Adaptive Random Forest - How many "experts" to ask before making a decision?

Alexander G. Schwing
ETH Zurich
aschwing@inf.ethz.ch

Christopher Zach
ETH Zurich

Yefeng Zheng
Siemens Corp. Research

Marc Pollefeys
ETH Zurich

Abstract

How many people should you ask if you are not sure about your way? We provide an answer to this question for Random Forest classification. The presented method is based on the statistical formulation of confidence intervals and conjugate priors for binomial as well as multinomial distributions. We derive appealing decision rules to speed up the classification process by leveraging the fact that many samples can be clearly mapped to classes. Results on test data are provided, and we highlight the applicability of our method to a wide range of problems. The approach introduces only one non-heuristic parameter, that allows to trade-off accuracy and speed without any re-training of the classifier. The proposed method automatically adapts to the difficulty of the test data and makes classification significantly faster without deteriorating the accuracy.

1. Introduction

Due to their inherent multi-class formulation, their simplicity and their very high accuracy, Random Forest classifiers [4] attract increasing attention within the computer vision community. Variants like Random Ferns [21] and Extremely Randomized Trees [13] are also well known. In general those methods construct a set of base classifiers (*e.g.* trees), also called "experts" subsequently, and a vote is computed to predict on unseen data.

The original problem targeted by those classifiers is supervised learning and it is formulated as follows: use the N members of the training set $\mathcal{S} = \mathcal{X} \times \mathcal{L} = \left\{ (\boldsymbol{x}_i, y_i)_{i=1}^N \right\}$ to learn a model $M(\boldsymbol{x})$ that generalizes well on unseen data when predicting the label $y \in \mathcal{L}$ using the F-dimensional feature or equivalently called attribute or covariate vector $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^F$. In case of above mentioned methods, the model is given by

$$M(\boldsymbol{x}) = f\left(\sum_{i=1}^{T} g_i(\boldsymbol{x}) \right) \tag{1}$$

with T base classifiers ("experts") g_i, $i \in \{1, \dots, T\}$ and a function $f(\cdot)$ casting the result obtained by summation into the final classifier output. Details on the base classifiers and the function $f(\cdot)$ for Random Forest are provided in Section 2.

The set of available "experts" can be relatively large. Looking at a Random Forest classifier the available "experts" are the T trees composing the forest. Several hundred or even up to some thousand are rather usual. In addition, detection tasks in computer vision tend to have a large number F of features. They range from Haar-like wavelets [27] and Scale-Invariant Feature Transform (SIFT) [20] descriptors to Histogram of Oriented Gradients (HOG) [9] to name just a few. Features specifically designed for a particular problem are in use as well. To save computational time during classification of unseen data, we generally compute only those elements of the feature vector $\boldsymbol{x}$ that are necessary, *i.e.* the required elements are obtained on demand. Time for computation heavily depends on the complexity of the particular feature assessed. In recent years computational complexity of the features increased from just a few additions for Haar-like features to more complex operations like filtering for gradient computation in HOG.

Those features are applied in sliding window based object detection or pixel based classification where we generally label an image (2D or 3D) with the classes $\mathcal{L}$ specified during training. Thus we often apply the classifier to every instance (pixel/voxel) of the unseen image. See [23] for a recent application applying a classifier pixel-wise. For decent sized images we already talk about hundred thousands to millions of samples that need to be classified. Depending on the problem, many of these instances are easy to classify, i.e. one "expert" can provide the answer. For some of them we better ask all available classifiers and for others even that is not sufficient for a fundamental base of a decision.

To get a feeling for the impact of the proposed approach, we view a Random Forest classifier as "walking" in the discrete $|\mathcal{L}|$-dimensional space before it concludes with a decision. Considering majority voting, every classifier votes for one of the $|\mathcal{L}|$ unit directions. For a binary classification scenario the proportion after asking $K = k_1 + k_2$ classifiers is denoted by (k_1, k_2) with k_1 and k_2 representing the num-

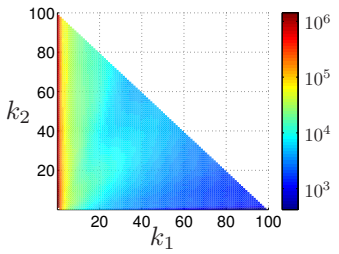


Figure 1. The number of test samples from our large scale data set that hit a certain proportion (k_1, k_2) during the classification process. The number of samples is color coded on a logarithmic scale.

ber of object and non-object guesses respectively. When classifying an easy sample we might only walk along one direction whereas for a difficult sample we end up close to $k_1 = k_2$. Fig. 1 illustrates, how often a certain proportion of positives and negatives, *i.e.* (k_1, k_2) was hit, when we apply a Random Forest classifier to test data of the two class classification problem presented in Section 5. For now it is sufficient to note that in many sliding window based classification problems like face and car detection, the number of negatives (non-object samples) exceeds the number of positives (object samples). In addition to the huge amount of easily classified negatives there are fewer (still several thousands) but easy to distinguish positives as well.

In the following, we propose an analysis based on the assumptions that 1) each **"expert" makes an independent prediction** and 2) the **"experts" are approximately equally knowledgeable**. The latter assumption holds well for Random Forest. However, the first assumption does not follow from the injected randomness which strives for uncorrelated trees [4]. Consequently, higher order moments do not necessarily drop out. Hence, the trees are not independent. Experiments nevertheless show, that we obtain encouraging results.

We describe how to integrate a statistical test, suitable for binary classification tasks into Random Forest. Next we derive an alternative method for the binomial case and extend it to multinomial problem instances. We base our analysis for the binomial and multinomial formulation on the well known beta or Dirichlet distribution.

For binary classification each "expert" is required to have a roughly equal probability p to favor a positive label and $1 - p$ for the negative label. If $p > 0.5$, *i.e.* the majority of the "experts" favor a positive label, we classify the sample to positive. Otherwise, it is negative. However, the probability p is unknown and can only be inferred from the binary decisions of the "experts." We derive a closed-form solution to make a consistent decision using the beta distribution of p given the number of "experts" consulted so far (K) and the number of "experts" preferring a specific class label (k_1 for the positive, k_2 for the negative, and $k_1 + k_2 = K$). We then decide if we can make a consistent classification (that means the classification after consulting K "experts" is consistent with that after consulting an infinite number of "experts") with a high probability, *e.g.*, larger than 95%.

For multi-class classification, it is hard to calculate the exact solution to make a consistent decision using a Dirichlet distribution for p. We therefore propose to calculate the confidence band of the estimate $\hat{p}$. If the class with the largest $\hat{p}$ is a clear winner (its lower confidence band is higher than the upper confidence bands of all other classes), we stop consulting more "experts."

- Thus, we propose a novel adaptive method for Random Forest that determines automatically when to stop classification. Experiments indicate a significant speed up for detection. In other words, we suggest to decrease T during the process of detection depending on the current state $M_t(\boldsymbol{x}) = f\left(\sum_{i=1}^{t} g_i(\boldsymbol{x})\right)$ of the model.
- For both, the binomial as well as the multinomial case, all the early termination criteria we obtain can be precomputed such that the only complexity added during classification is one table lookup and eventually comparisons.
- A further property of our method is the easiness to balance accuracy and speed for a given classifier. Neither re-training nor additional classifiers are required to enable tuning w.r.t. either speed or accuracy.

We provide details of our approach in Section 3 after reviewing Random Forest and related work in Section 2. To comfort the reader with the applicability of the method, we closely investigate the behavior of our algorithm on small multi-class data sets in Section 4 and provide insight into large scale behavior ($F \approx 100,000$, several millions of samples per image) in Section 5. We conclude with a summary in Section 6.

2. Related Work

Basically almost all the novel techniques in the object detection and recognition community apply classifiers in one or another way. A very popular choice are Random Forest classifiers [17, 29]. We subsequently provide a brief introduction.

A Random Forest Classifier is an ensemble/set of classification trees (*e.g.* CART [5]). Each leaf node of a tree i provides the probability $p_i(y \mid \boldsymbol{x})$ for $y \in \mathcal{L}$ which is obtained during training of the forest. We ask the interested reader to browse the respective literature [4] for details on the training procedure. We emphasize that in accordance with the original Random Forest formulation [4], $p_i(\boldsymbol{y} \mid \boldsymbol{x})$ is an $|\mathcal{L}|$-dimensional binary indicator variable, *i.e.* each tree casts a unit vote for the most popular class. To the best of our knowledge there is no evidence favoring

a distribution rather than an indicator variable for p_i. If, contrary to [4], p_i denotes a probability rather than a unit vote we use the label having the highest probability mass. As detailed later, prediction confidences are easily included into the presented methods. The final classification result is an averaging of unit votes $p_i \; \forall i$ and we assign the label resulting in the maximum vote,

$$M(\boldsymbol{x}) = \arg\max_{y \in \mathcal{L}} \frac{1}{T} \sum_{i=1}^{T} p_i(y \mid \boldsymbol{x}). \tag{2}$$

We obtain $f(\cdot) = \arg\max_{y \in \mathcal{L}} \frac{1}{T}(\cdot)$ and $g_i = p_i$ by simple comparison of Eq. (2) with the general model (1). The few comparisons denoted by $f(\cdot)$ are computationally less expensive than passing a sample down from the root of tree i using particular features for the binary decision.

To achieve improved performance w.r.t. time, several methods for Random Forest classifiers were proposed. The parallelized architecture of the graphics processing unit (GPU) can be leveraged. Sharp [26] shows the tremendous improvements achievable by a clever implementation. Combining this with our approach would facilitate real-time classification for even more complex tasks. It is however not our main intention to show real-time classification. We rather aim at indicating that further improvements are possible if we introduce smart decisions. Nevertheless we emphasize that our formulation allows for parallelization at the sample level, *i.e.* multiple samples can be classified at the same time. Another approach named Random Fern was proposed by Özuysal *et al.* [21]. As in Random Forest, the sample is passed down the set of ferns. Each node within a fern provides a result for the binary test which is used to access the leaf node containing the posterior probability $p_i(\boldsymbol{y} \mid \boldsymbol{x})$. When changing to log-space, the combination of posteriors is identical to Random Forest classifiers. Thus the model given in Eq. (2) is fully applicable and our method can be applied without any modification.

This emphasizes the general structure of our approach which can be combined with above two methods by Sharp and Özuysal *et al.* We underline that no particular hardware like a GPU is required in our case. Also note, that our approach is generally applicable to all classifiers that fulfill the assumptions highlighted in Section 1 and the structural requirement of Eq. (1).

Many more methods besides Random Forest were examined w.r.t. performance improvements. Among those are RANSAC [10] or other classifiers like Boosting [11, 24] or neural networks. Methods applied there are training of a rejection trace [3], genetic algorithms [30] and the sequential probability-ratio test (SPRT) [28] used *e.g.* in [7, 8] for Boosting and RANSAC. SPRT is a likelihood-ratio test which distinguishes two hypotheses. Thus it is applicable to binary classification problems but contrasting our approach, extensions to multiple classes are not straight forward.

We'll derive a decision rule suitable for binary problems in the following and compare to results obtained with SPRT. We then extend this alternative rule to multi-class problems. Note that the derived rules require only one threshold, whereas SPRT takes at least two parameters as shown in Section 3.1.

3. Early Termination of Classification

As stated in Section 1 and as we will derive shortly, the alternate rule is based on a purely statistical foundation using respective distributions and corresponding confidence intervals. Due to the discrete nature of the multinomial distribution, confidence intervals cannot be computed in a straight forward manner. Note that approximation of confidence intervals for multinomial distributions is still subject of current research [6]. In the following, we distinguish between binomial and multinomial classification, although the multinomial formulation is of course applicable to a binomial problem. For both cases all the early termination criteria can be pre-computed. Hence we just need to retrieve a value from the respective position in a table and our online pruning technique hardly adds any computational complexity to the classification task.

3.1. Binomial Formulation

Suppose given a sample, we ask K "experts" and k_1 of them tell us it is positive. An intuitive estimate of the probability for this sample to be positive is $p = \frac{k_1}{K}$. Is this an unbiased estimate? How reliable is this estimate? Suppose $k_1 \geq \frac{K}{2}$. If we stop consulting more "experts" and draw a conclusion that this sample is positive, what is the probability that we make a consistent decision (which means the decision we make so far is consistent with the decision after consulting an infinite number of "experts")? In this section, we will answer all these questions.

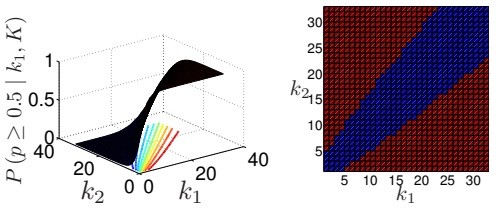
Suppose we consult K "experts" and each one independently casts a vote for the positive object class with probability p. The probability of observing k_1 (where $0 \leq k_1 \leq K$) positive tests follows the binomial distribution

$$b(k_1 \mid p, K) = \binom{K}{k_1} p^{k_1} (1-p)^{K-k_1}. \tag{3}$$

A positive result from the "expert" g_i is a vote for the positive class, *i.e.* g_i indicates a vote for label 1. As g_i is a binary indicator variable, this task is particularly simple such that k_1, k_2 and K are obtained at no extra cost. Note, that the derivation for the number of negative votes k_2 is analogous.

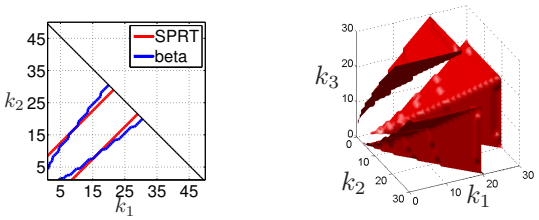
Using Bayes' rule we compute the distribution of p given the object votes k_1 and the number of trials K as

$$P(p \mid k_1, K) = \frac{b(k_1 \mid p, K)\, P(p \mid K)}{\int_0^1 P(k_1, p \mid K)\, \mathrm{d}p}. \tag{4}$$



(a) (b)

Figure 2. The binomial case: (a) illustrates the probability $P(p \geq 0.5 \mid k_1, K)$ for the current votes (k_1, k_2) and (b) shows the region of early termination obtained with a confidence of $1 - \alpha = 0.9$ when observing different votes (k_1, k_2).



(a) (b)

Figure 3. A comparison of SPRT and our approach is shown in (a). The boundary in the multinomial case with proportions (k_1, k_2, k_3) and confidence $1 - \alpha = 0.9$ is given in (b).

Here, $P(p \mid K) = P(p)$. Without a priori knowledge about the distribution of p we in general assume it to be least informative, *i.e.* uniform. By simplifying Eq. (4) we obtain

$$P(p \mid k_1, K) = \frac{b(k_1 \mid p, K)}{\int_0^1 b(k_1 \mid p, K)\,\mathrm{d}p}. \tag{5}$$

Plugging the binomial distribution given in Eq. (3) into the result (5) yields the conjugate prior of the binomial, *i.e.* the beta distribution

$$P(p \mid k_1, K) = \frac{(K+1)!}{k_1!\,(K-k_1)!} p^{k_1} (1-p)^{K-k_1}. \tag{6}$$

It is easy to verify that $\hat{p} = k_1/K$ is a maximum likelihood (ML) estimate. But the unbiased estimate for a beta distribution is computed as

$$\mathbb{E}_{P(p|k_1,K)}[p] = \frac{k_1+1}{K+2}. \tag{7}$$

The unbiased estimate is slightly smaller than the ML result. We will indicate the difference between this estimator and the ML result during our large scale experiments.

Having computed the distribution for our random variable p as given in Eq. (6), we calculate the probability to make a consistent decision for a positive label, *i.e.* $P(p \geq 0.5 \mid k_1, K) = \int_{0.5}^1 P(p \mid k_1, K)\,\mathrm{d}p$ to be

$$P(p \geq 0.5 \mid k_1, K) = 1 - \frac{(K+1)!0.5^{k_1+1}}{(k_1+1)!\,(K-k_1)!} {}_2F_1(\cdot) \tag{8}$$

with ${}_2F_1(k_1+1, k_1-K; k_1+2; 0.5)$ being the hypergeometric function.

The probability $P(p \geq 0.5 \mid k_1, K)$ is illustrated in Fig. 2(a) for the voting results (k_1, k_2) with $k_1 + k_2 = K$. If it exceeds a confidence $1 - \alpha$, we stop consulting more "experts."

Assume we obtain the current votes (k_1, k_2) to decide whether to ask further "experts." If either the probability for object samples or non-object samples exceeds a certain threshold $1 - \alpha$ we stop. The isolines for some thresholds are depicted in Fig. 2(a). Thus we get the region of early termination marked with red color in Fig. 2(b). Note, that the confidence $1 - \alpha$ does not necessarily need to be equal for different classes. A sample given to a binary classifier will "walk" in the (k_1, k_2) space till it either hits the area highlighted with red color or until no more "experts" are available. Interestingly our method inherently forces us to consult at least four "experts" to achieve a confidence of $1 - \alpha = 0.9$. This is intuitive and automatically incorporated.

In Fig. 3(a) we compare the "termination boundary" of our proposed approach with the one obtained by SPRT, which is $A \leq k_2 - k_1 \leq B$. The constants $A(\epsilon_1, \epsilon_2, \theta_1, \theta_2)$ and $B(\epsilon_1, \epsilon_2, \theta_1, \theta_2)$ depend on the error probabilities ϵ_1, ϵ_2 and the hypothesis probabilities $\theta_1 > 0.5$, θ_2. Assuming equal treatment of object and non-object samples we obtain at least two parameters with $\theta_2 = 1 - \theta_1$ and $\epsilon_1 = \epsilon_2$, contrasting one threshold α necessary for our proposed approach. In favor of experiments, we refer the interested reader to, *e.g.*, [16] for an in-depth discussion and examples regarding SPRT. Investigating Fig. 3(a) ($\alpha = \epsilon_1 = \epsilon_2 = 0.1, \theta_1 = 0.6$), the alternatively derived rule is more aggressive in making a decision when asking few "experts" and more cautious when the ensemble members cannot find a quick consensus.

Note that Eq. (8) depends only on the discrete values for the number of positive tests k_1 and total number of tests K. Additionally we know that the maximum amount of tests is bounded by the maximum amount of available "experts" ($K \leq T$) and that the number of positive tests k_1 is bounded by the total number of "experts" consulted so far. Summarizing those facts, we just need to store a polynomial amount of $\frac{(T+1)(T+2)}{2} - 1$ values which is tractable even for a large number of available base classifiers. Consequently the computation of (8) is replaced by a fast table lookup and almost no computational complexity is added. By extending the computed table to $\frac{(TN+1)(TN+2)}{2} - 1$ we can further incorporate confidences at the tree level. We note that many values of the enlarged table are close to 0 or 1 which can be used to decrease the size.

3.2. Multinomial Formulation

Current research is particularly interested in multi-class scenarios and our approach is applicable with minor modifications. Keep in mind that we do not want to add significant computational complexity to the classification process.

Let the vector $\boldsymbol{k} = \left[k_1, \ldots, k_{|\mathcal{L}|}\right]$ denote votes for classes with $\sum_{i=1}^{|\mathcal{L}|} k_i = K \leq T$ being the number of "experts" consulted so far. Note the simplex constraint $\sum_{i=1}^{|\mathcal{L}|} p_i = 1$. For the binomial formulation we derived usage of its conjugate prior (beta distribution). In the following, we apply the Dirichlet distribution $P\left(\boldsymbol{p} \mid \boldsymbol{k}, K\right) = \frac{1}{Z(\boldsymbol{k})} \prod_{i=1}^{|\mathcal{L}|} p_i^{k_i - 1}$ as the conjugate prior of the multinomial without showing the derivation. The multinomial beta function $Z\left(\boldsymbol{k}\right)$ is used for normalization of the distribution. Contrasting the binomial case shown in Eq. (8), exact computation of the probability to make a consistent decision for p_i is hard for $|\mathcal{L}| > 2$. The reason is the required marginalization over all variables except p_i and the one probability replaced via the constraint. This step, not necessary for the binomial formulation, results in an integral that cannot be expressed analytically. We therefore turn to the ML estimator of $\boldsymbol{p} = \left[p_1, \ldots, p_{|\mathcal{L}|}\right]$. For a Dirichlet distribution it is $\hat{\boldsymbol{p}} = \left[\hat{p}_1, \ldots, \hat{p}_{|\mathcal{L}|}\right]$ with $\hat{p}_i = \frac{k_i}{K}$.

Contrasting the binomial case and as mentioned above, it is hard to derive a closed-form solution for the probability to make a consistent decision, using a Dirichlet distribution for $\boldsymbol{p}$. Instead, we compute the confidence range for all the variables of the multinomial distribution of the probability $\boldsymbol{p} \in [0, 1]^{|\mathcal{L}|}$. Thus, we want the probability that p_i is within the range $[l_i, u_i]$ to be higher than the confidence $1 - \alpha$, *i.e.*

$$P\left(l_i \leq p_i \leq u_i\right) \geq 1 - \alpha. \tag{9}$$

Given the bounds, it is easy to judge whether to stop the classification process or consult further "experts." It is sufficient to compare the lower bound of the variable having the highest expected value $\hat{p}_i$ with the upper bound of the others. Formally, we first find $\gamma = \arg\max_{i \in \mathcal{L}} \hat{p}_i$ and stop the classification if $l_\gamma - u_i > 0 \; \forall i \in \mathcal{L} \backslash \gamma$. Thus we end up with $(|\mathcal{L}| - 1)$ comparisons.

To facilitate the comparison we need to compute confidence intervals for multinomial distributions in a fast manner. Moreover, we need a method providing reasonable results for a statistically small number of observations being at most the number of available "experts" T. This problem has a long history and it is well known that the confidence regions built from asymptotic statistics do not have good coverage for few observations.

The method proposed by Chafaï and Concordet [6] suggests conducting a comparison between two functions for each value of $\boldsymbol{p}$. Based on the result of the comparison they obtain the confidence region. It is generally not necessary in our case to compute the entire confidence region. Hence the number of comparisons used for this method can be reduced, but a simple computation is not possible. Another approach to construct simultaneous confidence intervals for multinomial proportions was described by Genz and Kwong [12]. Their method requires numerical solution of equations which will be too time consuming for our purposes. To really leverage the gain we prefer an analytic solution for the bounds. A number of different alternatives were proposed for asymptotic simultaneous confidence intervals for $\boldsymbol{p}$ [14, 22, 2]. Goodman [15] modified the methods which were later further improved by Kwong and Iglewicz [18]. Based on a small numerical study, we found the method proposed by Quesenberry and Hurst [22] to perform best. Let the current proportions be $\left[k_1, \ldots, k_{|\mathcal{L}|}\right]$ with $\sum_{i=1}^{|\mathcal{L}|} k_i = K$. The bounds $l_i\left(\alpha\right)$ and $u_i\left(\alpha\right)$ are given as

$$\frac{\chi^2 + 2k_i \pm \chi\left(\chi^2 + 4\frac{k_i}{K}(K - k_i)\right)^{\frac{1}{2}}}{2\left(K + \chi^2\right)} \tag{10}$$

with $\chi^2 = \chi_1^2\left(\frac{\alpha}{|\mathcal{L}|}\right)$ where $\chi_1^2\left(\frac{\alpha}{|\mathcal{L}|}\right)$ is defined as the $100\left(1 - \frac{\alpha}{|\mathcal{L}|}\right)$ percentage point of the chi-square distribution with 1 degree of freedom. (Note that those relations apply the Bonferroni correction.)

We illustrate the boundary between further consulting "experts" and stopping for a three-class classification problem in Fig. 3(b). Investigating the graph closely, we encounter that "experts" are consulted as long as we cannot clearly distinguish between any two or all three of the classes. Again the approach forces us to consult at least a certain number of "experts" before concluding with a decision. It is inherently built into our method without any heuristics.

Similar to the binomial case we pre-compute the bounds l_i and u_i necessary for Eq. (9) using the formula provided in Eq. (10) such that computation reduces to a simple table lookup given α, $|\mathcal{L}|$, k_i and K. The additional complexity introduced are $(|\mathcal{L}| - 1)$ comparisons and finding the most likely class. It remains to be shown in the following sections that we indeed achieve quite an improvement w.r.t. speed without degrading the accuracy.

4. Evaluation on Small Scale Data Sets

Due to the saturation nicely shown in [17] and observed in our own results (Fig. 4 and Fig. 6(d)) detailed below, we decide to use 100 "experts" as a base, similar to [4]. The work presented in [19] shows that nowadays the number of classifiers is decreased to a pre-determined threshold to obtain faster classification. We denote this as "Fixed Expert Set" in the following. Before targeting large scale data sets we evaluate our approach on a total of seven diverse data sets from the UCI repository [1] as detailed in Table 1.

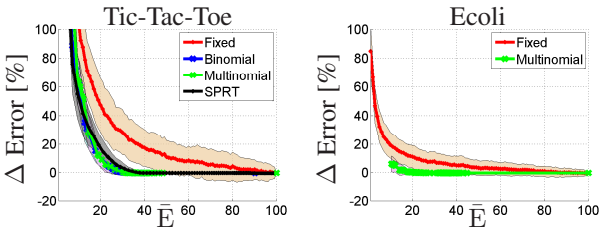


Figure 4. Comparison of fixed number of "experts" and our adaptive multinomial approach on the binary Tic-Tac-Toe and the eight class Ecoli data set plotted with identically scaled axis.

As those data sets are fairly small we test using the leave-one-out technique. Hence, we use all but one sample for training, and classify the left out sample during testing. For each leave-one-out experiment we train $T = 100$ "experts." Due to the randomness in the classifier we repeat each experiment 100 times. To study the trade off between speed and accuracy for the binomial and multinomial formulation, we tune $\alpha \in [0, 1]$. Hereby, $\alpha = 0$ means we always use all 100 "experts." For $\alpha > 0$, we adaptively decide for each sample how many of the available "experts" to consult. For binary classification problems we further apply the SPRT procedure and tune $\epsilon = \epsilon_1 = \epsilon_2 \in \left[0.5^{26}, 0.5\right]$ with the second specifiable parameter $\theta_1 = 1 - \theta_2 = 0.55$ being fixed. As ϵ sweeps through an interval, the particular choice for θ_1 and θ_2 is irrelevant for the figures to be shown. Given a specific value of α or ϵ, we obtain the average number of "experts" $\bar{\mathrm{E}}$ applied, which is a direct measure for the time spent during classification. We also compare our method with the approach of using a fixed number of "experts" E_X for all samples. The error is measured in deviation from the result obtained with all 100 classifiers. The representative results for the Tic-Tac-Toe and Ecoli data sets are shown in Fig. 4 together with the standard deviation. All three methods proposed to be applied for early termination, *i.e.* SPRT, the binomial and the multinomial formulation, clearly outperform the simple but very common approach of using a fixed number of "experts" [19]. The difference between the three proposed methods on binary data sets is minor. But most importantly, the standard deviation obtained with those approaches is far less than the one obtained when just decreasing the number of "experts." This shows the adaption to the difficulty of the data and makes testing results more reliable.

To provide the full picture, we summarize all results in Table 1. By inspection of graphs like the ones shown in Fig. 4, we find that value α_{100} for Eq. (9) and Eq. (10) or Eq. (8) as well as ϵ_{100} for SPRT such that we don't worsen the classification result compared to asking all 100 "experts." Consequently we obtain the average number of "experts" $\bar{\mathrm{E}}$ that were consulted after having classified all the samples. Next, we require consulting a fixed number of "experts" E_X that is approximately equal to the average number $\bar{\mathrm{E}}$. This will result in an approximately identical classification time. We read off the percentage of additional errors we make compared to using 100 classifiers. We also provide the minimal number of classifiers E_{min} necessary for the method named "Fixed Expert Set," such that we are not worse than asking 100 "experts."

Table 1. The UCI data set name together with α_{100} or ϵ_{100} achieving the same performance as 100 "experts." The average number of "experts" ($\bar{\mathrm{E}}$) that were evaluate when applying α_{100} or ϵ_{100} and how many additional misclassification errors (Error) made when requiring to consult a fixed number of classifiers ($\mathrm{E}_X \approx \bar{\mathrm{E}}$). The minimum number of classifiers E_{min} necessary to achieve the same accuracy as with 100 "experts." We compare binomial (b), multinomial (m) and SPRT (s) approach using the Ionosphere and the Tic-Tac-Toe data set (for further details see [25]).

Data set	α_{100} or ϵ_{100} ($\bar{\mathrm{E}}$)	Error (E_X)	E_{min}
Tic-Tac-Toe (b)	0.0002 (40.3)	16.5% (41)	100
Tic-Tac-Toe (m)	0.0018 (38.4)	18.0% (39)	100
Tic-Tac-Toe (s)	0.0185 (39.4)	17.5% (40)	100
Ionosphere (b)	0.0017 (19.6)	2.7% (20)	60
Ionosphere (m)	0.0065 (19.7)	2.7% (20)	60
Ionosphere (s)	0.0372 (25.1)	2.5% (26)	60
Iris (m)	0.0046 (15.4)	10.9% (16)	97
Wine (m)	0.0200 (15.2)	14.3% (16)	37
Glass (m)	0.0062 (49.7)	2.4% (50)	87
Ecoli (m)	0.0028 (35.6)	6.5% (36)	95
Yeast (m)	0.0019 (71.0)	0.5% (71)	100

Investigating Table 1, we observe that the average number of considered classifiers $\bar{\mathrm{E}}$ necessary to achieve a detection rate equivalent to the one obtained with 100 "experts" is smaller. We realize that we never manage to get the same accuracy by just asking a fixed number of "experts" $\mathrm{E}_X \approx \bar{\mathrm{E}}$. The minimum number of "experts" E_{min} necessary to achieve the same accuracy we obtain when applying the proposed approach is always larger. Savings depend on the particularities of the data, *i.e.* our method is only $1.4 \approx \frac{100}{71.0}$ times faster for the Yeast data but 6.3 times faster on the Iris set. The absolute error rates of our Random Forest settings (choose one among $\left\lfloor \sqrt{F} \right\rfloor$ features maximizing Gini entropy, 100 trees) for the Ionosphere, Glass and Ecoli set are 6.6%, 20.3% and 12.3% consistent with the 7.1%, 20.6% and 12.8% provided in [4].

Above results encourage to look at larger data sets, we obtain when considering typical detection problems in computer vision. Especially as the class distributions for above data sets are fairly uniform. Due to the skewed data, the possibilities to achieve the same accuracy by asking less "experts" should be given for imaging data sets as well.

5. Experiments on Large Scale Data Sets

As the usual computer vision data is too large for leave-one-out techniques we restrict ourselves to dividing the im-

Figure 5. Three zoomed in examples (size 100×100) of our wrist bone training database together with the overlaid annotations.

Table 2. The number of "experts" considered for different values of α together with the true positive rate (TPR) and the false positive rate (FPR) for biased (B) ML estimate and for the unbiased estimate (UB) given in Eq. (7).

α	$\bar{\mathrm{E}}$	$\mathrm{TPR_B}$	$\mathrm{FPR_B}$	$\mathrm{TPR_{UB}}$	$\mathrm{FPR_{UB}}$
0	100.0	0.948	0.045	0.948	0.045
0.0001	25.1	0.929	0.042	0.952	0.050
0.0057	15.0	0.930	0.042	0.960	0.056
0.03	10.2	0.929	0.042	0.954	0.053

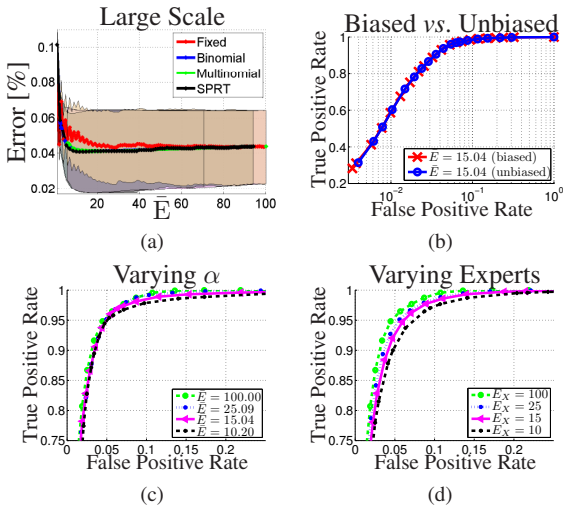


Figure 6. The three proposed methods are compared in (a). Comparison of biased and unbiased estimator in (b). Cutout of test set ROC curves using the binomial approach (Eq. (7)) for different α resulting in $\bar{\mathrm{E}}$ shown in (c) and for roughly equal number of fixed "experts" $\mathrm{E}_X \approx \bar{\mathrm{E}}$ in (d).

ages/volumes into two sets, a training part and a test set. We look at medical image classification as the images are usually fairly noisy. Specifically we decided to evaluate our method on 3D Magnetic Resonance Images (MRI) from the wrist and consider the task of detecting the eight carpal bones (binary classification). We show three sample images from a total of 23 3D training volumes together with annotations in Fig. 5. The volumes of left and right wrists are approximately parallel or anti-parallel to one of the two planar coordinate axes as visible in Fig. 5 and Fig. 7. The considered two-class voxel based classification task itself is similar to the work presented in *e.g.* [19].

The average size of the volumes is $240 \times 240 \times 30$ voxel and the number of Haar-like features is about $F \approx 100,000$. The five 3D volumes of the test set consist of approximately 1.5 million samples each. Having trained 100 base classifiers, we apply them to test data and count how often proportion (k_1, k_2) was hit during classification. We show the result in Fig. 1. Considering the logarithmic scale in Fig. 1 and recalling the region for termination, exemplarily shown for a confidence value $\alpha = 0.1$ in Fig. 2(b), indicates the potential to improve speed if we adaptively ask less "experts."

We apply a Random Forest classifier together with the proposed methods for different α and ϵ on the test set. The error rates are compared in Fig. 6(a). The three approaches outperform the common method of reducing the number of "experts" to a fixed number. This time, even 100 ensemble members cannot achieve the accuracy obtained with the SPRT, the binomial or the multinomial formulation. To measure the decrease in performance of the binomial formulation for a varying object classification threshold, we observe the receiver operator characteristic (ROC). Fig. 6(b) to (d) and Table 2 summarize the results. For the other methods, ROCs are provided in [25] together with notes on the practical applicability.

The result for varying α is shown in Fig. 6(c) and the effect introduced by simply decreasing the number of "experts" to a fixed number E_X is illustrated in Fig. 6(d). We observe that the true positive rate around a reasonable operating point (5% false positive rate) drops by about 5% if we reduce E_X to one tenth. Applying our proposed approach, we roughly maintain the true positive rate while decreasing the consulted number of "experts" to almost one tenth. Closer investigations [25] reveal that a fixed number of about 50 trees is necessary to achieve the same accuracy.

Besides the exact numbers for a usual operating point in Fig. 6(c), we indicate in Table 2 and via the markers in Fig. 6(b) ($\alpha = 0.0057$) the difference when using the unbiased estimator (UB) given in Eq. (7). The curves themselves are of course supposed to lie ontop of each other. The rates are provided for a threshold of 0.5, *i.e.* every sample having probability higher than 0.5 is considered an object sample. If we don't correct the probability, we move on the ROC curve towards slightly lower false positive rates which is avoided with the unbiased estimate. Of course we can alternatively tune the threshold to achieve the same operating point in ROC space. Applying the estimator given in Eq. (7) circumvents tuning.

The number of "experts" that were asked for each voxel of a test volume is shown color coded in Fig. 7. We observe the regions that are "easy" for the classifier, *i.e.* the interior of the bones and the exterior of the wrist.

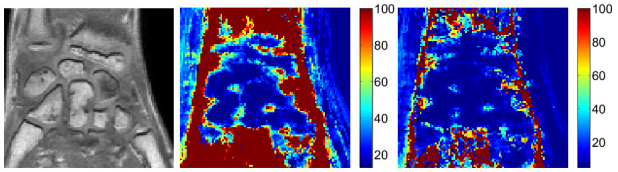


Figure 7. Number of "experts" asked on a zoomed in test image (left) for confidences resulting in an average number of "experts" $\bar{\mathrm{E}} \approx 25$ and $\bar{\mathrm{E}} \approx 10$ (right).

6. Conclusion

For small scale and large scale we emphasize that the common procedure of reducing the number of "experts" results in a drop of performance which is minimized by adaptively deciding *how many "experts" to ask before making a decision*. The detailed, parallelizable approach is based on a statistical formulation and we show in numerous leave-one-out experiments on publicly available data sets as well as on 3D images not used for training that we achieve faster classification without severely degrading the accuracy.

Acknowledgements: The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant#210806.

References

[1] A. Asuncion and D. Newman. UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007. 1381

[2] B. J. R. Bailey. Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of the cell frequencies. *Technometrics*, 22:583–589, 1980. 1381

[3] L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *Proc. CVPR*, volume 2, pages 236–243, 2005. 1379

[4] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. 1377, 1378, 1379, 1381, 1382

[5] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall, 1984. 1378

[6] D. Chafaï and D. Concordet. Confidence regions for the multinomial parameter with small sample size. *J. of the American Stat. Association*, 104(487):1071–1079, 2009. 1379, 1381

[7] J. Šochman and J. Matas. WaldBoost - Learning for Time Constrained Sequential Detection. In *Proc. CVPR*, pages 150–157, 2005. 1379

[8] O. Chum and J. Matas. Optimal Randomized RANSAC. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 30(8):1472–1482, 2008. 1379

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, pages 886–893, 2005. 1377

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications ACM*, 24(6):381–395, 1981. 1379

[11] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. Int'l Conf. on Machine Learning (ICML)*, pages 148–156, 1996. 1379

[12] A. Genz and K.-S. Kwong. Numerical evaluation of singular multivariate normal distributions. *J. of Stat. Computation and Simulation*, 68:1–21, 1999. 1381

[13] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 36(1):3–42, 2006. 1377

[14] R. Z. Gold. Tests auxiliary to χ^2 tests in a Markov chain. *Annals of Math. Statistics*, 30:56–74, 1963. 1381

[15] L. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7:247–254, 1965. 1381

[16] Z. Govindarajulu. *Sequential Statistics*. World Scientific Publishing, 2004. 1380

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009. 1378, 1381

[18] K.-S. Kwong and B. Iglewicz. On singular multivariate normal distribution and its applications. *Computational Statistics & Data Analysis*, 22(3):271–285, 1996. 1381

[19] V. Lempitsky, M. Verhoek, A. Noble, and A. Blake. Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. In *Proc. Int'l Conf. on Functional Imaging and Modeling of the Heart (FIMH)*, pages 447–456, 2009. 1381, 1382, 1383

[20] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999. 1377

[21] M. Özuysal, P. Fua, and V. Lepetit. Fast keypoint recognition in ten lines of code. In *Proc. CVPR*, 2007. 1377, 1379

[22] C. Quesenberry and D. Hurst. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6:191–195, 1964. 1381

[23] M. Reynolds, J. Doboš, L. Peel, T. Weyrich, and G. J. Brostow. Capturing time-of-flight data with confidence. In *Proc. CVPR*, 2011. 1377

[24] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistic*, 26(5):1651–1686, 1998. 1379

[25] A. G. Schwing. Adaptive Random Forest - How many "experts" to ask before making a decision? Supplementary Material. http://www.alexander-schwing.de, 2011. 1382, 1383

[26] T. Sharp. Implementing decision trees and forests on a GPU. In *Proc. ECCV*, pages 595–608, 2008. 1379

[27] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518, 2001. 1377

[28] A. Wald. *Sequential Analysis*. Dover, 1947. 1379

[29] Z. Yi, A. Criminisi, J. Shotton, and A. Blake. Discriminative, semantic segmentation of brain tissue in MR images. In *Proc. Int'l Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2009. 1378

[30] Z. Zhou, J. Wu, and W. Tang. Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1-2):239–263, 2002. 1379