 17.8	73.9; 22.8	66.3; 27.3
Example1; 0.6	4392.0; **3.5**	4387.8; 3.6	4371.6; 3.7	4371.6; 3.7	4257.9; 3.7	4402.1; 4.2
Prost.-Tumor; 0.6	6950.1; 8.6	6959.6; 8.6	6933.5; **7.6**	6933.5; **7.6**	7053.7; 9.6	7160.0; **7.6**
Leukemia1; 0.6	2740.4; 4.3	2774.3; **2.9**	2685.7; **2.9**	2685.7; **2.9**	2848.8; **2.9**	2774.9; **2.9**
ORL10P; 0.9	2924.4; 2.0	2596.0; **0.0**	2561.2; **0.0**	2561.2; **0.0**	3416.3; 1.0	3436.8; 2.0
Brain-Tumor2; 0.6	4767.7; **20.0**	4773.0; **20.0**	4527.6; 22.0	4527.6; 22.0	4316.4; 28.0	4585.5; 24.0
Dexter; 0.6	7287.0; 7.7	7288.4; 7.7	7285.5; 7.7	7285.5; 7.7	7197.5; 8.0	7289.9; **7.3**

References

1. I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, Morgan Kauffmann, 2005.
2. I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (Editors). *Feature Extraction, Foundations and Applications*. Springer, 2006.
3. S. Garcia, J. Luengo, J. Saez, V. Lopez, and F. Herrera. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013.
4. U. Fayyad and K. Irani. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Int. Joint Conf. on Art. Intell. (IJCAI)*, pages 1022–1027, 1993.
5. L. Kurgan and K. Cios. CAIM discretization algorithm. *IEEE Trans. on Know. and Data Engineering*, 16(2):145–153, Feb. 2004.
6. R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, 2nd edition, 2001.
7. T. Cover and J. Thomas. *Elements of Information Theory*. J. Wiley & Sons, 1991.
8. G. Brown, A. Pocock, M. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66, 2012.
9. B. Franay, G. Doquire, and M. Verleysen. Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. *Neurocomputing*, 112(0):64 – 78, 2013.
10. A. Ferreira and M. Figueiredo. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, 33(13):1794 – 1804, 2012.
11. N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *18th Annual Conference on Learning Theory*, pages 545–560. Springer-Verlag, 2005.
12. G. Strang and K. Borre. *Linear algebra, geodesy, and GPS*. Wellesley-Cambridge Press, 1997.
13. A. Frank and A. Asuncion. UCI machine learning repository, available at http://archive.ics.uci.edu/ml, 2010.