Abstract

In many computer vision classification and detection problems, the time to decision is no less important that the error rates. In the binary case, given the false positive and false negative rates, the optimal strategy minimizing the expected decision time is the Wald’s sequential probability ratio test (SPRT). Recently, Sochman and Matas proposed learning SPRT strategies from labeled examples by a boosting technique they dubbed WaldBoost [3]. WaldBoost was shown to outperform state-of-the-art methods in the average evaluation time while keeping comparable error rates. In this paper, we present the extension of WaldBoost to multi-hypothesis classification problems.

1 Introduction

In many computer vision problems such as detection and classification, the computational complexity of the classifiers plays a crucial role. This is especially true in real-time and mobile applications with stringent computation time and power budget. In the computer vision literature, overwhelming attention has been dedicated to the optimization of the classifier error rates, largely overlooking the complexity component. To the best of our knowledge,
the only study to explicitly address the optimal tradeoff between error rates and decision time is [3]. There, the authors addressed binary classification problems from the point of view of sequential decision making. In this framework, the optimal strategy in terms of expected decision time subject to a constraint on error rates is known to be Wald’s sequential probability ratio test (SPRT). The authors showed that an approximation of SPRT can be learned from labeled examples by a variant of AdaBoost.

In this paper, we present an extension of this approach to multi-class problems. Unlike the binary case, there exists no simple optimal sequential decision strategy comparable to SPRT. However, a particular multi-hypothesis strategy, MSPRT, has been shown to be nearly optimal among all simple strategies. We show how to map supervised construction of MSPRT onto a multi-class extension AdaBoost introduced in [5]. The rest of this paper is organized as follows: Section 2 presents the binary sequential decision formalism and overviews SPRT and WaldBoost. In Section 3, a multi-hypothesis SPRT is summarized. Section 4 introduces our extension of WaldBoost to multi-hypothesis classification problems. Finally, Section 5 concludes the paper.

2 Two-class sequential decision

Let \( x \) be a an object characterized by its class (hidden state) \( y \in \{-1, +1\} \), and let \( x_1, x_2, \ldots \) be an ordered sequence of measurements of \( x \). A sequential decision strategy \( S \) is a sequence of decision functions \( S = \{S_1, S_2, \ldots \} \) where \( S_k : (x_1, \ldots, x_k) \rightarrow \{-1, +1, \ast\} \). The sign \( \ast \) means that the object is undecided from the first \( k \) measurements. The strategy is evaluated sequentially starting from \( k = 1 \) until a decision of \( \pm 1 \) is made. Under mild and rather unrestrictive assumptions, it can be shown that \( S \) terminates in finite time with probability one. We denote by

\[
T_S(x) = \min\{k : S_k \neq \ast\}
\]

(1)
the decision time of an object, i.e., the smallest $k$ for which a decision is made about $x$. We will denote by $S(x)$ the result of such a decision.

A decision strategy can be characterized by the expected decision time

$$T_S = \mathbb{E}(T_S(x))$$

(2)

and the error rates

$$\alpha_S = \mathbb{P}(S(x) = -1|y = +1)$$
$$\beta_S = \mathbb{P}(S(x) = +1|y = -1);$$

(3)

$\alpha_S$ denotes false negative rate (i.e., the probability of $x$ belonging to $+1$ to be classified as $-1$), while $\beta_S$ denotes the false positive rate (i.e., the probability of $x$ belonging to $-1$ to be classified as $+1$).

Given some nominal error rates $\alpha$ and $\beta$, an optimal decision strategy is defined as

$$S^* = \arg\min_S T_S \text{ s.t. } \begin{cases} \alpha_S \leq \alpha \\beta_S \leq \beta \end{cases}$$

(4)

2.1 Wald’s sequential probability ratio test

Assuming that the joint conditional distributions $\mathbb{P}(x_1, \ldots, x_k|y = +1)$ and $\mathbb{P}(x_1, \ldots, x_k|y = -1)$ of the measurements given the class are known for all $k = 1, 2, \ldots$, Wald [4] defined the sequential probability ratio test (SPRT) as the decision strategy

$$S^*_k(x) = \begin{cases} +1 : & R_k \leq B \\ -1 : & R_k \geq A \\ \exists : & B < R_k < A, \end{cases}$$

(5)

where $R_k$ is the likelihood ratio

$$R_k = \frac{\mathbb{P}(x_1, \ldots, x_k|y = -1)}{\mathbb{P}(x_1, \ldots, x_k|y = +1)}$$

(6)

and the constants $A$ and $B$ are set to satisfy the required error rates.

Such a construction is remarkable due to the following property:
Theorem 1 (Wald). For every $\alpha$ and $\beta$, there exist such $A$ and $B$ that SPRT is an optimal test in the sense of (4).

Though the optimal thresholds $A$ and $B$ are difficult to compute in practice, the following approximation is available due to Wald:

Theorem 2 (Wald). The thresholds $A$ and $B$ in (5) are bounded by

$$A \leq A' = \frac{1 - \beta}{\alpha}$$
$$B \geq B' = \frac{\beta}{1 - \alpha}.$$  \hspace{1cm} (7)

Furthermore, when the bounds $A'$ and $B'$ are used in (5) instead of the optimal $A$ and $B$, the error rates of the sequential decision strategy change to $\alpha'$ and $\beta'$ for which

$$\alpha' + \beta' \leq \alpha + \beta.$$  

Stated differently, the approximate SPRT with the thresholds $A$ and $B$ replaced by the bounds $A'$ and $B'$ is Pareto-optimal (non-inferior) as is the optimal SPRT (5), in the sense that the approximation never compromises both error rates simultaneously.

2.2 WaldBoost

Though in his studies Wald limited his attention to the simplified case of i.i.d. measurements, the SPRT is valid in general cases. However, the practicality of such a test is limited to a very modest $k$, as the evaluation of the likelihood ratios $R_k$ in (5) involves multi-dimensional density estimation. A remedy to this problem was proposed by Sochman and Matas in [3] who used AdaBoost for ordering the measurements and for approximation of the likelihood ratios, dubbing the resulting learning procedure as WaldBoost.

Specifically, the authors examined the real AdaBoost, in which a strong classifier

$$H_k(x) = \sum_{i=1}^{k} h_i(x)$$  \hspace{1cm} (8)
is constructed as a sum of some real-valued weak classifiers $h_i$. The algorithm is a greedy approximation to the minimizer of the exponential loss

$$L(H) = \mathbb{E}(e^{-yH(x)}).$$

(9)

One of the powerful properties of AdaBoost is the fact that selecting the weak classifier at each iteration $k$ to be even slightly better than a random coin toss leads to the following asymptotic behavior [2]:

$$H_\infty(x) = \lim_{k \to \infty} H_k(x) = \arg \min_H L(H) = \frac{1}{2} \log \frac{\mathbb{P}(y = +1|x)}{\mathbb{P}(y = -1|x)},$$

(10)

which using the Bayes theorem can be rewritten as

$$H_\infty(x) = -\frac{1}{2} \log \frac{\mathbb{P}(x|y = -1)}{\mathbb{P}(x|y = +1)} + \frac{1}{2} \log \frac{\mathbb{P}(y = +1)}{\mathbb{P}(y = -1)}.$$ 

(11)

If we are free in the selection of an arbitrary weak classifier, the fastest convergence is achieved by

$$h_{k+1}(x) = \arg \min_h L(H_k + h) = \frac{1}{2} \log \frac{\mathbb{P}(y = +1|x, w_k(x, y))}{\mathbb{P}(y = -1|x, w_k(x, y))},$$

(12)

where $w_k(x, y) = e^{-yH_k(x)}$ is the (unnormalized) weight of the sample $x$.

Arguing that the asymptotic relation (11) holds approximately for a finite $k$, one gets

$$H_k(x) \approx -\frac{1}{2} \log \frac{\mathbb{P}(h_1(x), \ldots, h_k(x)|y = -1)}{\mathbb{P}(h_1(x), \ldots, h_k(x)|y = +1)} + \frac{1}{2} \log \frac{\mathbb{P}(y = +1)}{\mathbb{P}(y = -1)}$$

$$= -\frac{1}{2} \log R_k(x) + \text{const}$$

(13)

where

$$R_k(x) = \frac{\mathbb{P}(h_1(x), \ldots, h_k(x)|y = -1)}{\mathbb{P}(h_1(x), \ldots, h_k(x)|y = +1)}$$

(14)
is the likelihood ratio of the $k$ observations $h_1(x), \ldots, h_k(x)$ of $x$.

This relation between the strong classifier $H_k(x)$ and the log-likelihood ratio is fundamental, as it allows to replace the $k$-dimensional observation $(h_1(x), \ldots, h_k(x))$ of $x$ in $R_k(x)$ by its one-dimensional projection $H_k(x) = h_1(x) + \cdots + h_k(x)$, resulting in the following one-dimensional approximation of the likelihood ratio

$$
\hat{R}_k(x) = \frac{\mathbb{P}(H_k(x)|y = -1)}{\mathbb{P}(H_k(x)|y = +1)}. 
$$

Replacing $R_k$ in the SPRT (5) by $-H_k(x)$ yields the following sequential decision strategy

$$
S_k(x) = \begin{cases}
+1 & : H_k(x) \geq \tau_k^+ \\
-1 & : H_k(x) \leq \tau_k^- \\
\tau_k^- < H_k(x) < \tau_k^+ & : \tau_k^+ < H_k(x) < \tau_k^- 
\end{cases}
$$

which approaches the true SPRT as $k$ grows. The thresholds $\tau_k^+$ and $\tau_k^-$ are obtained from the estimated densities of $H_k(x)|y = +1$ and $H_k(x)|y = -1$.

Since for small $k$’s $\hat{R}_k$ approximates $R_k$ rather inaccurately, special precaution has to be taken to establish the thresholds. Sochman and Matas propose to use a separate validation set (independent of the training set), on which the densities are estimated using oversmoothed Parzen sums. The authors remark that such a conservative approach may result in a suboptimal decision strategy, yet it allows to reduce the risks of wrong irreversible decisions.

Another advantage of WaldBoost compared to AdaBoost is the fact that samples for which a decision has been made at iteration $k$ are excluded from the training set at subsequent iterations. New samples can be added to the training set replacing the removed ones. This allows WaldBoost to explore a potentially very large set of negative examples while keeping a modestly sized training set at each iteration.
3 Multi-class sequential decision

We now assume that an object $x$ may belong to $M$ distinct classes, $y \in \{1, \ldots, M\}$. A multi-class sequential decision strategy readily generalizes to a sequence of functions $S_k : (x_1, \ldots, x_k) \rightarrow \{1, \ldots, M, \top\}$. For $M = 2$, $S_k$ reduces to the two-class decision strategy from the previous section.

In the case of $M > 2$, a natural way of generalizing the error rates $\alpha_S$ and $\beta_S$ associated with $S$ is by defining

$$\alpha^m_S = \mathbb{P}(S(x) = m | y = n),$$  \hspace{1cm} (17)

for $m \neq n$, measuring the probability of incorrectly assigning class $m$ to an object from class $n$. We also denote by

$$\alpha^m_S = \sum_{n \neq m} \alpha^m_S \mathbb{P}(y = n)$$  \hspace{1cm} (18)

the probability of incorrectly assigning class $m$, and with some abuse of notation, by

$$\alpha_S = \sum m \alpha^m_S$$  \hspace{1cm} (19)

the probability of an incorrect decision.

The optimality criterion (4) can be generalized either as

$$S^* = \arg \min_S T_S \text{ s.t. } \alpha^m_S \leq \alpha^m, m = 1, \ldots, M$$  \hspace{1cm} (20)

with the nominal error rates $(\alpha^1, \ldots, \alpha^M)$, or as

$$S^* = \arg \min_S \mathcal{T}_S \text{ s.t. } \alpha_S \leq \alpha$$  \hspace{1cm} (21)

with the nominal error rate $\alpha$.

3.1 MSPRT

Unfortunately, the optimality property of the SPRT is lost in the generalization to multiple classes. Unlike in the binary case, there exists no simple
sequential decision strategy minimizing (20) or (21). Researchers in the field pursued either the truly optimal recursive Bayesian test, which is impractical due to its complexity, or devised numerous *ad hoc* tests, mainly based on repeated pair-wise application of the two-class SPRT. For most of such sequential tests, bounds relating them to the truly optimal strategy are either very loose, or do not exist.

A remarkable multi-class generalization of the SPRT was introduced by Baum *et al.* in [1]. For the set of parameters $0 < A_1, \ldots, A_M < 1$, they defined the following multi-class sequential decision strategy

\[
S_k(x) = \begin{cases} 
  m & : \Pr(y = m|x_1, \ldots, x_k) > \frac{1}{1 + A_m} \\
  \uparrow & : \text{otherwise.}
\end{cases}
\]  

(22)

Note that the procedure is well-defined when all $A_m < 1$, as there can be only one $m$ for which $\Pr(y = m|x_1, \ldots, x_k) > 1/(1 + A_m) > \frac{1}{2}$ since the probabilities must sum to one.

The authors refer to this strategy as to MSPRT because using Bayes’ theorem, the posterior probabilities $\Pr(y|x_1, \ldots, x_k)$ can be expressed as

\[
\Pr(y|x_1, \ldots, x_k) = \frac{\Pr(x_1, \ldots, x_k|y)\Pr(y)}{\sum_{n=1}^{M} \Pr(x_1, \ldots, x_k|y = n)\Pr(y = n)}.
\]  

(23)

It is straightforward to show that for $M = 2$, the MSPRT (22) is identical to the SPRT (5) with

\[
A = \frac{\Pr(y = +1)A_{+1}}{\Pr(y = -1)},
\]

\[
B = \frac{\Pr(y = +1)}{\Pr(y = -1)A_{-1}};
\]  

(24)

however, for $M > 2$, the prior probabilities $\Pr(y)$ cannot be absorbed into the parameters $A_m$. 

8
Though not having the optimality property of the two-class SPRT, the authors show that the MSPRT is still optimal among sequential tests restricted to the simple SPRT-like form, and provide the following bounds:

**Theorem 3** (Baum et al.). Let $S$ be an MSPRT with the parameters $A_1, \ldots, A_M$. Then,

a. The decision time of $S$ is exponentially bounded, i.e., there exist constants $c > 0$ depending on the $A_m$, and $P(y)$, and $\rho < 1$ depending only on the conditional distributions $P(x|y)$, such that $P(T_S > n) \leq cp^n$.

b. $\alpha_S^m \leq P(y = m)A_m$ for all $m = 1, \ldots, M$, and $\alpha_S \leq \sum_m P(y = m)A_m$. If in addition $A_1 = \cdots = A_M = A$, then

$$\alpha_S \leq \frac{A}{1 + A}.$$  

Consequently, given a nominal error rate $\alpha$, the MSPRT with $A_1 = \cdots = A_m = \alpha/(1 - \alpha)$ appears to be a reasonable approximation to the solution of (21).

### 4 Boosted MSPRT

In [5], Zhu et al. introduced a multi-class generalization of AdaBoost which they dubbed as stagewise additive modeling using a multi-class exponential loss (SAMME) algorithm. SAMME is very similar in its structure to the classical two-class AdaBoost, with a few exceptions. First, instead of using a scalar $y \in \{1, \ldots, M\}$, the class is encoded as an $M$-dimensional vector $y = (y_1, \ldots, y_M)^T$ where $y_m = 1$ if $y = m$ and $y_m = -\frac{1}{M-1}$ otherwise. The classifier itself becomes a vector valued function

$$H_k(x) = \sum_{i=1}^k h_i(x)$$  

constructed progressively as the sum of vector-valued weak classifiers. Informally, $H(x)$ can be thought of as a fuzzy classifier with its $m$-th element
representing the confidence of \( x \) belonging to the class \( m \). “Crisp” classification result can be obtained by selecting \( m \) for which \( (H(x))_m \) has the maximum value.

As in standard AdaBoost, the algorithm aims at greedily minimizing the exponential loss, which now assumes the following multi-variate form

\[
L(H) = \mathbb{E} \left( e^{-\frac{1}{M} y^T H(x)} \right).
\]  

(26)

Note that there is some degree of arbitrariness in the selection of \( H \); in order to make it estimable, the authors impose the zero-sum constraint \( \mathbf{1}^T H = 0 \), where \( \mathbf{1} \) is an \( M \)-dimensional vector of ones. The minimizer of the multi-class exponential loss is given by

\[
H^* = \arg\min_H L(H) \text{ s.t. } \mathbf{1}^T H = 0
\]  

(27)

where \( p(\cdot|x) = (\mathbb{P}(y = 1|x), \ldots, \mathbb{P}(y = M|x))^T \) is the vector of posterior probabilities, and

\[
M = (M - 1) \mathbf{I} + \left( \frac{1}{M} - 1 \right) \mathbf{1}^T
\]  

(28)

is a matrix having \( \frac{(M-1)^2}{M} \) on the diagonal and \( \frac{1-M}{M} \) off the diagonal. In other words, the \( m \)-th element of \( H^* \) is given by

\[
(H^*)_m = e_m^T H^* =
(M - 1) \left( \log \mathbb{P}(y = m|x) - \frac{1}{M} \sum_{n=1}^{M} \log \mathbb{P}(y = n|x) \right)
= \frac{M - 1}{M} \sum_{n=1}^{M} \frac{\mathbb{P}(y = m|x)}{\mathbb{P}(y = n|x)},
\]  

(29)

where \( e_m \) is the \( m \)-th vector of the standard Euclidean basis. Zhu et al. show that similarly to AdaBoost, if at each iteration the weak classifier performs better than a random \( M \)-sided dice, SAMME converges to this optimal solution, \( H_\infty = H^* \). The optimal weak classifier at iteration \( k \) is given by

\[
h_{k+1} = \arg\min_h L(H_k + h) \text{ s.t. } \mathbf{1}^T h = 0,
\]  

(30)
whose $m$-th element is given by

$$(h_{k+1})_m = \frac{M-1}{M} \sum_{n=1}^{M} \log \frac{\mathbb{P}(y = m|x, w_k(x, y))}{\mathbb{P}(y = n|x, w_k(x, y))},$$

where $w_k(x, y) = e^{\frac{1}{M} y^T H_k(x)}$ is the unnormalized weight of the sample $x$. Note that such $h_{k+1}$ reduces to (12) for $M = 2$.

Rewriting (27) as

$$p(\cdot|x) = \frac{\exp\left(\frac{1}{M-1} H_k\right)}{\mathbf{1}^T \exp\left(\frac{1}{M-1} H_k\right)}$$

it becomes evident that asymptotically the strong classifier related directly to the posterior probability $\mathbb{P}(y|x)$, based on which the decision is made in MSPRT. Arguing that this asymptotically exact relation holds approximately for a finite $k$, we arrive at the following sequential decision strategy

$$S_k(x) = \begin{cases} 
m : \frac{e^T \exp\left(\frac{1}{M-1} H_k(x)\right)}{\mathbf{1}^T \exp\left(\frac{1}{M-1} H_k(x)\right)} > \tau^m_k \\
\hat{z} : \text{otherwise.}
\end{cases}$$

For the nominal error rate $\alpha$, the decision threshold $\tau^m_k$ should be $1 - \alpha$. However, since the posterior probability estimated from $H_k$ at first iterations is rather inaccurate, we follow Sochman and Matas’ methodology and select the threshold by estimating oversmoothed Parzen sums.

### 5 Conclusion

In this study, we addressed supervised learning of multi-hypothesis classification strategies minimizing expected decision time. We showed that similarly to mapping the optimal binary decision strategy learning to the two-class AdaBoost algorithm (resulting in WaldBoost), a nearly-optimal multi-hypothesis decision strategy can be mapped to a multi-class variant of AdaBoost. This effectively extends WaldBoost to the multi-hypothesis case. In future studies, we intend to evaluate the performance of the proposed algorithm on different multi-class classification problems in computer vision.
References


