Bagging Evolutionary ROC-based Hypotheses Application to Terminology Extraction

Jérôme Azé AZE@LRI.FR
Mathieu Roche ROCHE@LRI.FR
Michèle Sebag SEBAG@LRI.FR
LRI CNRS UMR8623
Université Paris Sud
91450 Orsay Cedex France

Abstract

The claim of the paper is that Evolutionary Learning is a source of diverse hypotheses "for free", and this specificity can be used to combine in an ensemble the hypotheses learned in independent runs. The aim of our algorithm named BROGER (Bagging-ROC GEnetic LEarneR) consists of optimizing the Area Under the ROC Curve using Evolutionary Learning. This paper first presents the theoretical framework of BROGER and then its application to a Term Extraction task in Text Mining.

1. Introduction

The Area Under the ROC curve (AUC) has been used as a learning criterion in many works (Mozer et al., 2001, Ferri et al., 2002, Rosset, 2004, Cortes & Mohri, 2005) since Bradley's seminal paper (Bradley, 1997).

Although this criterion, shown equivalent to the Wilcoxon ranking test, offers good stability properties after the theoretical and experimental studies conducted by (Rosset, 2004), it suffers from the fact that it induces an ill-posed learning optimisation problem.

This learning problem has mostly been tackled using greedy optimization, for learning decision trees (Ferri et al., 2002), or evolutionary optimization (Goldberg, 1989, Bäck, 1995, Goldberg, 2002), for learning Neural Nets (Fogel, 1998) or linear hypotheses (Mozer et al., 2001, Sebag et al., 2003b, Sebag et al., 2003a). Other attempts have been done to convert the underlying (discontinuous) optimisation problem into a continuous one, amenable to gradient-based optimisation (Herschtal & Raskutti, 2004, Rakotomamonjy, 2004).

This paper, inspired by Ensemble Learning (Breiman, 1998, Freund et al., 2003), investigates the free source of diverse hypotheses offered by evolutionary AUC-based learning.

Specifically, the aggregation (bagging) of the diverse hypotheses produced along independent evolutionary learning runs (with different random seeds, everything else being equal) is defined according to the AUC criterion.

The approach extends earlier work, devoted to evolutionary learning of linear hypotheses (Sebag et al., 2003b, Sebag et al., 2003a) or (a restricted form of) non-linear hypotheses (Jong et al., 2004) optimizing the AUC criterion.

Based on the above algorithm named ROGER (*ROc-based GEnetic LearneR*), the BROGER (*Bagging* ROGER) algorithm outputs the aggregation of the T hypotheses constructed along T independent runs of ROGER.

The comparative validation of BROGER is conducted on two Terminology extraction applications (involving different domains, biology *vs* human resources) and different languages, English *vs* French). Terminology Extraction, a key step in Text Mining (Bourigault & Jacquemin, 1999, Daille, 1996), is concerned with extracting the relevant collocations of tagged words. Given a few collocations, manually labelled as relevant/irrelevant, terminology extraction can be formalized as a classification problem (Vivaldi et al., 2001).

The AUC-based approach can be viewed as particularly relevant to terminology extraction. On one hand, terminology extraction can be viewed as a rank-

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

ing problem (term t_1 is more relevant than term t_2), rather than a classification problem (term t_1 is relevant/irrelevant). On the other hand, the Wilcoxon ranking test, equivalent to the AUC, is naturally suited to the evaluation of ranking functions.

The experimental study, comparing BROGER with standard Support Vector Machines (Collobert & Bengio, 2001) and the classical statistical relevance measures used in terminology extraction (see Section 4), shows a significant edge of the proposed approach.

This paper is organized as follows. Section 2 briefly introduces and discusses the AUC criterion. For the sake of self-containedness, section 3 describes the ROGER algorithm, first presented in (Sebag et al., 2003a) and extended in (Jong et al., 2004).

2. The Area Under the ROC Curve

This section describes and discusses the state-of-the-art in ROC-based learning.

2.1. The ROC Curve

The ROC (Receiver Operating Characteristics) curve (Jin et al., 2003), intensively used in medical data analysis, shows the trade-off between the true positive rate (the fraction of positive examples that are correctly classified, aka recall) and the false positive rate (the fraction of negative examples that are misclassified) achieved by a given hypothesis/classifier/learning algorithm. Therefore, the Area Under the ROC curve (AUC) does not depend on the imbalance of the training set (Kolcz et al., 2003), as opposed to other measures such as F score (Caruana & Niculescu-Mizil, 2004). The ROC curve also shows the misclassification rates achieved depending on the error cost coefficients (Domingos, 1999). For these reasons, (Bradley, 1997) argues the comparison of the ROC curves attached to two learning algorithms to be more fair and informative, than comparing their misclassification rates only.

2.2. Wilcoxon ranking test

Using the standard notations for binary concept learning, the dataset $\mathcal{E}$ is noted as:

$$\mathcal{E} = \{(\mathbf{x}_i, y_i), i = 1..n, \mathbf{x}_i \in X, y_i \in Y = \{-1, +1\}\}$$

As shown in (Jin et al., 2003), the Area Under the ROC Curve is equivalent to the Wilcoxon ranking test, measuring the probability that a hypothesis h ranks $\mathbf{x}_i$ lower than $\mathbf{x}_j$ when $\mathbf{x}_i$ is a positive and $\mathbf{x}_j$ is a negative example:

$$\mathcal{W}(h) = Pr(h(x_i) < h(x_j) \mid y_i > y_j) \qquad (1)$$

This criterion, with quadratic complexity in the number n of examples[1] offers an increased stability compared to the misclassification rate ($Pr(h(x_i).y_i < 0)$, with linear complexity in n); see (Rosset, 2004) and references therein.

Another tentative explanation for the good behavior of the above criterion, is that its formulation is equally suited to binary classification or regression problems.

2.3. AUC-based learning

Accordingly, the Area Under the ROC curve defines a new learning criterion, used e.g. for the evolutionary optimization of neural nets (Fogel, 1998), or the greedy search of decision trees (Ferri et al., 2002).

It must be noted that the AUC optimization problem is intrinsically ill posed, in the following sense.

Consider the space of linear continuous hypotheses $\mathcal{H} = \mathbb{R}^d$. The AUC criterion maps this continuous space onto a finite number of values, at most $n!$ if n is the number of examples: two hypotheses inducing the same ranking on the training set have the same AUC. Confidence intervals on AUC values have been proposed accordingly by (Cortes & Mohri, 2005).

In other words, the fitness landscape defined by the AUC criterion is made of a collection of plateaus, as AUC(h) is almost surely continuous wrt h coefficients, and discontinuities occur when some positive examples outpass a negative example, and *vice versa*.

3. Overview of BROGER

For the sake of self containedness, this section first recalls the ROGER algorithm, before describing the bagging of evolutionary ROC-based hypotheses, achieved in BROGER. In the remainder of the paper, the instance space X is that of d-dimensional real-valued vectors $X = \mathbb{R}^d$.

3.1. ROGER

ROGER implements the optimisation of the AUC criterion on the hypothesis search space $\mathcal{H}$. Two types of hypothesis space have been considered: linear hypotheses (Sebag et al., 2003b, Sebag et al., 2003a, Roche et al., 2004a); and a restricted form on non-linear hypotheses (Jong et al., 2004).

Specifically, ROGER uses an evolution strategy; the interested reader is referred to (Bäck, 1995) for a

[1] Actually, the computational complexity is in $\mathcal{O}(n \log n)$ since $\mathcal{W}(h)$ is proportional to the sum of ranks of the positive examples.

comprehensive presentation. Evolution strategies are among the evolutionary computation algorithms best suited to continuous optimization, due to the specific variation operators developed for evolution of real-valued genotypes (e.g. self adaptive mutation and extended versions thereof (Auger et al., 2004)).

In the rest of the paper, ROGER employs a $(\mu+\lambda)$-ES, involving the generation of λ offspring from μ parents through uniform crossover and self-adaptive mutation, and deterministically selecting the next μ parents from the best μ parents + λ offspring.

In a first step (Sebag et al., 2003b), the search space $\mathcal{H}$ considered is that of linear hypotheses ($\mathcal{H} = \mathbb{R}^d$, where d is the number of real-valued or boolean features). To genotype w (vector in $\mathbb{R}^d$) is associated the phenotype h_w, hypothesis defined on the instance space $X = \mathbb{R}^d$ as:

$$h_w(x) = \ < w, x >$$

Hypothesis h_w defines an order on the training set $\mathcal{E}$, which is evaluated after the Wilcoxon rank test (Eq. 1): the fitness of h_w is given as:

$$\begin{aligned} \mathcal{F}(h_w) \ &= \frac{\#\{(i,j) \ s.t. \ (\ (h_w(x_i)>h_w(x_j))\wedge(y_i>y_j)\ )\}}{\#\{(i,j) \ s.t. \ (y_i>y_j)\}} \\ &\propto \sum_{y_i>0} rank(h_w(x_i)) \end{aligned} \tag{2}$$

3.2. Non linear hypotheses

Linear hypotheses are intrinsically ill-suited to many application domains, particularly pertaining to preference learning and medical domains, where *more* does not mean *better*; e.g., physiological indicators such as blood pressure, should neither be too high nor too low for a good health state; the gastronomic taste requires enough but not too much salt.

For this reason, a limited kind of non-linear hypotheses was considered in (Jong et al., 2004, Roche et al., 2004a). The new hypothesis search space $\mathcal{H}$ achieves a tradeoff between the non-linear function complexity, and the computational complexity, as follows.

A genotype individual is made of a pair (w, c) (in $\mathbb{R}^{2d}$), where $w \in \mathbb{R}^d$ is a weight vector as in the linear case, and c is a center point in $\mathbb{R}^d$. The associated phenotype $h_{w,c}$ is defined on the instance space X as the weighted L_1 distance between the current example $x = (x_1, \ldots, x_d)$ and the center $c = (c_1, \ldots, c_d)$

$$h_{w,c}(x) = \sum_{i=1}^{d} w_i |x_i - c_i| \tag{3}$$

Likewise, $h_{w,c}$ defines an order on the training set, and the fitness $\mathcal{F}(h_{w,c})$ is defined as above.

It must be noted that this representation allows ROGER for searching (a limited kind of) non linear hypotheses, by (only) doubling the size of the linear search space. Previous work has shown that non-linear ROGER significantly outperforms linear ROGER for some text mining applications (Roche et al., 2004a).

3.3. Ensemble for free with Evolutionary Learning

In this paper, inspired from (Breiman, 1998, Imamura et al., 2002), the free source of diversity offered by Evolutionary Computation is exploited to construct ensemble ranking functions.

Ensemble learning, one of the prominent domains of Machine Learning since (Schapire, 1990), includes the bagging of independently learned hypotheses (Breiman, 1998), and the boosting of sequentially learned hypotheses (Freund et al., 2003).

The BROGER algorithm aggregates the ranking functions $h_1, \ldots, h_T$ learned by T independent ROGER runs (with different random seeds, everything else being equal).

Several aggregation procedures have been considered: to each example x, BROGER associates the average or median value in $\{h_1(x), \ldots, h_T(x)\}$ (after normalization of the h_is); or its average or median rank. Interestingly, preliminary experiments have shown no significant difference between the above aggregation procedures, with respect to the AUC criterion of the bagged hypothesis (measured by cross validation).

In the rest of the paper, BROGER associates to example x the median value in $\{h_1(x), \ldots, h_T(x)\}$ (Alg. 1).

Algorithm 1 BROGER

Require:
$h_1, \ldots, h_T$: T hypotheses
x: example
Ensure:
$H(x) = \{\}$
Begin
for $(t \leftarrow 1; t \leq T; t++)$ **do**
compute $h_t(x)$
$H(x) = H(x) \bigcup \{h_t(x)\}$
end for
$H'(x) = sort(H(x))$
Return the median value in $H'(x)$
End

4. Application to Terminology Extraction

Besides the known difficulties of Data Mining, Text Mining presents specific difficulties due to the structure of natural language. In particular, the polysemy and synonymy effects are dealt with by constructing ontologies or terminologies (Bourigault & Jacquemin, 1999), structuring the words and their meanings in the domain application. Terminology Extraction (TE), a preliminary step for ontology construction, is concerned with extracting the relevant terms, or word collocations, attached to the expert's concepts (Bourigault & Jacquemin, 1999, Smadja, 1993).

Terminology extraction can be formalized as a classification problem (Vivaldi et al., 2001); it can also be formalized as a ranking problem (Cohen et al., 1999).

The literature presents a variety of *a priori* ranking criteria, mostly based on statistical measures about the word occurrences (see e.g., (Daille et al., 1998, Xu et al., 2002, Roche et al., 2004b)). (Vivaldi et al., 2001) used three of these criteria as term features; an extended set of features is used in the following. Specifically, a term example is described as a vector of values computed after the following 13 statistical measures:

- Mutual Information (MI) (Church & Hanks, 1990)
- Mutual Information with cube (MI^3) (Daille et al., 1998)
- Dice Coefficient ($Dice$) (Smadja et al., 1996)
- Log-likelihood (L) (Dunning, 1993)
- Number of occurrences + Log-likelihood (Occ_L)[2] (Roche et al., 2004a)
- Association Measure (Ass) (Jacquemin, 1997)
- Sebag-Schoenauer ($SeSc$) (Sebag & Schoenauer, 1988)
- J-measure (J) (Goodman & Smyth, 1988)
- Conviction ($Conv$) (Brin et al., 1997)
- Least contradiction (LC) (Azé & Kodratoff, 2004)
- Cote multiplier (CM) (Lallich & Teytaud, 2004)
- Khi2 test used in text mining ($Khi2$) (Manning & Schütze, 1999)
- T-test used in text mining ($Ttest$) (Manning & Schütze, 1999)

[2] Occ_L is defined by ranking collocations according to their number of occurrences, and breaking the ties based on the term Log-likelihood.

5. Goals of Experiments and Experimental Setting

The goal of experiments is twofold. On one hand, the ranking efficiency of BROGER will be assessed and compared to that of state-of-the-art supervised learning algorithms, specifically Support Vector Machines with linear, quadratic and Gaussian kernels, using SVM-Torch implementation[3] with default options.

On the other hand, the results provided by BROGER will be examined with respect to their generality and intelligibility.

The experimental setting involves 5-fold stratified cross-validation, averaged over 10 independent stratifications. On each fold, hypotheses learned by SVM and BROGER are evaluated on the test set and the corresponding ROC curves are constructed.

The ROGER parameters are as follows: $\mu = 20; \lambda = 200$; the self adaptative mutation rate is 1.0; the uniform crossover rate is 0.6.

6. Empirical validation

After describing the datasets, this section reports on the comparative performances of the algorithms, and inspects the results actually provided by BROGER.

6.1. Datasets

In both domains, the data preparation step (Roche et al., 2004b) allows for categorizing the word collocations depending on the grammatical tag of the words (e.g. Adjective, Noun).

A first corpus related to Molecular Biology involves 6119 paper abstracts in English (9,4 Mb) gathered from queries on Medline[4]. The 1028 Noun-Noun collocations occurring more than 4 times are labelled by the expert; the dataset includes a huge majority of relevant collocations.

A second corpus related to Curriculum Vitæ (CV)[5] involves 582 CVs in French (952 Kb). The "Frequent CV" dataset includes the 376 Noun-Adjective collocations with at least 3 occurrences (two hours labelling required), with a huge majority of relevant collocations. The "Infrequent CV" dataset includes the 2822 Noun-Adjective collocations occurring once or twice (two days labelling required), with a significantly different distribution of relevant/irrelevant collocations.

[3] http://www.idiap.ch/machine_learning.php?content=Torch/en_OldSVMTorch.txt

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

[5] Courtesy of the VediorBis Foundation.

Collocations	#	Relevant	Irrelevant
Molecular Biology (MB)	1028	90.9%	9.1%
Frequent CV (F-CV)	376	85.7%	14.3%
Infrequent CV (I-CV)	2822	56.6%	43.4%

Table 1. Relevant and irrelevant collocations.

Corpus	BROGER (∼ 17s/fold)	SVM (∼ 1.5s/fold) Linear	Gaussian	Quadratic
MB	0.73 ± 0.05	0.50 ± 0.08	0.46 ± 0.08	0.59 ± 0.08
F-CV	0.64 ± 0.08	0.48 ± 0.08	0.48 ± 0.08	0.50 ± 0.10
I-CV	0.73 ± 0.01	0.72 ± 0.01	0.72 ± 0.02	0.71 ± 0.02

Table 2. Ranking accuracy (Area under the ROC curve) of learning algorithms. Computational times are given for a Pentium 4 (3GHz, 512 Mb of RAM).

Table 1 presents these two corpora with details on distribution of relevant/irrelevant collocations.

6.2. Ranking accuracy

After the experimental setting described in Section 5, Table 2 compares the average AUC achieved for BROGER and SVMTorch with linear, Gaussian and quadratic kernels. On these domain applications, both supervised learning approaches significantly improve on the standalone statistical criteria (Table 3). Further, BROGER improves significantly on SVM using any kernel, excepted on the *Infrequent CV* dataset. A tentative interpretation for this result is based on the fact that this dataset is the most balanced one; SVM has some difficulties to cope with imbalanced datasets.

A more detailed picture is provided by Fig. 1, showing the ROC curve associated to SVM, BROGER and the Occ_L and J measures on the *Frequent CV* dataset on a **R**epresentative **F**old corresponding to the function the most used by BROGER (termed *RF* in this paper). Interestingly, the major differences between BROGER and the other measures are seen at the beginning of the curve, i.e. they concern the top ranked collocations. Typically, a recall (True Positive Rate) of 50% is obtained for 18% false positive with BROGER, against 23% with Occ_L, 31% with J measures and 68% for quadratic SVM[6].

In summary, BROGER improves the accuracy of the top-ranked collocations, and therefore the satisfaction and productivity of the expert if he/she only examines the top ranked terms.

6.3. Analysis of a ranking function

As shown in (Jong et al., 2004), the weights associated to distinct features by ROGER can provide some insights into the relevance of the features. Accordingly, the hypotheses constructed by BROGER are examined, focusing on the features (statistical criteria) with high weights.

[6]SVM ROC Curves is not significant as its AUC is lower than 0.5 on this test fold.

Corpus	MI	MI^3	$Dice$	L	Occ_L	Ass	J
MB	0.30	0.35	0.31	0.42	0.57	0.31	**0.59**
F-CV	0.31	0.40	0.39	0.43	**0.58**	0.32	**0.58**
I-CV	0.29	0.30	0.33	0.30	0.37	0.29	**0.50**

Corpus	$Conv$	$SeSc$	CM	LC	$Ttest$	$Khi2$
MB	0.35	0.43	0.31	0.46	0.31	0.31
F-CV	0.39	0.40	0.31	0.44	0.36	0.36
I-CV	0.40	0.39	0.30	0.45	0.30	0.30

Table 3. Ranking accuracy (Area under the ROC curve) of statistical criteria.

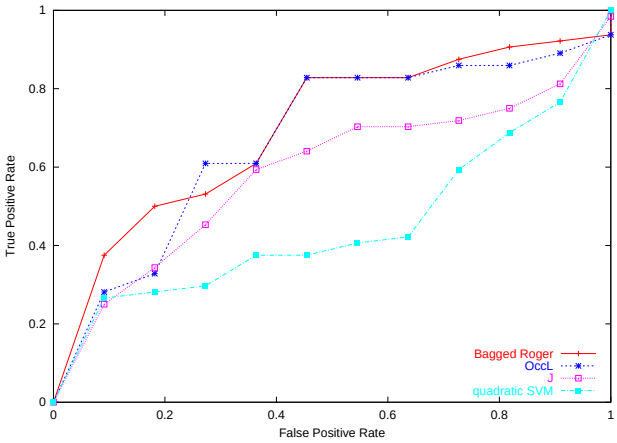

Figure 1. ROC Curves on Frequent Collocations of CV corpus (for the test set of *RF*).

As expected, ROGER detects, on the Frequent CV dataset (F-CV), that the mutual information (MI) criterion does badly (AUC(MI)= 0.31, Table 3), with a high center $c_{MI} = 0.59$ and weight $w_{MI} = 0.68$ values (collocations with high MI are less relevant, everything else being equal). Inversely, as the Occ_L criterion does well (AUC(Occ_L) = 0.58), the center $c_{Occ_L} = 0.65$ is high associated with a highly negative weight $w_{Occ_L} = -0.41$ (collocations with low Occ_L are less relevant, everything else being equal).

Although these tendencies could have been exploited by linear hypotheses, this is no longer the case for the J criterion (AUC(J) = 0.58): interestingly, the center c_J takes on a medium value, with a high negative weight w_J. This is interpreted as collocations with either *too low or too high* values of J, are less relevant everything else being equal.

Figure 2 shows the weights associated by BROGER to the 13 measures.

A more detailed analysis, including comparison of col-

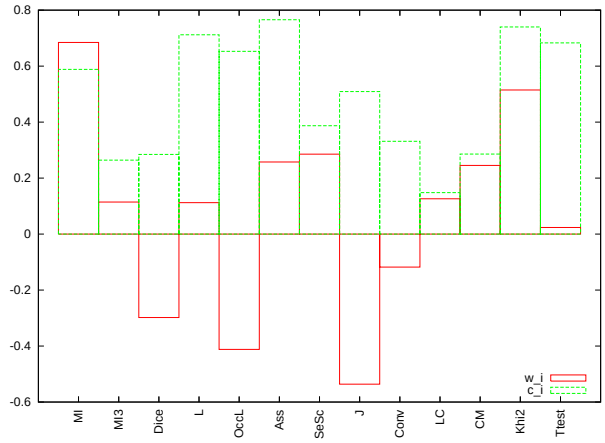

Figure 2. Weights (w_j, c_j) on the Frequent CVs (for the learn set of RF).

locations ranked according to different measures, can be find in (Azé et al., 2005).

7. Discussion and Perspectives

A first claim of the paper is that Evolutionary Learning is a source of diverse hypotheses "for free", and this specificity can be used to combine in an ensemble the hypotheses learned in independent runs. Previous attempts have been done for incorporating ensemble-based approaches in Evolutionary Computation (Iba, 1999, Imamura et al., 2002), but to our best knowledge, the simple combination of hypotheses derived from independent runs is new.

A second claim of the paper is that supervised learning can significantly contribute to the Term Extraction task in Text Mining. Based on a domain- and language- independent description of the terms along a set of standard statistical criteria, and on a few collocations manually labelled as relevant/irrelevant by the expert, a ranking hypothesis is learned.

Further research is concerned with incorporating multi-modal optimization (Deb, 2001) in AUC-based evolutionary learning, in the line of (Lee & Yao, 2004); the advantage of the approach would be to extract several different and complementary hypotheses from a single run.

Another on-going work is concerned with enriching the description of terms, e.g. adding typography-related indications (e.g. distance to the closest typographic signs) or distance to the closest Noun, possibly providing additional cues on the role of relevant collocations.

A long-term goal is to study along a variety of domain applications and expert goals, the eventual regularities associated to i) the (domain and language independent) description of the relevant collocations; ii) the ranking hypotheses.

Acknowledgment

We thank Yves Kodratoff for his knowledge and expertise on text-mining, Oriane Matte-Tailliez for her expertise and labelling of the Molecular Biology dataset, and Mary Felkin who did her best to improve the readability of this paper. The authors are partially supported by the PASCAL Network of Excellence, IST-2002-506778.

References

Auger, A., Schoenauer, M., & Vanhaecke, N. (2004). LS-CMA-ES: a Second-order algorithm for Covariance Matrix Adaptation. *Proceedings of Eighth International Conference on Parallel Problem Solving from Nature PPSN VIII* (pp. 182–191).

Azé, J., & Kodratoff, Y. (2004). Extraction de "pépites" de connaissances dans les données : une nouvelle approche et une étude de la sensibilité au bruit. *Revue RNTI, numéro spécial "Mesures de qualité pour la fouille de données"*, *E-1*, 247–270.

Azé, J., Roche, M., Kodratoff, Y., & Sebag, M. (2005). Preference Learning in Terminology Extraction: A ROC-based approach. *Proceeedings of Applied Stochastic Models and Data Analysis* (pp. 209–219). ISBN 2-908849-15-1.

Bäck, T. (1995). *Evolutionary Algorithms in Theory and Practice*. New-York:Oxford University Press.

Bourigault, D., & Jacquemin, C. (1999). Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. *Proceedings of EACL'99, Bergen.* (pp. 15–22).

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Breiman, L. (1998). Arcing Classifiers. *Annals of Statistics*, *26*, 801–845.

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. *Proceedings of ACM SIGMOD'97* (pp. 265–276).

Caruana, R., & Niculescu-Mizil, A. (2004). Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria. *Proceed-*

ings of "ROC Analysis in AI" Workshop (ECAI) (pp. 9–18).

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual information, and Lexicography. *Computational Linguistics*, *16*, 22–29.

Cohen, W., Schapire, R., & Singer, Y. (1999). Learning to Order Things. *Journal of Artificial Intelligence Research*, *10*, 243–270.

Collobert, R., & Bengio, S. (2001). SVMTorch: Support Vector Machines for Large-Scale Regression Problems. *Journal of Machine Learning Research*, *1*, 143–160.

Cortes, C., & Mohri, M. (2005). Confidence Intervals for the Area Under the ROC Curve. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, 305–312. Cambridge, MA: MIT Press.

Daille, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *P. Resnik and J. Klavans (eds). The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press* (pp. 49–66).

Daille, B., Gaussier, E., & Lang, J. (1998). An Evaluation of Statistical Scores for Word Association. *Proceedings of The Tbilisi Symposium on Logic, Language and Computation, CSLI Publications* (pp. 177–188).

Deb, K. (2001). *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, Chichester.

Domingos, P. (1999). Meta-cost: A general method for making Classifiers Cost Sensitive. *Proceedings of Knowledge Discovery from Databases* (pp. 155–164).

Dunning, T. E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*, 61–74.

Ferri, C., Flach, P., & Hernandez-Orallo, J. (2002). Learning decision trees using the area under the ROC curve. *Proceedings of ICML'02* (pp. 139–146).

Fogel, D. (1998). *Evolutionary Computing: The Fossile Record*. IEEE Press.

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research*, *4*, 933–969.

Goldberg, D. (2002). *The Design of Innovation: Lessons from and for Genetic and Evolutionary Algorithms*. MIT Press.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison Wesley.

Goodman, M., & Smyth, P. (1988). Information-theoretic rule induction. *Proceedings of ECAI'88* (pp. 357–362).

Herschtal, A., & Raskutti, B. (2004). Optimising area under the ROC curve using gradient descent. *Proceedings of ICML'04: Twenty-First International Conference on Machine Learning*. ACM Press.

Iba, H. (1999). Bagging, Boosting, and Bloating in Genetic Programming. *Proceedings GECCO'99* (pp. 1053–1060). Morgan Kaufmann.

Imamura, K., Heckendorn, R., Soule, T., & Foster, J. (2002). Abstention reduces errors; Decision abstaining N-version Genetic Programming. *Proceedings of the Genetic and Evolutionary Conference* (pp. 796–803). Morgan Kaufmann.

Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. *Mémoire d'Habilitation à Diriger des Recherches, Université de Nantes*.

Jin, R., Liu, Y., Si, L., Carbonell, J., & Hauptmann, A. (2003). A New Boosting Algorithm Using Input-Dependent Regularizer. *Proceedings of ICML 2003*. AAAI Press.

Jong, K., Mary, J., Cornuéjols, A., Marchiori, E., & Sebag, M. (2004). Ensemble Feature Ranking. *Proceedings of ECML-PKDD'04* (pp. 267–278).

Kolcz, A., Chowdhury, A., & Alspector, J. (2003). Data duplication: An Imbalance Problem ? *Workshop on Learning from Imbalanced Data Sets II (ICML)* (pp. 1–6).

Lallich, S., & Teytaud, O. (2004). Évaluation et validation de l'intérêt des règles d'association. *Revue RNTI, numéro spécial "Mesures de qualité pour la fouille de données"*, *E-1*, 193–217.

Lee, C. Y., & Yao, X. (2004). Evolutionary programming using the mutations based on the Lévy probability distribution. *IEEE Transactions on Evolutionary Computation*, *8*, 1–13.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, chapter Collocations, 165–184. Cambridge, MA: MIT Press.

Mozer, M., Dodier, R., Colagrosso, M. C., Guerra-Salcedo, C., & Wolniewicz, R. (2001). Prodding the ROC curve: Constrained optimization of classifier performance. *Proceedings NIPS-13*. MIT Press.

Rakotomamonjy, A. (2004). Quadratic programming for AUC optimization. *Proceedings of Modelling, Computation and Optimization in Information Systems and Management Sciences* (pp. 603–610). Hermès Publishing.

Roche, M., Azé, J., Kodratoff, Y., & Sebag, M. (2004a). Learning Interestingness Measures in Terminology Extraction. A ROC-based approach. *Proceedings of "ROC Analysis in AI" Workshop (ECAI)* (pp. 81–88).

Roche, M., Azé, J., Matte-Tailliez, O., & Kodratoff, Y. (2004b). Mining texts by association rules discovery in a technical corpus. *Proceedings of IIPWM'04, Springer Verlag* (pp. 89–98).

Rosset, S. (2004). Model Selection via the AUC. *Proceedings of the Twenty-First International Conference on Machine Learning (ICML'04)*.

Schapire, R. (1990). The Strength of Weak Learnability. *Machine Learning*, *5*, 197.

Sebag, M., Azé, J., & Lucas, N. (2003a). Impact studies and sensitivity analysis in medical data mining with ROC-based genetic learning. *Proceedings of ICDM 2003* (pp. 637–640).

Sebag, M., Lucas, N., & Azé, J. (2003b). ROC-based Evolutionary Learning: Application to Medical Data Mining. *Proceedings of EA 2003* (pp. 384–396).

Sebag, M., & Schoenauer, M. (1988). Generation of Rules with Certainty and Confidence Factors from Incomplete and Incoherent Learning Bases. *Proceedings of the European Knowledge Acquisition Workshop (EKAW'88)* (pp. 28–1 – 28–20).

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, *19*, 143–177.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, *22*, 1–38.

Vivaldi, J., Màrquez, L., & Rodríguez, H. (2001). Improving Term Extraction by System Combination Using Boosting. *Proceedings of ECML* (pp. 515–526).

Xu, F., Kurz, D., Piskorski, J., & Schmeier, S. (2002). A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. *Proceedings of the 3rd International Conference on Language Resources an Evaluation (LREC'02)*.