Pattern Recognition 43 (2010) 2068–2081

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Supervised feature selection by clustering using conditional mutual information-based distances

José Martínez Sotoca *, Filiberto Pla

Institute of New Imaging Technologies, Dept. Llenguatges i Sistemes Informátics, Universitat Jaume I, Campus de Riu Sec, 12071 Castellón, Spain

ARTICLE INFO

Article history:
Received 19 June 2009
Received in revised form
5 December 2009
Accepted 17 December 2009

Keywords:
Supervised feature selection
Clustering
Conditional mutual information

ABSTRACT

In this paper, a supervised feature selection approach is presented, which is based on metric applied on continuous and discrete data representations. This method builds a dissimilarity space using information theoretic measures, in particular conditional mutual information between features with respect to a relevant variable that represents the class labels. Applying a hierarchical clustering, the algorithm searches for a compression of the information contained in the original set of features. The proposed technique is compared with other state of art methods also based on information measures. Eventually, several experiments are presented to show the effectiveness of the features selected from the point of view of classification accuracy.

1. Introduction

Feature selection and construction are key steps in designing pattern recognition and machine learning systems, because of the need to identify and/or extract the most relevant attributes for a given classification or recognition task. Feature construction, sometimes called feature extraction, is referred to the process of extracting new features by transforming the original sample data set representation, in order to have the problem represented in a more discriminative (informative) space that makes the classification task more efficient.

We would refer to feature selection, or variable selection, to the process of selecting the most relevant features (attributes) from an initial set of variables the data set is represented [14,16,21]. In this framework, the term relevant refers to the influence of a given feature or feature set to get the minimum possible error in the classification or recognition problem.

There exist several trends to focus the feature selection problem (see Guyon et al.'s book [12] for a recent comprehensive review). They can be roughly classified into univariate or multivariate methods. The former considers one variable at the time, and the later considers jointly subsets of variables. It is well known that considering individual variables in an independent way do not generally lead to satisfactory classification results, due to the conditional dependencies among different subsets of features. That is, combinations of m best individual features do not imply they are the best m best subset of features [6]. Thus, a multivariate method is considered as better approach.

Regarding the selection strategy, filter methods rank features or feature subsets independently of the classifier, while wrapper methods use a classifier to assess feature subsets, training one learning machine for every feature subset considered, thus, these methods are usually computationally heavy and they are conditioned to the type of classifier used [15,3].

Filters use another evaluation criterion different from the target classification scheme, therefore, it does not usually involve any learning machine in the feature selection process [11]. Eventually, embedded methods try to include the feature selection as a part of the classifier training process, like inherently binary decision tree classifiers do [4].

In order to avoid the combinatorial search problem to find an optimal subset of m features, the most popular variable selection methods apply forward, backward or floating sequential schemes, which always provide a sub-optimal solution. Forward strategies usually provide a nested rank of variables, with the drawback of conditioning the m selected features given the previous $m-1$ selected ones. However, we can often find the case where the m' best subset of features is not included in the m'' best subset of variables, with $m' < m''$ [7]. The backward strategy is analogous to the forward one, but starting from the whole set of variables and discarding one at the time to get to the subset of m desired features. Floating search is an attempt to overcome this problem [25], although, like in forward and backward strategies, you still need to specify a criterion function, and the computational cost may easily increase to evaluate all the possible subsets of features to be tested. The reader is addressed to [13] for a detailed comparison of these strategies.

* Corresponding author. Tel.: +31 964728355.
E-mail addresses: sotoca@lsi.uji.es (J. Martínez Sotoca), pla@lsi.uji.es (F. Pla).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.12.013

The work here presented focuses on filter strategies, in order to meet a twofold objective filter methods pursue, that is, to be computationally efficient and to provide a solution regardless the classification scheme. Filter methods have proven to work better than wrapper using forward and backward strategies for different classification schemes [23]. Moreover, the criterion functions used by most filter methods are driven by some basic and common fundamental feature selection criteria. However, these methods mainly use heuristic approaches.

As already pointed out, the fundamental idea of feature relevancy means minimum classification error. In the supervised feature selection problems, this usually requires the subset of selected features have maximal statistical dependency with respect to the target class variable in the data set, leading to the so-called *maximal dependency* criterium [23]. Further on, we will provide some information theoretical hints that support this criterion in order to achieve minimum classification error.

A drawback when using maximal dependency criterion is that it is usually unpracticable to be assessed. There exist some approaches to approximate the problem, most of them are from a heuristic nature. This is the case of *maximal relevance* criterion, looking for the subset of features that individually have more correlation with the target class variable in the data set. However, this approximation cannot deal with the problem of correlations among features, and provides poor results when original features have significant redundancies. Other works try to tackle this problem by reducing redundancy among the selected features using the *minimal redundancy* principle [23,7], also mostly used in a heuristic manner.

This paper uses an information theoretic framework to try to understand what are the goals pursued by feature selection in a decision problem context, in order to support with a theoretical background the criteria used to explain, define and, possibly obtain a satisfactory solution to some still open issues in feature selection. Thus, a theoretical analysis based on the minimum classification error will lead us to a particular interpretation of how to approximate the maximal dependency principle generally pursued in feature selection processes.

The technique here proposed is based on a novel strategy on feature clustering, inspired by works on distributional word clustering in text classification [27,28], overcoming some of the drawbacks of wrapper feature selectors, using no classification scheme in the selection process, and therefore, able to work in a satisfactory way on several classification frameworks. The feature clustering strategy will also provide a way to overcome the nesting problem of forward and backward methods, widely used as a variable selection strategies, and to show through an extensive experimental comparison its generally better performance than other state of the art feature selection methods based on information theoretic approaches using filter strategies.

2. Foundations and related work

As stated in the introduction, the method here presented is based on a filter strategy, using a criterion function different from any classification rule to be used. The proposed criterion is based on some relationships between information theory concepts when dealing with decision problems.

Let the data set be represented in a feature space denoted, in principle, by a discrete random variable, usually multivariate $\mathbf{X}=(X_1,\ldots,X_n)$, then, the Shannon entropy with probability distribution $p(\mathbf{x})$ for all possible events $\mathbf{x}\in\Omega$ is defined as

$$H(\mathbf{X}) = -\sum_{\mathbf{x}\in\Omega} p(\mathbf{x})\log p(\mathbf{x}) \tag{1}$$

The Shannon entropy can be interpreted as an estimation of the quantity of information represented in random variable $\mathbf{X}$. Analogously, the Shannon entropy of the relevant variable Y representing the class labels, with distribution $p(y)$, is expressed as

$$H(Y) = -\sum_{i=1}^{C} p(y_i)\log p(y_i) \tag{2}$$

where C is the numbers of classes, having possible values of $Y=\{y_1,\ldots,y_c\}$.

The mutual information $I(\mathbf{X};Y)$ between the variable representing the data set $\mathbf{X}$ and the class labels Y is defined as the Kullback–Leibler (KL) divergence [6] between the joint probability distribution $p(\mathbf{x},y)$ and the product distribution $p(\mathbf{x})p(y)$:

$$I(\mathbf{X};Y) = \sum_{\mathbf{x}\in\Omega}\sum_{i=1}^{C} p(\mathbf{x},y_i)\log\frac{p(\mathbf{x},y_i)}{p(\mathbf{x})p(y_i)} \tag{3}$$

The mutual information is a measure of generalized correlation between two random variables, and can also be interpreted as the amount of information shared by two random variables or the quantity of information one variable can predict about the other one.

The conditional mutual information $I(\mathbf{X};Y/\mathbf{Z})$ can also be defined given two random variables $\mathbf{X}$ and $\mathbf{Z}$, corresponding to two feature spaces over the same data set, and the relevant variable Y, representing the class labels. In this case, $I(\mathbf{X};Y/\mathbf{Z})$ can be interpreted as how much information the feature space $\mathbf{X}$ can predict about the relevant variable Y that the feature space $\mathbf{Z}$ cannot.

The conditional entropy $H(\mathbf{X}/Y)$ represents the amount of independent information that data set $\mathbf{X}$ has with respect to the class labels variable Y, and it can be expressed in terms of class probability distributions $p(\mathbf{x}/y)$ as

$$H(\mathbf{X}/Y) = -\sum_{\mathbf{x}\in\Omega}\sum_{i=1}^{C} p(\mathbf{x},y_i)\log p(\mathbf{x}/y_i) \tag{4}$$

In a decision problem, the probability of Bayes error is one of the parameters that characterize the classification problem, assuming the true probability distributions are known. The following property relates the Jensen–Shannon (JS) divergence of the class probability distributions $p(\mathbf{x}/y)$ and the Bayes error defining a lower and upper bound of the Bayes error [20],

$$\frac{1}{4(C-1)}(H(Y)-JS[p(\mathbf{x}/y)])^2 \le P_{bayes}(e) \le \frac{1}{2}(H(Y)-JS[p(\mathbf{x}/y)]) \tag{5}$$

where $JS[p(\mathbf{x}/y)]$ denotes the JS-divergence of probability distribution $p(\mathbf{x}/y)$, which is defined as

$$\begin{aligned} JS[p(\mathbf{x}/y)] &= H\left[\sum_{i=1}^{C} p(y_i)p(\mathbf{x}/y_i)\right] - \sum_{i=1}^{C} p(y_i)H[p(\mathbf{x}/y_i)] \\ &= H(\mathbf{X}) - H(\mathbf{X}/Y) = I(\mathbf{X};Y) \end{aligned}$$

Thus, expression (5) can be written as

$$\frac{1}{4(C-1)}(H(Y)-I(\mathbf{X};Y))^2 \le P_{bayes}(e) \le \frac{1}{2}(H(Y)-I(\mathbf{X};Y)) \tag{6}$$

The JS-divergence is an information measure adequate when there is a set of likelihoods that have associated a set of priors, like in the case of a data partition in a classification problem. It could be interpreted as the difference between the information content when considering the whole distribution and the weighted information content of the likelihoods, that is, the data partition. As pointed out in previous expression, in a decision problem context, it is equivalent to the mutual information among the random variable representing the data distribution and the partition (classes) distribution.

Given a certain classification problem Y, also the so-called the relevant variable, since $I(\mathbf{X};Y) \leq H(Y)$, if we can define the class probability distributions in such a way we maximize $I(\mathbf{X};Y)$, the more such a mutual information is increased, the more the Bayes error is tighten in that particular decision problem.

Given a subset of features $\tilde{\mathbf{X}} \subset \mathbf{X}$, representing a multivariate random variable $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_m)$, $m < n$, it can easily be shown that

$$I(\tilde{\mathbf{X}};Y) \leq I(\mathbf{X};Y)$$

Therefore, when selecting m features, the higher $I(\tilde{\mathbf{X}};Y)$, the closer to $I(\mathbf{X};Y)$. Consequently, using the above relationship and expression (6), taking into account that mutual information between two variables is always non-negative,

$$P_{bayes}(e) \leq \tfrac{1}{2}(H(Y) - I(\mathbf{X};Y)) \leq \tfrac{1}{2}(H(Y) - I(\tilde{\mathbf{X}};Y)) \tag{7}$$

Note that the higher $I(\tilde{\mathbf{X}};Y)$, the more it approaches the Bayes error, which also leads to the subset of selected features that better represent to the original set with respect to the relevant variable Y.

This is the underlying principle that has motivated different approaches for supervised feature selection, named also *max-dependency criterion* [23] and using different optimization strategies [18,2]. All these works differ in the way they approximate two practical issues:

- The estimation of $I(\tilde{\mathbf{X}};Y)$ and $I(\tilde{\mathbf{X}};\mathbf{X})$ become very complex and highly computational expensive, due to the complexity in calculating the joint distributions in high dimensional spaces, because the number of samples is often limited and insufficient, overall when dealing with continuous variables. When having either large number of features or small sample sets, the joint probability distribution can not be estimated reliably.
- The search strategy to overcome the combinatorial problem of finding the optimal solution, for instance by applying forward, backward or floating sequential schemes, which always provide a sub-optimal solution, such as it has been commented in the introduction. Feature ranking and sequential greedy algorithms often do not provide adequate estimates, mainly when selecting few features out of a high dimensional original space.

Distributional clustering of words in text classification problems was first introduced by Pereira et al. [24], and then formalized as a more general technique, the information bottleneck method [28]. This method is based on the application of the variational principle used in the rate distortion theory; formulating the problem of quantizing the values of a random variable $\mathbf{X}$ into a codebook represented by the random variable $\tilde{\mathbf{X}}$ preserving the relevant information about another variable Y, by introducing the variational principle

$$R(D) = \min_{p(\tilde{\mathbf{x}}/\mathbf{x}):I(\tilde{\mathbf{X}};Y) > D} I(\mathbf{X};\tilde{\mathbf{X}})$$

This variational principle leads to interpret the distortion function as the Kullback–Leibler (KL) divergence between the conditional posterior probability of the relevant variable given the original signal $p(y/\mathbf{x})$, and the conditional posterior probability of the compressed one $p(y/\tilde{\mathbf{x}})$, that is,

$$d(\mathbf{X},\tilde{\mathbf{X}}) = KL[p(y/\mathbf{x})|p(y/\tilde{\mathbf{x}})]$$

In general, vector quantization for lossy signal compression is in fact a clustering process, where the possible values of a signal $\mathbf{x} \in \mathbf{X}$ are mapped into a set of clusters $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$, through the stochastic mapping $p(\tilde{\mathbf{x}}/\mathbf{x})$. For instance, the compressed signal values (codewords) could be chosen as the mean value of each cluster; which becomes in a soft version of the well-known k-means or central clustering.

However, the information bottleneck principle does not provide a straightforward way to select features about a relevant variable. It rather provides a criterion to find a codebook of values that live in the same feature space as the original signal or data values. Moreover, it does not provide an algorithm to find the codebook values $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$. A practical approach was then developed in [27], where an agglomerative clustering strategy is proposed, using a single link strategy, merging at each step in a greedy procedure the two clusters of words whose distance was a minimum. The distance is based on measuring the decrease in the mutual information $I(\tilde{\mathbf{X}};Y)$ of clusters $\tilde{\mathbf{x}} \in \tilde{\mathbf{X}}$ with respect to the relevant variable before and after the merging, following an interpretation of the Information Bottleneck Method based on a decision problem as pointed out in expression (7). A similar agglomerative algorithm about distributional clustering of words was previously proposed in [1], using the same distance among clusters.

A closer approach to the problem of supervised feature selection using clustering of words is the work in [9]. They propose to cluster features in the original space (words in the text classification problem), using the information about a relevant variable (the document classes), suggesting a divisive clustering algorithm, with a k-means-like structure, which also pursues forming feature clusters that minimize the already introduced principle of keeping the difference $I(\mathbf{X};Y) - I(\tilde{\mathbf{X}};Y) \geq 0$ as small as possible. To this end, they propose the following distance based in KL-divergence between a feature $X_i; \mathbf{X} = (X_1, \ldots, X_n)$, and a cluster mean feature $\tilde{X}_j; \tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_m)$,

$$d(X_i, \tilde{X}_j) = KL[p(y/X_i)|p(y/\tilde{X}_j)]$$

However, estimating $p(y/X_i)$ in any problem is not a trivial issue, since it expresses the probability that the relevant variable has a certain value (text class) given a feature (word). Notice that this is not a usual posterior, that is, the probability that, given a sample x, belongs to class y, $p(y/x)$. That is, the conditionality was expressed over the feature, not the data values. In [9], they provide a way to estimate these conditional probabilities for text classification, assuming a generative multinomial model, class conditional independence of words and the use of Laplace's rule of succession to relate the original features (words) and the relevant variable values (text classes), although this approach is clearly not extensible to any classification problem. Due to this drawback, for the best of our knowledge, it has not been used in other application fields.

In the present work, we propose a feature clustering-based method for the general feature selection problem, inspired in the previous works on distributional clustering of words, but using a strategy based on minimizing an information measure that approximates the general criterion pointed out in expression (7) for feature selection, that is, to find the subset of m features out of the original set of n features that minimizes as much as possible the decrease of mutual information, $I(\mathbf{X};Y) - I(\tilde{\mathbf{X}};Y) \geq 0$.

3. Clustering features as a selection process

3.1. Introducing the minimal-relevant-redundancy criterion function

As discussed before, the criterion function has to be linked to the concept of selecting the features that keep as much mutual information as possible about the relevant variable, that is, either maximizing $I(\tilde{\mathbf{X}};Y)$, which represents the *max-dependency* criterion with respect to the relevant variable, or minimizing

$I(\mathbf{X};Y)-I(\tilde{\mathbf{X}};Y)\geq 0$. To this end, let us introduce the following properties.

Lemma. *If $\tilde{\mathbf{X}}=(\tilde{X}_1,\ldots,\tilde{X}_m)$ is a subset of m random variables out of the original set of n random variables $\mathbf{X}=(X_1,\ldots,X_n)$, that is, $\tilde{\mathbf{X}}\subset\mathbf{X}$, then, the decrease in mutual information about a relevant variable Y can be expressed as*

$$I(\mathbf{X};Y)-I(\tilde{\mathbf{X}};Y)=I(\mathbf{X};Y/\tilde{\mathbf{X}}) \tag{8}$$

Proof. If $\tilde{\mathbf{X}}\subset\mathbf{X}$, then $H(\mathbf{X})=H(\mathbf{X},\tilde{\mathbf{X}})$, therefore,

$$I(\mathbf{X},\tilde{\mathbf{X}};Y)=H(\mathbf{X},\tilde{\mathbf{X}})-H(\mathbf{X},\tilde{\mathbf{X}}/Y)=H(\mathbf{X})-H(\mathbf{X}/Y)=I(\mathbf{X};Y) \tag{9}$$

On the other hand, by the mutual information chaining rule,

$$I(\mathbf{X},\tilde{\mathbf{X}};Y)=I(\tilde{\mathbf{X}};Y)+I(\mathbf{X};Y/\tilde{\mathbf{X}})$$

$$I(\tilde{\mathbf{X}};Y)=I(\mathbf{X},\tilde{\mathbf{X}};Y)-I(\mathbf{X};Y/\tilde{\mathbf{X}}) \tag{10}$$

Introducing equality (9) in (10), we obtain

$$I(\tilde{\mathbf{X}};Y)=I(\mathbf{X};Y)-I(\mathbf{X};Y/\tilde{\mathbf{X}})$$

Thus,

$$I(\mathbf{X};Y)-I(\tilde{\mathbf{X}};Y)=I(\mathbf{X};Y/\tilde{\mathbf{X}}) \qquad \square$$

Proposition. *Let $\tilde{\mathbf{X}}=(\tilde{X}_1,\ldots,\tilde{X}_m)$ be a subset of m random variables out of the original set of n random variables $\mathbf{X}=(X_1,\ldots,X_n)$, that is, $\tilde{\mathbf{X}}\subset\mathbf{X}$, then, the decrease of mutual information of the original and the reduced set with respect to a relevant variable Y is upper bounded by,*

$$I(\mathbf{X};Y)-I(\tilde{\mathbf{X}};Y)=I(\mathbf{X};Y/\tilde{\mathbf{X}})\leq\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}I(X_i;Y/\tilde{X}_j) \tag{11}$$

Proof. For any three discrete random variables, the following property holds:

$$I(X,Y;Z)=I(Y;Z)+I(X;Z/Y)\leq I(Y;Z)+I(X;Z)$$

Applying successively the previous inequality,

$$\begin{aligned}
&I(\mathbf{X};Y/\tilde{\mathbf{X}})\leq I(X_1;Y/\tilde{\mathbf{X}})+I(\mathbf{X}^{n-1};Y/\tilde{\mathbf{X}})\\
&I(\mathbf{X};Y/\tilde{\mathbf{X}})\leq I(X_1;Y/\tilde{\mathbf{X}})+I(X_2;Y/\tilde{\mathbf{X}})+I(\mathbf{X}^{n-2};Y/\tilde{\mathbf{X}})\\
&\qquad\vdots\\
&I(\mathbf{X};Y/\tilde{\mathbf{X}})\leq\sum_{i=1}^{n}I(X_i;Y/\tilde{\mathbf{X}})
\end{aligned} \tag{12}$$

where $\mathbf{X}^{n-i}=(X_{i+1},X_{i+2},\ldots,X_n)$.

On the other hand, taking into account that, for any set of discrete random variables, it holds that

$$I(X;Y/(Z_1,\ldots,Z_k))\leq I(X;Y/Z_i);\quad \forall i=1,\ldots,k$$

and applying it m times to expression $I(X_i;Y/\tilde{\mathbf{X}})$ then,

$$\begin{aligned}
&I(X_i;Y/\tilde{\mathbf{X}})\leq I(X_i;Y/\tilde{X}_1)\\
&\qquad\vdots\\
&I(X_i;Y/\tilde{\mathbf{X}})\leq I(X_i;Y/\tilde{X}_m)
\end{aligned}$$

Thus,

$$I(X_i;Y/\tilde{\mathbf{X}})\leq\frac{1}{m}\sum_{j=1}^{m}I(X_i;Y/\tilde{X}_j)$$

and inserting it in expression (12),

$$I(\mathbf{X};Y/\tilde{\mathbf{X}})\leq\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}I(X_i;Y/\tilde{X}_j)$$

proving the proposition. $\square$

As we will see later on, the bound in Proposition 1 can also be interpreted as a *minimal relevant redundancy—mRR* criterion, meaning that the selected features will tend to be as much independent as possible, but with respect to the information content of the relevant variable they are trying to explain. Therefore, the max-dependency criterion would be equivalent to a minimal-relevant-redundancy of the selected variables.

3.2. Feature dissimilarities in the mRR criterion

If we look for m features $\tilde{\mathbf{X}}=(\tilde{X}_1,\ldots,\tilde{X}_m)$ to be selected out the n original ones $\mathbf{X}=(X_1,\ldots,X_n)$, the more the m selected features minimize the bound in Proposition 1, the more these features are close to what the original n features can predict about the relevant variable Y.

In order to find a solution to this minimization problem, let us approximate that bound interpreting it as a clustering process. That is, let us assume that we would like to group the n features $\mathbf{X}=(X_1,\ldots,X_n)$ into m clusters, being $\tilde{\mathbf{X}}=(\tilde{X}_1,\ldots,\tilde{X}_m)$ the cluster means or representatives. Therefore, expression

$$F(\tilde{X}_1,\ldots,\tilde{X}_m)=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{m}I(X_i;Y/\tilde{X}_j)=\sum_{i=1}^{n}\sum_{j=1}^{m}p(\tilde{X}_j/X_i)I(X_i;Y/\tilde{X}_j) \tag{13}$$

could be approximated as a k-means-like clustering criterion function, where the conditional posteriors $p(\tilde{X}_j/X_i)=1/m$ correspond to a uniform distribution, instead of a delta distribution the usual k-means algorithm uses:

$$p(\tilde{X}_j/X_i)=\begin{cases}1 & \text{if } X_i\in C(\tilde{X}_j)\\ 0 & \text{otherwise}\end{cases}$$

Moreover, we could associate $I(X_i;Y/\tilde{X}_j)$ to the dissimilarity between feature X_i and the cluster representative (mean) $\tilde{X}_j$. Note that $I(X_i;Y/\tilde{X}_j)$ is the conditional mutual information and can be interpreted as how much information variable X_i can predict about relevant variable Y that variable $\tilde{X}_j$ cannot, that is, in terms of cost, what would be the cost if feature X_i is represented by feature $\tilde{X}_j$, with respect to predicting variable Y.

Therefore, from a clustering point of view, minimizing expression (13) would mean that features in the same cluster would have been grouped around a representative, having features small dissimilarity values with respect to such a representative, meaning that the cost of substituting the variables in the cluster by their representative would be small, that is, the amount of information the features can predict about Y that the representative cannot, would be small. Obviously, the cluster representative would become the selected feature for such a group. Thus, the problem of minimizing expression (13) has been cast as a clustering problem looking for m clusters of n original features, and choosing a cluster representative for each cluster, becoming the m selected features.

Expression (13) may suggest using a k-means-like clustering, but it would involve some practical difficulties when trying to work out the mean feature of each cluster. In order to avoid such drawback, instead of using a k-means-like clustering strategy, an agglomerative hierarchical strategy is proposed, starting with the n original features as single feature clusters, and then merging at each step the two nearest clusters.

The conditional mutual information $I(X_i;Y/\tilde{X}_j)$ is not symmetric, that is, $\forall X_i,X_j: I(X_i;Y/\tilde{X}_j)\neq I(\tilde{X}_j;Y/X_i)$. Thus, a symmetric dissimilarity function would be more adequate for most of clustering processes based on distance computations. Thus, let us define the symmetric version of the dissimilarity function, and

we will proof that this expression is a metric distance

$$D(X_i,\tilde{X}_j)=I(X_i;Y/\tilde{X}_j)+I(\tilde{X}_j;Y/X_i) \tag{14}$$

Let us show that $\forall X_i,X_j : D(X_i,X_j)$ defined previously fulfils all the properties of a metric:

1. It is always non-negative $D(X_i,X_j)\geq 0$.

   Proof. Consider the first term of the right hand part of expression (14)

   $I(X_i;Y/X_j)=H(X_i/X_j)-H(X_i/Y,X_j)$

   but $H(X_i/X_j)\geq H(X_i/Y,X_j)$, then $I(X_i;Y/X_j)\geq 0$. Analogously with the second term of the sum. □
2. If $X_i=X_j$ then $D(X_i,X_j)=0$.

   Proof. Consider

   $I(X_i;Y/X_j)+I(X_j;Y/X_i)=2I(Y;X_i,X_j)-I(Y;X_i)-I(Y;X_j)$

   If $X_i=X_j$, then: $I(Y;X_i,X_j)=I(Y;X_i)=I(Y;X_j)$. Therefore,

   $I(X_i;Y/X_j)+I(X_j;Y/X_i)=0$ □
3. It is commutative: $D(X_i,X_j)=D(X_j,X_i)$. This is direct from the commutative property of the sum in expression (14).
4. Triangular inequality: $\forall X_i,X_j,X_k$: $D(X_i,X_j)+D(X_j,X_k)\geq D(X_i,X_k)$

   Proof. Consider the following properties of mutual information [30]

   $I(X;Y,Z)=I(X;Y)+I(X;Z/Y)$ (a)

   $I(X;Y,Z/T)=I(X;Y/T)+I(X;Z/Y,T)$ (b)

   and the inequality

   $I(X;Z/Y,T)\geq 0$ (c)

   From the first term of the right hand part of expression (14)

   $I(X_i;Y/X_j)+I(X_j;Y/X_k)\geq I(X_i;Y/X_j,X_k)+I(X_j;Y/X_k)$

   and from the identity (b)

   $I(X_i;Y/X_j,X_k)+I(X_j;Y/X_k)=I(Y;X_i,X_j/X_k)$

   Using the identity (a) and the inequality (c)

   $I(Y;X_i,X_j/X_k)=I(Y;X_i/X_k)+I(Y;X_j/X_i,X_k)\geq I(X_i;Y/X_k)$

   Changing in the previous proof X_i and X_k, we obtain

   $I(X_k;Y/X_j)+I(X_j;Y/X_i)\geq I(X_k;Y/X_i)$

   and adding the inequalities,

   $I(X_i;Y/X_j)+I(X_j;Y/X_k)+I(X_k;Y/X_j)+I(X_j;Y/X_i)\geq I(X_i;Y/X_k)+I(X_k;Y/X_i)$

   Then, proving the triangular inequality

   $D(X_i,X_j)+D(X_j,X_k)\geq D(X_i,X_k)$ □

Let us define the normalized version of the metric distance as

$$D_{CMI}(X_i,\tilde{X}_j)=\frac{I(X_i;Y/\tilde{X}_j)+I(\tilde{X}_j;Y/X_i)}{2H(Y)} \tag{15}$$

where the entropy of the relevant variable $2H(Y)$ is a normalization factor, making the function range between 0 and 1, since

$$I(X_i;Y/\tilde{X}_j)\leq H(Y/\tilde{X}_j)\leq H(Y)$$

$$I(\tilde{X}_j;Y/X_i)\leq H(Y/X_i)\leq H(Y)$$

$$I(X_i;Y/\tilde{X}_j)+I(\tilde{X}_j;Y/X_i)\leq H(Y/\tilde{X}_j)+H(Y/X_i)\leq 2H(Y)$$

The interpretation of such a distance function would now be what is the amount of information feature X_i can predict about the relevant variable Y and feature $\tilde{X}_j$ cannot, and vice versa. This is in fact a measure of independency between variables X_i and $\tilde{X}_j$ with respect to variable Y, that is, what is the amount of information they can predict about Y, but they do not share.

This is the concept to what we refer as *relevant independency*, that is, the amount of information two random variables can predict about a relevant one and it is not shared by each other. The complementary concept would be *relevant redundancy*, that is, the amount of information they share that can predict about a relevant one.

Note that, when clustering features using metric distance (15), the intra-cluster feature distances would tend to be small around the cluster feature representatives, that is, they share maximum relevant redundancy (minimum relevant independency), but inter-cluster feature dissimilarities will tend to be as large as possible, leading to a minimum relevant redundancy (maximum relevant independency) among cluster representatives, that is, the selected features.

When using a clustering strategy that minimizes intra-cluster variances while increasing inter-cluster distances, the max-dependency criterion, or equivalently, minimizing the bound in proposition, tends to provide minimal relevant redundancy selected features, leading to what we refer as to mRR criterion.

3.3. Conditional mutual information estimations

To calculate the distance $D_{CMI}(X_i,\tilde{X}_j)$ based on the conditional Mutual Information between pair of variables X_i and $\tilde{X}_j$ with respect to the relevant variable Y, the *pdf* s (probability density functions) have to be estimated from the data set. In the case of discrete data, the joint probability distribution $p(x_i,x_j,y_k)$ is estimated for all different pair of features X_i and X_j and the relevant variable Y by means of joint histograms of the combinations of these three variables in the data set values.

The conditional mutual information $I(X_i;Y/X_j)$ between the variable X_i and the relevant variable Y with respect to the variable X_j, can be calculated as

$$I(X_i;Y/X_j)=\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{C}p(x_i,x_j,y_k)\log\frac{p(x_i,y_k/x_j)}{p(x_i/x_j)p(y_k/x_j)} \tag{16}$$

where N is the number of samples in the database, and C is the number of classes. The conditional mutual information between variable X_j and relevant variable Y with respect to variable X_i, $I(X_j;Y/X_i)$ can be calculated in an analogous way.

For continuous-valued databases, the data distribution can be calculated using a density estimator such as Parzen Windows, which have proven to be effective in real problems [17]. An alternative approach to work with continuous data sets is converting continuous features into discrete ones by sampling, but this usually leads to a loss of information.

Given a set of N d-dimensional training samples defined by the set of variables (features) $\mathbf{X}=\{X_1,\ldots,X_d\}$, an approximation to the density function $p(x)$ can be expressed as

$$p(x)=\frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i,h) \tag{17}$$

where $\delta(\cdot)$ is the Parzen window function, and h is the window width. Usually, $\delta(\cdot)$ is chosen as the Gaussian window:

$$\delta(w,h)=\frac{1}{(2\pi)^{d/2}h^d|\Sigma|^{1/2}}\exp\left\{-\frac{w^T\Sigma^{-1}w}{2h^2}\right\} \tag{18}$$

where $w=x-x_i$, d is the number of dimensions, and Σ is the covariance matrix of w.

Thus, any class conditional likelihood $p(x/y_k)$ with respect to the class y_k can be approximated as

$$p(x/y_k) = \frac{1}{N_k} \sum_{x_i \in y_k} \delta(x - x_i, h) \tag{19}$$

where N_k is the number of the training samples belonging to class y_k. Class conditional posteriors $p(y_k/x)$ can be obtained applying Bayes's Theorem.

Using the estimations of class conditional likelihoods and posteriors, conditional entropy $H(Y/\mathbf{X})$ in a d-dimensional space can be estimated as

$$H(Y/\mathbf{X}) = -\sum_{i=1}^{N} \frac{1}{N} \sum_{k=1}^{C} p(y_k/x_i) \log p(y_k/x_i) \tag{20}$$

Eventually, from the estimation of the conditional entropies, the sum of the conditional mutual information of the variables X_i and X_j with respect to the relevant variable Y, $I(X_i; Y/X_j) + I(X_j; Y/X_i)$ can be calculated as

$$\begin{aligned} I(X_i; Y/X_j) + I(X_j; Y/X_i) &= 2I(X_i, X_j; Y) - I(X_i; Y) - I(X_j; Y) \\ &= H(Y/X_i) + H(Y/X_j) - H(Y/X_i, Xj) \end{aligned} \tag{21}$$

3.4. Hierarchical clustering

The algorithm here proposed uses an agglomerative strategy, that is, it starts with n initial clusters and, at each step, it merges the two most similar groups to form a new cluster. Thus, the number of groups is reduced at each iteration until m clusters are reached. In particular, a hierarchical clustering algorithm based on a Ward's linkage method is used, where the dissimilarity measure consists of the distance D_{CMI} described in the previous section (see expression (15)).

Ward's linkage method [29] has the property of producing minimum variance partitions. Therefore, this method is also called minimum variance clustering because it tries to form each group in a manner that would minimize the loss of intra-cluster variance associated with each grouping. To this end, the hierarchical grouping merges at each step the pair of clusters that minimize the increment intra-cluster dispersion of the whole partition.

Let us suppose that clusters C_r and C_s are merged. The general expression for the distance between the new cluster (C_r, C_s) and another cluster C_k is

$$\begin{aligned} D[(C_k), (C_r, C_s)] = \alpha \cdot D(C_k, C_r) + \beta \cdot D(C_k, C_s) + \gamma \cdot D(C_r, C_s) \\ + \delta \cdot |D(C_k, C_r) - D(C_k, C_s)| \end{aligned} \tag{22}$$

where α, β, γ and δ are coefficients. Ward's inter-cluster distance results from the following coefficients:

$$\alpha = \frac{n_r + n_k}{n_r + n_s + n_k}, \quad \beta = \frac{n_s + n_k}{n_r + n_s + n_k}, \quad \gamma = \frac{-n_k}{n_r + n_s + n_k}, \quad \delta = 0 \tag{23}$$

where n_i is the number of instances in group C_i.

The clustering algorithm starts with the disjoint partition where each feature is a cluster. Next, the algorithm searches for the two nearest clusters that have the minimum distance. Finally, the two clusters selected are merged and the distances among clusters are updated using expression (22).

This process is repeated until $m+1$ clusters are obtained. For the resulting clusters, the representative feature $\tilde{X}_j$ for each cluster C_j is selected as the feature X_j with the highest value of the mutual information with respect to the relevant variable Y,

$$\tilde{X}_j = \{X_j \in C_j; I(X_j; Y) \geq I(X_i; Y);\ \forall X_i \in C_j\} \tag{24}$$

In most practical problems, some of the features from the original set tend to be grouped in a cluster with low significance with respect to the relevant variable. This cluster is considered formed by what we could call residual features in the original set. These features are sometimes random noise features, which have high correlation with other noisy features and share little information with the relevant variable.

The representative feature of the cluster with lowest value of the mutual information is considered as the representative feature belonging to the group of residual features, and it is eliminated obtaining the $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_m)$ subset of the m desired features.

4. Empirical results

The experimental results showed in this section consist of a comparison of the method here presented with similar state of the art techniques that use information measures from the estimation of the *pdf* s. Databases, methods and classifiers used in the validation process are described in the following sections.

4.1. Database description

To test the proposed approach, twelve artificial and real databases have been selected, in order to represent a wide variety of problems. In our experiments, six databases with continuous attributes and six databases with discrete attributes have been used, aiming at analyzing the goodness of the different information measures proposed from the estimation of the *pdf* in both types of data. These databases are known to have irrelevant and redundant features besides attributes with varying relevance.

Two of the databases proposed are multispectral images, called *92AV3C* and *DAISEX'99*. The 92AV3C source of data corresponds to a spectral image (145×145 pixels, 220 bands, and 17 classes composed of different crop types, vegetation, man-made structures, and regions with unknown areas) acquired with the airborne visible/infrared imaging spectrometer (AVIRIS) in June 1992 over the Indian Pine Test site in Northwestern Indiana (http://dynamo.ecn.purdue.edu/~biehl/MultiSpec). In this multispectral image several bands are discarded due to the effect of atmospheric absorption. Thus, 185 out of the 220 bands were used discarding the lowest signal-to-noise ratio (SNR) bands.

The DAISEX'99 project provides useful aerial images about the study of the variability in the reflectance on natural surfaces. This source corresponds to a spectral image (700×670 pixels and seven classes that are composed of crops and an unknown class) acquired with the 128-bands HyMap spectrometer during the DAISEX'99 campaign (http://io.uv.es/projects/daisex/).

The rest of databases included in Table 1 are from the UCI repository[22]. All these databases are well known benchmarks data used in the literature, thus, a brief description of the three databases with higher dimensionality is only provided. *Arr* database is a medical study to determine the type of arrhythmia from the EGC recordings. The authors proposed 16 classes to treat the problem, but only 13 classes are represented in the data set. This database contains 279 attributes, 206 are continuous values and the rest are discrete. All features in this database have been considered as continuous.

Mfeat database consists of features of handwritten numbers (' 0' ,...,' 9') extracted from a collection of Dutch utility maps. The data contains 10 classes, corresponding to binary images where each pattern is described from six feature sets. Gisette database is a two-class classification problem with sparse data applied to handwritten digit recognition. This database contains a relatively small number of samples in a space with high

Table 1
Characteristics of the data sets continuous and discrete.

	Features	Classes	Samples	Data
Waveform40	40	3	5000	Continuous
Vehicle	18	4	846	Continuous
Sonar	60	2	208	Continuous
Arr	279	13	452	Continuous
Hillvalley	100	2	1212	Continuous
Mfeat	679	10	2000	Continuous
Satimage	36	6	6435	Discrete
Dna	180	3	3186	Discrete
92AV3C	185	17	21 025	Discrete
DAISEX'99	128	6	469 000	Discrete
Optdigits	64	10	5620	Discrete
Gisette	5000	2	6000	Discrete

dimensionality. In the UCI repository appears the label classes for training and validation set (6000 and 1000 instances respectively), but the class labels of the test set are not provided. Therefore, for this database, only the samples of the training set have been used, and the number of features were reduced from 5000 to 500. The 500 best features were chosen with respect to the individual relevant information $I(\tilde{X}_i;Y)$ when they are considered independently. Note that most of the 5000 were not relevant, since the average classification accuracy using a KNN3 classifier in a 10-fold cross-validation using the 5000 features was 77.87%, while using the 500 pre-selected features was 78.25%.

4.2. Comparison with other methods

In order to assess the performance of the proposed method, an experimental comparison has been done with respect to other three supervised feature selection approaches [2,18,23]. All these methods are based on information measures, and the techniques used to estimate the class-conditional *pdf* s have been the same for all algorithms. A brief description of the methods used in this comparison is as follows:

- *MIFS algorithm* [2]: The mutual information feature selection (MIFS) is a greedy selection algorithm that uses the mutual information $I(\tilde{X}_i;Y)$ between the feature $\tilde{X}_i$ and the relevant variable Y, and all relations of the mutual information between variable $\tilde{X}_i$ and the rest of selected features so far $\tilde{\mathbf{X}}$. At each step, the best feature is selected, which maximizes the following criterion:

  $$R = I(\tilde{X}_i;Y) - \beta \sum_{\tilde{X}_s \subset \tilde{\mathbf{X}}} I(\tilde{X}_i;\tilde{X}_s)$$

  where β is the redundance parameter; which has to be set up to consider the redundancy among features. If $\beta = 0$, the algorithm chooses features in the order of maximizing the mutual information with respect to the relevant variable, thus mutual information between input features is not taken into consideration, and the redundancy between features is never reflected. The more β grows, the more it excludes redundant features more efficiently.
- *MIFS-U algorithm* [18]: This greedy algorithm tries to maximize the criterion $I(Y;\tilde{X}_i,\tilde{X}_s)$. In order to avoid computational complexity, the following approximation is used:

  $$R = I(\tilde{X}_i;Y) - \beta \sum_{\tilde{X}_s \subset \tilde{\mathbf{X}}} \frac{I(\tilde{X}_s;Y)}{H(\tilde{X}_s)} I(\tilde{X}_i;\tilde{X}_s)$$

  This approximation proposed by the authors leads to a slight change in computational load with respect to the original MIFS, and only the entropy of each feature $H(\tilde{X}_i)$ is calculated when computing the mutual information. Parameter β offers flexibility to the algorithm MIFS and MIFS-U. The authors prove that MIFS-U algorithm performs well when $\beta = 1$. We will use this value in the comparison in MIFS and MIFS-U methods.
- *mRMR algorithm* [23]: This approach combines the constrains based on the *maximal relevance* criterion

  $$\max D(\tilde{\mathbf{X}},Y) = \frac{1}{m} \sum_{\tilde{X}_i \subset \tilde{\mathbf{X}}} I(\tilde{X}_i;Y)$$

  and the *minimal redundancy* criterion

  $$\min R(\tilde{\mathbf{X}}) = \frac{1}{m^2} \sum_{\tilde{X}_i\tilde{X}_j \subset \tilde{\mathbf{X}}} I(\tilde{X}_i;\tilde{X}_j)$$

  where m is the number of features included in $\tilde{\mathbf{X}}$. In practice, for incremental search methods, the criterion called "minimal-redundancy-maximal-relevance" (mRMR) can be defined as

  $$R = I(\tilde{X}_i;Y) - \frac{1}{r-1} \sum_{\tilde{X}_s \subset \tilde{\mathbf{X}}_{r-1}} I(\tilde{X}_i;\tilde{X}_s)$$

  where $r-1$ is the number of features selected so far at each step.

4.3. Classifiers for validation

In order to compare the significance of the subsets of features obtained when using different classification schemes, the quality of the rankings or subsets of features obtained for the different methods is validated using three widely used classifiers:

1. *K-nearest neighbor (KNN3)* [5]: The methods based on distances as the k-nearest neighbor rule, achieve high performance when the number of samples is sufficiently large. This classifier only uses the spatial distributions of the empirical samples without *a priori* assumptions about the distributions of classes. A new sample is classified by calculating the distance to the k-nearest training samples. The class label of the k-nearest neighbor then determines the classification of the new sample by a majority-voting scheme. In the experimental results a value of $k = 3$ neighbors has been used.
2. *Support vector machines (SVM)* [8]: This is a classifier that has become very popular for the last decade, and it uses a kernel to transform the original data in a transformed space where a linear classifier is applied. In this work, the LIBSVM package [19] has been used, which supports both 2-class and multiclass classification. The kernel used is a radial basis function (RBF). LIBSVM provides well performance in RBF kernels, estimating the parameters γ and c.
3. *Decision tree C4.5* [26]: C4.5 is an algorithm based on decision trees developed by Ross Quinlan. The algorithm builds decision trees from a set of training data using an entropy based criterion. Thus, each attribute of the data can be used to make a decision that splits de data into two smaller subsets. The algorithm examines the normalized *information gain* to make the decision. To decrease the level of over-fitting, the tree can be further pruned.

In order to increase the statistical significance of the results when using data sets with a limited number of samples, the average values over 10-fold cross-validation have been used to obtain the classification rates.

For every 10-fold partitions, nine were used as training set and one as test set to assess the classification accuracy. The feature

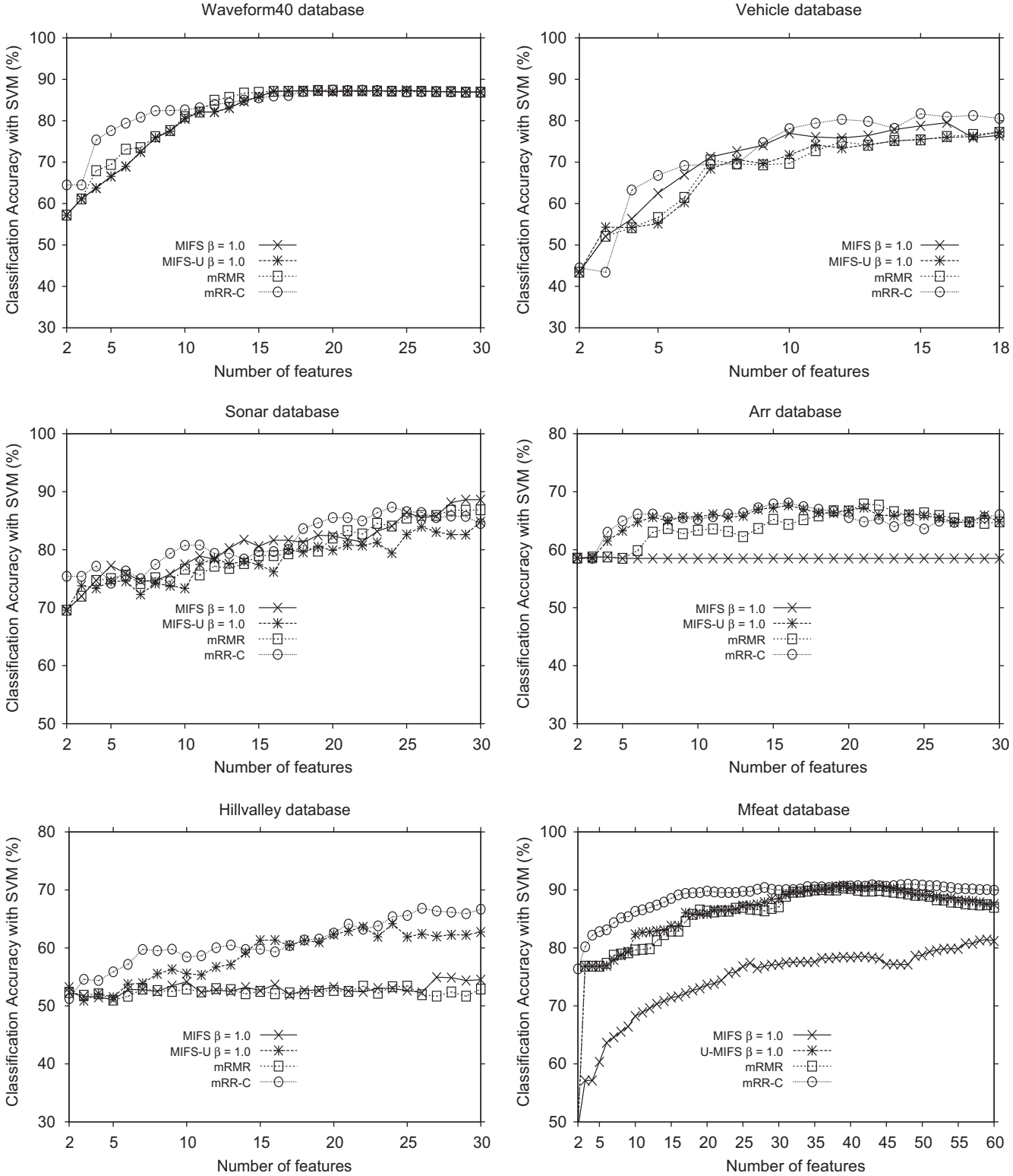

Fig. 1. Results of the classification accuracy with support vector machines in the data sets (left to right, top to bottom): Waveform40, Vehicle Sonar, Arr, Hillvalley, Mfeat. In all cases, classification accuracy has been plotted using an SVM classifier with respect to the number of features obtained by MIFS ($\beta = 1$), MIFS-U ($\beta = 1$), mRMR, mRR-C.

selection algorithm was run on the training set, and for every subset of selected features, the classifiers tested were trained using the same training set where features were selected. Eventually, the average classification results for the 10-fold cross-validation and for every subset of features extracted were obtained.

In the case of data from multispectral images (92AV3C and DAISEX'99), given the huge size of the data sets, twenty independent partitions were randomly extracted (10 as training sets and 10 as test sets). This setup satisfies that the sum of the elements from the different partitions constitutes the whole original set, and the *a priori* probabilities for each class in the data

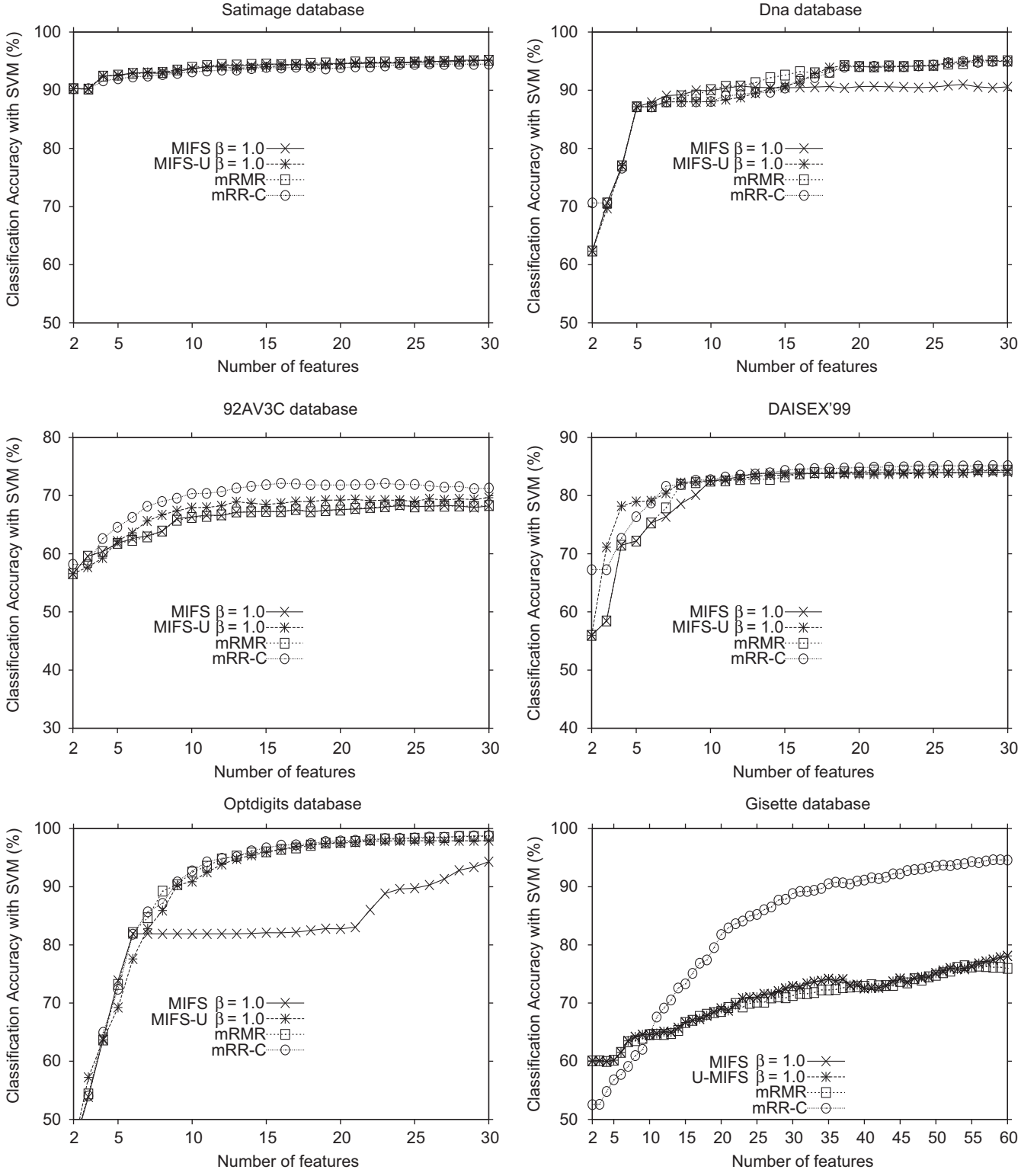

Fig. 2. Results of the classification accuracy with support vector machines in the data sets (left to right, top to bottom): Satimage, Dna, 92AV3C, DAISEX'99, Optdigits and Gisette. In all cases, classification accuracy has been plotted using an SVM classifier with respect to the number of features obtained by MIFS ($\beta = 1$), MIFS-U ($\beta = 1$), mRMR, mRR-C.

sets have been preserved, as well as the statistical independence between the training and test sets of every partition.

Eventually, given the high computational cost that the SVM have if the training set has a significant number of samples, like in the case of DAISEX'99, in this database the instances of the training set were reduced to 1000 samples. Several tests were carried out in order to evaluate how the number of samples used to train the SVM affects both classification rate and computational cost. We found that a size of the order of 1000 samples in the training set obtains very satisfactory classification accuracy with respect to larger amount of samples, and the computational cost is kept affordable.

Table 2
Average classification accuracy (% class) with KNN3, SVM and C4.5 classifiers with different filter methods.

Waveform40 database					Vehicle database			
KNN3	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	56.48	56.48	59.80	**67.50**	53.15	48.25	51.10	**61.43**
$K=10$	64.54	64.54	66.08	**72.71**	59.39	55.07	53.68	**62.28**
$K=15$	69.84	69.84	71.62	**74.91**	61.13	56.81	56.13	**62.84**
$K=30$	75.72	75.72	76.75	**78.12**	61.18	57.74	57.06	**62.73**
SVM	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	65.09	65.09	67.90	**74.22**	72.55	72.25	70.80	**78.05**
$K=10$	72.08	72.08	73.55	**78.70**	76.18	75.78	74.84	**79.40**
$K=15$	76.54	76.54	78.08	**80.93**	77.85	77.21	76.64	**80.40**
$K=30$	82.00	82.00	82.75	**84.09**	72.14	69.58	77.41	**80.74**
C4.5	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	71.40	71.40	73.60	**78.63**	59.45	56.00	70.80	**78.05**
$K=10$	75.98	75.98	77.03	**81.06**	67.65	64.29	74.84	**79.40**
$K=15$	78.25	78.25	79.48	**81.74**	71.23	68.05	76.64	**80.40**
$K=30$	80.71	80.71	81.29	**82.36**	78.54	77.98	77.41	**80.74**
Sonar database					Arr database			
KNN3	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	**72.53**	69.03	69.83	69.83	47.58	51.40	47.80	**52.88**
$K=10$	74.28	71.08	73.73	**74.39**	47.92	53.27	51.92	**55.71**
$K=15$	76.28	74.29	75.98	**76.86**	48.02	54.32	53.59	**57.05**
$K=30$	77.13	78.60	78.19	**78.96**	48.24	55.58	55.16	**56.43**
SVM	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	74.96	74.07	74.36	**75.79**	58.62	62.02	58.96	**63.16**
$K=10$	75.69	74.15	74.83	**77.39**	58.55	64.01	61.36	**64.52**
$K=15$	77.42	75.36	75.93	**78.07**	58.53	64.96	62.21	**65.45**
$K=30$	78.93	76.12	77.68	**79.83**	58.52	65.40	64.12	**65.46**
C4.5	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	78.38	78.80	78.95	**81.25**	66.50	70.20	66.60	**72.30**
$K=10$	79.86	78.71	78.18	**81.00**	66.72	71.30	69.49	**73.80**
$K=15$	80.51	80.33	78.54	**81.62**	66.79	71.65	70.25	**74.06**
$K=30$	80.87	82.97	80.77	**81.68**	66.80	71.40	71.37	**73.20**

4.4. Performance evaluation

Using the previously described experimental setup, the methods that were described for the comparison and the proposed approach were applied in order to obtain a ranking of relevance of the selected features with respect to classification accuracy. The different methods have been validated using three different classifiers. Note that the nature of the decision and the learning process of each classifier is different. Thus, we are interested in checking the goodness of the subsets of selected features, independently from the type of classification rule applied.

Figs. 1 and 2 represents the classification rates using the SVM classifier with respect to the subset of K features selected by each method. The method here proposed uses the mRR criterion with the hierarchical clustering described previously, and it is denoted hereafter as mRR-C. The x-axis represents the subset of features selected, whereas the y-axis shows the average classification accuracy obtained by each method. The results for the k-NN and binary tree classifiers had a similar behavior to the SVM.

In general, the learning rate with regard to the number of features varies depending on the total number of features and the nature of the distribution of classes in the database. In some databases (Waveform40, Sonar, Arr, Dna, 92AV3C, DAISEX'99 and Optdigits) the classifiers reach their maximum classification performance around $K=15$ features. In other databases (Vehicle, Hillvalley, Satimage), they reach their maximum accuracy with very few features, and adding more features tend to maintain or even loose some classification performance. Finally, in the two databases (Mfeat and Gisette) with a high number of features, only the first 60 features are plotted, with the aim of analyzing the behavior of the feature selection methods when they have to select fewer features, where the performance of the feature selection methods become more evident and noticeable.

Tables 2–4 summarize the classification rates for continuous and discrete databases. Results in rows $K=5$, 10, 15, and 30 show the average classification rate in ranges from 2 to 5, from 2 to 10, from 2 to 15, and from 2 to 30 features, respectively. Vehicle database has 18 features, thus the average up to 18 features is shown in row $K=30$. These four average classification rates have been considered to be the approximate transitory period to reach a stable performance for most of the databases and classifiers used. The transitory zone of classification rates with respect to the number of selected features (see Figs. 1 and 2) is considered to be the most critical stage of the feature learning curve, where the feature selection methods show their potential to really select relevant features.

In order to analyze the statistical significance of the results from all methods used in the comparison, a Friedman test [10] has been performed. This is a non-parametric technique to measure

Table 3
Average classification accuracy (% class) with KNN3, SVM and C4.5 classifiers with different filter methods.

Hillvalley database					Mfeat database			
KNN3	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	**52.80**	50.55	52.40	50.85	47.08	63.00	63.00	**75.93**
$K=10$	52.94	51.37	**53.16**	51.29	55.06	69.38	69.18	**78.67**
$K=15$	52.68	52.01	**52.99**	51.78	58.97	72.39	71.69	**80.44**
$K=30$	52.70	**52.96**	52.76	52.29	63.92	77.87	77.10	**83.32**
SVM	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	51.64	51.93	51.80	**54.02**	55.85	69.83	69.83	**80.48**
$K=10$	52.47	53.64	52.20	**56.74**	61.32	74.96	74.79	**82.89**
$K=15$	52.55	55.17	52.33	**57.81**	64.50	77.84	77.08	**84.52**
$K=30$	52.92	58.79	52.40	**60.99**	69.92	82.42	81.77	**87.21**
C4.5	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	63.48	**64.98**	63.88	63.68	81.43	85.88	85.73	**86.88**
$K=10$	62.32	**65.14**	62.67	64.70	87.56	**89.68**	89.61	89.64
$K=15$	62.39	**64.35**	62.35	64.05	89.39	91.15	**91.16**	90.71
$K=30$	61.67	**63.70**	61.38	63.22	91.36	93.16	**93.18**	92.77
Satimage database					Dna database			
KNN3	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	**90.33**	90.05	**90.33**	89.85	68.38	61.95	68.86	**68.90**
$K=10$	91.38	91.12	**91.47**	90.74	77.76	73.83	**78.86**	77.12
$K=15$	92.04	91.66	**92.05**	91.35	80.95	78.56	**82.38**	80.71
$K=30$	92.89	92.71	**92.92**	92.36	82.31	83.56	**85.50**	84.49
SVM	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	**92.04**	**92.04**	92.02	91.53	**80.67**	80.25	80.48	80.39
$K=10$	92.77	92.84	**92.85**	92.29	**85.69**	84.60	85.44	84.74
$K=15$	93.33	93.25	**93.41**	92.78	87.41	86.58	**87.78**	86.67
$K=30$	93.52	93.44	**93.62**	92.97	87.96	87.83	**88.78**	87.80
C4.5	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	**93.08**	92.98	**93.08**	92.65	80.48	**80.58**	80.23	80.15
$K=10$	**93.76**	93.68	93.73	93.37	**86.00**	85.13	85.99	85.01
$K=15$	93.98	93.92	**94.03**	93.64	88.29	87.41	**88.58**	87.31
$K=30$	94.09	**94.14**	94.12	94.00	90.91	91.43	**91.97**	91.31

the significance of the statistical difference of several algorithms that provide results on the same problem, using rankings of results obtained by the algorithms to be compared. For each subset of features, the different accuracies are ranked from one to the number of methods. In this case, the comparison is made over four methods. The method with highest classification accuracy will have rank 1, while the worst method will have rank 4. In case two or more methods have the same accuracy, an average of the ranks is assigned to them. The Friedman estimator F_F follows a Fisher distribution that allows to analyze the statistical significance of the results. This estimator is expressed as

$$\chi_F^2 = \frac{12N_B}{N_M(N_M+1)}\left(\sum_{j=1}^{N_M} R_j^2 - \frac{N_M(N_M+1)^2}{4}\right)$$

$$F_F = \frac{(N_B-1)\chi_F^2}{N_B(N_M-1)-\chi_F^2}$$

where N_M is the number of methods, N_B is the number of databases compared and R_j is the average of the ranks for each method. F_F follows a Fisher distribution with N_M-1 and $(N_M-1)*(N_B-1)$ degrees of freedom. We will use a confidence level $\alpha=0.05$, 95% confidence, to set up the critical value of the Fisher distribution for the four methods $N_M=4$ that appear in the comparative and 12 database $N_B=12$, with degrees of freedom $N_M-1=3$ and $(N_M-1)*(N_B-1)=33$, obtaining a critical value of the Fisher distribution $F(3,33)=2.89$.

Table 5 presents the average classification accuracy of each method for the twelve databases. Note that, in all cases, the proposed method gets the best results for KNN3, SVM and C4.5 classifiers. Nevertheless, the differences in classification accuracy between the methods in some cases are not significant enough according to the Friedman test. In the case of the average classification rate the range from 2 to 30 ($K=30$), the Friedman test is positive for the three classifiers tested, which means that statistical differences are considered significant.

From the rest of the results in the comparison, other interesting points deserve our attention:

1. Notice that the proposed mRR-C method obtained better performance with respect to the rest of methods in the case of databases of Vehicle, Waveform40, Arr, Mfeat, 92AV3C and Gisette and similar accuracy that the best of the other approaches for Sonar, Hillvalley, Satimage, Dna, DAISEX'99 and Optdigits. Thus, the criterion function introduced and the way it is approximated by clustering features with the dissimilarity distance proposed, has the ability to group features decreasing the conditional mutual information between pairs of features in the same group with respect to the relevant variable Y.

Table 4
Average classification accuracy (% class) with KNN3, SVM and C4.5 classifiers with different filter methods.

92AV3C database					DAISEX'99 database			
KNN3	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	50.05	49.75	49.95	**53.45**	57.63	**67.60**	57.13	62.30
$K=10$	52.47	53.51	52.26	**56.92**	70.32	**75.74**	70.75	73.98
$K=15$	54.04	55.24	53.94	**58.33**	75.45	**78.83**	75.60	78.17
$K=30$	56.42	57.68	56.44	**60.32**	80.64	82.31	80.93	**82.84**
SVM	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	61.09	60.69	61.01	**62.89**	69.34	**76.84**	69.34	73.77
$K=10$	63.36	64.26	63.25	**66.55**	75.31	**79.73**	76.04	78.58
$K=15$	64.69	65.81	64.63	**68.32**	78.31	**81.08**	78.55	80.52
$K=30$	66.36	67.59	66.35	**70.07**	81.23	82.53	81.53	**82.81**
C4.5	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	62.40	61.90	62.35	**63.80**	74.65	**82.10**	74.65	80.20
$K=10$	63.76	63.33	62.64	**64.92**	80.86	**84.72**	81.10	84.02
$K=15$	62.98	63.83	62.89	**65.33**	83.37	**85.85**	83.39	85.48
$K=30$	63.40	64.45	63.31	**65.82**	85.76	86.97	85.86	**87.17**
Optdigits database					Gisette database			
KNN3	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	47.93	45.50	47.18	**48.28**	**55.23**	**55.23**	54.80	54.03
$K=10$	63.91	65.42	68.18	**68.72**	**58.86**	**58.86**	58.21	57.11
$K=15$	68.49	75.16	77.37	**77.81**	60.21	60.21	59.81	**60.22**
$K=30$	76.46	86.36	87.59	**87.87**	67.07	67.09	66.99	**79.59**
SVM	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	59.18	58.81	59.08	**59.25**	**60.07**	**60.07**	**60.07**	55.48
$K=10$	71.80	73.83	**75.02**	**75.02**	**62.07**	**62.07**	61.91	59.58
$K=15$	75.43	81.08	82.08	**82.31**	63.28	63.27	63.08	**64.06**
$K=30$	81.64	89.63	90.32	**90.49**	74.70	74.72	74.08	**92.58**
C4.5	MIFS	U-MIFS	mRMR	mRR-C	MIFS	U-MIFS	mRMR	mRR-C
$K=5$	**66.50**	65.35	65.80	66.45	**70.95**	**70.95**	69.99	68.18
$K=10$	76.61	77.43	78.42	**78.89**	**73.88**	**73.88**	72.55	71.79
$K=15$	79.53	82.39	83.19	**83.64**	**74.66**	**74.66**	73.58	74.06
$K=30$	83.99	87.77	88.20	**88.38**	77.74	77.74	77.74	**90.84**

Table 5
Average accuracy (% class) for all databases with KNN3, SVM and C4.5 classifiers, and the result of the Friedman test.

KNN3	MIFS	U-MIFS	mRMR	mRR-C	$F_F(2.89)$	
$K=5$	58.26	59.07	59.35	**62.94**	3.55	Positive
$K=10$	64.07	65.27	65.62	**68.30**	0.93	Negative
$K=15$	66.51	68.28	68.20	**70.87**	1.60	Negative
$K=30$	69.56	72.35	72.28	**74.94**	3.49	Positive
SVM	MIFS	U-MIFS	mRMR	mRR-C	F_F	
$K=5$	65.95	67.70	68.35	**71.20**	2.08	Negative
$K=10$	69.90	71.69	72.17	**74.70**	0.81	Negative
$K=15$	71.93	74.08	74.32	**76.82**	3.76	Positive
$K=30$	74.99	77.51	77.57	**81.25**	5.06	Positive
C4.5	MIFS	U-MIFS	mRMR	mRR-C	F_F	
$K=5$	73.48	74.78	73.81	**76.02**	1.04	Negative
$K=10$	76.87	77.90	77.19	**78.79**	1.42	Negative
$K=15$	78.17	79.25	78.67	**80.17**	1.19	Negative
$K=30$	79.65	81.02	80.63	**82.62**	4.09	Positive

2. In the technique proposed here, the clustering process plays an important role, which can be interpreted as a global strategy to approximate the mRR criterion function introduced in Section 3.1. This strategy allows to explore the different subsets of features in the dissimilarity space built from the dissimilarity measure proposed using conditional mutual information. This strategy is particularly effective and significative in databases with large number of features, obtaining

better results than the greedy selection algorithms used in the comparison (e.g. see the graphics results corresponding to Mfeat and Gisette databases).

3. It is worthwhile mentioning (see Tables 2–4) that the proposed mRR-C algorithm has a good behavior in all cases when choosing smaller sets of features (first two to fifteen features). Only in the case of Gisette database using SVM classifier, the proposed method does not have a learning capability as good as the greedy strategies.
4. Excluding the proposed method, MIFS-U provides higher classification accuracy than the other methods, except in Vehicle, Sonar and Optdigits, where other methods overcome it.
5. mRMR seeks for a balance between maximum relevance and minimum redundance. This allows a slight improvement over MIFS-U in Waveform40, Hillvalley, Satimage, Dna and Optdigits, although from the viewpoint of global ranking of features, it does not overcome the others methods except in Dna database.
6. MIFS performs poorer with than the other approaches in some databases such Arr, Dna and Optdigits although in Vehicle the performance is almost as good as the proposed mRR-C.

5. Conclusions

In this work, a supervised feature selection method has been presented, which uses a feature clustering as an approximation to minimize the proposed minimal relevant redundancy criterion. This criterion is based on minimizing the conditional mutual information between the initial features and the relevant variable denoting the class assigned to each sample, given the selected variables. It has been proven (see Proposition leading to expression (11)) that there exist a direct relationship between the minimal relevant redundancy criterion and the classification error in decision problems.

The feature clustering strategy introduced to minimize the minimal relevant redundancy criterion uses a metric space defined by the conditional mutual information proposed, allowing to compare two variables, overcoming the problem of estimating complex n-dimensional co-joint probabilities, making the method affordable from the computational point of view.

The method has been tested using three different classification schemes, and it has been proven to perform satisfactorily. In addition to the advantages of filter strategies, which can be applied independently of any classification approach, the clustering strategy used also avoids the nesting drawback of most conventional ranking and sequential search methods.

The comparison with other state of art methods based on information theoretic criteria has shown the satisfactory performance of the proposed mRR-C method, which performs better than the other methods for most of databases and classifiers tested. In addition, the Friedman test used to assess the statistical significance of the differences from all tested methods, reveals that for all comparative experiments where the Friedman test is positive, the proposed mRR-C method was the one that performed the best, considering in these cases the differences with the other methods statistically significant.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Education under Project CSD2007-00018, AYA2008-059665-C04-04 from the Spanish CICYT, P1-1B2007-48 and P1-1B2009-45 (Fundació Caixa-Castello).

References

[1] L.D. Baker, A. McCallum, Distributional clustering of words for text classification, in: 21th Annual International ACM SIGIR, ACM, August 1998, pp. 96–103.
[2] R. Battiti, Using mutual information for selection features in supervised neural net learning, IEEE Trans. Neural Networks 5 (4) (1994) 537–550.
[3] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1–2) (1997) 245–271.
[4] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, Boca Raton, 1984.
[5] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory IT-13 (1) (1967) 21–27.
[6] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[7] T.M. Cover, The best two independent measurements are not the two best, IEEE Trans. Syst. Man Cybern. 4 (1974) 116–117.
[8] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, UK, 2000.
[9] I. Dhillong, S. Mallela, R. Kumar, A divisive information-theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res. 3 (2003) 1265–1287.
[10] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (200) (1937) 675–701.
[11] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Special issue on variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[12] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), Feature extraction, foundations and applications, in: Series Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.
[13] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[14] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: 11th International Conference on Machine Learning, Morgan Kaufmann, Los Altos, CA, 1994, pp. 121–129.
[15] R. Kohavi, G.H. John, Wrapper for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[16] M. Kudo, J. Sklansky, Comparison of algorithms that select features for pattern classifiers, Pattern Recognition 33 (2000) 25–41.
[17] N. Kwak, C.-H. Choi, Input feature selection by mutual information based on parzen window, IEEE Trans. Pattern Anal. Mach. Intell. 24 (12) (2002) 1667–1671.
[18] N. Kwak, C.-H. Choi, Input feature selection for classification problems, IEEE Trans. Neural Networks 13 (1) (2002) 143–159.
[19] C.W. Hsu, C.J. Lin, A comparison of methods for multi-class support vector machines, IEEE Trans. Neural Networks 13 (2002) 415–425 [Online]. Available: 〈www.csie.ntu.edu.tw/∼cjlin/libsvm〉.
[20] J. Li, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1) (1991) 145–151.
[21] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Anal. 1 (1997) 131–156.
[22] P.M. Murphy, UCI Repository of Machine Learning 〈http://archive.ics.uci.edu/ml/〉, Department of Information and Computer Science, University of California, Irvine, CA, 1995.
[23] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundance, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[24] F.C. Pereira, N. Tishby, L. Lee, Distributional clustering of English words, in: 30th Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993, pp. 183–190.
[25] P. Pudil, F.J. Ferri, J. Novovicova, J. Kittler, Floating search methods for feature selection with nonmonotonic criterion functions, Pattern Recognition 2 (1994) 279–283.
[26] J.R. Quinlan, Improved use of continuous attributes in C4.5, J. Artif. Intell. Res. 4 (1996) 77–90.
[27] N. Slonim, N. Tishby, Agglomerative information bottleneck, in: Proceedings of Neural Information Processing Systems (NIPS99), 1999, pp. 617–623.
[28] N. Tishby, F. Pereira, W. Bialek, The information bottleneck method, in: B. Hajek, R.S. Sreenivas (Eds.), 37th Annual Allerton Conference Urbana, University of Illinois, 1999.
[29] J.H. Ward, Hierarchical grouping to optimize an objective function, Am. Stat. Assoc. 58 (301) (1963) 236–244.
[30] R.W. Yeung, A First Course in Information Theory, Information Technology: Transmision, Processing, and Storage (series), Springer Science + Business Media, LLC, 2002.

About the Author—JOSÉ MARTÍNEZ SOTOCA received the B.Sc. degree in Physics from the Universidad Nacional de Educación a Distancia, Madrid, Spain, in 1996 and the M.Sc. and Ph.D. degrees in Physics from the University of Valencia, Valencia, Spain, in 1999 and 2001, respectively. His Ph.D. work was on surface reconstructions with structured light.

He is currently an Assistant Lecturer in the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana, Spain. He has collaborated in different projects, most of which are in the medical application of computer science. He has published more than 50 scientific papers in national and international conference proceedings, books and journals. His research interests include pattern recognition and biomedical applications, including image pattern recognition, hyperspectral data, structured light, and feature extraction and selection.

Dr. José Martínez is a member of the International Association for Pattern Recognition.

About the Author—FILIBERTO PLA received the B.Sc. and Ph.D. degrees in Physics from the University of Valencia, Valencia, Spain, in 1989 and 1993, respectively.

He is currently a Full Professor in the Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellón de la Plana, Spain. He has been a Visiting Scientist at several universities and research centers in the UK, France, Italy, Portugal and Switzerland. He has authored more than 100 scientific papers in the fields of computer vision and pattern recognition. He has also been a coeditor of two books and acted as reviewer for several international journals in the field of computer vision and pattern recognition. His current research interests are color and spectral image analysis, visual motion analysis, active vision, and pattern recognition techniques applied to image processing.

Prof. Pla is a member of the International Association for Pattern Recognition.