ROC-based Evolutionary Learning: Application to Medical Data Mining

Michèle Sebag, Jérôme Azé, and Noël Lucas

TAO : Thème Apprentissage et Optimisation
Laboratoire de Recherche en Informatique, CNRS UMR 8623
Université Paris-Sud Orsay, 91405 Orsay Cedex
sebag,aze,lucas@lri.fr

Abstract. A novel way of comparing supervised learning algorithms has been introduced since the late 90's, based on Receiver Operating Characteristics (ROC) curves.
From this approach is derived a NP complete optimization criterion for supervised learning, the area under the ROC curve.
This optimization criterion, tackled with evolution strategies, is experimentally compared to the structural risk criterion tackled by quadratic optimization in Support Vector Machines. Comparable results are obtained on a set of benchmark problems in the Irvine repository, within a fraction of the SVM computational cost.
Additionally, the variety of solutions provided by evolutionary computation can be exploited for visually inspecting the contributing factors of the phenomenon under study. The impact study and sensitivity analysis facilities offered by *ROGER* (ROC-based Genetic LearneR) are demonstrated on a medical application, the identification of Atherosclerosis Risk Factors.

1 Introduction

Supervised machine learning (ML) is interested in estimating a nominal or numerical variable based on some set of labeled examples, or training set.

The learning performance is usually measured from the predictive accuracy of the estimator or hypothesis, i.e. the percentage of correctly identified labels in another set of examples, the test set [Die98].

Though predictive accuracy was commonly used to compare learning algorithms, it suffers from several shortcomings regarding skewed example distributions (e.g. discriminating a 1% positive examples from 99% negative examples) and asymmetric mis-classification costs [Dom99].

A remedy to these limitations was offered by Receiver Operating Characteristics (ROC) analysis [Bra97,PFK98,LHZ03], as will be detailed in the next Section. ROC curves, originated from the signal theory, are popular in Medical Data Analysis as they offer a synthetic representation of the trade-off between the true positive rate (TP) and the false positive rate (FP) depending on how a medical test is interpreted, e.g. which thresholds are used to tell pathological from normal cases (Fig. 1).

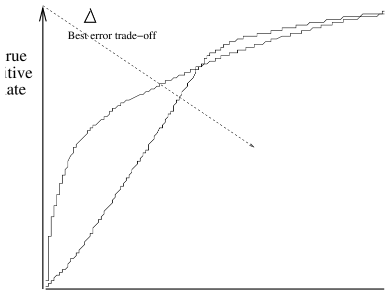


Fig. 1. Comparative evaluation of two hypotheses from their ROC curves

Along these lines, it came naturally to consider the area under the ROC curve (AUC) as a learning criterion [MDC$^+$01,FFHO02]. As this criterion induces a mixed, NP complete optimization problem [CSS98], its optimization was tackled within greedy search [FFHO02] or genetic algorithms [MDC$^+$01].

Independently, the foundational approach of Statistical Learning offered a variety of learning optimization criteria, drastically renewing the Machine Learning perspective [Vap98,SBS98,CST00]. Such criteria derive well-posed optimization problems, such that the optimum solution offers statistical guarantees of learning performance; for instance, Support Vector Machines (SVMs) are determined by quadratic minimization of the structural risk criterion.

The present paper presents the *ROGER* (*ROC-based Genetic Learning*) algorithm, implementing the evolution-strategy based optimization of the AUC criterion; *ROGER* is compared to a state-of-art SVM algorithm known as *SVM-Torch* [CB01].

Both algorithms demonstrate comparable learning performance on a subset of benchmark problems in the Irvine repository [BKM98]; however, *ROGER* requires a fraction of *SVMTorch* computational effort. The differences in the learning behavior of both algorithms are discussed and some interpretations are proposed.

Finally, this paper presents a novel exploitation of the variety of hypotheses provided by evolutionary optimization, through impact study and sensitivity analysis visual facilities. As noted by [CMS99], visual representations can provide the expert with a wealth of easy-to-understand and yet precise information. These visual facilities are illustrated and discussed on a medical data mining application, the identification of Atherosclerosis Risks, presented in the PKDD 2002 Challenge [LAS02].

This paper is organized as follows. Section 2 presents ROC analysis and reviews related work. Section 3 describes the *ROGER* algorithm. Section 4 reports on comparative experiments, using *SVMTorch* as reference algorithm. Section 5 finally describes and discusses how to exploit the variety of evolution-based solutions in a Visual Data Mining perspective.

2 ROC curves and Machine Learning

This section briefly situates the use of ROC curves from a machine learning perspective.

2.1 Robustness of ROC curves

As mentioned in the introduction, predictive accuracy might be a poor performance indicator when the problem domain suffers from skewed class distributions and involve asymmetric mis-classification costs. For instance, medical or text retrieval applications commonly present negative examples outnumbering positive examples by a factor 100, with a false positive cost (mistaking a negative example for a positive one) usually much lower than the false negative cost (missing a true positive).

Specific heuristics are devised to resist such characteristics, e.g. through oversampling the rare class, incorporating the mis-classification costs in the learning criteria, and/or relabeling the examples [Dom99].

An alternative would be to see learning as a multi-objective optimization problem (see [Deb01] and references therein), simultaneously maximizing the true positive and true negative rates. From this perspective, ROC curves simply depict the Pareto front associated to a set of hypotheses and/or learners (Fig. 1) [Bra97,PFK98,SKB99].

Three particular points in the ROC space correspond to the *All positive* hypothesis (point $(1, 1)$), *All negative* hypothesis (point $(0, 0)$), and discriminant hypotheses (point $O = (1, 0)$).

By construction this representation does not depend on the class distribution, since it uses normalized coordinates, the true positive and false positive rates.

Furthermore, this representation immediately allows one to select the best hypothesis depending on the error costs. Assuming that a false negative error costs r times more than a false positive one, the best hypothesis lies at the intersection of the ROC curve with the straight line Δ of slope $\frac{-1}{r}$ (Fig. 1).

2.2 Related work

As argued in [Bra97,MDC$^+$01,FWB$^+$98], ROC curves also allow one to deal with practical requirements on the minimal TP or maximal FP rates. For instance, the detection of churners within a given sensitivity (TP) and precision (1-FP) range was achieved in [MDC$^+$01], with a GA-based optimization of the area under the ROC curve in the desired ranges. The detection of malignant mammograms with a minimum sensitivity and precision was similarly tackled by EP-based optimization of ANN and linear hypotheses in [FWB+98].

When such desired ranges are unknown, the whole area under the ROC curve can be taken as learning criterion (AUC), with two warnings. Firstly, the AUC criterion is not "better" than the predictive accuracy [LHZ03]; rather, both criteria define distinct optimization landscapes. Secondly, the AUC criterion defines an NP complete, combinatorial optimization problem; to our best knowledge, this optimization problem was only addressed through greedy or evolutionary search, respectively learning decision trees [FFHO02], or linear hypotheses [MDC+01].

Nevertheless, optimizing the AUC criterion enforces the learning stability with respect to the misclassification costs. Learning stability is most generally desirable, for the target hypothesis should be independent as much as possible from fortuitous effects in the problem domain. To our best knowledge, learning stability has mostly been investigated in relation with the training example distribution [BE02]. But stability wrt misclassification costs is desirable as well, for two reasons. On one hand, the expert usually sets the misclassification costs by trials and errors; optimizing the ROC curve provides optimal hypotheses for various misclassification costs, which allows the expert for a more informed and better choice.

On the other hand, the appropriate misclassification costs might vary, depending on additional information on the case at hand (e.g. the "normal" range for a bio-chemical exam might depend on the age and mobility of a patient). The decision making based on a ROC curve can thus be locally adjusted depending on the case at hand.

Conversely, the use of ROC spaces offers a geometrical and intuitive representation for the behavior and dynamics of a learning strategy on a given domain [Fla03]; for instance, experimentations with different learning criteria (m-estimate, Gini criterion, entropy) offer new insights into how they trade-off the FP and TP rates [FF03].

Less related to ROC analysis, the *Learning to Order Things* approach developed by Cohen et al. [CSS98] searches for ranking hypotheses, compatible with the preferences of some experts in a Web-based and text retrieval context. Indeed, any ordering hypothesis that would rank positive examples first, would reach the optimum of the AUC criterion.

3 Genetic ROC-based learning

This section describes the *ROGER* algorithm, implementing an evolution-strategy based optimization of the AUC criterion, and discusses its limitations.

3.1 Overview

Let the data set be given as $\mathcal{E} = \{(x_i, y_i), i = 1 \ldots n,\ x_i \in X, y_i \in \{1, -1\}\}$, where X denotes the instance space. In the following, we restrict ourselves to attribute-value learning, e.g. $X = \mathbb{R}^d$.

The hypothesis space considered in the following is the set of real-valued functions on X.

For the sake of comparison, only linear functions are considered in the following; the hypothesis space $\mathcal{H}$ is set to $\mathbb{R}^d$. The extension to richer function spaces using kernel-based representations is planned for further research.

Any real-valued hypothesis h on X induces by thresholding a family of binary classifiers $\{h_t,\ t \in \mathbb{R}\}$, with

$$h_t(x) = \begin{cases} 1 & \text{if } h(x) > t \\ -1 & \text{otherwise} \end{cases}$$

It is straightforward to see that the true positive and the false positive rates monotonically increase as t decreases: the curve defined by $(FP(h_t), TP(h_t), t \in \mathbb{R})$ is a ROC curve.

The fitness of h is set to the area under the above ROC curve, computed with complexity $n \log n$ where n is the number of examples (Table 1). Normalization is omitted as of no effect on the optimization problem.

Fitness function of hypothesis h

Input
Data set $\mathcal{E} = \{(x_i, y_i), i = 1 \ldots n,\ x_i \in X, y_i \in \{1, -1\}\}$
Hypothesis $h : X \mapsto \mathbb{R}$
Init
Sort $\mathcal{E} = \{(x_i, y_i)\}$ by decreasing order, where $i > j$
iff $(h(x_i) > h(x_j))$ or $((h(x_i) = h(x_j)$ and $(y_i > y_j))$.
$p = 0$
$\mathcal{F} = 0$
For $i = 1$ to n
if $y_i = 1$, increment p;
else $\mathcal{F} = \mathcal{F} + p$
EndFor
Return $\mathcal{F}$

Table 1. Fitness of h = Area Under the Roc Curve of h

The optimization of fitness function $\mathcal{F}$ on the search space $\mathcal{H} = \mathbb{R}^p$ is achieved using evolution strategies (ES) with self-adaptive mutation [Sch81], well suited to parametric optimization [Bäc95].

We use the $(\mu+\lambda)-ES$ selection/replacement mechanism; μ parents generate λ offspring, and the best individuals among the μ parents + λ offspring are selected as parents for the next generation.

3.2 Discussion and Limitations

Since the 90's, the use of real-valued hypotheses has been investigated in two major areas of machine learning, namely statistical learning [Vap98] and ensemble learning [SFPL97]. The efficiency of both approaches is explained from the optimization of the minimal margin (the diagnostic confidence, or distance to the discrimination threshold t).

Based on structural risk minimization, SVMs actually depend on a few selected examples, the support vectors; they achieve stable learning as they pay no attention to (the distribution of) other examples.

The difference with AUC optimization is twofold. On one hand, the AUC criterion depends on the whole example distribution. On the other hand, AUC is an order-based criterion, reputed more stable under statistical noise than real-valued criteria. In counterpart, AUC minimization defines an ill-behaved optimization landscape, as it maps a continuous search space onto a finite integer set, while structural risk defines a convex, quadratic optimization landscape.

However, AUC minimization achieves learning stability wrt misclassification cost, as desired and discussed in Sect. 2; a single real-valued hypothesis h is learned, and the misclassification cost only governs the discrimination threshold t. In opposition, modifying the misclassification cost would require to retrain SVMs and result in a different hypothesis.

In conclusion, AUC-based learning presents two main drawbacks. One, shared with SVMs, is that it does not provide an intelligible hypothesis, though experts are often willing to sacrifice some accuracy for more intelligible hypotheses. The other drawback results from the fact that AUC defines an under-specified optimization problem, admitting infinitely many solutions.

We shall see in Sect. 5 that this multiplicity of solutions can be exploited and provide the expert with some facilities for inspecting the hypotheses, "for free".

4 Comparative validation

ROGER is experimentally validated on eight problems from the Irvine repository [BKM98], and compared to a state-of-art SVM, *SVMTorch* [CB01].

4.1 Experimental setting

On each problem, the dataset is splitted into a training set and test set with same class distribution as the global dataset, and 11 independent splits are considered. The split is done with 2/3 of the data in the training set, and 1/3 in the test set; in some cases (see Table 3), the size of the training set has been reduced such that *SVMTorch* learning computational cost be less than 15 minutes on Pentium-II, 400 MHz.

On each training set, *SVMTorch* is run with default parameters, and computes the SRM-optimal hyperplane on the training set, $h(\boldsymbol{x}) = \boldsymbol{w}.\boldsymbol{x} + b$. Hypothesis h is assessed from the ROC curve associated to $\boldsymbol{w}.\boldsymbol{x}$ on the test set.

On the same training set, *ROGER* is run 21 times, providing 21 independent solutions h to the AUC maximization problem (Table 1). Same parameters, summarized in Table 2, are used for all runs and all problems.

For each training set, we consider the median of the ROC curves on the test set, over the 21 runs. As already noted by [PFK98], the representativity of the median ROC curve is difficult to assess since different portions of the curve correspond to distinct hypotheses.

Finally, we take the mean of the above medians over the 11 splits for each dataset.

population size	# parents μ	10
	# offspring λ	50
max nb evaluations		10,000
crossover	uniform	rate .6
mutation	self-adaptive	rate 1

Table 2. *ROGER* parameters

The experiment goal is to compare AUC-based learning with SVMs in terms of predictive accuracy and learning stability. At the moment, only linear hypotheses are considered; *SVMTorch* is run with a linear kernel.

4.2 Experiments

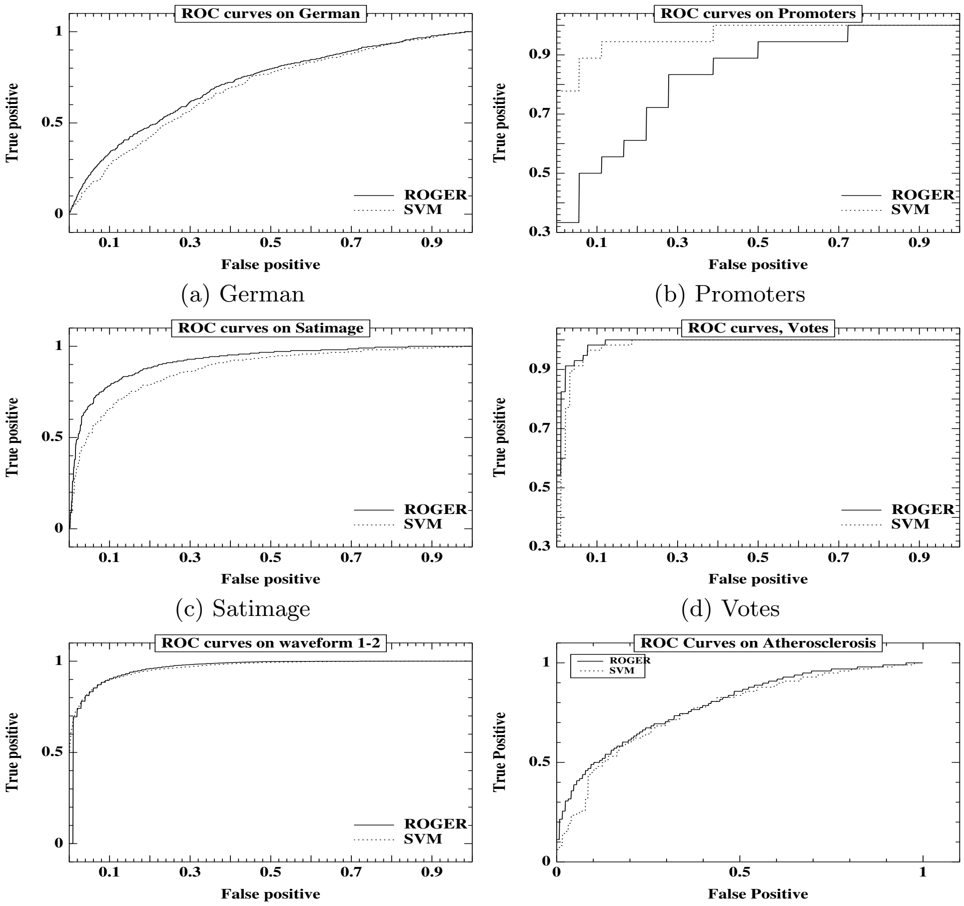
Table 3 summarizes the datasets considered, the size of the training and test sets, the initial number of attributes and the size of the hypothesis search space, being reminded that a nominal attribute with k modalities is expressed as k boolean attributes, and accounts for k coefficients in the linear hypothesis h. As already mentioned, the size of the training set was limited to restrict the

					ROGER		*SVMTorch*	
	nb att	nb att lin	#Train	#Test	AUC	time	AUC	time
Breast Cancer	9	42	189	97	.674 ± .05	7”	.672 ± .05	1”
Crx	15	47	70	620	.816 ± .06	7”	.839 ± .04	886”
German	25	25	100	900	.712 ± .03	6”	.690 ± .02	96”
Promoters	59	229	70	36	.863 ± .07	2”	.974 ± .02	< 1”
Satimage	36	36	139	1237	.918 ± .01	4”	.876 ± .02	14”
Vehicle	18	18	125	291	.994 ± .005	1”	.993 ± .007	< 1”
Votes	16	32	287	148	.993 ± .004	7”	.989 ± .005	> 1,000
Waveform 1-2	22	22	211	3321	.971 ± .004	4”	.963 ± .008	2”

Table 3. The AUC values and computational time of *ROGER* and *SVMTorch* on eight datasets from Irvine Repository

computational cost of *SVMTorch* to a maximum of 15 minutes on Pentium II, 400 MHz. The computational cost of *ROGER* is lower by one or several orders of magnitude than *SVMTorch* in the worst cases.

The average and standard deviation of the areas under the ROC curve, averaged over 11 runs for *SVMTorch* and 11 $\times$ 21 runs for *ROGER*, are reported in Table 3. Similar AUCs values are obtained. Likewise, *ROGER* and *SVMTorch* have similar predictive accuracies, as can be observed from Fig. 2. *SVMTorch* significantly outperforms *ROGER* on Promoters; *ROGER* significantly outperforms *SVMTorch* on Satimage. In most other cases, the median curves are almost indistinguishable, except sometimes in the beginning of the curve.



(a) German (b) Promoters

(c) Satimage (d) Votes

(e) Waveform (f) Atherosclerosis

Fig. 2. Comparison of the median ROC curves obtained with *SVMTorch* and *ROGER*

These experiments suggest that AUC maximization is competitive with respect to *SVMTorch*, which is much encouraging given the maturity and the strong mathematical foundations of SVMs. A fortiori, *ROGER* compare favorably to more traditional learners such as C4.5, naive Bayes, and k-NN on these same problems, after [PFK98].

The scalability of the approach with respect to the size of the dataset, in $n \log n$, is quite satisfactory. Empirically, the computational cost of AUC evolutionary minimization is much lower on average than for *SVMTorch*, with a very low standard deviation.

However, the AUC scalability with respect to the number of attributes is questionable. The worst performances are obtained for the Promoters problem, with 229 attributes. On-going experiments are underway to investigate and understand this limitation.

5 Impact Studies and Sensitivity Analysis

This section shows how the multiplicity of solutions provided by AUC evolutionary optimization can be exploited by the expert to gain some insights into the phenomenon at hand. The approach is illustrated on the PKDD 2002 Challenge dataset[1], concerned with the identification of risk factors for atherosclerosis and cardio-vascular diseases (CVD).

5.1 The data and learning goal

Two databases have been made publicly available for the PKDD2002 Challenge. The Entry database describes the personal and family case for 1419 middle aged men.

This database involves 219 attributes, which have been manually compressed into 28 boolean and numerical attributes [LAS02].

The Control database presents the longitudinal study over 20 years of a sample of men. Using the medical expertise of the third author, this sample was divided into three classes, depending whether their health after 20 years is good, bad, or other (the later class includes in particular all men who disappeared from the study)[2].

5.2 Experimental setting

The goal is to predict from the individual description given in the Entry database, his health state after twenty years.

The dataset is splitted into a 2/3 training set and 1/3 test set with same class distribution as the global dataset, and 11 independent splits are considered.

On each training set, *SVMTorch* is run with default parameters, and the optimal hypotheses are again evaluated from their median ROC curve.

[1] http://lisp.vse.cz/challenge/ecmlpkdd2002/

[2] The prepared dataset is available at `http://www.lri.fr/~ aze/PKDD2002/`.

On each training set, 21 independent *ROGER* runs are launched with parameters given in Table 2). The median ROC curves over 21 runs are averaged over the 11 splits of the data, and the mean ROC curve is displayed in Fig. 2.(f).

ROGER shows good performances, with an average AUC of .79 ±.012 to be compared with .76 ±.045 for *SVMTorch*.

Interestingly, the main difference between the two curves occurs close to the origin. It appears that some negative examples are classified as positive with high confidence by *SVMTorch*. Indeed, SVMs make no difference between misclassified examples provided that their confidence is above the cost threshold; and one would not increase the cost threshold too much, as this would increase the sensitivity to noise of the algorithm. In contrast, the AUC criterion offers a finer-grained evaluation of mis classifications, as the cost of an error actually depends on its rank; improving the example order, even in extreme regions, is rewarded. Accordingly, the *True Positive* rate increases abruptly at the beginning of the *ROGER* curve: individuals classified as the most at risk are on average in bad shape. In medical terms, the sensibility of the *ROGER* hypothesis is better compared to *SVMTorch* on this problem.

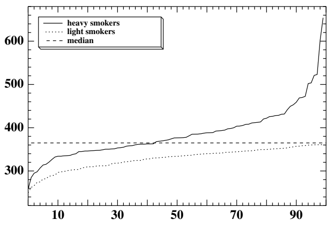
5.3 Impact Studies

A well known limitation of SVMs (also incurred by *ROGER*) is that it does not provide an easy-to-read hypothesis. An alternative to the analytic inspection of hypotheses is offered by diagrammatic representations, as investigated in Visual Data Mining [CMS99]. Along these lines, we explore some graphical interpretations of the *ROGER* hypotheses.

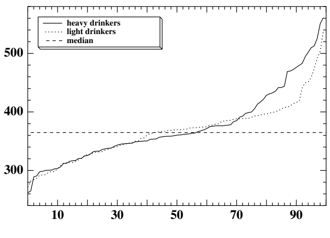
A first graphical exploitation concerns the impact study, analyzing the contribution of a given feature on the concept under examination; classically, this contribution is measured using correlation, chi-square or entropy.

However, *ROGER* hypotheses (and more generally, any ordered hypothesis) provide a more detailed, intuitive and yet precise picture, about the contribution of a feature (attribute, function of attributes). As an example, let us investigate the impacts of the tobacco and alcohol intoxication on atherosclerosis risk factors.

These impacts are graphically assessed, using the following protocol.



(a) Tobacco



(b) Alcohol

Fig. 3. Tobacco and Alcohol Impacts on Atherosclerosis Risks

For each feature (here, an attribute), the 10% individuals in the test set with maximal (resp. minimal) values for this attribute, are considered. In both subsets, the individuals are ranked by increasing value of h, and the curves $(i, h(x_i))$ are displayed.

Each curve shows globally the risk range for the individuals with high (resp. low) intoxication (though the risk might be due to other factors, correlated with the intoxication). It is believed that such curves convey a lot more information than the correlation factor or quantity of information.

Furthermore, they allow for an intuitive comparison of the factors, by superposing the curves. For instance, the impact of tobacco can be argued from the fact that the non-smoking individuals all lie in the better half of the population (their risk is less than the median risk). The heavy smoker risk is always higher than for non-smokers; 2/3 of the heavy smokers show an above-average risk and the risk rises sharply for the worst 20% of the heavy smokers.

The apparently lesser impact of alcohol must be taken with care. On one hand, it is true that a small amount of red wine was found beneficial against some CVD. On the other hand, it appeared from the data that the men considered "light drinkers"... were not drinking so lightly.

5.4 Sensitivity Analysis

The multiplicity of optimal solutions for the AUC criterion and/or the variability of stochastic optimization, can also be exploited for sensitivity analysis.

Let us represent a model h as a curve i, w_i, where i stands for the index of the attribute and w_i is the associated weight. Fig. 4 displays 21 models learned from the total dataset, showing that some attributes play a major role for the target concept (typically the tobacco factor, attribute 9). Conversely, some other attributes can be considered as weakly relevant at best. Last, the inspection of the curves suggests that some attributes might be inversely correlated, hinting at the creation of compound attributes.

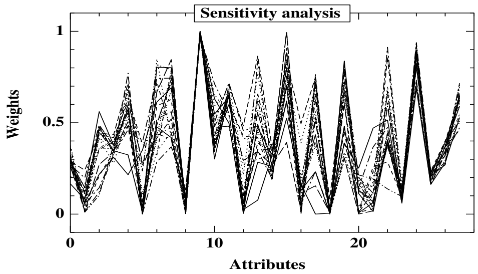


Fig. 4. Sensitivity analysis

6 Conclusion and Perspectives

This paper investigates a recent learning criterion, the maximization of the area under the ROC curve. A simple ES-based maximization of this criterion appears to be competitive with well-founded statistical learning algorithms, SVMs [Vap98].

The real-valued nature of the hypotheses allows for visual impact studies, assessing the contribution of any attribute to the concept at hand; as shown in Sect. 5, such visual representations provide much richer information than a correlation factor.

Moreover, the intrinsic variability of evolutionary results can be exploited to provide "for free" a sensitivity analysis.

Further research is concerned with extending *ROGER* to more complex instance and hypothesis languages, using for instance kernel representations. In parallel, the sensitivity analysis will be exploited for feature selection and construction.

Acknowledgment Our thanks go to Dr Maria Temeckova and R. Collobert, for providing very valuable free data and algorithmic resources.

We also warmly thank M.-C. Jaulent, Dr. I. Colombet, Dr. F. Gueyffier and Pr. G. Chatellier, for strong multidisciplinary interactions.

References

[Bäc95] T. Bäck. *Evolutionary Algorithms in theory and practice.* New-York:Oxford University Press, 1995.

[BE02] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[BKM98] C. Blake, E. Keogh, and C.J. Merz. *UCI Repository of machine learning databases.* http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[Bra97] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 1997.

[CB01] R. Collobert and S. Bengio. Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[CMS99] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Information Visualization: Using vision to think.* Morgan Kaufmann, 1999.

[CSS98] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. In *Advances in Neural Information Processing Systems*, volume 10. The MIT Press, 1998.

[CST00] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines.* Cambridge University Press, 2000.

[Deb01] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms.* John Wiley, 2001.

[Die98] T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 1998.

[Dom99] P. Domingos. Meta-cost: A general method for making classifiers cost sensitive. In *Knowledge Discovery from Databases*, pages 155–164. Morgan Kaufmann, 1999.

[FF03] J. Furnkranz and P. Flach. An analysis of rule evaluation metrics. In *Proc. of the 20^{th} Int. Conf. on Machine Learning*. Morgan Kaufmann, 2003.

[FFHO02] C. Ferri, P. A. Flach, and J. Hernández-Orallo. Learning decision trees using the area under the ROC curve. In Morgan Kaufmann, editor, *Proceedings of the 19^{th} International Conference on Machine Learning*, pages 179–186, 2002.

[Fla03] P. Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proc. of the 20^{th} Int. Conf. on Machine Learning*. Morgan Kaufmann, 2003.

[FWB+98] D.B. Fogel, E.C. Wasson, E.M. Boughton, V.W. Porto, and P.J. Angeline. Linear and neural models for classifying breast cancer. *IEEE Trans. Medical Imaging*, 17:3:485–488, 1998.

[LAS02] N. Lucas, J. Azé, and M. Sebag. Atherosclerosis risk identification and visual analysis. In *Discovery Challenge ECML-PKDD 2002*. http://lisp.vse.cz/challenge/ecmlpkdd2002/, 2002.

[LHZ03] C.X. Ling, J. Hunag, and H. Zhang. AUC: a better measure than accuracy in comparing learning algorithms. In *Proc. of 16th Canadian Conference on AI 2003*, 2003. to appear.

[MDC+01] M.C. Mozer, R. Dodier, M. C. Colagrosso, C. Guerra-Salcedo, and R. Wolniewicz. Prodding the ROC curve: Constrained optimization of classifier performance. In *Advances in Neural Information Processing Systems*, volume 13. The MIT Press, 2001.

[PFK98] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing classifiers. In *Proc. of the 15^{th} Int. Conf. on Machine Learning*, pages 445–553. Morgan Kaufmann, 1998.

[SBS98] B. Schölkopf, C. Burgess, and A. Smola. *Advances in Kernel Methods*. MIT Press, 1998.

[Sch81] H.-P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.

[SFPL97] R.E. Shapire, Y. Freund, P.Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proc. of the 14^{th} Int. Conf. on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997.

[SKB99] A. Srinivasan, R.D. King, and D.W. Bristol. An assessment of submissions made to the Predictive Toxicology Evaluation Challenge. In *Proc. of Int. Joint Conf. on Artificial Intelligence, IJCAI'99*, pages 270–275. Morgan Kaufmann, 1999.

[Vap98] V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.