# Speeding up Feature Selection by Using an Information Theoretic Bound

Patrick E. Meyer Olivier Caelen[1] Gianluca Bontempi

*ULB Machine Learning Group, Computer Science Department,*
*Université Libre de Bruxelles,CP 212, 1050 Brussels*

**Abstract**

The paper proposes a technique for speeding up the search of the optimal set of features in classification problems where the input variables are discrete or nominal. The approach is based on the definition of an upper bound on the mutual information between the target and a set of $d$ input variables. This bound is derived as a function of the mutual information of its subsets of $d-1$ cardinality. The rationale of the algorithm is to proceed to evaluate the mutual information of a subset only if the respective upper bound is sufficiently promising. The computation of the upper bound can thus be seen as a pre-estimation of a subset. We show that the principle of pre-estimating allows to explore a much higher number of combinations of inputs than the classical algorithm of forward selection by preserving the same computational complexity. Some preliminary results showing the effectiveness of the proposed technique with respect to the classical forward search are reported.

## 1 Introduction

In recent years, there has been an increasing demand for effective feature selection due to the need of analyzing datasets with large dimensionality, like in the case of bioinformatics or text mining applications. The benefits of feature selection are well known in the literature [9]: by selecting subsets of relevant features we can increase the accuracy and the intelligibility of a model while at the same time decreasing measurements, storage and time requirements. Also, a lot of experimental results have shown that removing irrelevant and redundant features can dramatically improve the predictive accuracy of learning algorithms [10, 5, 12].

Given a supervised learning problem with $n$ input features, feature subset selection can be seen as the problem of searching the best subset of features in a space containing all the $2^n$ possible combinations. While early approaches to feature selection were based on exhaustive search [3], more recent techniques can be classified according to three main strategies:

**Filter methods:** they are preprocessing methods that attempt to assess the merits of features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm [11, 13]. Examples are methods that select variables by ranking them or by computing correlation with the output.

**Wrapper methods:** these methods assess subsets of variables according to their usefulness to a given predictor [5, 1, 10, 6]. A wrapper method conducts a search for a good subset by using the learning algorithm itself as part of the evaluation function. An example could be a greedy search techniques combined with a neural network assessed by cross-validation.

**Embedded methods:** these techniques perform variable selection as part of the learning procedure and are usually specific to given learning machines. Examples are classification trees, regularization techniques (e.g. lasso).

[1] The work of the researcher was supported by the the F.I.R.S.T. project "Data mining prédictif en anesthésie intraveineuse informatise en vue de certification" (project number EP1A320501R059F/415717 ) of the Région Wallonne, Belgium.

We propose here a filter method based on information theoretic criteria. Several filter approaches to the feature selection problem which rely on empirical estimates of the mutual information between the target and the inputs have been proposed in literature [4, 8]. Their main justification is that the measure of mutual information does not depend on any specific learner but provides a bound on the error rate using any learning technique.

In this context, our contribution is a filter approach based on an information theoretic result that appears as a formal justification of a heuristic commonly used in searching the feature space: *the relevance of a feature set of size $d$ is dependent on the relevance of its subsets of size $d-1$*. This heuristics underlies, implicitly or explicitly, most of the search approaches in the feature selection domain. Think for example to all the greedy techniques that search in a neighborhood of the best feature sets in order to improve the performance.

Here we discuss under which assumption on the input variables it is possible to define an upper bound on the mutual information between the target and a subset $S$ of $d$ variables. This bound is expressed as a function of the mutual information of all the subsets $S_{-(i)}$ of $S$ obtained by removing the $i$th input variable, $i = 1, \ldots, d$. When this assumption holds, the theoretical formulation of the upper bound makes possible a fast estimation of the mutual information of a set on the basis of the mutual information of its subsets.

The resulting algorithm is based on the following idea: since the computation of the mutual information of all the subsets of variables would be too expensive, a shortcut can be provided by first assessing the information of a set of candidates by using the proposed bound (hence-after we call this step *pre-estimation*) and then evaluating only the most promising feature sets. The proposed algorithm presents some similarities with respect to a state of the art approach like Beam [2, 1]. For example, like the Beam algorithm, it keeps a list of the most promising candidates and searches in a neighborhood of them. Unlike Beam, however, an information theoretic bound is used to rapidly scan the neighborhood and focusing on interesting regions.

In order to assess the quality of our algorithm we compare it to the classical forward selection strategy. First, we show that, thanks to the reduced cost of pre-estimation, once we set a computational complexity equal to forward search, our algorithm is able to explore a much higher number of combinations of inputs. Secondly, we perform experimental tests to assess the deterioration of the resulting accuracy when a complexity lower than forward selection is imposed by using a more aggressive approach in discarding non promising subsets. The numerical experiments show that in this case no significant deterioration is visible.

## 2 Upper Bound on the Mutual Information of Set of Variables

Let us consider a classification problem where the target $Y$ is an unidimensional discrete random variable with alphabet $\mathcal{Y}$ and $X = \{X_1, X_2, ..., X_n\}$ is a $n$-dimensional discrete random input variable having alphabet $\mathcal{X}$. Let $S = \{X_i : \ i \in \omega\}, \omega \subset \{1, 2, ..., n\}$ be a subset composed of $d < n$ distinct terms of $X$, whose alphabet is denoted by $\mathcal{S}$. Note that like most of the existing publications on feature selection based on information theoretic approaches we only consider here the case where all the variables are discrete or nominal.

Let $D_N$ be a training dataset made of $N$ random realizations of the joint random variable $(X, Y)$ having distribution $p(x, y)$.

Hence-after we will focus on the following problem: *Given a dataset $D_N$ and an integer number $d \leq n$, find the subset $S = \{X_{(1)}, X_{(2)}, ..., X_{(d)}\}$ of $X$ that maximizes the mutual information $I(S;Y)$ between the feature subset and the output.*

In other words, the goal consists in finding the subset $S$ of $d$ variables such as:

$$S = \arg \max_{S:|\omega|=d} I(S;Y) \tag{1}$$

where $|\omega|$ denotes the cardinality of $\omega$ and the mutual information is [7]:

$$I(S;Y) = \sum_{s \in \mathcal{S}} \sum_{y \in \mathcal{Y}} p(s, y) \log(\frac{p(s, y)}{p(s)p(y)}) \tag{2}$$

We present now a theorem describing the conditions which have to be satisfied in order to derive an upper bound on the mutual information of a subset $S$ of $d$ feature variables.

**THEOREM**: Let $S = \{X_{(1)}, ..., X_{(d)}\}$ be a set of variables and $S_{-(i)} = S \setminus X_{(i)}$ a subset of $S$ that does not contain the variable $X_{(i)}$, $i = 1, \ldots, d$. If we assume that

$$\sum_{i=1}^{d} I(X_{(i)}; Y | S_{-(i)}) \leq I(S; Y) \tag{3}$$

then:

$$\frac{I(S;Y)}{d} \leq \frac{1}{d} \sum_{i=1}^{d} \frac{I(S_{-(i)}; Y)}{d-1} \tag{4}$$

**PROOF**: According to the chain rule for mutual information [7] we obtain:

$$I(X_{(1)}, ..., X_{(d)}; Y) = \sum_{i=1}^{d} I(X_{(i)}; Y | X_{(i-1)}, ..., X_{(1)}) \tag{5}$$

and equivalently for the $d-1$ variables:

$$I(X_{(1)}, ..., X_{(d-1)}; Y) = \sum_{i=1}^{d-1} I(X_{(i)}; Y | X_{(i-1)}, ..., X_{(1)}) \tag{6}$$

Combining (6) and (5) we have:

$$I(X_{(1)}, ..., X_{(d)}; Y) = I(X_{(1)}, ..., X_{(d-1)}; Y) + I(X_{(d)}; Y | X_{(1)}, ..., X_{(d-1)}) \tag{7}$$

which can be written as

$$I(S;Y) = I(S_{-(d)}; Y) + I(X_{(d)}; Y | S_{-(d)}) \tag{8}$$

In the same manner , the following relations hold for the variables $X_{(d-1)}, \ldots, X_{(1)}$

$$I(S;Y) = I(S_{-(d-1)}; Y) + I(X_{(d-1)}; Y | S_{-(d-1)}) \tag{9}$$

$$\ldots$$

$$I(S;Y) = I(S_{-(1)}; Y) + I(X_{(1)}; Y | S_{-(1)}) \tag{10}$$

Summing the $d$ above identities we obtain:

$$dI(S;Y) = \sum_{i=1}^{d} I(S_{-(i)}; Y) + \sum_{i=1}^{d} I(X_{(i)}; Y | S_{-(i)}) \tag{11}$$

If the assumption (3) holds, we obtain:

$$dI(S;Y) \leq \sum_{i=1}^{d} I(S_{-(i)}; Y) + I(S;Y) \Longleftrightarrow (d-1) I(S;Y) \leq \sum_{i=1}^{d} I(S_{-(i)}; Y)$$

$$\Longleftrightarrow I(S;Y) \leq \sum_{i=1}^{d} \frac{I(S_{-(i)}; Y)}{d-1} \tag{12}$$

and dividing by $d$:

$$\frac{I(S;Y)}{d} \leq \frac{1}{d} \sum_{i=1}^{d} \frac{I(S_{-(i)}; Y)}{d-1} \tag{13}$$

□

According to this theorem, if the assumption (3) holds then the *average mutual information per variable* between the target variable $Y$ and a set $S$ is upper bounded by the average of the same quantity on all the sub-subsets $S_{-i}$ of $S$.

In plain words, this theorem justifies the principle according to which, during the exploration of the feature space, in absence of any additional knowledge, we should build new candidate feature sets by combining the most interesting subsets found so far.

Let us discuss now what means and how strong is the assumption (3) in two simple cases: $d = 1$ and $d = 2$.

In the case $d = 1$, we have $S = X_{(1)}$ and the assumption (3) takes the form

$$I(X_{(1)}; Y) \geq I(X_{(1)}; Y|\emptyset) \tag{14}$$

which is always true.

In the case $d = 2$ , we have $S = \{X_{(1)}, X_{(2)}\}$ and the assumption (3) takes the form

$$I(S; Y) \geq I(X_{(1)}; Y|X_{(2)}) + I(X_{(2)}; Y|X_{(1)}) \tag{15}$$

Since according to [7]

$$I(S; Y) = I(X_{(1)}; Y|X_{(2)}) + I(X_{(2)}; Y|X_{(1)}) + I(X_{(1)}; X_{(2)}; Y) \tag{16}$$

where

$$I(X_{(1)}; X_{(2)}; Y) = I(X_{(1)}; Y) - I(X_{(1)}; Y|X_{(2)}) \tag{17}$$

is the information common to $Y$, $X_{(1)}$ and $X_{(2)}$, we obtain that the assumption (15) is equivalent to

$$I(X_{(1)}; X_{(2)}; Y) \geq 0 \tag{18}$$

or

$$I(X_{(1)}; Y) \geq I(X_{(1)}; Y|X_{(2)}) \tag{19}$$

In other words, the theorem assumes that conditioning reduces mutual information. We expect this to be true in general, although some exceptions exist, like when a variable $X_{(1)}$ has more information on an other variable $Y$ once we condition on a third variable $X_{(2)}$. An example is provided in [7]: let $X_{(1)}$ and $Y$ be drawn independently and let $X_{(2)} = X_{(1)} + Y$. This means that although $I(X_{(1)}; Y) = 0$ the quantity $I(X_{(1)}; Y|X_{(2)})$ is maximal. In plain words, this means that when $X_{(2)}$ is a common consequence of $X_{(1)}$ and $Y$, the information on $Y$ deriving from $X_{(1)}$ through $X_{(2)}$ can be higher than the information coming from $X_{(1)}$ directly.

This points out one of the major problems of feature selection: the possible non monotonic behavior of combinations of variables. This means, for example, that a set of two variables, containing an individually irrelevant one, can have more information than a subset composed of two variables, which appear to be significant once taken one at the time.

Notwithstanding, we expect that the proposed theorem will provide a formal framework to characterize the cases when the procedure of combining the most promising subsets is well justified.

# 3 The pre-estimation algorithm

The theoretical results described in the previous section inspired us a new search algorithm for feature selection, hence-after denoted by *PreEST*. In this method the upper bound (4) takes the role of estimator of the mutual information of a subset.

The *PreEST*algorithm relies on two different ways of assessing the relevance of a subset of variables: (i) a *pre-estimation* returned by the upper bound (4) and (ii) an *evaluation* obtained by replacing probability by frequency counts in (2). Note that the pre-estimation step is very fast since it simply requires the computation of an average. It follows that for a set of size $d$ the pre-estimation has a computational complexity of order $O(d)$ whereas the evaluation of mutual information demands a cost of order $O(N \times d)$ with $N$ the number of training samples.

The rationale of the *PreEST*algorithm (see pseudo-code) is that the evaluation assessment of a subset is carried out only if its pre-estimated value is sufficiently interesting. Two parameters are required: the number of the $p$ best subsets returned by the previous step which has to be considered for further exploration and the number $e$ of the most promising subsets to be evaluated among the pre-estimated ones.

At each step, the algorithm first selects the $p$ best subsets assessed so far and obtains $p \times n$ new candidates by combining the $p$ best subsets with the $n$ input variables. The $p \times n$ sets are then pre-estimated by computing the quantity (4). Eventually, the $e$ best subsets according to the pre-estimation assessments are evaluated by directly computing the mutual information (2).

It follows that the algorithm carries out $p \times n$ pre-estimations and $e$ evaluations at each step. As a result, the complexity of each step has order $O(\max(p \times n \times d,\ e \times N \times d))$. Let us note that a conventional greedy strategy using mutual information would require evaluating about $n$ subsets per step, thus featuring a $O(n \times N \times d)$ complexity per step.

Choosing $p = N$ and $e = n$, we can shape the algorithm such that it has the same complexity as a classical greedy search (e.g. forward search). However, in this case, our algorithm, besides evaluating $e = n$ subsets, has the advantage of pre-estimating, for an equal computational complexity, $N$ times more subsets than the forward search. It follows that the region of the feature space which is explored by our algorithm is larger than the one considered by the greedy search.

Note that a drawback of the algorithm resides in how to pre-estimate a feature set when its subsets of size $d-1$ have not been evaluated in the previous step. This is expected to be a rare event for small $d$ but becomes exponentially more frequent for large $d$. Our solution is to assign by default to these subsets the value of the worst evaluation. This amounts to assume that the sets which have not been evaluated are presumably non relevant subsets.

This leads to a progressive bias of the pre-estimation assessment which for increasing $d$ tends to weigh more the best subsets of the previous evaluation. Consequently, the method starts by exploring larger regions than the forward search for small $d$ but ends up converging to a classical forward search for increasing $d$.

PSEUDO-CODE of the preEST search algorithm

d: maximum size of the subset
X: vector of inputs
Y: output
p: number of best subsets
e: number of estimation

```
Improved_Forward(Y,X,d,e,p)
{
subsetlist = X
evaluationlist = evaluate(subsetlist,Y)
subsets = select_best(p,evaluationlist,subsetlist)
while(size(subsets)<d)
    {
    pre-subsetlist = combinate(subsets, X)
    pre-estimationlist = pre-estimate(pre-subsetlist,evaluationlist, subsetlist)
    subsetlist = select_best(e, pre-estimationlist, presubsetlist)
    evaluationlist = evaluate(subsetlist,Y)
    subsets = select_best(p,evaluationlist,subsetlist)
    return(subsets)
    }
}
```

Evaluate(subsetlist,Y) = returns the mutual information between Y
and each subset of the subsetlist.

Select_best(p,evaluationlist,subsetlist) = selects the p best subset from
the subsetlist following their evaluations in the evaluationlist.

Combinate(list1, list2) = returns a list made of the combination of all
the elements of list1 and of list2.

```
Pre-estimate(pre-subsetlist, evaluationlist, subsetlist)
{
worst = minimum(evaluationlist)
for all subset s in the pre-subsetlist
    {
    pre-estim(s)=0
    for all sub-subset ss of the subset s
        {
        if(ss belong to subsetlist)
            pre-estim(s) = pre-estim(s) + evaluationlist(ss)
        else
            pre-estim(s) = pre-estim(s) + worst
        }
    }
return(pre-estim)
}
```

| dataset | n | N | forward | *PreEST* 1 | *PreEST* 2 |
|---|---|---|---|---|---|
| Wisconsin | 32 | 97 | 0.79 | 0.79 | 0.77 |
| Covertype3 | 54 | 5000 | 0.68 | 0.68 | 0.68 |
| Lung-Cancer | 57 | 16 | 0.83 | 0.84 | 0.84 |
| Musk2 | 169 | 3299 | 0.78 | 0.75 | 0.72 |
| Arrythmia | 280 | 226 | 0.73 | 0.74 | 0.71 |
| Isolet | 618 | 780 | 0.78 | 0.79 | 0.75 |
| Multi-features | 649 | 1000 | 0.88 | 0.88 | 0.88 |

Table 1: The first column indicates the name of the dataset coming from the UCI ML repository, the two following columns indicates respectively the number $n$ of variables and the number $N$ of samples of both selection set and test set. The three last columns report the percentage of mutual information (averaged over the 10 test sets for $d = 1, \ldots, 10$) between the target and the subsets returned by the three algorithms.

# 4 Experiments

We carried out an experimental session based on seven datasets provided by the UCI ML repository. The name of the datasets together with the number of variables and samples are reported in Table 1. The datasets are divided into a training set, used to perform the feature selection, and an equal-sized test set. The continuous attributes are discretized by first partitioning their domain into seven equal-sized intervals and then associating a different class to each interval.

We compare the accuracy of the classical forward search and of two versions of the *PreEST*algorithm. The two versions are obtained by setting two set of values for the parameters $p$ and $e$ in order to shape properly their computational cost. In the first version (*PreEST*1) we set $e = n$ and $p = N$ in order to impose the same complexity $O(n \times N \times d)$ as the forward search. In the second version (*PreEST*2) we set $e = \frac{n}{d}$, $p = \frac{N}{d}$ in order to obtain a complexity equal to $O(n \times N)$.

The accuracy of the three algorithms is assessed by performing ten selection tasks of type 1 for $d = 1, \ldots, 10$ and measuring the mutual information (the higher, the better) in the test sets. The value of the normalized mutual information $0 \leq \frac{I(S;Y)}{H(Y)} \leq 1$ averaged over the ten test sets for the three algorithms is reported in Table 1. We have chosen this quantity instead of the usual classification accuracy in order to avoid the bias deriving from the adoption of a specific learning machine.

The column *PreEST* 1 of Table 1 shows that the performance of the improved technique is equivalent to that of classical forward search when we set the parameters in order to obtain the same computational complexity.

At the same time the column *PreEST* 2 shows that a drastic reduction of the complexity of the search can be obtained without causing any major deterioration in the generalization accuracy.

# 5 Conclusion

The paper proposes an information theoretic formalization of the heuristic strategy based on obtaining good feature sets by combining promising subsets. The theoretical result is then exploited in the novel *PreEST*algorithm in order to speed up the search in large feature space. Experimental results show that the *PreEST*algorithm performs as well as the classic forward search for the same complexity. Also, by reducing the *PreEST*complexity no significant deterioration is remarked in our experimental benchmarks. The assumption on conditional information does not appear to be problematic, supposedly because there are few features in datasets that are a common cause (common to the output) of other interesting variables (selected in the subsets).

Future work will focus on extending the pre-estimation approach to other feature selection algorithms (e.g. wrappers) and to study the sensitivity of the performance to the parameters of the algorithm.

# References

[1] D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. *Artificial Intelligence and Statistics*, V, 1996.

[2] D.W. Aha and R. L. Bankert. Feature selection for case-based classification of cloud types: An empirical comparison. In *Proceedings of the Conference on Artificial Intelligence (AAAI-94)*. AAAI Press, 1994.

[3] H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, volume 2. AAAI Press, 1991.

[4] Ron Bekkerman, Ran El-Yaniv, Naftali Tishby, and Yoad Winter. Distributional word clusters vs. words for text categorization. *J. Mach. Learn. Res.*, 3:1183–1208, 2003.

[5] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[6] R. Caruana and D. Freitag. Greedy attribute selection. In *International Conference on Machine Learning*, pages 28–36, 1994.

[7] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wiley, 1990.

[8] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information theoretic feature clustering algorithm for text classification. *J. Mach. Learn. Res.*, 3:1265–1287, 2003.

[9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[10] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[11] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[12] G. Provan and M. Singh. Learning bayesian networks using feature selection, 1995.

[13] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.