# A Novel Information Theory Method for Filter Feature Selection

Boyan Bonev, Francisco Escolano, and Miguel Angel Cazorla

Department of Computer Science and Artificial Intelligence,
Alicante University, P.O.B. 99, E-03080 Alicante, Spain
{boyan, sco, miguel}@dccia.ua.es

**Abstract.** In this paper, we propose a novel filter for feature selection. Such filter relies on the estimation of the mutual information between features and classes. We bypass the estimation of the probability density function with the aid of the entropic-graphs approximation of Rényi entropy, and the subsequent approximation of the Shannon one. The complexity of such bypassing process does not depend on the number of dimensions but on the number of patterns/samples, and thus the curse of dimensionality is circumvented. We show that it is then possible to outperform a greedy algorithm based on the maximal relevance and minimal redundancy criterion. We successfully test our method both in the contexts of image classification and microarray data classification.

## 1 Introduction

Dimensionality reduction of the raw input variable space is a fundamental step in most pattern recognition tasks. Focusing on the most relevant information in a potentially overwhelming amount of data is useful for a better understanding of the data, for example in genomics [2][21][22]. A properly selected features set significantly improves classification performance. Thus, the removal of the noisy, irrelevant and redundant features is a challenging task.

There are two major approaches to dimensionality reduction: Feature Selection and Feature Transform. Whilst Feature Selection reduces the feature set by discarding the features which are not useful for some purpose (generally for classification), Feature Transform methods (also called feature extraction) build a new feature space from the original variables.

The literature differentiates among three kinds of Feature Selection: *Filter* methods [4], *Wrapper* methods [5], and *On-line* [6]. Filter Feature Selection does not take into account the properties of the classifier (it relies on statistical tests to the variables), while Wrapper Feature Selection tests different feature sets by building the classifier. Finally, On-line Feature Selection incrementally adds new features during the selection process.

Feature Selection is a combinatorial computational complexity problem. Algorithms must be oriented to find suboptimal solutions in a feasible number of iterations. Nevertheless, when there are thousands of features, Wrapper approaches become unfeasible. Among the Filter approaches, a fast way to evaluate

individual features is given by their relevance to the classification, by maximizing the mutual information between each variable and the classification output. As Guyon and Elisseeff state in [4], this is usually suboptimal for building a predictor, particularly if the variables are redundant. Conversely, a subset of useful variables may exclude many redundant, but relevant, variables. To overcome this limitation Peng et al.[7] minimize redundancy among the selected features set. Still a problem remains in the fact that these criteria are based on individual features, and this is due to the fact that estimating mutual information (and entropy) in a continuous multi-dimensional feature space is a hard task.

In this work, we overcome the latter problem by using Entropic Spanning Graphs to estimate Mutual Information [15]. The method's complexity does not depend on the number of dimensions, but on the number of samples. It allows us to estimate mutual information and thus maximize dependency between combinations of thousands of features and the class labels. We compare classification results to another Filter Feature Selection approach and perform an experiment on gene patterns with thousands of features.

This paper is structured as follows. In Section 2, the estimation of Mutual Information (Subsection 2.1) and Entropy (Subsection 2.2) are detailed. Then in Section 3, Feature Selection criteria and algorithms are explained and complexity is discussed. Finally experimental results are presentend in Section 4, and conclusions and future work are stated in Section 5.

## 2 Estimation of Mutual Information and Entropy

### 2.1 Mutual Information Estimation

Mutual Information (MI) is used in Filter Feature Selection as a measure of the dependency between a set of features $S$ and the classification prototypes $C$. MI can be calculated in different ways. In [1], Neemuchwala et al. study the use of entropic graph for MI estimation. In our approach we calculate MI based on entropy estimation:

$$I(S;C) = \sum_{s\in S}\sum_{c\in C} p(s,c) log \frac{p(s,c)}{p(s)p(c)} \tag{1}$$

$$= H(S) - H(S|C) \tag{2}$$

$$= H(S) + H(C) - H(S,C) \tag{3}$$

Using the Eq. 2 the conditional entropy $H(S|C)$ has to be calculated. To do this, $\sum(X|C = c)p(C = c)$ entropies have to be calculated, and this is feasible insofar C is discrete (C consists of the class labelling). On the other hand, using Eq. 3 implies estimating the joint entropy. In our experiments we used Eq. 2 because it is faster, due to the complexity of the entropy estimator, which depends on the number of samples as we will see in the following subsection.

### 2.2 Entropy Estimation

Entropy is a basic concept in information theory [8]. For a discrete variable $Y$ with $y_1, ..., y_N$ (the set of values), we have:

$$\begin{aligned} H(Y) &= -E_y[\log(p(Y))] \\ &= -\sum_{i=1}^{N} p(Y = y_i) \log p(Y = y_i). \end{aligned} \tag{4}$$

The estimation of the Shannon entropy of a probability density given a set of samples has been studied widely in the past [9][10][11][12][13][14]. Most current nonparametric entropy and divergence estimation techniques are based on estimation of the density function followed by the substitution of these estimates into the expression for entropy. This method has been widely applied to estimation of the Shannon entropy and it is called "plug-in" estimation [9]. Other methods of Shannon entropy estimation include sample spacing estimators, restricted to $d = 1$, and estimates based on nearest neighbor distances.

In [15] an alternative method for entropy and divergence estimation based on using entropic spanning graphs is presented. It is considered as a "non plug-in" method, because the entropy is directly estimated from a set of samples of the pdf, by-passing the non-parametric density estimation.

Among the "plug-in" methods, a widely used one is the Parzen's Window. Each method has its own advantages and drawbacks: On the one hand in the Parzen's Windows approach problems arise due to the infinite dimension of the spaces in which the unconstrained densities lie. Specifically: density estimator performance is poor without stringent smoothness conditions; no unbiased density estimators generally exist; density estimators have high variance and are sensitive to outliers; the high dimensional integration required to evaluate the entropy might be difficult. In contrast, the entropic graphs method does not estimate Shannon entropy directly and a new technique to obtain it must be developed. The main advantage of this approach is the possibility to work in a very high-dimensional space, in contrast to Parzen's Windows, the complexity of which is quadratic with respect to the number of dimensions.

Entropic Spanning Graphs obtained from data to estimate Renyi's $\alpha$-entropy [15] belong to the "non plug-in" methods of entropy estimation. Renyi's $\alpha$-entropy of a probability density function $f$ is defined as:

$$H_\alpha(p) = \frac{1}{1-\alpha} \ln \int_z p^\alpha(z) dz \tag{5}$$

for $\alpha \in (0, 1)$. The $\alpha$ entropy converges to the Shannon entropy $-\int p(z) \ln p(z) dz$ as $\alpha \to 1$, so it is possible to obtain the second one from the first one.

A graph $G$ consists of a set of vertices $X_n = \{x_1, ..., x_n\}$, with $x_n \in R^d$ and edges $\{e\}$ that connect vertices in graph: $e_{ij} = (x_i, x_j)$. If we denote by $M(X_n)$ the possible sets of edges in the class of acyclic graphs spanning $X_n$ (spanning trees), the total edge length functional of the Euclidean power weighted Minimal

Spanning Tree is:

$$L_\gamma^{MST}(X_n) = \min_{M(X_n)} \sum_{e \in M(X_n)} \mid e \mid^\gamma \tag{6}$$

with $\gamma \in (0, d)$ y $\mid e \mid$ the euclidian distance between graph vertices.

The MST has been used as a way to test for randomness of a set of points. In [16] it was showed that in d-dimensional feature space, with $d \geq 2$:

$$H_\alpha(X_n) = \frac{d}{\gamma}\left[\ln \frac{L_\gamma(X_n)}{n^\alpha} - \ln \beta_{L_\gamma,d}\right] \tag{7}$$

is an asymptotically unbiased, and almost surely consistent, estimator of the $\alpha$-entropy of $p$ where $\alpha = (d-\gamma)/d$ and $\beta_{L_\gamma,d}$ is a constant bias correction depending on the graph minimization criterion, but independent of $p$. Closed form expressions are not available for $\beta_{L_\gamma,d}$; only known approximations and bounds: (i) Monte Carlo simulation of uniform random samples on unit cube $[0,1]^d$; (ii) Large d approximation: $(\gamma/2)\ln(d/(2\pi e))$ in [17].

We can estimate $H_\alpha(p)$ for different values of $\alpha = (d-\gamma)/d$ by changing the edge weight exponent $\gamma$. As $\gamma$ modifies the edge weights monotonically, the graph is the same for different values of $\gamma$, and only the total length in expression 7 needs to be recomputed.

Entropic spanning graphs are suitable for estimating $\alpha$-entropy with $\alpha \in [0,1[$, so Shannon entropy can not be directly estimated with this method. In [18] relations between Shannon entropy and Rényi entropies of integer order are discussed. For any discrete probability n-points distribution for which the Rényi entropies of order two and three are known he provides a lower and an upper bound for the Shannon entropy. In [19] Mokkadem constructed a nonparametric estimate of the Shannon entropy from a convergent sequence of $\alpha$-entropy estimates.

The value of $H_\alpha$ for $\alpha = 1$ is approximated by means of a continuous function that captures the tendency of $H_\alpha$ in the environment of 1. Such a function is a monotonous decreasing one, and by means of a dichotomic search we find the $\alpha^*$ value that is used for extrapolating the correct entropy value. In [20] the process is explained in more detail, and it is experimentally verified that $\alpha^*$ is constant for a fixed number of samples and dimensions, and for different covariance matrixes.

## 3 Feature Selection criteria and algorithms

There are different Filter Feature Selection criteria for selecting or discarding a feature or a feature set. In [7], Peng et al. study the possibility maximize the dependency between the feature set $S$ and the prototypes $C$: $\max I(S;C)$, called Max-Dependency criterion. Eq. 8 formulates the maximization objective for selecting the $m$-th feature from the $X - S_{m-1}$ set of features which still are not selected.

$$\max_{x_j \in X - S_{m-1}} I(S_{m-1} \cup \{x_j\}; c) \tag{8}$$

In [7] this criterion is found to be unfeasible because the entropy estimation in high-dimensional feature spaces is very hard, and yields poor results, due to the way they estimate entropy. So instead of estimating mutual information in a multidimensional space, they maximize the relevance $I(x_j; c)$ of each individual feature $x_j$ and at the same time minimize the redundancy between $x_j$ and the rest of selected features $x_i \in S, i \neq j$. This is the Max-Relevance Min-Redundance (mRMR) criterion, Eq. 9.

$$\max_{x_j \in X - S_{m-1}} \left[ I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right] \tag{9}$$

In this work, we state that the Max-Dependency criterion is feasible, even in very high-dimensional feature spaces. The complexity of MI estimation, explained in Section 2, depends on the Entropic Spanning Graph construction, which has a $O(s \log(s))$ order, where $s$ is the number of samples. Also, the accuracy of the estimations does not depend on the number of dimensions.

Another issue, is the way features combinations are generated. An exhaustive search among the features set combinations would present a $O(n!)$ combinatorial complexity, where $n$ is the total number of features. For our experiments we have used a Greedy Forward Feature Selection algorithm, which starts from a small feature set, and adds one feature at a time.

Therefore, with the mRMR criterion each iteration would consist of calculating the MI between a feature and the prototypes, as well as the MI between that feature and each one of the already selected ones (see Eq. 9). Such a search performs $\sum_{i=1}^{n} i(n-i+1)$ estimations of the MI, which has a $O(n^3+n^2+n)$ computational complexity. Using the Max-Dependency criterion instead, requires just one MI calculus per iteration. The total number of MI estimations is $\sum_{i=1}^{n} n-i+1$, which has a $O(n^2+n)$ computational complexity.

## 4 Experiments

### 4.1 Image Data

Two experiments on real data are presented in this paper. The first of them compares the Max-Dependency and the mRMR criteria. The data set consists of a set of 721 images (samples), labeled with 6 different classes. Each sample has 48 features, which come from some basic filters responses, like color filters, corner and edge detectors, and range information. Such a configuration is useful for image classification and image registration purposes. On Fig. 1 we have represented image registration results for a few outdoor images.

In Fig. 2 we show the classification errors of the feature sets selected with both criteria. A Nearest Neighbour classification was evaluated because the number of


**Fig. 1.** Image registration experiment on different test images. The first row contains test images. Each one of them is associated to some image of the training set, shown on the second row. The training set contains 721 images taken during an indoor-outdoor walk. The test set has not been used for the feature selection process and it is taken during a different walk following a similar trajectory. The amount of low-level filters selected for building the classifier is 13, out of 48 in total.

samples is not very high. Only 20 selected features are represented, as for larger features set the error does not decrease. The 10-fold Cross Validation error is represented, as well as a test-error, which was calculated using an additional test-set of images, containing 470 samples.

With mRMR, the lowest CV error (8.05%) is achieved with a set of 17 features, while with Max-Dependency a CV error of 4.16% is achieved with a set of 13 features. The test-errors have similar tendencies.

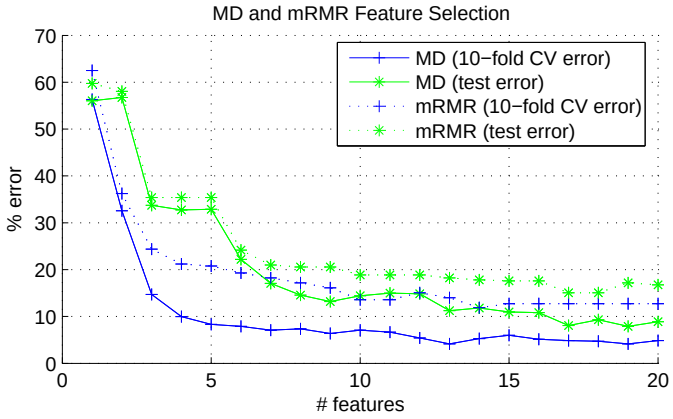


**Fig. 2.** Feature Selection performance on image histograms data with 48 features. Comparison between the Maximum-Dependency (MD) and the Minimum-Redundancy Maximum-Relevance (mRMR) criterions.

In Peng et al.[7], the experiments yielded better results with the mRMR criterion than with the Max-Dependency criterion. Contrarily, we obtain better

performance using Max-Dependency. This may be explained by the fact that the entropy estimator we use does not degrade its accuracy as the number of dimensions increases.

### 4.2 Microarray Data

In order to illustrate the dimensionality independence of the the entropy estimator we use, we performed another experiment on the well-known NCI60 DNA microarray dataset. It contains only 60 samples, labeled with 14 different classes of human tumor diseases. Each sample has 6380 dimensions (genes). The purpose of Feature Selection is to select those genes which are useful for disease classification/prediction. An example can be seen on Fig 4 where we have represented the first 36 features selected by both criteria discussed in this paper. The rows represent the samples, or diseases. The columns represent features. The intensity of the colour is the level of expression of a gene for a given disease. In this figure we can see that MD and mRMR select different genes.

In Fig. 3, we show the Leave One Out Cross Validation errors [1] for the selected feature subsets, using the Max-Dependency criterion. Only the best 220 genes (out of 6380) are on the X axis, and it took about 24 hours on a PC with Matlab to select them. During this time MI was calculated $\sum_{i=1}^{220}(6380-i+1) = 1,385,670$ times.

In [3] an evolutionary algorithm is used for feature selection and the best LOOCV error achieved is $23,77\%$ with a set of 30 selected features. In our experiment we achieve a $10,94\%$ error with 39 selected features.

## 5 Conclusions and future work

In this paper we presented a Filter Feature Selection approach based on Mutual Information. The Mutual Information estimation does not depend on the number of features, but it depends n-logarithmically on the number of samples. Therefore this approach is useful for high-dimensional patterns, such as DNA microarray data. In contrast to Wrapper approaches, this Filter approach does not rely on minimizing the classification error, but on maximizing MI of sets of features and class labels. However as a consequence of this, the classification error actually decreases.

Finally, we obtain better results by evaluating MI (Max-Dependency) for the entire feature subsets, than the criterion of Min-Redundancy Max-Relevance.

In the future we want to explore Feature Selection algorithms different than FFS. This algorithm starts selecting small feature subsets, but with our approach it would not be hard to start from larger feature subsets and remove the less informative ones.

[1] LOOCV measure is used when the number of samples is so small that a test set cannot be built. It consists of building all possible classifiers, each time leaving out only one sample for test.

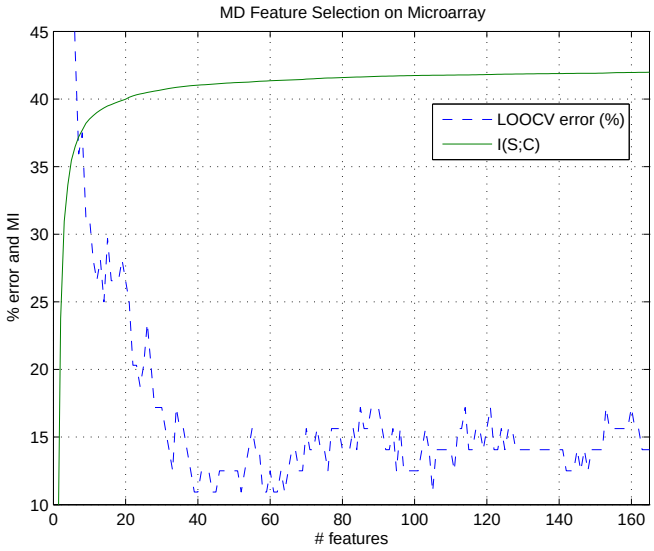


**Fig. 3.** Feature Selection performance on microarray data with 6380 features. The FFS algorithm using the Maximum Dependency criterion obtained the lowest LOOCV error with a set of 39 features. The function that is maximized (the mutual information) is also represented.

**Acknowledgments.** This research is funded by the project DPI2005-01280 from the Spanish Government.

## References

1. H. Neemuchwala, A. Hero, P. Carson: *Image registration methods in high dimensional space*, International Journal on Imaging, 2006
2. C. Sima, E.R. Dougherty: *What should be expected from feature selection in small-sample settings*, Bioinformatics, Vol.22, No.19, pages 2430-2436, 2006
3. T. Jirapech-Umpai, S. Aitken: *Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes*, BMC Bioinformatics, 6:148, 2005
4. I. Guyon, A. Elisseeff: *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, 3(2003):1157-1182
5. A.L. Blum, P. Langley: *Selection of Relevant Features and Examples in Machine Learning*, Artificial Intelligence, 1997
6. S. Perkins, J. Theiler: *Online Feature Selection using Grafting*, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003
7. H. Peng, F. Long, C. Ding: *Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.8, August 2005

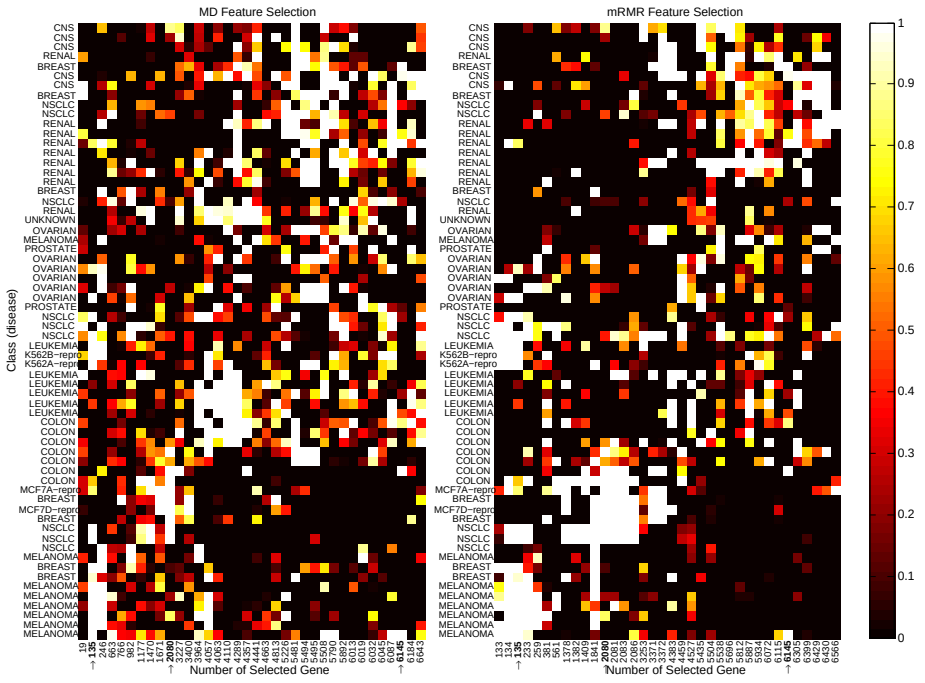


**Fig. 4.** Feature Selection on the NCI DNA microarray data. The MD (on the left) and mRMR (on the right) criteria were used. Features (genes) selected by both criteria are marked with an arrow.

8. T. Cover, J. Thomas: *Elements of Information Theory*, Publisher J. Wiley and Sons, 1991
9. E. Beirlant, E. Dudewicz, L. Gyorfi, E. Van der Meulen: *Nonparametric Entropy Estimation*, International Journal on Mathematical and Statistical Sciences, Vol.5, No.1, pages 17-39, 1996
10. I. Paninski: *Estimation of Entropy and Mutual Information*, Neural Computation, Vol.15, No.1, 2003
11. P. Viola, W. M. Wells-III: *Alignment by Maximization of Mutual Information*, 5th Intern. Conf. on Computer Vision, IEEE, 1995
12. P. Viola, N. N. Schraudolph, T. J. Sejnowski: *Empirical Entropy Manipulation for Real-World Problems*, Adv. in Neural Infor. Proces. Systems, Vol.8, No.1, 1996
13. A. Hyvarinen, E. Oja: *Independent Component Analysis: Algorithms and Applications*, Neural Networks, Vol.13, No.4-5, pages 411-430, 2000
14. D. Wolpert, D. Wolf: *Estimating Function of Probability Distribution from a Finite Set of Samples*,Los Alamos National Laboratory Report LA-UR-92-4369, Santa Fe Institute Report TR-93-07-046, 1995
15. A.O. Hero, O. Michel: *Applications of spanning entropic graphs*, IEEE Signal Processing Magazine, Vol.19, No.5, pages 85-95, 2002
16. A.O. Hero, O. Michel: *Asymptotic theory of greedy aproximations to minnimal k-point random graphs*, IEEE Trans. on Infor. Theory, Vol.45, No.6, pages 1921-1939, 1999
17. D.J. Bertsimas, G. Van Ryzin: *An asymptotic determination of the minimum spanning tree and minimum matching constants in geometrical probability*, Operations Research Letters, Vol.9, No.1, pages 223-231, 1990
18. K. Zyczkowski: *Renyi Extrapolation of Shannon Entropy*, Open Systems and Information Dynamics, Vol.10, No.3, pages 298-310, 2003
19. A. Mokkadem: *Estimation of the entropy and information of absolutely continuous random variables*, IEEE Trans. on Inform. Theory, Vol.35, No.1, pages 193-196, 1989
20. A. Peñalver, F. Escolano, J.M. Sáez: *EBEM: An Entropy-based EM Algorithm for Gaussian Mixture Models*, ICPR, 451-455, 2006
21. E.P. Xing, M.I. Jordan, R.M. Karp: *Feature selection for high-dimensional genomic microarray data*, Proceedings of the Eighteenth International Conference on Machine Learning, pages 601-608, 2001
22. C. Gentile: *Fast Feature Selection from Microarray Expression Data via Multiplicative Large Margin Algorithms*, In Proceedings NIPS, 2003