Automatic Model Selection in Cost-sensitive Boosting

Stefano Merler, Cesare Furlanello*, Barbara Larcher, and Andrea Sboner

ITC-irst,
v. Sommarive 18,
38050 Trento, Italy
Tel +39 0461 314592
fax +39 0461 302040
{merler,furlan,larcher,sboner}@itc.it
http://mpa.itc.it

Automatic Model Selection in Cost-sensitive Boosting

Abstract. This paper introduces SSTBoost, a predictive classification methodology designed to target the accuracy of a modified boosting algorithm towards required sensitivity and specificity constraints. The SSTBoost method is demonstrated in practice for the automated medical diagnosis of cancer on a set of skin lesions (42 melanomas and 110 naevi) described by geometric and colorimetric features. A cost-sensitive variant of the AdaBoost algorithm is combined with a procedure for the automatic selection of optimal cost parameters. Within each boosting step, different weights are considered for errors on false negatives and false positives, and differently updated for negatives and positives. Given only a target region in the ROC space, the method also completely automates the selection of the cost parameters ratio, tipically of uncertain definition. On the cancer diagnosis problem, SSTBoost outperformed in accuracy and stability a battery of specialized automatic systems based on different types of multiple classifier combinations and a panel of expert dermatologists. The method thus can be applied for the early diagnosis of melanoma cancer or in other problems in which an automated cost-sensitive classification is required.
KEYWORDS: cost-sensitive classification, melanoma, boosting, classification trees, automated medical diagnosis.

1 Introduction

Most of the predictive classification tools that we may expect to apply in the future for real-world applications, e.g. for automated diagnosis systems from biomedical data in a distributed setting, will have to incorporate into the learning process the appropriate cost parameters in order to drive the system towards the optimal performance in terms of sensitivity and specificity. Control of both these error measures is critical [1, 26].

However the usefulness of simply adapting a cost-sensitive mechanism within a good predictive data mining tool, in our case the Adaboost algorithm [16], is limited. Given a data set, a specific choice of the cost parameters will determine a model with a specific pair of sensitivity and specificity values, e.g. a point in the ROC space. The costs of a false negative or of a false positive in a binary medical classification problem are often estimated only approximately: as they may influence significantly the classifier accuracy one is often left with the doubt that, in order to reach the minimum specified performance, altering the costs is more effective than refining the model. At the same time, the learning procedure will also depend on the prior probabilities of the classes, thus adding training material at fixed costs may produce a model with different sensitivity and specificity.

A further complication arises whenever the classification process is split in several phases. For example, in a screening phase a high sensitivity test is required in order to recognize the highest number of positive cases, while later a more specific test (e.g. a visit by a more experienced practitioner) may be administered to these positives. It is unclear whether the cost parameters should be defined for the whole process or differently for the two tests.

What remains in any case effective is the definition of pairs of sensitivity and specificity constraints as a target of the classification process.

In this paper we discuss how to develop a good cost-sensitive classification algorithm, which is independent as much as possible from a precise definition of cost parameters and from class imbalance. We complete our methodology with a practical search procedure to get into, or as close as possible to, a target region in the sensitivity-specificity space. The aim is to wrap all of the cost-sensitive boosting learning cycle with a model selection procedure. As a cost-sensitive algorithm, we will present in this paper a variant of the AdaBoost algorithm [16]. The basic AdaBoost algorithm allows to develop systems with high accuracy, but misclassification analysis on different output classes were not originally included within the training mechanism. However, it is still possible to build a good cost-sensitive variant of AdaBoost which differently optimizes the model for the two classes. In our variant, cost-sensitive boosting is achieved by (A) weighting the model error function with separate costs for false negative and false positives errors, and (B) updating the weights differently for negatives and positives at each boosting step.

Similar approaches have been described elsewhere. In particular, a cost-sensitive variant of AdaBoost was adopted for AdaCost [14]: based on the assumption that a misclassification cost factor has to be assigned for each training data, the weights are increased in case of misclassification or decreased otherwise according to a non negative function of the costs. A different model error function than in (A) is considered, as we focus on explicit weighting in terms of sensitivity and specificity. Karakoulas and Shawe-Taylor [25], have also introduced a similar approach based on misclassification costs constant for all the samples in a class. Their procedure leads to increase the weights of false nega-

tives more than false positives and, differently from our approach, to decrease the weights of true positives more than true negatives.

We applied the procedure in a medical diagnosis task: a classification model for assisting the screening of skin lesions was developed on real data, requiring sensitivity greater than 0.95 and specificity greater than .50 . Our AdaBoost variant (SSTBoost: Sensitivity-Specificity Tuning Boosting) allowed a remarkable improvement over previous results on the same data set which had been obtained with a combination of classifiers specifically designed for the task [5]. An improvement was also found in the control of variability (standard deviation of error estimates). The combined strategy of SSTBoost resulted more effective than applying an external cost criterion to AdaBoost, as documented in [35].

The paper is organized as follows. The next Section 2 briefly introduces the classification problem which inspired our approach. The SSTBoost method is described in Section 3. The approach is evaluated on the melanoma data in Section 4. Section 5 concludes the paper.

2 Automatic Melanoma Classification

Melanoma is one of the most dangerous skin cancers. About 91% of the skin cancer deaths are due to this tumor. Its incidence is constantly increasing worldwide. The early diagnosis is the key factor for its prognosis, but the early melanoma can have a benign appearance. Digital epiluminescence microscopy (D-ELM) is a non-invasive clinical technique that allows the visualization of several colorimetric and morphological characteristics of the skin lesions, providing additional diagnostic criteria in the dermatological assessment. D-ELM is effective in increasing the diagnostic accuracy of a dermatologist, but it requires a well-trained specialist to be correctly exploited. Therefore, in order to support physicians in the melanoma diagnosis, a number of computerized systems were developed with different approaches: Shindewolf et al. [33] used a decision tree to classify images of skin lesions digitized from ELM slides. Binder et al. [2] applied an artificial neural network to classify dermatological images, using features provided by a physician. Green et al. [24] used a discriminant analysis as classification system. Ercal et al. [13] applied an artificial neural network on features extracted by photographic images with different films. Takiwaki et al. [34] used a decision tree to discriminate among the lesions. Seidenari et al. [32] applied a discriminant analysis describing the most significant features. Bishof et al. [4] used a decision tree to classify images form a D-ELM system. Binder et al. [3] used a neural network and showed that clinical data could improve the classification accuracy, especially when the system has to classify also dysplastic lesions. Dreiseitl et al. [10] compared several machine learning methods for the diagnosis of pigmented skin lesions.

Following this approach, we developed MEDS, a computerized system whose aim is to support physician in the early diagnosis of melanoma. As described in [5], the MEDS database is composed by 152 digital epiluminescence microscopy images (D-ELM) of skin lesions, acquired at the Department of Dermatology

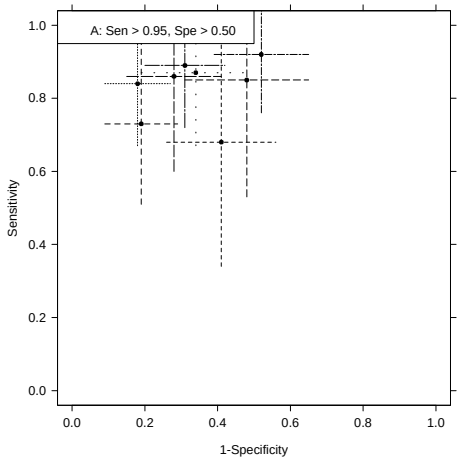


Fig. 1. Sensitivity and specificity (indicated by dots) with standard deviations (represented with the crosses) achieved by a panel of 8 dermatologists. The top-left rectangle represents the region of the ROC space defined by the constraints (sensitivity ≥ 0.95 and specificity ≥ 0.5).

of Santa Chiara Hospital, Trento. Image processing of D-ELM data produces 5 geometric-morphologic and 33 colorimetric features for each image, for a total of 38 features. All the lesions were excised and submitted to the histological analysis, whose results represent our gold standard. According to this protocol, the MEDS database includes 42 malignant lesions (melanoma: positive examples) and 110 nevi (negative examples).

In [5], different classifiers and a panel of 8 dermatologists were compared. The 152 D-ELM images were displayed randomly on a monitor and the dermatologist assessed a diagnosis for each image, reproducing a tele-dermatology setting. This procedure was repeated for every dermatologist. The performances of the panel of dermatologists are shown in Fig.1 and collected for comparison with system results in Tab.1. The variability of the physicians' performance is mainly due to their different expertise in the epiluminescence analysis.

In this paper, we simulated a model to support a screening campaign of skin lesions. In this case, a non-expert physician, i.e. a general practitioner (GP), examinates a great number of patients in order to recognize as many early malignant lesions as possible. As noted by Menzies [27], the benign/melanoma ratio of excised lesions for GPs is 30:1; this evidence supports the design of a classifier system which warrants very high levels of sensitivity at the cost of a moderate accuracy in specificity. While the ultimate decision is demanded to the physician, this strategy may limit the number of unnecessary inspections by a specialist. Hence, we simulated the development of a model for early diagnosis with sensitivity greater than 0.95 and specificity greater than 0.50.

3 The SSTBoost Cost-sensitive Procedure

3.1 AdaBoost and SSTBoost

In this section we describe first the basic Adaboost learning procedure [16]. Given a training data set $L = \{(x_i, y_i)\}$, with $i = 1, ..., N$, where the x_i are input vectors (numerical, categorical or mixed) and the y_i are class labels taking values -1 or 1, the discrete Adaboost classification model outputs the sign of an incremental linear combination of different realizations of a base classifier. Each realization is trained on a weighted version of L, and it is obtained increasing the weights for the samples currently misclassified. Alternatively, if the base model does not accept internally weights, it can be trained over weighted bootstrap versions of L. The AdaBoost procedure for a combination H of T base classifiers is summarized in Box 1.

- Given $L = \{(x_i, y_i)\}_{i=1,...,N} \subset X \times \{-1, +1\}$
- Initialize $D_1(i) = 1/N$
- For $t = 1, ..., T$:
  1. Train the base classifier h using distribution D_t.
  2. Get hypothesis $h_t : X \rightarrow \{-1, +1\}$
  3. Compute model error $\epsilon_t = \sum_i D_t(i)\Theta[y_i h_t(x_i) = -1]$ where $\Theta[P]$ returns 1 if predicate P = true, 0 otherwise.
  4. Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$
  5. Update $D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$ where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution
- Output the final hypothesis: $H(x) = sign \left(\sum_{t=1}^{T} \alpha_t h_t(x) \right)$

Box 1: The **AdaBoost** algorithm

Maximal decision trees can be considered as base classifiers h_t. Decision trees were implemented following the classic reference [6]. Using unpruned trees avoids the need of introducing the regularization metaparameter in the system; moreover, maximal trees give best results with boosting when there is enough interaction between variables, as discussed in [8, 20, 23].

The model error ϵ defined in AdaBoost (Box 1) does not differentiate the costs of misclassification for different output classes. In Box 2 we introduce a variant of AdaBoost (Sensitivity-specificity Tuning Boosting: SSTBoost) which takes into account costs at two different levels.

- Given $L = \{(x_i, y_i)\}_{i=1,...,N} \subset X \times \{-1, +1\}$
- Given cost parameter $w \in [0, 2]$
- Define $c_i = \begin{cases} w \text{ if } y_i = +1 \\ 2 - w \text{ if } y_i = -1 \end{cases}$
- Initialize $D_1(i) = 1/N$
- For $t = 1, ..., T$:
  1. Train base classifier h using distribution D_t.
  2. Get hypothesis $h_t : X \to \{-1, +1\}$
  3. Compute model error $\epsilon_t = (1 - Se)\pi_{+1}w + (1 - Sp)\pi_{-1}(2 - w)$
  4. Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$
  5. Update $D_{t+1}(i) = \begin{cases} \frac{D_t(i)e^{-\alpha_t(2-c_i)}}{Z_t} \text{ if } y_i h_t(x_i) = +1 \\ \frac{D_t(i)e^{\alpha_t c_i}}{Z_t} \text{ if } y_i h_t(x_i) = -1 \end{cases}$
     where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution
- Output the final hypothesis: $H_w(x) = sign \left(\sum_{t=1}^{T} \alpha_t h_t(x) \right)$

Box 2: The **SSTBoost** algorithm: internal learning procedures

Firstly, given class priors π_i and costs (or losses) c_i of a misclassification for class $i \in \{-1, +1\}$, we propose to consider the following weighted version of the model error:

$$\epsilon = (1 - Se)\pi_{+1}c_{+1} + (1 - Sp)\pi_{-1}c_{-1}. \quad (1)$$

As discussed in [1], rather than considering separately the values of the two c_{-1} and c_{+1}, it is more convenient to consider the cost ratio $\frac{c_{+1}}{c_{-1}}$ or to impose a constraint $c_{+1} + c_{-1} = cost$. Imbalance between classes may also play an important function, not necessarily correlated with the cost ratio: in these cases one should consider the extended cost ratio $\frac{c_{+1}\pi_{+1}}{c_{-1}\pi_{-1}}$, not discussed in this study. In Box 2, a cost parameter $w \in [0, 2]$ is defined such that $c_{+1} = w$ and $c_{-1} = 2 - w$: clearly, $w = 1$ corresponds to the classical AdaBoost model, while values of $w > 1$ will increase contribution to error by misclassification of positive cases, and vice versa for $w < 1$. In particular, suppose $w > 1$: a greater weight α_t will be therefore assigned to the models with higher sensitivity.

On a more local scale, Step 5 in Box 2 introduces a second variation to AdaBoost in the weight updating procedure. For $w > 1$, the weights of the misclassified positive samples will be increased more than those of misclassified negatives, and the weights of the correctly classified negative samples will be decreased more than those positive and correctly classified (Figure 2). In order to induce

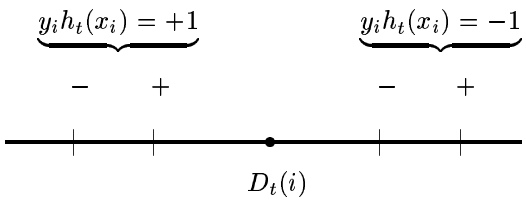


Fig. 2. Weight updates.

higher sensitivity, the procedure therefore puts more attention on the hardest positive examples. In terms of the margin $y_i H(x_i)$, the result of the procedure is to increase the margin of the positive samples more than for the negative ones. According to the results in [31], it follows that the measure of confidence in the prediction is higher for the positive samples, i.e. for $w > 1$ the final SSTBoost model has been trained for generalizing with higher sensitivity.

This property was tested on the MEDS melanoma data base: in the left panel of Figure 3, the cumulative margin distribution (data from both classes) is shown for three different values of the misclassification costs. For $w = 1$ (equivalent to the AdaBoost algorithm), we can see that the margins are approximately concentrated between 0.5 and 0.8. For values of w different from 1, a gap in the margin distribution is observed. In particular, for $w = 1.34375$ the cumulative distribution remains flat approximately from 0.3 to 0.8 (solid curve in the left panel of Figure 3). The right panel of Figure 3 clarifies how the gap is originated for this value of the cost parameter: training has aggressively increased the margin of the positive samples (always greater than 0.8), while the margin of the negative samples remains lower than 0.3.

3.2 The SSTBoost Search Procedure

The cost-sensitive procedure discussed in Section 3.1 supports the development of classification models H_w differently polarized towards sensitivity or specificity. Here we discuss a method for the automatic selection of an optimal cost parameter w^* in order to satisfy or to get as close as possible to admissible sensitivity and specificity constraints. The idea is to take advantage of the cost-sensitive learning variant described in Box 2 and at the same time to avoid a manual tuning of w or an extensive tabulation of the possible H_w in order to reach the minimal operative requirements. If A is a target region in the ROC space, i.e. A is a compact subset of $[0,1] \times [0,1]$, the constraints are satisfied for w^* such that $(1 - \widehat{Sp(w^*)}, \widehat{Se(w^*)}) \in A$, where $\widehat{Se(w)}$ and $\widehat{Sp(w)}$ are predictive estimates (e.g. from cross-validation over the training data L) of the sensitivity and specificity of the model H_w computed according to Box 2. The goal is then the minimization of the distance between the ROC curve and the compact set A, where the ROC curve is defined as $\phi_H : [0,2] \rightarrow \Re^2$

$$\phi_H(w) = (1 - \widehat{Sp(w)}, \widehat{Se(w)}). \tag{2}$$

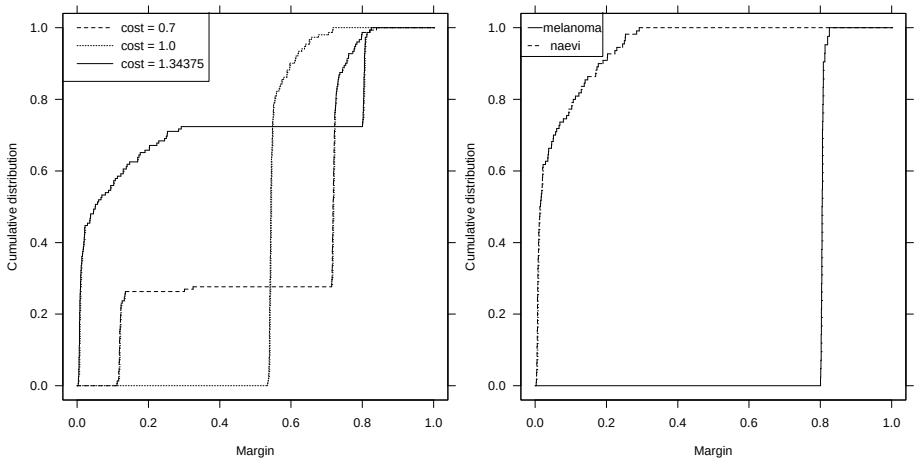


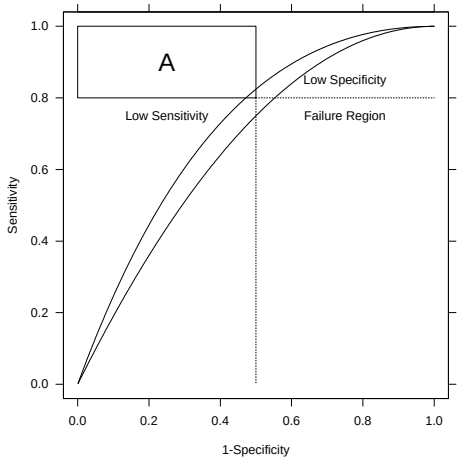
Fig. 3. Left panel: cumulative margin distribution for different values of the misclassification costs. For values of the misclassification cost w different from $+1$, a gap in the margin distribution is observable. Right panel: cumulative margin distribution of positive and negative samples for $w = 1.34375$.

The problem can be addressed as a minimization problem of a function of one real variable. Let $\Delta : [0, 2] \to \Re^+$ be defined as

$$\Delta(w) = dist(\phi_H(w), A) = \min_{a \in A} \|\phi_H(w) - a\|. \tag{3}$$

The problem admits a solution, not necessarily unique: the possible optimal cost parameters are selected by $\tilde{w} = \text{argmin}_w \Delta(w)$. In practice, constraints are likely to be of the type ($Se \geq a$ AND $Sp \geq b$). In this case, A is a rectangular subset and the two components of ϕ_H are increasing, so numerous search algorithms can be applied to quickly individuate an optimal cost parameter $\tilde{w}$. A simple but effective bisection method is described in Box 3. The algorithm fails when $\phi_H([0, 2]) \cap A = \emptyset$, otherwise one has to choose one of the w such that $\phi_H(w) \in A$ according to some super-optimality criterion, or just stopping at the first admissible $\tilde{w}$. Several effective alternatives for the search procedure are available, all leading to a fast convergence towards A, or at least as near to A as possible, particulary without strict hypotheses over the cost parameter. However, it must be taken into account that the $\phi_H(w)$ is only estimated (by cross-validation, in our example), and thus its smoothness is not necessarily ensured. Following [1], in case information about the c_i costs resulted available, the search can be constrained within a smaller interval $I \subset [0, 2]$. A further improvement might be introduced in the procedure by considering a non-euclidean distance for the $dist$ function in Eq. 3.

– Given a target region A
– Set $w_1 = 1$, $w_{min} = 0$, $w_{max} = 2$
– Train $H_{w_1}(x)$ as in Box 2 and use cross-validation to estimate $\phi_H(w_1)$
– For $i = 1, ..., M$
  1. If $\phi_H(w_i) \in$ Failure Region
     Abort
     ElseIf $\phi_H(w_i) \in$ Low Specificity
     $w_{i+1} = 1/2(w_{max} - w_i)$
     $w_{min} = w_i$
     ElseIf $\phi_H(w_i) \in$ Low Sensitivity
     $w_{i+1} = 1/2(w_i - w_{min})$
     $w_{max} = w_i$
     ElseIf $\phi_H(w_i) \in$ A
     Return $\tilde{w} = w_i$
     EndIf
  2. Train $H_{w_{i+1}}(x)$ and use cross-validation to estimate $\phi_H(w_{i+1})$
  EndFor



Box 3: The **SSTBoost** tuning procedure

4 Application to the melanoma data

We applied the procedure described in Boxes 2 and 3 to develop an effective model for early melanoma diagnosis. The goal was the development of a tool for supporting the discrimination between malignant and benign lesions in accordance with application-specific constraints based on the MEDS data set described in Section 2.

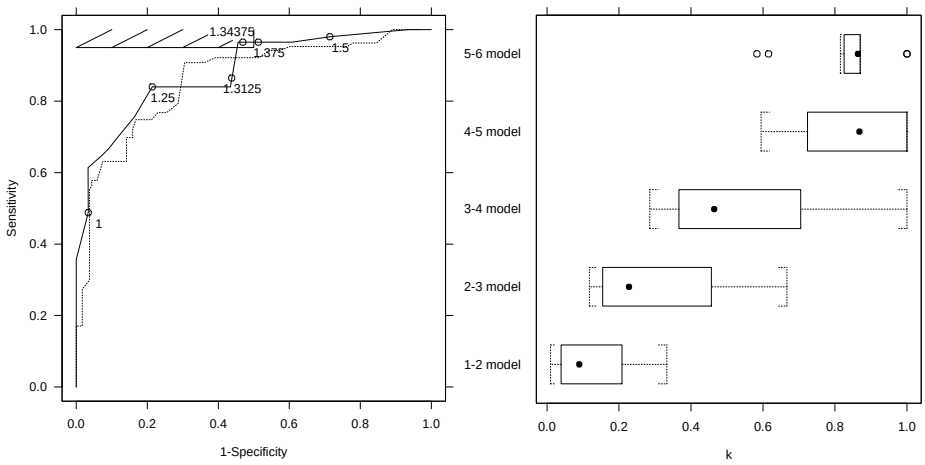


Fig. 4. Left panel: ROC curve on the melanoma MEDS data for the SSTboost model (solid line) compared with the ROC curve (dotted) of the th-Adaboost models obtained shifting the classification threshold. The dashed rectangle on the top left represents the target region and the points indicate the value of the cost w during the SSTBoost optimization as described in Box 3. Right panel: distribution of the κ statistic for pairs (m_i, m_{i+1}) of models from the tuning procedure.

The system was designed to support early diagnosis in a screening modality: it was required to recognize the maximum possible number of malignant lesions, accepting a specificity of at least 0.5. The target region A in the ROC space is therefore defined as $Se \geq 0.95$ and $1 - Sp \leq 0.5$. The target region corresponds to the shaded rectangle in the left panel of Fig. 4. The ROC curve, estimated by cross-validation for different H_w models, is also plotted: the curve is obtained by tabulation of $\phi_H(w)$, following the SSTBoost procedure in Box 2. The 6 circles indicate the performance for models $m1, \ldots, m6$: the models were obtained as steps of the tuning procedure in Box 3 (the SSTBoost steps alternate along the curve in correspondance to the values w=1, 1.5, 1.375, 1.3125, 1.34375). The complete ROC curve of the th-Adaboost classification model performance is also included in the same plot. The experiment shows that we can avoid computing a dense estimate of the ROC curve and leave the algorithm self-tune in order to reach the target region in 6 convergence steps, while improving over the baseline th-Adaboost procedure.

Measuring the effective difference between different proposed classifiers is an important issue in model selection procedures. Given two models and a common test set, the κ statistic can be computed in order to test the difference between the models [9]. For $\kappa = 0$ the agreement between classifiers equals that expected by chance, while $\kappa = 1$ indicates complete agreement between the two

models. The distribution of κ statistic at each step of the tuning algorithm, i.e. $\kappa(m_i, m_{i+1})$ is shown in the right panel of Fig. 4. For each pair of models (i.e. of pairs of cost parameters), the κ statistic is computed on each of the 10 cross-validation test sets. It can be observed that diversity between models progressively reduces at each step: at the end, the median κ value is greater than 0.8 indicating very small changes in model performance.

Table 1. For each classifier and combination of classifiers, sensitivity and specificity, with the standard deviation, are shown. The asterisk indicates results from [5]. The last row of the table concerns with the averaged performances of 8 dermatologists (see Section 2).

Classifier	*Sens.* $\pm$ *SD*	*Spec.* $\pm$ *SD*
Discr. Ana.*	0.65 ± 0.30	0.83 ± 0.11
C4.5*	0.64 ± 0.28	0.84 ± 0.05
1-NN*	0.68 ± 0.30	0.90 ± 0.10
9-NN*	0.41 ± 0.25	0.96 ± 0.04
Discr. Ana. + C4.5 + 1-NN*	0.86 ± 0.32	0.64 ± 0.11
Discr. Ana. + C4.5 + 9-NN*	0.84 ± 0.32	0.71 ± 0.12
Bagging	0.48 ± 0.28	0.96 ± 0.07
AdaBoost	0.49 ± 0.32	0.97 ± 0.04
th-AdaBoost	0.92 ± 0.12	0.70 ± 0.14
SSTBoost	0.97 ± 0.07	0.54 ± 0.18
Dermatologists	0.83 ± 0.23	0.66 ± 0.13

The classification results on the MEDS data set are summarised in Table 1. The first group of rows includes results from a previous study [5] in the same experimental condition; results for bagging and the different variants of AdaBoost studied in this paper, including SSTBoost, are then reported. The machine learning results may be compared with the average performance over a panel of 8 dermatologists again in the same experimental conditions. In [5] the most interesting results were obtained by combination of classifiers. In particular, the performance closest to the constraints was obtained with a combination of three models: Discriminant Analysis, C4.5 and Nearest Neighbors.

The value for AdaBoost reported in Table 1 is slightly better than bagging and clearly unbalanced towards specificity. First a family of models was obtained from the AdaBoost models by thresholding the margin distributions for the two output classes and then choosing an optimal model as a function of the threshold. The results for this procedure are listed as the th-AdaBoost model. The SSTBoost procedure yielded the overall best results (see also Fig. 4). The target region A was reached in only 6 steps. Moreover, the model variability was

very moderate in comparison with the other models developed in this and in the previous study. It is interesting to note that the improvement of SSTBoost over th-AdaBoost seems to confirm the comparison between the MetaCost architecture for boosting and AdaCost reported in [35].

5 Conclusions

We have developed a methodology for cost-sensitive classification which extends boosting into a classification tool for automated diagnosis with self-tuning properties. Given a required minimal performance, in terms of a target region for sensitivity and specificity, we have indicated and tested a procedure for selecting the optimal w, i.e. such that the corresponding model reaches or goes as close as possible to the accuracy goals.

The introduction of a cost parameter w both within the estimated error function as well as within the weight updating of AdaBoost (as in [14]) allowed to effectively increase the margin of the predictions of one class with respect to the other.

In the skin cancer diagnosis task (cfr. Fig. 1 and Tab. 1), the model achieved a better sensitivity than each of the dermatologists. Both the model and 7 of the 8 dermatologists performed more than 0.5 specificity. On the average the dermatologists obtained a better specificity than SSTBoost, but none reached the 0.95 value in sensitivity.

Acknowledgments

The authors wish to thank B. Caprile (ITC-irst) for illuminating discussions on boosting magic, and S. Forti and C. Eccher (ITC-irst), M. Cristofolini and P. Bauer (Department of Dermatology, S. Chiara Hospital – Trento) for their significant collaboration on the melanoma diagnosis application. CF thanks M. Bauer and A. Bergamo for correcting minor errors in a previous version.

References

1. N. Adams and D. Hand, "Classifier performance assessment," *Neural Computation*, vol. 12, no. 2, pp. 305–311, 2000.
2. M. Binder, A. Steiner, M. Schwarz, S. Knollmayer, K. Wolff, and H. Pehamberger, "Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesion: a pilot study," *Brit J Dermatol*, vol. 130, pp. 460–5, 1994.
3. M. Binder, H. Kittler, S. Dreiseitl, H. Ganster, K. Wolff, and H. Pehamberger, "Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process," *Melanoma Research*, vol. 10, pp. 556–61, 2000.
4. L. Bischof, H. Talbot, E. Breen, D. Lovell, D. Chan, G. Stone, S. Menzies, A. Gutenev, and R. Caffin, "An automated melanoma diagnosis system," in *New Approaches in Medical Image Analysis* (B. Pham, M. Braun, A. Maeder, and M. Eckert, eds.), SPIE, 1999.

5. E. Blanzieri, C. Eccher, S. Forti, and A. Sboner, "Exploiting classifier combination for early melanoma diagnosis support," in *Proceedings of ECML-2000* (R. de Mantaras and E. Plaza, eds.), (Berlin), pp. 55–62, Springer-Verlag, 2000.
6. L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Pacific Grove CA: Wadsworth and Brooks/Cole, 1984.
7. L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
8. L. Breiman, "Combining predictors," in *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems* (A. Sharkey, ed.), (London), Springer-Verlag, 1999. pages 31–50.
9. T. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–158, 2000.
10. S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *J Biomed Inform*, vol. 34, pp. 28–36, 2001.
11. B. Efron and R. Tibshirani, *An introduction to the bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*. New York: Chapman & Hall, Inc., 1993.
12. B. Efron and R. Tibshirani, "Cross-validation and the bootstrap: estimating the error rate of a prediction rule," tech. rep., Standford University, 1995.
13. F. Ercal, A. Chawla, W. Stoecker, H. Lee, and R. Moss, "Neural network diagnosis of malignant melanoma from color images," *IEEE Transaction on Biomedical Engineering*, vol. 4, pp. 837–45, September 1994.
14. W. Fan, S. Stolfo, J. Zhang, and P. Chan, "Adacost: Misclassification cost-sensitive boosting," in *Proceedings of ICML-99*, 1999.
15. Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996. pages 148–156.
16. Y. Freund and R. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
17. Y. Freund and R. Schapire, "Discussion of the paper arcing classifiers by leo breiman," *The Annals of Statistics*, vol. 26, no. 3, pp. 824–832, 1998.
18. Y. Freund and R. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.
19. J. Friedman and P. Hall, "On bagging and nonlinear estimation," tech. rep., Stanford University, 2000.
20. J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," tech. rep., Stanford University, 1999.
21. C. Furlanello, S. Merler, and C. Chemini, "Tree-based classifiers and GIS for biological risk forecasting," in *Advanced in Intelligent Systems* (F. Morabito, ed.), (Amsterdam), IOS Press, 1997. pages 316–323.
22. C. Furlanello, S. Merler, A. Rizzoli, C. Chemini, and C. Genchi, "Bagging as a predictive method for landscape epidemiology of Lyme disease," *Giornale Italiano di Cardiologia*, vol. 29, no. 5, pp. 143–147, 1999.
23. C. Furlanello and S. Merler, "Boosting of tree-based classifiers for predictive risk modeling in gis," in *Multiple Classifier Systems* (J. Kittler and F. Roli, eds.), vol. 1857, (Amsterdam), Springer, 2000. pages 220-229.
24. A. Green, N. Martin, J. Pfitzner, M. O'Rourke, and N. Knight, "Computer image analysis in the diagnosis of melanoma," *J Am Acad Dermatol*, vol. 31, pp. 958–64, 1994.

25. G. Karakoulas and J. Shawe-Taylor, "Optimizing classifiers for imbalanced training sets," in *Advances in Neural Information Processing Systems 11* (M. Kearns, S. Solla, and D. Cohn, eds.), MIT Press, 1999.
26. D. Margineantu and T. Dietterich, "Bootstrap methods for the cost-sensitive evaluation of classifiers," in *Proceedings of ICML-2000*, pp. 583–590, Morgan Kaufmann, 2000.
27. S. Menzies, "Epiluminescence microscopy diagnostic criteria with follow-up computer-based monitoring of "less suspicious" lesion may increase sensitivity for the diagnosis of melanoma while maintaining adequate specificity," *Arch Dermatol*, vol. 137, no. 3, pp. 378–379, 2001.
28. S. Merler and C. Furlanello, "Selection of tree-based classifiers with the bootstrap 632+ rule," *Biometrical Journal*, vol. 39, no. 2, pp. 1–14, 1997.
29. S. Merler, C. Furlanello, C. Chemini, and G. Nicolini, "Classification tree methods for analysis of mesoscale distribution of ixodes ricinus (acari: ixodidae) in Trentino, Italian Alps," *Journal of Medical Entomology*, vol. 33, no. 6, pp. 888–893, 1996.
30. J. Quinlan, "Bagging, boosting, and c4.5," in *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996. pages 725–730.
31. R. Schapire, Y. Freund, P. Bartlett, and W. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.
32. S. Seidenari, G. Pellacani, and P. Pepe, "Digital videomicroscopy improves diagnostic accuracy for melanoma," *J Am Acad Dermatol*, vol. 39, pp. 175–81, 1998.
33. T. Shindenwolf, W. Stolz, R. Albert, W. Abmayr, and H. Harms, "Classification of melanocytic lesions with colour and texture analyses using digital image processing," *The International Academy of Cytology; Analytical and Quantitative Cytology and Histology*, vol. 15, no. 1, pp. 1–11, 1993.
34. H. Takiwaki, S. Shirai, Y. Watanabe, K. Nakagawa, and S. Arase, "A rudimentary system for automatic discrimination among basic skin lesions on the basis of color analysis of video images," *J Am Acad Dermatol*, vol. 32, pp. 600–3, 1995.
35. K. Ting, "A comparative study study of cost-sensitive boosting algorithms," in *Proceedings of ICML-2000* (M. Kaufmann, ed.), pp. 983–990, 2000.
36. K. Ting, "An empirical study of metacost using boosting algorithms," in *Proceedings of ECML-2000* (R. de Mantaras and E. Plaza, eds.), (Berlin), pp. 983–990, Springer-Verlag, 2000.