Impact Studies and Sensitivity Analysis in Medical Data Mining with ROC-based Genetic Learning

Michèle Sebag Jérôme Azé Noël Lucas
PCRI, CNRS UMR 86-23, Université Paris-Sud Orsay, 91405 France
{ Michele.Sebag, Jerome.Aze, Noel.Lucas} @lri.fr

Abstract

ROC curves have been used for a fair comparison of machine learning algorithms since the late 90s. Accordingly, the area under the ROC curve (AUC) is nowadays considered a relevant learning criterion, accommodating imbalanced data, misclassification costs and noisy data.

This paper shows how a genetic algorithm-based optimization of the AUC criterion can be exploited for impact studies and sensitivity analysis.

The approach is illustrated on the Atherosclerosis Identification problem, PKDD 2002 Challenge.

1 Introduction

A novel criterion has been proposed in the late 90's for evaluating and comparing classifiers, based on the Receiver Operating Characteristics (ROC) curve [2, 11, 9].

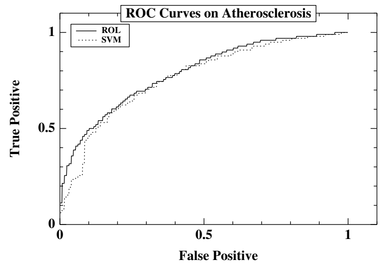

Figure 1. ROC Curves of *ROGER* and SVM-Torch on Atherosclerosis Identification

ROC curves, originated from signal processing and popularized from Medical Data Analysis, represent the tradeoff between the true positive rate *vs* the false positive rate over the (interpretations of) hypotheses (Fig. 1). Though ROC curves have been extended to multi-class concepts [8], only binary concepts will be considered in the rest of the paper.

It comes naturally to define the best hypothesis, as the hypothesis with maximal area under the ROC curve (AUC). Note that the optimization of the AUC criterion is a mixed, continuous and combinatorial, optimization problem. This optimization problem has been tackled using greedy search, learning decision trees [4], and genetic algorithms, learning neural nets [7] or linear hypotheses [12].

This paper focuses on providing the expert with Impact Study and Sensitivity Analysis facilities, describing how the various factors/descriptors of the problem contribute to the concept at hand. The point is to build a easy-to-interpret and yet precise picture of the factor impacts, as a first step toward Visual Data Mining.

To this aim, we exploit a genetic ROC-based learner named *ROGER*, first described in [12], which is briefly summarized in Section 2 for the sake of self-containedness. The Impact Study facilities are illustrated in Section 3, on the Atherosclerosis Identification problem presented for PKDD Challenge 2002, showing the influence of the tobacco and alcohol factors on the risk of atherosclerosis and cardio-vascular diseases.

The paper ends with some perspectives for further research. The approach is believed to be particularly appropriate to Medical Data Mining for two reasons. On one hand, by construction ROC curves do not suffer from difficulties such as imbalanced class distributions, noisy examples, and cost-sensitive problems, which are commonly encountered in medical domains (and in other domains as well, see [6]). On the other hand, medical experts are used to handle and interpret ROC curves.

2 Learning with AUC

This section presents the *ROGER* algorithm, referring the interested reader to [12] for more details.

Let the dataset be noted $\mathcal{E} = \{(x_i, y_i), i = 1 \ldots n,\ x_i \in X, y_i \in \{1, -1\}\}$, where X denotes the instance space.

The hypothesis space $\mathcal{H}$ considered in the following, inspired from margin-oriented learners such as Support Vector Machines [14] or bagging [3] and boosting [13], is the set of real-valued hypotheses, mapping the instance space X onto $\mathbb{R}$. Each hypothesis h in $\mathcal{H}$ induces by thresholding a family of binary classifiers $\{h_t,\ t \in \mathbb{R}\}$, with $h_t(x) = 1$ if $h(x) > t$ and -1 otherwise.

It is straightforward to see that the true positive and the false positive rates of h_t monotonically increase as t decreases. Let us define the performance of the real-valued hypothesis h as the area under the ROC curve (AUC) associated to the set of binary classifiers h_t.

Learning can thus be achieved by optimizing the AUC criterion, which is a mixed, continuous and combinatorial optimization problem as the ROC curve depends on the order induced by h on the examples.

This optimization problem is therefore tackled using evolutionary computation techniques, which are population-based stochastic optimization algorithms, crudely inspired from the natural evolution of biological populations and Darwinian principles [1].

Assuming the reader's familiarity with canonical GAs, we only detail *ROGER* specificities for the sake of reproducibility.

The fitness function computing the AUC criterion is described in Table 1 (normalization is omitted as of no effect on the optimization problem).

Fitness function of hypothesis h

```
Input
  Data set E = {(x_i, y_i), i = 1...n, x_i ∈ X, y_i ∈ {1, −1}}
  Hypothesis h : X ↦ R
Init
  Sort 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
  F = 0
For i = 1 to n
  if y_i = 1, increment p;
  else F = F + p
EndFor
Return F
```

Table 1. Area Under the Roc Curve of h

At the moment, *ROGER* is restricted to attribut-value languages and linear hypotheses. Its extension to more complex instance and hypothesis languages is a perspective for further research. Formally, the instance space X is set to $\mathbb{R}^d$, where any nominal attribute with k modalities is handled as k boolean attributes. The hypothesis search space $\mathcal{H}$ is similarly set to $\mathbb{R}^d$; hypothesis h, a weighted sum of the attributes, is a vector in $\mathbb{R}^d$.

The optimisation of the fitness function $\mathcal{F}$ on $\mathcal{H} = \mathbb{R}^d$ is performed using evolution strategies (ES) [1]. The main difference compared to canonical GAs regards the mutation operator, specifically tailored to numerical optimization. Mutation, used with probability 1, perturbs the i-th component of individual h by addition of a Gaussian noise $\mathcal{N}(0, \sigma_i)$. We used self-adaptive mutation, where the genetic material of the individual incorporates the standard deviations σs (the individual thus belongs to $\mathbb{R}^{2d}$); this way, evolution automatically adjusts the mutation amplitude along generations, for each individual. The reader is referred to [1] for an exhaustive presentation of self adaptive mutation.

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.

For space limitations, the interested reader is referred to [12] for a comparative validation on benchmark problems from the Irvine repository; compared to Support Vector Machines, *ROGER* show similar performances within a fraction of the computational effort.

3 Impact Studies and Sensitivity Analysis

This section focuses on exploiting the by-products of *ROGER* for a visual inspection of the data, particularly the impact of various factors on the concept at hand. The approach is illustrated on the PKDD 2002 Challenge[1], concerned with the identification of risk factors for atherosclerosis and cardio-vascular diseases (CVD).

3.1 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. These men are classified into three groups; the pathologic group includes all men with manifested cardio-vascular or other malignant diseases; other men are respectively classified in the normal or high risk group, depending on their (family-related or personal) risk factors.

This database involves 219 attributes, most of which are sparingly informed (e.g. a Myocardial Infarction for the patient's Fourth Sister appears 4 times in the whole dataset). The data preparation, summarizing the family case, toxicological (alcohol and tobacco) ingestion, and bio-chemical exams into 28 boolean and numerical attributes, has been described in detail in [10].

The Control database presents the longitudinal study over 20 years of a sample of men belonging to the normal and high risk groups.

3.2 Experimental setting

Using the medical expertise of the third author, all men were divided into three classes according to their health

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

state at the end of the medical campaign: healthy, ill, and other (the later includes all men not considered in the Control database, and those who disappeared from the study)[2].

Eventually 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, the SVMTorch software with default parameters determines the SRM-optimal separating hyperplane $h(\vec{x}) = \vec{w}.\vec{x} + b$, again evaluated from the associated ROC curve.

On each training set, 21 independent *ROGER* runs are launched (with same parameters as for all benchmark problems, given in Table 2). The ROC curve of the best hypothesis extracted by *ROGER* is evaluated on the test set, and the median of all ROC curves is displayed in Fig 1. A baseline reference is given by the state-of-the art SVMTorch algorithm, with default parameters and linear kernel. The separating hyperplane $h(\vec{x}) = \vec{w}.\vec{x}+b$ is also evaluated and represented from its median ROC curve on the test sets (Fig. 1).

As already noted by [11], the representativity of the median ROC curve is difficult to assess since different portions of the curve correspond to distinct hypotheses. However, the general trend is confirmed from the comparison of the min and max ROC curves obtained for both algorithms.

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**

ROGER shows good performances, with an average AUC of .79 $\pm$.012 to be compared with .76 $\pm$.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 misclassifications, 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.

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

3.3 Impact Studies

A well known limitation of SVMs (also incurred by *ROGER*) is that it does not provide an easy-to-read hypothesis, although in many domains the expert needs a readable hypothesis about the concept at hand even more than an accurate predictor.

An alternative to the analytic inspection of hypotheses is offered by diagrammatic representations, as investigated in Visual Data Mining [5]. 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; this contribution is classically measured by its correlation with the concept or its quantity of information.

It is shown that *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.

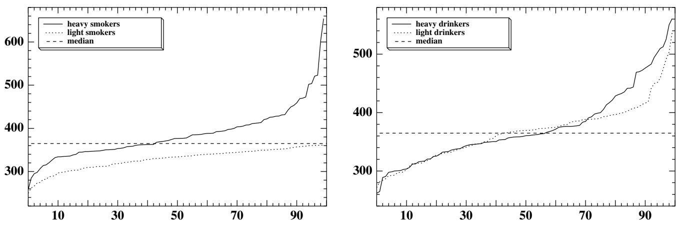

Figure 2. Comparative influence of Tobacco and Alcohol on Atherosclerosis

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, for instance by superposing the curves. For instance, the impact of tobacco can be argumented 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, the population considered as "light drinker"... was not.

3.4 Sensitivity Analysis

The multiplicity of optimal solutions for the AUC criterion and/or the variability of stochastic optimization, can also be exploited for feature selection and more generally 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. 3 displays the 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.

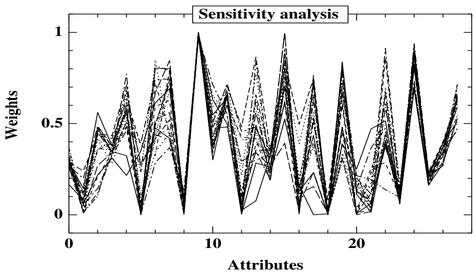

Figure 3. Sensitivity analysis

4 Conclusion and Perspectives

This paper presents some exploitations of the genetic ROC-based learner *ROGER*, allowing for a visual inspection of the hypotheses. The inspection is at the moment limited to the impact studies and sensitivity analysis. It has been argumented that such a visual inspection provides a harvest of precise information in a compact and readable way.

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 their valuable data and algorithm free resources.

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

References

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

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

[3] L. Breiman. Arcing classifiers. *Annals of Statistics*, 26(3):801–845, 1998.

[4] P. Flach, C. Ferri and J. Hernández-Orallo. Learning decision trees using the area under the ROC curve. In *Proc. of the 19th Int. Conf. on Machine Learning*, pages 179–186, Morgan Kaufmann, 2002.

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

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

[7] 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.

[8] D. Hand and R. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.

[9] C. 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.

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

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

[12] M. Sebag, J. Azé, and N. Lucas. Roc-based evolutionary learning. application to medical data mining. In P. Liardet et al., editors, *Artificial Evolution'03*. Springer Verlag, 2004.

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

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