Ensemble Feature Ranking

Kees Jong[1], Jérémie Mary[2], Antoine Cornuéjols[2],
Elena Marchiori[1], and Michèle Sebag[2]

[1] Department of Mathematics and Computer Science
Vrije Universiteit Amsterdam, The Netherlands
{cjong,elena}@cs.vu.nl
[2] Laboratoire de Recherche en Informatique, CNRS-INRIA
Université Paris-Sud Orsay, France
{mary,antoine,sebag}@lri.fr

Abstract. A crucial issue for Machine Learning and Data Mining is Feature Selection, selecting the relevant features in order to focus the learning search. A relaxed setting for Feature Selection is known as Feature Ranking, ranking the features with respect to their relevance.
This paper proposes an ensemble approach for Feature Ranking, aggregating feature rankings extracted along independent runs of an evolutionary learning algorithm named *ROGER*. The convergence of ensemble feature ranking is studied in a theoretical perspective, and a statistical model is devised for the empirical validation, inspired from the complexity framework proposed in the Constraint Satisfaction domain. Comparative experiments demonstrate the robustness of the approach for learning (a limited kind of) non-linear concepts, specifically when the features significantly outnumber the examples.

1 Introduction

Feature Selection (FS) is viewed as a major bottleneck of Supervised Machine Learning and Data Mining [13, 10]. For the sake of the learning performance, it is highly desirable to discard irrelevant features prior to learning, especially when the number of available features significantly outnumbers the number of examples, as is the case in Bio Informatics. FS can be formalized as a combinatorial optimization problem, finding the feature set maximizing the quality of the hypothesis learned from these features. Global approaches to this optimization problem, referred to as wrapping methods, actually evaluate a feature set by running a learning algorithm [20, 13]; for this reason, the wrapping approaches hardly scale up to large size problems. Other approaches combine GA-based feature selection with ensemble learning [9].

A relaxed formalization of FS, concerned with feature ranking (FR) [10], is presented in section 2. In the FR setting, one selects the top ranked features, where the number of features to select is specified by the user [11] or analytically determined [19].

A new approach, inspired from bagging and ensemble learning [4] and referred to as ensemble feature ranking (EFR) is introduced in this paper. EFR aggregates several feature rankings independently extracted from the same training set; along the same lines as [4, 5], it is shown that the robustness of ensemble feature ranking increases with the ensemble size (section 3).

In this paper, EFR is implemented using GA-based learning. Practically, we used the *ROGER* algorithm (*ROC-based Genetic Learner*) first presented in [17], that optimizes the so-called AUC criterion. The AUC, the area under the Receiver Operating Characteristics (ROC) curve has been intensively studied in the ML literature since the late 90's [3, 6, 14, 16]. The ensemble feature ranking aggregates the feature rankings extracted from hypotheses learned along independent *ROGER* runs.

The approach is validated using a statistical model inspired from the now standard Constraint Satisfaction framework known as Phase Transition paradigm [12]; this framework was first transported to Machine Learning by [8]. Seven order parameters are defined for FS (section 4); the main originality of the model compared to previous ones [10] is to account for (a limited kind of) non-linear target concepts. A principled and extensive experimental validation along this model demonstrates the good performance of Ensemble Feature Ranking when dealing with non linear concepts (section 5). The paper ends with a discussion and perspectives for further research.

2 State of the art

Without aiming at an exhaustive presentation (see [10] for a comprehensive introduction), this section introduces some Feature Ranking algorithms. *ROGER* is then described for the sake of completeness.

Notations used throughout the paper are first introduced. Only binary concept learning is considered in the following. The training set $\mathcal{E}$ includes n examples, $\mathcal{E} = \{(\mathbf{x}_i, y_i),\ \mathbf{x}_i \in \mathbb{R}^d,\ y_i \in \{-1, 1\},\ i = 1 \ldots n\}$. The i-th example is described from d continuous feature values; label y_i indicates whether the example pertains to the target concept (positive example) or not (negative example).

2.1 Univariate Feature Ranking

In univariate approaches, a score is associated to each feature independently from the others. In counterpart for this simplicity, univariate approaches are hindered by feature redundancy; indeed, features correlated to the target concept will be ranked first, no matter whether they offer little additional information.

The feature score is computed after a statistical test, quantifying how well this feature discriminates positive and negative examples. For instance the Mann-Whitney test, reported to support the identification of differentially relevant features [15], associates to the k-th feature the score defined as $Pr(x_{i,k} > x_{j,k} \mid y_i > y_j)$, i.e. the fraction of pairs of (positive, negative) examples such that feature k ranks the positive example higher than the negative one. This criterion coincides with the Wilcoxon rank sum test, which is equivalent to the AUC criterion [21].

2.2 Univariate FR + Gram Schmidt orthogonalization

A sophisticated extension of univariate approaches, based on an iterative selection process, is presented in [19]. The score associated to each feature is proportional to its cosine with the target concept:

$$score(k) = \frac{\sum_{i=1}^{n} x_{i,k} \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_{i,k}^2}}$$

The two-step iterative process i) determines the current feature k maximizing the above score; ii) projects all remaining features and the target concept on the hyperplane perpendicular to feature k. The stopping criterion is based on an analytic study of the random variable defined as the cosine of the target concept with a random uniform feature.

Though this approach addresses the limitations of univariate approaches with respect to redundant features, it still suffers from the myopia of greedy search strategies (with no backtrack).

2.3 ML-based Approaches

As mentioned earlier on, an alternative to univariate approaches is to exploit the output of a machine learning algorithm, which assumedly takes into account every feature one by one in relation with the other ones [13].

When learning a linear hypothesis ($h(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i\ [+b]$), one associates a score to each feature k, namely the square of the weight w_k; the higher the score, the more relevant the feature is *in combination with the other features*.

A two-step iterative process, termed SVM-Recursive Feature Elimination, is proposed by [11]. In each step, i) a linear SVM is learned, the features are ranked by decreasing absolute weight; ii) the worst features are removed.

Another approach, based on linear regression [1], uses a randomized approach for better robustness. Specifically, a set of linear hypotheses are extracted from independent subsamples of the training set, and the score of the k-th feature averages the feature weight over all hypotheses learned from these subsamples. However, as subsamples must be significantly smaller than the training set in order to provide diverse hypotheses, this approach might be limited in application to domains with few available examples, e.g. DNA array mining.

Another work, more loosely related, is concerned with learning an ensemble of GA-based hypotheses extracted along independent runs [9], where: i) the underlying GA-inducer looks for good feature subsets; and ii) the quality of a feature subset is measured from the accuracy of a k-nearest neighbor or euclidean decision table classification process, based on these features.

2.4 *ROGER* (ROC-based Genetic learneR)

ROGER is an evolutionary learning algorithm first presented in [18, 17]. Using elitist evolution strategies (($\mu + \lambda$)-ES), it determines hypotheses maximizing the Area Under the ROC curve (AUC) [3, 14]. As already mentioned, the AUC criterion was shown equivalent to the Wilcoxon statistics [21].

ROGER allows for constructing a limited kind of non linear hypotheses. More precisely, a hypothesis h measures the weighted L_1 distance to some point $\mathbf{c}$ in the instance space $\mathbb{R}^d$. Formally, to each genetic individual $Z = (w_1, \ldots, w_d, c_1, \ldots, c_d)$ is associated the hypothesis h_Z defined as:

$$h_Z : \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d \mapsto \mathbb{R}, \qquad h_Z(\mathbf{x}) = \sum_{i=1}^{d} w_i \times |x_i - c_i|$$

This way, *ROGER* explores search space $\mathbb{R}^{2d}$, with size linear in the number of features while possibly detecting some non-linearities in the data. *ROGER* maximizes the fitness function $\mathcal{F}$, where $\mathcal{F}(Z)$ is computed as the Wilcoxon statistics associated to h_Z ($\mathcal{F}(Z) = Pr(h_Z(\mathbf{x}_i) > h_Z(\mathbf{x}_j) | y_i > y_j)$).

3 Ensemble Feature Ranking

This section describes an ensemble approach to feature ranking which will be implemented using the *ROGER* algorithm above. The properties of ensemble feature ranking are first examined from a theoretical perspective.

3.1 Notations

Inspired from ensemble learning [4] and randomized algorithms [5], the idea is to combine independent feature rankings into a hopefully more robust feature ranking. Formally, let O_t denote a feature ranking (permutation on $\{1, ..d\}$). With no loss of generality, we assume that features are enumerated with decreasing relevance (e.g. feature i is more relevant than feature j iff $i < j$).

Let $O_1, \ldots, O_T$ be T independent, identically distributed feature rankings. For each feature pair (i, j) let $N_{i,j}$ denote the number of O_t that rank feature i before feature j and let $Y_{i,j}$ be true iff $N_{i,j} > \frac{T}{2}$. We start by showing that a feature ranking can be constructed from variables $Y_{i,j}$, referred to as ensemble feature ranking (EFR); the EFR quality is then studied.

3.2 Consistent Ensemble Feature Ranking

In order to construct an ensemble feature ranking, variables $Y_{i,j}$ must define a transitive relation, i.e. $Y_{i,k}$ is true if $Y_{i,j}$ and $Y_{j,k}$ are true; when this holds for all i, j, k, feature rankings $O_1, \ldots O_T$ are said *consistent*.

The swapping of feature pairs (i, j) is observed from the boolean random variables $X_{i,j}$ ($X_{i,j}(O_t) = ((O_t(i) < O_t(j)) \neq (i < j))$. Inspired from [16], it is assumed that variables X_{ij} are independent Bernoulli random variables with same probability p. Although the independence assumption is certainly not valid (see discussion in [16]), it allows for an analytical study of EFR, while rigorously combining permutations raises more complex mathematical issues.

Lemma. *Let $p = Pr((O_t(i) < O_t(j)) \neq (i < j))$ denote the swapping rate of feature rankings O_t, and assume that $p = \frac{1}{2} - \varepsilon,\ \varepsilon > 0$, (that is, each ranking does a little better than random guessing wrt every pair of attributes). Then*

$$Pr(Y_{i,j} \text{ false } \mid i < j) \ \leq\ e^{-2\varepsilon^2 T}$$

Proof. Follows from Hoeffding's inequality.

Proposition 1. *Under the same assumption, $O_1, \ldots O_T$ are consistent with probability 1 as T goes to infinity.*
Proof. It must be noted first that from Bayes rule, $Pr(i < j \mid Y_{i,j} \text{ true }) = Pr(Y_{i,j} \text{ true } \mid i < j)$ (as $Pr(i < j) = Pr(Y_{i,j} \text{ true }) = \frac{1}{2}$).
Assume that $Y_{i,j}$ and $Y_{j,k}$ are true. After the working assumption that the $X_{i,j}$ are independent,

$$Pr(i < j, j < k \mid Y_{i,j} \wedge Y_{j,k} \text{ true }) = Pr(i < j \mid Y_{i,j} \text{ true}) \cdot Pr(j < k \mid Y_{j,k} \text{ true})$$

Therefore after the lemma, $Y_{i,j}$ and $Y_{j,k}$ true imply that $i < k$ and hence that $Y_{i,k}$ is true, with probability going exponentially fast to 1 as T goes to infinity.

3.3 Convergence

Assuming the consistency of the feature rankings $O_1, \ldots, O_T$, the ensemble feature ranking O^* is naturally defined, counting for each feature i the number of features j that are ranked before i by over half the O_t ($O^*(i) = \#\{Y_{j,i} \text{ true},\ j = 1..d\}$, where $\#A$ is meant for the size of set A).

The convergence of ensemble feature ranking is studied with respect to the probability of misranking a feature i by at most τ indices ($Pr(|O^*(i) - i| \geq \tau)$). Again, for the simplicity of this preliminary analytical study, it is assumed that the misranking probability does not depend on the "true" rank of feature[1] i.

Proposition 2. *Let p^* denote the probability for the ensemble feature ranking to to swap two features, $p^* = Pr(Y_{ij} \neq (i < j))$, and let $\tau = (d-1)p^* + \varepsilon,\ \varepsilon > 0$.*

$$\text{Then} \qquad Pr(|O^*(i) - i| \geq \tau) \ \leq\ e^{-\frac{2\varepsilon^2}{d-1}}$$

Proof. Feature i is misranked by at least τ indices if there exists at least τ features j in the remaining $d-1$ features, such that $Y_{ij} \neq (i < j)$.
Let $\mathcal{B}(d-1, p^*)$ denote the binomial distribution of parameters $d-1$ and p^*, then $Pr(|O^*(i) - i| \geq \tau) < Pr(\mathcal{B}(d-1, p^*) \geq \tau)$, where after Hoeffding's inequality,

$$Pr(\mathcal{B}(d-1, p^*) - (d-1)p^* > \varepsilon) \ \leq\ e^{-\frac{2\varepsilon^2}{d-1}}$$

[1] Clearly, this assumption does not hold, as the probability of misranking top or bottom features is biased compared to other features. However, this preliminary study focuses on the probability to largely misrank features, e.g. the probability of missing a top 10 feature when discarding the 50% features ranked at the bottom.

The good asymptotic behavior of ensemble feature ranking then follows from the fact that: i) the swapping rate p^* of the EFR decreases with the size T of the ensemble, *exponentially amplifying the advantage of the elementary* feature ranking over the random decision [5]; ii) the distribution of the EFR misranking error is centered on $p^* \times (d-1)$, d being the total number of features.

4 Statistical Validation Model

Before proceeding to experimental validation, it must be noted that the performance of a feature selection algorithm is commonly computed from the performance of a learning algorithm based on the selected features, which makes it difficult to compare standalone FS algorithms.

To sidestep this difficulty, a statistical model is devised, enabling the direct evaluation of the proposed FR approach. This model is inspired from the statistical complexity analysis paradigm developed in the Constraint Satisfaction community [12], and first imported in the Machine Learning community by Giordana and Saitta [8]. This model is then discussed wrt [10].

4.1 Principle

In the statistical analysis paradigm, the problem space is defined by a set of order parameters (e.g. the constraint density and tightness in CSPs [12]). The performance of a given algorithm is viewed as a random variable, observed in the problem space. To each point in the problem space (values of the order parameters), one associates the average behavior of the algorithm over all problem instances with same value of the order parameters.

This paradigm has proved insightful in studying the scalability of prominent learning algorithms, and detecting unexpected "failure regions" where the performance abruptly drops to that of random guessing [2].

4.2 Order parameters

Seven order parameters are defined for Feature Selection:

- The number n of examples.
- The total number d of features.
- The number r of relevant features. A feature is said to be relevant iff it is involved in the definition of the target concept, see below.
- The type l of target concept, linear ($l = 1$) or non-linear ($l = 2$), with

$$
\begin{array}{llll}
l = 1: & y(\mathbf{x}) = 1 & \text{iff} & (\sum_{i=1}^{r} x_i > s) \qquad (1.1) \\
l = 2: & y(\mathbf{x}) = 1 & \text{iff} & (\sum_{i=1}^{r} (x_i - .5)^2 < s) \qquad (1.2)
\end{array}
$$

– The redundancy ($k = 0$ or 1) of the relevant features. Practically, redundancy ($k = 1$) is implemented by replacing r of the irrelevant features, by linear random combinations of the r relevant ones[2].
– The noise rate e in the class labels: the class associated to each example is flipped with probability e.
– The noise rate σ in the dataset feature values: each feature value is perturbed by adding a Gaussian noise drawn after $\mathcal{N}(0, \sigma)$.

4.3 Artificial problem generator

For each point $(n, d, r, l, k, e, \sigma)$ in the problem space, independent instances of learning problems are generated after the following distribution.

All d features of all n examples are drawn uniformly in $[0, 1]$. The label of each example is computed as in equation (1.1) (for $l = 1$) or equation (1.2) (for $l = 2$)[3]. In case of redundancy ($k = 1$), r irrelevant features are selected and replaced by linear combinations of the r relevant ones. Last, the example labels are randomly flipped with probability e, and the features are perturbed by addition of a Gaussian noise with variance σ.

The above generator differs from the generator proposed in [10] in several respects. [10] only considers linear target concepts, defined from a linear combination of the relevant features; this way, the target concept differentially depends on relevant features, while all features have the same relevance in our model. In contrast, the proposed model investigates linear as well as a (limited kind of) non-linear concepts.

4.4 Format of the results

Feature rankings are evaluated and compared using a ROC-inspired setting. To each index $i \in \{1, d\}$ is associated the fraction of true relevant features (respectively, the fraction of irrelevant, or falsely relevant, features) with rank higher than i, denoted $TR(i)$ (resp. $FR(i)$). The curve $\{(FR(i), TR(i)), i = 1, \ldots, d\}$ is referred to as ROC-FS curve associated to $\mathcal{O}$.

The ROC-FS curve shows the trade-off achieved by the algorithm between the two objectives of setting high ranks (resp. low ranks) to relevant (resp. irrelevant) features. The ROC-FS curve associated to a perfect ranking (ranking all relevant features before irrelevant ones), reaches the global optimum $(0, 1)$ (no irrelevant feature is selected, $FR = 0$, while all relevant features are selected, $TR = 1$).

[2] Since any subset of r features selected among the r relevant ones plus the r redundant ones is sufficient to explain the target concept, the true relevance rate is set to 1. when at least r features have been selected among the true $2r$ ones.

[3] The threshold s referred to in the target concept definition is set to $r/2$ in equation (1.1) (respectively $r/12$ in equation (1.2)), guaranteeing a balanced distribution of positive and negative examples. The additional difficulties due to skewed example distributions are not considered in this study.

The inspection of the ROC-FS curves shows whether a Feature Ranking algorithm consistently dominates over another one. The curve also gives a precise picture of the algorithm performance; the beginning of the curve shows whether the top ranked features are actually relevant, suggesting an iterative selection approach as in [19]; the end of the curve shows whether the low ranked features are actually irrelevant, suggesting a recursive elimination procedure as in [11].

Finally, three indicators of performance are defined on a feature ranking algorithm. The first indicator measures the probability for the best (top) ranked feature to be relevant, noted p_b, reflecting the FR potential for a selection procedure. The second indicator measures the worst rank of a relevant feature, divided by d, noted p_w, reflecting the FR potential for an elimination procedure.
A third indicator is the area under the ROC-FS curve (AUC), taken as global indicator of performance (the optimal value 1 being obtained for a perfect ranking).

5 Experimental Analysis

This section reports on the experimental validation of the EFR algorithm described in section 3. The results obtained are compared to the state of the art [19] using the cosine criterion. Both algorithms are compared using the ROC-FS curve and the performance measures introduced in section 4.4.

5.1 Experimental setting

A principled experimental validation has been conducted along the formal model defined in the previous section. The number d of features is set to 100, 200 and 500. The number r of relevant features is set to $d/20, d/10$ and $d/5$. The number n of examples is set to $d/2$, d and $2d$. Linear and non-linear target concepts are considered ($l = 1$ or 2), with redundant ($k = 1$) and non-redundant ($k = 0$) feature sets. Last, the label noise e is set to 0, 5 and 10%, and the variance σ of the feature Gaussian noise is set to 0., .05 and .10.

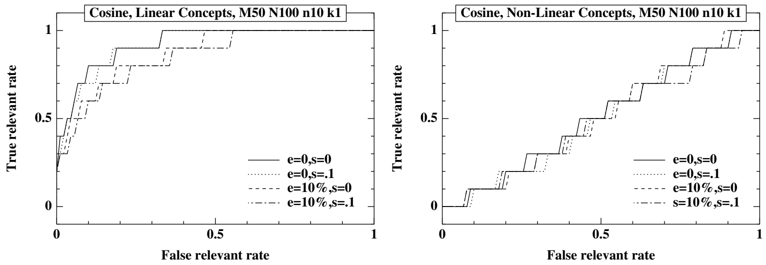
In total 972 points $(d, r, m, l, k, e, \sigma)$ of the problem space are considered. For each point, 20 datasets are independently generated. For each dataset, 15 independent *ROGER* runs are executed to construct an ensemble feature ranking $\mathcal{O}$; the associated indicators p_b, p_w and the AUC are computed, and their median over all datasets with same order parameters is reported.

The reference results are obtained similarly from the cosine criterion [19]: for each point of the problem space, 30 datasets are independently generated, the cosine-based feature ranking is evaluated from indicators p_b, p_w and the AUC, and the indicator median over all 30 datasets is reported.

Computational runtimes are measured on PC Pentium-IV; the algorithms are written in C++. *ROGER* is parameterized as a (20+200)-ES with self adaptive mutation, uniform crossover with rate .6, uniform initialization in $[0, 1]$, and a maximum number of 50,000 fitness evaluations[4].

[4] All datasets and *ROGER* results are available at `http://www.lri.fr/~sebag/EFRDatasets` and `http://www.lri.fr/~sebag/EFResults`.

5.2 Reference results



(a) Linear concepts (b) Non-linear concepts

Fig. 1. Cosine criterion: Median ROC-FS curves over 30 training sets on Linear and Non-Linear concepts, with $d = 100$, $n = d/2$, $r = d/10$, Non redundant features.

Table 1. The cosine ranking criterion: Probability p_b of top ranking a relevant feature, Median relative rank p_w of the worst ranked relevant feature, Area under the ROC-FS curve.

					Linear concepts						Non Linear concepts					
					No redundancy			Redundancy			No redundancy			Redundancy		
n	d	r	e	σ	p_b	p_w	AUC	p_b	p_w	AUC	p_b	p_w	AUC	p_b	p_w	AUC
50	100	10	0	0	0.87	.33	0.920	0.97	.10	0.97	0.03	.93	0.49	0.17	.42	0.74
50	100	10	0	0.1	0.9	.33	0.916	0.9	.10	0.97	0.03	.94	0.49	0.1	.45	0.74
50	100	10	10%	0	0.87	.47	0.87	0.77	.16	0.95	0.1	.93	0.49	0.1	.44	0.73
50	100	10	10%	0.1	0.8	.56	0.848	0.83	.17	0.95	0.03	.93	0.51	0.1	.44	0.75
100	100	10	0	0	1	.18	0.97	1	.10	0.99	0	.91	0.53	0.23	.42	0.76
100	100	10	0	0.1	1	.22	0.966	1	.10	0.99	0.03	.90	0.52	0.17	.41	0.75
100	100	10	10%	0	0.93	.29	0.944	0.97	.10	0.98	0.17	.92	0.52	0.27	.47	0.76
100	100	10	10%	0.1	0.93	.36	0.934	0.97	.10	0.97	0.1	.92	0.52	0.3	.46	0.74

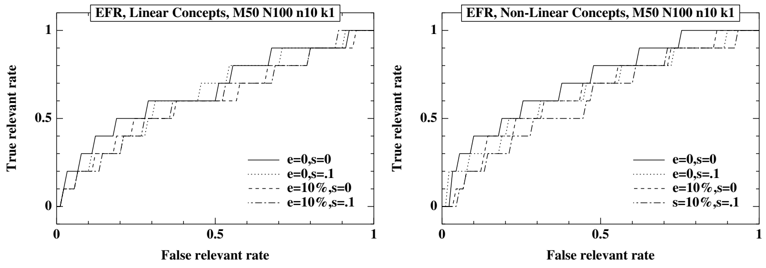
The performance of the cosine criterion for linear and non-linear concepts is illustrated on Fig. 1, where the number d of features is 100, the number n of examples is 50, and the number of relevant features r is 10. The performance indicators are summarized in Table 1; complementary results, omitted due to space limitations, show similar trends for higher values of d.

An outstanding performance is obtained for linear concepts. With twice as many features as examples, the probability p_b of top ranking a relevant feature is around 90%. A graceful degradation of p_b is observed as the noise rate increases, more sensitive to the label noise than to the feature noise. The relevant features are in the top p_w features, where p_w varies from 1/3 to roughly 1/2.

The performance steadily improves when the number of examples increases, p_b reaching 100% and p_w ranging from 1/5 to 1/3 for $n = d$.

In contrast, the cosine criterion behaves no better than random ranking for non-linear concepts; this is visible as the ROC-FS curve is close to the diagonal, and the situation does not improve by doubling the number of examples. The seemingly better performances for redundant features is explained as the true relevance rate models the probability of extracting at most r features among $2r$ ones.

5.3 Evolutionary Feature Ranking



(a) Linear concept (b) Non-linear concept

Fig. 2. EFR performance: Median ROC-FS curves over 20 training sets on Linear and Non-Linear concepts, with $d = 100$, $n = d/2$, $r = d/10$, Non redundant features.

Table 2. Ensemble Feature Ranking with ROGER: Probability p_b of top ranking a relevant feature, Median relative rank p_w of the worst ranked relevant feature, Area under the ROC-FS curve.

					Linear concepts: No redundancy			Linear concepts: Redundancy			Non Linear concepts: No redundancy			Non Linear concepts: Redundancy		
n	d	r	e	σ	p_b	p_w	AUC	p_b	p_w	AUC	p_b	p_w	AUC	p_b	p_w	AUC
50	100	10	0	0	0.5	.92	0.67	0.65	.29	0.86	0.20	.75	0.71	0.50	.26	0.88
50	100	10	0	0.1	0.5	.80	0.63	0.75	.30	0.85	0.45	.82	0.68	0.25	.33	0.84
50	100	10	10%	0	0.35	.94	0.61	0.45	.31	0.85	0.25	.81	0.68	0.30	.32	0.83
50	100	10	10%	0.1	0.35	.89	0.62	0.60	.40	0.82	0.25	.88	0.61	0.35	.28	0.83
100	100	10	0	0	0.85	.79	0.79	0.90	.23	0.92	0.55	.63	0.81	0.80	.20	0.92
100	100	10	0	0.1	0.50	.74	0.77	0.95	.21	0.92	0.60	.72	0.78	0.50	.22	0.90
100	100	10	10%	0	0.55	.77	0.72	0.65	.27	0.89	0.65	.78	0.77	0.50	.20	0.91
100	100	10	10%	0.1	0.65	.82	0.75	0.45	.28	0.88	0.40	.72	0.75	0.40	.26	0.87

The performance of EFR is measured under the same conditions (Fig. 2, Table 2). EFR is clearly outperformed by the cosine criterion in the linear case. With twice as many features as examples, the probability p_b of top ranking a relevant feature ranges between 35 and 50% (non redundant features), against 80 and 90% for the reference results. When the number of examples increases,

p_b increases as expected; but p_b reaches 55 to 85% against 93 to 100% for the reference results.

In contrast, EFR does significantly better than the reference criterion in the non-linear case. Probability p_b ranges around 30%, compared to 3% and 10% for the reference results with $n = 50$ and p_b increases up to circa 55% when n increases up to 100.

With respect to computational cost, the cosine criterion is linear in the number of examples and in $d \log d$ wrt the number of features; the runtime is negligible in the experiment range.

The computational complexity of EFR is likewise linear in the number of examples. The complexity wrt the number of features d is more difficult to assess as d governs the size of the *ROGER* search space ($[0,1]^{2d}$). The total cost is less than 6 minutes (for 20 data sets $\times$ 15 *ROGER* runs) for $n = 50, d = 100$ and less than 12 minutes for $n = 100, d = 100$. The scalability is demonstrated in the experiment range as the cost for $n = 50, d = 500$ is less than 23 minutes.

6 Discussion and Perspectives

The contribution of this paper is based on the exploitation of the diverse hypotheses extracted along independent runs of evolutionary learning algorithms, here *ROGER*. This collection of hypotheses is exploited for ensemble-based feature ranking, extending the ensemble learning approach [4] to Feature Selection and Ranking [10].

As should have been expected, the performances of the Evolutionary Feature Ranker presented are not competitive with the state of the art for linear concepts. However, the flexibility of the hypothesis search space explored by *ROGER* allows for a breakthrough in (a limited case of) non-linear concepts, even when the number of examples is a fraction of the number of features.

These results are based on experimental validation over 9,000 datasets, conducted after a statistical model of Feature Ranking problems. Experiments on real-world data are underway to better investigate the EFR performance, and the limitations of the simple model of non-linear concepts proposed.

Further research will take advantage of multi-modal evolutionary optimization heuristics to extract diverse hypotheses from each *ROGER* run, hopefully reducing the overall computational cost of the approach and addressing more complex learning concepts (e.g. disjunctive concepts).

Acknowledgments

The second, third and last authors are partially supported by the PASCAL Network of Excellence, IST-2002-506778.

References

1. J. Bi, K.P. Bennett, M. Embrechts, C.M. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *J. of Machine Learning Research*, 3:1229–1243, 2003.
2. M. Botta, A. Giordana, L. Saitta, and M. Sebag. Relational learning as search in a critical region. *J. of Machine Learning Research*, 4:431–463, 2003.
3. A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 1997.
4. L. Breiman. Arcing classifiers. *Annals of Statistics*, 26(3):801–845, 1998.
5. R. Esposito and L. Saitta. Monte Carlo theory as an explanation of bagging and boosting. In *Proc. of IJCAI'03*, pages 499–504. 2003.
6. C. Ferri, P. A. Flach, and J. Hernández-Orallo. Learning decision trees using the area under the ROC curve. In *Proc. ICML'02* , pages 179–186. Morgan Kaufmann, 2002.
7. Y. Freund and R.E. Shapire. Experiments with a new boosting algorithm. In L. Saitta, editor, *Proc. ICML'96* , pages 148–156. Morgan Kaufmann, 1996.
8. A. Giordana and L. Saitta. Phase transitions in relational learning. *Machine Learning*, 41:217–251, 2000.
9. C. Guerra-Salcedo and D. Whitley. Genetic approach to feature selection for ensemble creation. In *Proc. GECCO'99* , pages 236–243, 1999.
10. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. of Machine Learning Research*, 3:1157–1182, 2003.
11. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.
12. T. Hogg, B.A. Huberman, and C.P. Williams (Eds). *Artificial Intelligence: Special Issue on Frontiers in Problem Solving: Phase Transitions and Complexity*, volume 81(1-2). Elsevier, 1996.
13. G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proc. ICML'94* , pages 121–129. Morgan Kaufmann, 1994.
14. C.X. Ling, J. Hunag, and H. Zhang. AUC: a better measure than accuracy in comparing learning algorithms. In *Proc. of IJCAI'03*, 2003.
15. M. S. Pepe, G. Longton, G. L. Anderson, and M. Schummer. Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59:133–142, 2003.
16. S. Rosset. Model selection via the AUC. In *Proc. ICML'04* . Morgan Kaufmann, 2004, to appear.
17. M. Sebag, J. Azé, and N. Lucas. Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning. In *IEEE-ICDM03* , pages 637–640, 2003.
18. M. Sebag, J. Azé, and N. Lucas. ROC-based evolutionary learning: Application to medical data mining. In *Artificial Evolution VI*, pages 384–396. Springer Verlag LNCS 2936, 2004.
19. H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *J. of Machine Learning Research*, 3:1399–1414, 2003.
20. H. Vafaie and K. De Jong. Genetic algorithms as a tool for feature selection in machine learning. In *Proc. ICTAI'92* , pages 200–204, 1992.
21. L. Yan, R. H. Dodier, M. Mozer, and R. H. Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proc. of ICML'03* , pages 848–855. Morgan Kaufmann, 2003.