JMLR: Workshop and Conference Proceedings 25:459–474, 2012 Asian Conference on Machine Learning

Multi-Stage Classifier Design

Kirill Trapeznikov KTRAP@BU.EDU
Venkatesh Saligrama SRV@BU.EDU
David Castañón DAC@BU.EDU
Boston University, Boston, MA

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

In many classification systems, sensing modalities have different acquisition costs. It is often *unnecessary* to use every modality to classify a majority of examples. We study a multi-stage system in a prediction time cost reduction setting, where the full data is available for training, but for a test example, measurements in a new modality can be acquired at each stage for an additional cost. We seek decision rules to reduce the average measurement acquisition cost. We formulate an empirical risk minimization problem (ERM) for a multi-stage reject classifier, wherein the stage k classifier either classifies a sample using only the measurements acquired so far or rejects it to the next stage where more attributes can be acquired for a cost. To solve the ERM problem, we factorize the cost function into classification and rejection decisions. We then transform reject decisions into a binary classification problem. We construct stage-by-stage global surrogate risk, develop an iterative algorithm in the boosting framework and present convergence results. We test our work on synthetic, medical and explosives detection datasets. Our results demonstrate that substantial cost reduction without a significant sacrifice in accuracy is achievable.

Keywords: multi-stage classification, sequential decision, boosting, cost sensitive learning

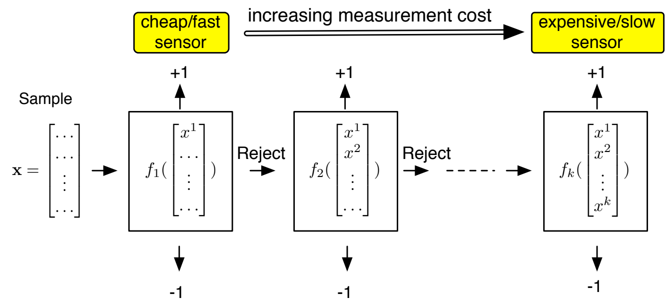
1. Introduction

In many applications including homeland security and medical diagnosis, decision systems are composed of an ordered sequence of stages. Each stage is associated with a sensor or a physical sensing modality. Typically, a less informative sensor is cheap (or fast) while a more informative sensor is either expensive or requires more time to acquire a measurement. In practice, a measurement budget (or throughput constraint) does not allow all the modalities to be used simultaneously in making decisions. The goal in these scenarios is to attempt to classify examples with low cost sensors and limit the number of examples for which more expensive or time consuming informative sensor is required.

For example, in explosives detection, in the first stage, an infra-red (IR) imager is used. The second stage is a more expensive and time consuming active millimeter wave (AMMW) scanner. The final third stage is a time consuming human inspection. In medical applications, first stages are typically non-invasive procedures (such as a physical exam) followed by more expensive tests (blood test, CT scan etc) and the final stages are invasive (surgical) procedures.

Many such examples share a common structure (see Fig. 1), and we list some of its salient aspects below:

Figure 1: Multi-Stage System consists of K stages. Each stage is a binary classifier with a reject option. The system incurs a penalty of δ_k at kth stage if it rejects to seek more measurements. The kth classifier only sees the first k sensing modalities in making a decision.



(A) *Sensors & Ordered Stages:* Each stage is associated with a new sensor measurement or a sensing modality. Multiple stages are an ordered sequence of sensors or sensor modalities with later stages corresponding to expensive or time-consuming measurements. In many situations, there is often some flexibility in choosing a sensing modality from a collection of possible modalities. In these cases, the optimal choice of sensing actions also becomes an issue. While our methodology can be modified to account for this more general setting, we primarily consider a fixed order of stages and sensing modalities in this paper. This is justified on account of the fact that many of the situations we have come across consist of a handful of sensors or sensing modalities. Consequently, for these situations, the problem of choosing sensor ordering is not justified since one could by brute force enumerate and optimize over the different possibilities.

(B) *Reject Classifiers:* Our sequential decision rules either attempt to fully classify an instance at each stage or "reject" the instance on to the next stage for more measurements in case of ambiguity. For example, in explosives detection, a decision rule in the first stage, based on IR scan, would attempt to detect whether or not a person is a threat and identify the explosive type/location in case of a threat. If the person is identified as a threat at the first stage it is unnecessary (and indeed dangerous – the explosive could be detonated) to seek more information. Similarly in medical diagnosis if a disease is diagnosed at an early stage, it makes sense to begin early treatment rather than waiting for more conclusive tests.

(C) *Information vs. Computation*: Note that our set up can only use the partial measurements acquired up to a stage in making a decision. In other methods such as detection cascades the full measurement and therefore all the information is available to every stage. Therefore, any region in the feature space can be carved out with more complex regions in the measurement space, or equivalently complex features can be extracted but with higher costs. In contrast, we have only partial measurements (or information) and so any feature or classifier that we employ has to be agnostic to unavailable measurements at that stage.

The two stage toy example in Fig. 2 illustrates some of the advantages of our scheme over the alternative scheme that first acquires measurements from all the sensing modalities, which we refer to as the centralized classifier. A reject classifier utilizes the 2nd stage sensor only for a fraction of the data but achieves the same performance as the centralized classifier.

Our approach is based on the so called *Prediction Time Cost Reduction approach* (Kanani and Melville (2008)). Specifically, we assume a set of training examples in which measurements from all the sensors or sensing modalities as well as the ground truth labels are available. Our goal is to derive *sequential reject classifiers* that reduces cost of measurement acquisition and error in the *prediction (or testing) phase.*

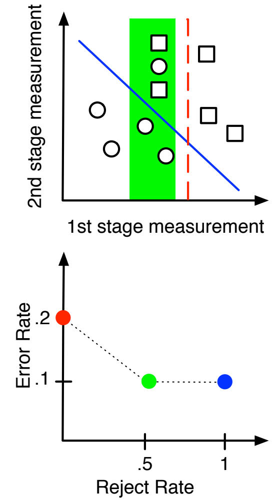


Figure 2: (Advantage of a 2 stage classifier: 10 samples, binary (squares, circles). The red line is the optimal decision when using only 1st stage modality. The blue line is optimal if using both. (2nd stage) The curve is classification error vs. samples rejected (cost) The red point corresponds to classifying everything at stage 1. The blue corresponds to rejecting everything and classifying using both modalities.(Stage 2) The green is a partial reject strategy. The samples outside the green region are classified using only the first modality, and samples inside the region are rejected to stage 2 and are classified using both modalities. Note that blue and green have the same error, while the reject strategy (green) has to use 2nd stage sensor only for $\frac{1}{2}$ of examples, reducing the cost by a factor of 2.

We show that this sequential reject classifier problem can be formulated as an instance of a *Markov Decision Problem* **(MDP)** when the class-specific probability models for the different sensor measurements are known. In this case the optimal sequential classifier can be cast as a solution to a Dynamic Program (DP). The DP solution is a sequence of *stage-wise optimization* problems, where each stage problem is a combination of the cost from the current stage and the cost-to-go that is carried on from later stages.

Nevertheless, class probability models are typically unknown; our scenarios produce high-dimensional sensor data (such as images). Consequently, unlike some of the conventional approaches (Ji and Carin (2007)), where probability models are first estimated to solve MDPs, we have to adopt a non-parametric *discriminative learning* approach. We formulate a novel *multi-stage expected risk minimization (ERM) problem.* This ERM formulation closely emulates limiting stage-wise-optimization suggested by the Dynamic Programming solution to the Markov Decision Problem (MDP). We solve this ERM problem at each stage by first factorizing the cost function into classification and rejection decisions. Then we transform reject decisions into a binary classification problem. Specifically, we show that the optimal reject classifier at each stage is a combination of two binary classifiers, one biased towards positive examples and the other biased towards negative examples. The disagreement region of the two then defines the reject region.

We then approximate this empirical risk with a global surrogates. We present an iterative solution and demonstrate local convergence properties. The solution is obtained in a boosting framework. We tested our methods on synthetic, medical and explosives datasets. Our results demonstrate an advantage of multistage classifier: cost reduction without a significant sacrifice in accuracy.

1.1. Related Work

Active Feature Acquisition (AFA): The subject of this paper is not new and has been studied in the Machine Learning community as early as MacKay (1992a). Our work is closely related to the so called prediction time active feature acquisition (AFA) approach in the area of cost-sensitive learning. The goal there is to make sequential decisions of whether or not to acquire a new feature to improve prediction accuracy. A natural approach is to formalize a problem as an MDP. Ji and Carin (2007); Kapoor and Horvitz (2009) model the decision process and infer feature dependencies while taking acquisition costs into account. Sheng and Ling (2006); Bilgic and Getoor (2007); Zubek and Dietterich (2002) study strategies for optimizing decision trees while minimizing acquisition costs. The construction is usually based on some purity metric such as entropy. Kanani and Melville (2008) propose a method that acquires an attribute if it increases an expected utility. However, all these methods require estimating a probability likelihood that a certain feature value occurs given the features collected so far. While surrogates based on classifiers or regressors can be employed to estimate likelihoods, this approach requires discrete, binary or quantized attributes. In contrast, our problem domain deals with high dimensional measurements (images consisting of million of pixels), so we develop a discriminative learning approach and formulate a multi-stage empirical risk optimization problem to reduce measurement costs and misclassification errors. At each stage, we solve the reject classification problem by factorizing the cost function into classification and rejection decisions. We then embed the rejection decision into a binary classification problem.

Single Stage Reject Classifiers: Our paper is also closely related to the topic of reject classifiers, which has also been investigated. However, in the literature reject classifiers have been primarily considered in a single stage scenario. In the Bayesian framework, Chow (1970) introduced Chow's rule for classification. It states that given an observation x and a reject cost δ and J classes, reject x if the maximum of the posteriors for each class is less than the reject cost: $\max_{k=1..J} \mathrm{P}(y = j|x) < \delta$. In the context of machine learning, the posterior distributions are not known, and a decision rule is estimated directly. One popular approach is to reject examples with a small margin. Specifically, in the context of support vector machine classifiers, Yuan and Casasent (2003); Bartlett and Wegkamp (2008); Rodríguez-Díaz and Castañón (2009); Grandvalet et al. (2008), define a reject region to lie within a small distance (margin) to the separating hyperplane and embed this in the hinge loss of the SVM formulation. El-Yaniv and Wiener (2011) propose a reject criteria motivated by active learning but its implementation turns out to be computationally impractical. In contrast, we consider multiple stages of reject classifiers. We assume an error prone second stage which occurs in such fields as threat detection and medical imaging. In this scenario, rejecting in the margin is not always meaningful. Figure 3 illustrates that thresholding the margin to reject can lead to significant degradation. This usually happens when stage measurements are complimentary; then examples within a small margin of the 1st stage boundary may not be meaningful to reject. Multiple stages of margin based reject classifiers have been considered by Liu et al. (2008) using SVMs in image classification. The method does not take into account the cost of later stages and is similar to the myopic method that we compare in the Experiments section.

Detection Cascades: Our multi-stage sequential reject classifiers bears close resemblance to detection cascades. There is much literature on cascade design (see Zhang and Zhang (2010); Chen et al. (2012) and references therein) but most cascades roughly follow the set-up introduced by Viola and Jones (2001) to reduce computation cost during classification. At each stage in a cascade, there is a binary classifier with a very high detection rate and a mediocre false alarm rate. Each stage makes a partial decision; it either detects an instance as negative or passes it on to the next stage. Only the last stage in the cascade makes a full decision, namely, whether the example belongs to a positive or negative class.

There are several fundamental differences between detection cascades and the multi-stage reject classifiers (MSRC). A key difference is the system architecture. Detection cascades are primarily concerned with binary classification problems. They make partial decisions, delaying a positive decision until the final stage. In contrast, MSRCs can deal with multi-class problems and can make classification decisions at any stage. Conceptually, this distinction requires a fundamentally new approach; detection cascades work because their focus is on unbalanced problems with few positives and a large number of negatives; and so the goal at each stage is to admit large false positives with negligible missed detections. Consequently, each stage can be associated with a binary classification problem that is acutely sensitive to missed detections. In contrast, our scheme at each stage is a composite scheme composed of a multi-class classifier as well as a rejection decision. The rejection decision is itself a binary classification problem. In practice, MSRCs arise in important areas such as medical diagnosis and explosives detection as we argued in Sec 1, item **(B)**. As a performance metric detection cascades tradeoff missed detections at the final stage with average computation. MSRC's tradeoff average misclassification errors against number of examples that reached later stages (i.e. required more sensors or sensing modalities). For these reasons it is difficult to directly compare algorithms developed for MSRCs to those developed for detection cascades. Nevertheless, our goals and resulting algorithms are similar to some of the issues that arise in cascade design (see Chen et al. (2012) and references therein), namely, perform a joint optimization for all the stages in a cascade given a cost structure for different features.

Other Cost Sensitive Methods: Network intrusion detection systems (IDS) is an area where sequential decision systems have been explored. (see Fan et al. (2000); Lee et al. (2002); Cordella and Sansone (2007)). In IDS, features have different computation costs. For each cost level, a ruleset is learned. The goal is to use as many low cost rules as possible. In a related set-up, Fan et al. (2002); Wang et al. (2003) consider a more general ensemble of base classifiers and explore how to minimize the ensemble size without sacrificing performance. In the test phase, for a sample, another classifier is added to the ensemble if the confidence of the current classification low. Here, similar to detection cascades, the goal is to reduce computation time. As we described in Sec 1, item **(C)**, the important distinction is that, in our setting, a decision is based only on the partial information acquired up to a stage. In a computation driven method, a stage (or base classifier) decides using a feature computed from the full measurement vector.

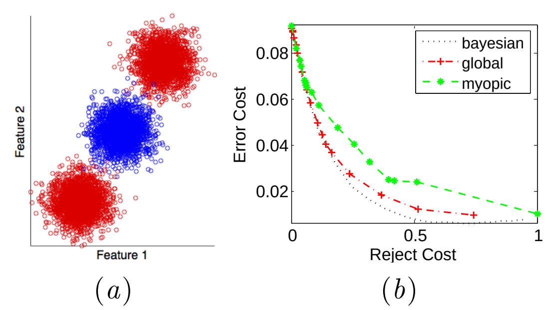


Figure 3: (a) Gaussian Mixture (binary). (b) Error rate vs reject rate on complimentary measurements. 1st stage uses only dim 1. 2nd stage uses only dim. 2. Myopic strategy (green) is thresholding the margin of the classifier, our method is global surrogate; Bayesian classifier (best performance). Thresholding the margin performs significantly worse than our method.

2. Problem Statement

Let $(x, y) \in \mathcal{X} \times \{+1, -1\}$ be distributed according to $\mathcal{D}$. A data point has K features: $x = \{x(1), x(2), \ldots, x(K)\}$. A kth feature is extracted from a measurement acquired at kth stage. Define a truncated feature vector at kth stage: $x^k = \{x(1) \ldots x(k)\}$.

The system has K stages, the order of the stages is fixed, and kth stage acquires a kth measurement. At each stage, k, there is a binary classifier with a reject option, f^k. It can either classify an example, $f^k(x^k) = +1/-1$ or delay the decision until the next stage, $f^k(x^k) = r$ and incur a penalty of δ_k. f^k has to make a decision using only the first k sensing modalities. The last stage K is terminal, a standard binary classifier. Define the system risk to be,

$$R(f^1, f^2, \ldots, f^K, x^k, y) = \sum_{k=1}^{K} R_k(x^k, y, f^k, S^k) \tag{1}$$

Here, R_k is the cost of classifying at kth stage, and $S^k(x) \in \{0, 1\}$ is the state variable indicating whether x has been rejected up to kth stage. [1]

$$R_k(x^k, y, f^k, S^k) = \begin{cases} S^k(x^k)\delta_k, \;\; f^k(x^k) = r \\ S^k(x^k)w_p, \;\; f^k(x^k) = -1, y = +1 \\ S^k(x^k)w_n, \;\; f^k(x^k) = +1, y = -1 \end{cases} \tag{2}$$

If example is active and is misclassified, the penalty is either w_n or w_p depending on error type. If it is rejected then the system incurs a penalty of δ_k and the state variable for that example remains at 1 (active).

$$S^{k+1}(x^{k+1}) = \begin{cases} 0, & S^k(x^k) = 0 \vee f^k(x^k) \neq r \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

2.1. Bayesian Setting

If the distribution $\mathcal{D}$ is known then the problem is to minimize the expected risk,

$$\min_{f^1, \ldots, f^K} \mathbf{E}_{\mathcal{D}} \left[R(f^1, \ldots, f^K, x^k, y) \right] \tag{4}$$

If we allow arbitrary decision functions then we can equivalently minimize conditional risk,

$$\min_{f^1, \ldots, f^K} \mathbf{E} \left[R(f^1, \ldots, f^K, x^k, y) | x \right] \tag{5}$$

This problem—by appealing to dynamic programming—remarkably reduces to a single stage optimization problem for a modified risk function. To see this, we denote,

$$\tilde{\delta}(x^k) = \min_{f^k, \ldots, f^K} \mathbf{E} \left[\sum_{t=k+1}^{K} R_t(x^t, y, f^t, S^t) \mid x^k, S^k = 1 \right] + \delta_k \tag{6}$$

1. ”r” denotes one of 3 possible outcomes; f^k can either classify x^k as $+1, -1$ or reject ”r”

and the modified risk function:

$$\tilde{R}_k(x^k, y, f^k, S^k, \tilde{\delta}) = \begin{cases} S^k(x^k)\tilde{\delta}(x^k), \; f^k(x^k) = r \\ S^k(x^k)w_p, \; f^k(x^k) = -1, y = 1 \\ S^k(x^k)w_n, \; f^k(x^k) = 1, y = -1 \end{cases} \tag{7}$$

We now claim the following: (for proof see the Trapeznikov et al. (2012))

Claim 1 *The optimal solution f^k minimizing $\mathbf{E}\left[\tilde{R}_k(\cdot) \mid x^k\right]$ is equal to the optimal kth stage classifier, f^k obtained in Eq. 5. Furthermore, f^k is obtained by suitably thresholding the posterior.*

$$f^k(x^k) = \begin{cases} +1, & \mathrm{P}(y=1|x^k) \geq 1 - \frac{\tilde{\delta}(x^k)}{w_n} \\ -1, & \mathrm{P}(y=1|x^k) \leq \frac{\tilde{\delta}(x^k)}{w_p} \\ r, & \frac{\tilde{\delta}(x^k)}{w_p} \leq \mathrm{P}(y=1|x^k) \leq 1 - \frac{\tilde{\delta}(x^k)}{w_n} \end{cases} \tag{8}$$

Finally, $f^k(\cdot)$ also minimizes the unconditional risk function:

$$f^k(\cdot) = \arg\min \mathbf{E}_{x^k,y}\left[\tilde{R}_k(y, x^k, f^k, S^k, \tilde{\delta})\right] \tag{9}$$

Note that the modified risk functional, $\tilde{R}_k$, is remarkably similar to R_k in Eq. 2 except for the modified reject cost. The main implication is that if the cost-to-go function $\tilde{\delta}(x^k)$ is known then we can ignore all of the other stages and minimize a single stage risk.

Reject Classifier As Two Binary Decisions: Classifier in Claim 1 is cumbersome. It is clear from the expression that we can express the decision region in terms of two binary classifiers f_n and f_p. Observe that for a given reject cost $\tilde{\delta}(x)$, the reject region is an intersection of two binary decision regions. To this end we further modify the risk function in terms of agreement and disagreement regions of the two classifiers, f_n, f_p, namely,

$$L_k(x^k, y, f_n, f_p, S^k, \tilde{\delta}) = \begin{cases} S^k(x^k)\tilde{\delta}(x^k), \; f_n \neq f_p \\ S^k(x^k)w_p, \; f_n = f_p = -y = -1 \\ S^k(x^k)w_n, \; f_n = f_p = -y = 1 \end{cases} \tag{10}$$

Note that the above loss function is symmetric between f_n and f_p and so any optimal solution can be interchanged. Nevertheless, we claim: (for proof see Trapeznikov et al. (2012))

Claim 2 *Suppose f_n and f_p are two binary classifiers that minimize $\mathbf{E}\left[L_k(x^k, y, f_n, f_p, S^k) \mid x^k\right]$ over all binary classifiers f_n and f_p. Then following resulting reject classifier, f^k:*

$$f^k(x^k) = \begin{cases} f_p(x^k), & \text{if } f_n(x^k) = f_p(x^k) \\ r, & \text{if } f_n(x^k) \neq f_p(x^k) \end{cases} \tag{11}$$

is the minimizer for $\mathbf{E}\left[\tilde{R}_k \mid x^k\right]$ in Claim 1 and the kth stage minimizer in Eq. 5. Also, $f^k(\cdot)$ minimizes the unconditional risk function:

$$f_n(x^k), f_p(x^k) = \arg\min \mathbf{E}_{x^k,y}\left[L_k(y, x^k, f_n, f_p, S^k, \tilde{\delta})\right] \tag{12}$$

We refer to Fig 4 for an illustration. We can express the new loss L_k compactly as follows:

$$L_k(x^k) = S^k(x^k)\left\{\mathbb{1}_{\left[f_p(x^k)\neq y \bigwedge f_n(x^k)\neq y\right]} + \tilde{\delta}(x^k)\mathbb{1}_{\left[f_p(x^k)\neq f_n(x^k)\right]}\right\} \tag{13}$$

Note that in arriving at this expression we have used $\mathbb{1}_{[a\neq c]}\mathbb{1}_{[a=b]} = \mathbb{1}_{[a\neq c]}\mathbb{1}_{[b\neq c]}$. To simplify the derivations, we assume equal false positive and false negative penalties. $(w_n = w_p)$

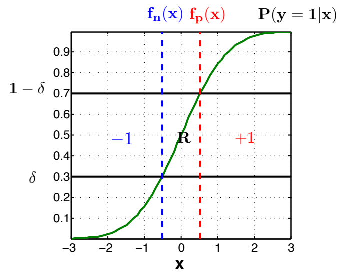


Figure 4: Optimal Reject Region (denoted by R) can be expressed as the disagreement region of two binary classifiers (f_n and f_p).

2.2. Empirical Setting

We are now given training data: $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Let the empirical distribution of the samples be given by $\hat{\mathcal{D}}_N$. Our task is to find multi-stage decision rules based on training data. To this end we consider the empirical version of Eq. 4, namely,

$$\min_{f^1,\ldots,f^K} \mathbf{E}_{\hat{\mathcal{D}}_N}\left[R(f^1,\ldots,f^K,x^k,y)\right] \tag{14}$$

Observe that, as in standard setting, we need to constrain the class of decision rules $\{f^1, \ldots f^K\} \in \{\mathcal{F}_1 \times \ldots \times \mathcal{F}_K\}$ here. This is because with no constraints the minimum risk is equal to zero and can be achieved in the first stage itself. This issue poses a problem because decisions on different examples cannot be arbitrary. So we cannot proceed from Eq. 4 to the conditional minimization of Eq 5.

To address this issue we proceed as follows. We assume that an estimate of the cost-to-go at the kth stage is available, i.e.,

$$\tilde{\delta}_i^k = \tilde{\delta}(x_i^k),\, \forall i = 1,\, 2, \ldots, N \tag{15}$$

Then, we can approximate f_p and f_n by a surrogate risk function that mimics Eq. 12,

$$f_n(x^k), f_p(x^k) = \arg\min \frac{1}{N}\sum_{i=1}^{N}\left[L_k(y_i, x_i^k, f_n, f_p, S_i^k, \tilde{\delta}_i^k)\right] \tag{16}$$

We can then compute the kth stage classifier exactly as in Eq. 11. This development then settles most of the issues except for the lack of knowledge of $\tilde{\delta}_i^k$. However, an estimate of $\tilde{\delta}_i^k$ can be obtained iteratively. Note by definition δ_i^k is actually a function of $f_p^{k+1}, f_n^{k+1}, \ldots, f^K$. So we define the recursion:

$$\tilde{\delta}_i^{k-1} = \delta_{k-1} + L_k(y_i, x_i^k, f^k, S_i^k, \tilde{\delta}_i^k) \tag{17}$$

Note that by definition there is an update only when state is active, i.e, $S_i^k = 1$

Given δ_i^k and all the stages but the kth, we can solve the subproblem 16 by iterating between f_p and f_n. To solve for f_p, we fix f_n and minimize a weighted error

$$f_p = \arg\min_f \sum_{i=1}^N w_i \mathbb{1}_{[f(x_i^k)\neq y_i]}, \;\; w_i = S_i^{k-1}\left[\mathbb{1}_{[f_n(x_i^k)\neq y_i]} + \tilde{\delta}_i^k - 2\mathbb{1}_{[f_n(x_i^k)\neq y_i]}\tilde{\delta}_i^k\right] \tag{18}$$

We can solve for f_n in the same fashion by fixing f_p. To derive this expression, we used another identity, $\mathbb{1}_{[a\neq b]} = \mathbb{1}_{[a\neq c]} + \mathbb{1}_{[b\neq c]} - 2\mathbb{1}_{[a\neq c]}\mathbb{1}_{[b\neq c]}$.

3. Algorithm

Minimizing the indicator loss is a hard problem. Instead, we take the usual ERM (empirical risk minimization) approach and replace it with a surrogate. We introduce an algorithm in the boosting framework based on the analysis from the previous section. Boosting is just one of our many possible machine learning approaches that can be used to solve it. We use boosting because it is easy to implement and is known to have good performance.

Boosting is a way to combines simple classifiers to form a strong classifier. We are given a set of such weak classifiers $\mathcal{H} = \{h_1(x), h_2(x)\ldots h_M(x)\},\; h_j(x) \in \{-1,+1\}$. The strong classifier is the linear combination: $sign(\sum_{h_j\mathcal{H}} q_j h_j(x))$ This set of weak classifiers need not be finite. Also, denote $\mathcal{H}_k \subset \mathcal{H}$ as a subset of weak classifiers that operate only on the first k measurements of x. $h_j(x) = h_j(x^k)\ if\ h_j \in \mathcal{H}_k$.

Global Surrogate: In our algorithm, we use the sigmoid loss function $\mathbf{C}(z) = \frac{1}{1+exp(z)}$ to approximate the indicator. Similar sigmoid based losses have been used in boosting before (Masnadi-Shirazi and Vasconcelos (2009)). Each subproblem (18) reduces to boosting a weighted loss[2].

To solve for stage k, we keep the rest of the stages constant. To find $f_p^k = \sum q_j h_j(x)$, we fix f_n^k and solve:

$$f_p^k = \arg\min_{q_1,q_2,\ldots} \sum_{i=1}^N w_i \mathbf{C}\left(y_i \sum_{h_j\in\mathcal{H}^k} q_j h_j(x_i)\right) \tag{19}$$

Note that the weights w_i are also expressed in terms of the $\mathbf{C}(z)$ instead of $\mathbb{1}_{[z]}$:

$$w_i = S_i^{k-1}\left[\mathbf{C}(yf_n^k(x_i)) + \tilde{\delta}_i^k - 2\mathbf{C}(yf_n^k(x_i))\tilde{\delta}_i^k\right] \tag{20}$$

To solve for f_n^k, we solve the same problem but keep f_p^k constant instead:

$$f_n^k = \arg\min_{q_1,q_2,\ldots} \sum_{i=1}^N w_i \mathbf{C}\left(y_i \sum_{h_j\in\mathcal{H}^k} q_j h_j(x_i)\right),\;\; w_i = S_i^{k-1}\left[\mathbf{C}(yf_p^k(x_i)) + \tilde{\delta}_i^k - 2\mathbf{C}(yf_p^k(x_i))\tilde{\delta}_i^k\right] \tag{21}$$

2. To reduce overtraining, we introduce a simple but effective regularization. For any loss $\mathbf{C}(z)$ and a parameter λ, we introduce a multiplicative term to the cost function:$\min_q exp(\lambda|\mathbf{q}|)\sum_{i=1}^N \mathbf{C}(y_i \sum_{h_j\in\mathcal{H}} q_j h_j(x_i))$ The term $exp(\lambda|q|)$ limits how large a step size for a weak hypothesis can become. It also introduces a simple stopping criteria: abort if $\frac{\sum_{i=1}^n w_i y_i h_{t+1}(x_i)}{\sum_{i=1}^n \mathbf{C}(y_i f_t(x_i))} \leq \lambda$. This corresponds to a situation when no descent directions (read weak hypothesis h_{t+1}) can be found to minimize the cost function

Note that the terms $\tilde{\delta}_i^k$ and S_i^{k-1} do not depend on stage k and remain constant when solving for f_p^k and f_n^k. For the ease of notation, we define a new term $\mathbf{C}_R$ that indicates if x_i is rejected at a kth stage. The term is close to one if f_p and f_n disagree (reject) and small if they agree.

$$\mathbf{C}_R(f_p^k, f_n^k, x_i^k, y_i) = \mathbf{C}(y_i f_p^k(x_i)) + \mathbf{C}(y_i f_n^k(x_i)) - 2\mathbf{C}(y_i f_p^k(x_i))\mathbf{C}(y_i f_n^k(x_i))$$

The expressions for state variables and cost-to-go are now simplified.

$$S_i^k = S_i^{k-1}\mathbf{C}_R(f_p^k, f_n^k, x_i^k, y)$$

The state variable remains greater than zero as long as x_i is rejected at every stage. The expression for cost-to-go at kth stage is:

$$\tilde{\delta}_i^k = \underbrace{\delta_k}_{\text{meas. cost}} + \underbrace{\mathbf{C}(y_i f_p^{k+1}(x_i^{k+1}))\mathbf{C}(y_i f_n^{k+1}(x_i^{k+1}))}_{\text{err. penalty if not rejected at stage } k+1} + \underbrace{\tilde{\delta}_i^{k+1}\mathbf{C}_R(f_p^{k+1}, f_n^{k+1}, x_i^{k+1}, y)}_{\text{cost-to-to if rejected at stage } k+1}$$

For the last stage (a standard binary classifier), we fix the first $K-1$ stages and solve:

$$f^K = \arg\min_{q_1,q_2,\ldots} \sum_{i=1}^{N} S_i^K \mathbf{C}\left(y_i \sum_{h_j \in \mathcal{H}^K} q_j h_j(x_i)\right) \tag{22}$$

We perform alternating minimization over one stage at a time. We start with the last stage and make our way backwards to the first stage. Then do a forward pass from 1st stage to last. This is repeated it until convergence [3] See Algorithm 1.

Our formulation allows us to form a surrogate for the entire risk in Equation 16, not just for each subproblem. This enables us to prove the following theorem,

Theorem 1 *Our global surrogate algorithm converges to a local minimum.*

This is simply due to a fact that we are minimizing a global smooth cost function by coordinate descent. However, since the global loss and the loss for each subproblem are non-convex programs, there is no global optimality guarantee. Theorem 1 ensures that our algorithm terminates. For proof see Trapeznikov et al. (2012).

Our system is composed of margin maximizing classifiers, therefore it is appropriate to derive generalization error bounds based on margins. It turns out that we can employ maximum margin generalization techniques from Bartlett et al. (1998) to derive error bounds for a two stage version of the system. We refer the reader to Trapeznikov et al. (2012) for theorem statement, the proof and explanation.

4. Experiments

The goal is to demonstrate that a large fraction of data can be classified at an early stage using a cheap modality. In our experiments, we use four real life datasets with measurements arising from meaningful stages.

3. To initialize f_n^k and f_p^k, we simply hard code f_p^k to classify any x as $+1$ and f_n^k as -1 so that all x's are rejected to the last stage.

Algorithm 1 Global Algorithm

INPUT: $\{x_i, y_i\}_{i=1}^N$, $\{\mathcal{H}_k\}_{k=1}^K$ {Weak Learners for each stage}, $\{\delta_k\}_{k=1}^{K-1}$ {reject costs}, D { Loop Iterations}
INITIALIZE: $f_n^k(x) \leftarrow +1, f_p^k(x) \leftarrow -1,\ for\ k = 1 \dots K-1$ {first $K-1$ stages reject everything}
for $d = 1, \dots, D$ **do**
  for $k = K, \dots, 1, 2, \dots K-1$ **do**
    {Start from the last stage then iterate to the first stage and then back to last stage}
    if $k < K$ **then**
      FInd f_p^k by solving boosting subproblem in 19
      Find f_n^k by solving boosting subproblem to 21
    else if $k = K$ **then**
      {Last Stage}
      Find $f^K(x)$ by solving boosting subproblem in 22
    end if
  end for
end for

$$F^k(x) \leftarrow \begin{cases} sign(f_p^k(x)), & \text{if } f_p^k(x) = f_n^k(x) \\ \text{reject}, & \text{if } f_p^k(x) \neq f_n^k(x) \end{cases}$$

OUTPUT: $F^1, F^2, \dots, F^K$

4.1. Related Algorithms:

We compare our algorithm to two methods:

Myopic: An absolute margin of a classifier is a measure of how confident a classifier is on an example. Examples with small margin have low confidence and should be rejected to the next stage to acquire more features. This approach is based on reject classification (Bartlett and Wegkamp (2008)). We know from Claim 1 that the optimal classifier is a threshold of the posterior. For each stage, we obtain a binary boosted classifier, $f^k(\cdot)$, trained on all the data. We then threshold the margin of the classifier, $|f^k(x)|$. It is known that given an infinite amount of training data, boosting certain losses (sigmoid loss in our case) approaches the log likelihood ratio, $f(x) = \frac{1}{2}\log\frac{\mathrm{P}(y=1|x)}{\mathrm{P}(y=-1|x)}$(Masnadi-Shirazi and Vasconcelos (2009)). So a reject region for a given threshold t_k is defined: $\{x \mid |f^k(x)| \leq t_k\}$. This is a completely myopic approach as the rejection does not take into account performance of later stages. This method is very similar to TEFE (Liu et al. (2008)) which also uses absolute margin as a measure for rejection. The difference is that our myopic strategy is a boosting classifier not an SVM as used in TEFE.

Expected Utility/Margin: An expected margin difference measures how a new attribute, if acquired, would be useful for an example. If this expected utility for an example is large then a new attribute should be acquired. This approach is based on the by Kanani and Melville (2008). We train boosted binary classifiers on all the data for each stage: $f^k(x^k)$. Given the measurement at the current stage x^k, we compute an expected utility (change in normalized margin) of acquiring the next measurement x_{k+1}:

$$U(x^k) = \sum_{x_{k+1} \in \mathcal{X}_{k+1}} \left| f^k(x^k) - f^{k+1}([x^k, x_{k+1}]) \right| \mathrm{P}(x_{k+1}|x_k)$$

An x^k is rejected to the next stage if its utility $U(x^k) \geq t_k$ is greater than a threshold. Here, $\mathcal{X}_{k+1}$ denotes the possible values that x_{k+1} can take. Note this approach requires estimating

$P(x_{k+1}|x^k)$[4], therefore the $(k+1)$th measurement has to be discrete or distribution needs to be parametrized. Due to this limitation, we only compare this method on two datasets.

4.2. Simulations

Performance Metric: A natural performance metric is the trade off between system error and measurement cost. Note, for utility and myopic methods, it is unclear how to set a thresholds t_k for each stage given a measurement cost δ_k. For this reason, we only compare them in a two stages system. More than two stages is not-practical because we would need to test every possible t_k for every stage k. In a two stage setting, measurement cost is proportional to the fraction of examples rejected to the second stage. For our algorithm, we vary a reject cost δ to generate a system error vs reject rate plot. For margin and utility, we sweep a threshold t_k. System error is the sum of 1st stage and 2nd stage errors. Reject rate is the fraction of examples rejected to the 2nd stage and require additional measurements. Low reject rate (cost) corresponds to higher error rate as most of the data will be classified at the first stage using less informative measurements. High reject rate will have performance similar to a centralized classifier, as most examples will be classified at the 2nd stage.

Set Up: In all our experiments, we use stumps [5] as weak learners. For each dataset and experiment, we randomly split the data 50/50 for training and testing. The results are evaluated on a separate test set, and the simulations are averaged over 50 monte-carlo trials. The number of iterations for each boosting subproblem is set to $T = 50$. In our global surrogate algorithm, the number of outer loop iterations is set to $P = 10$

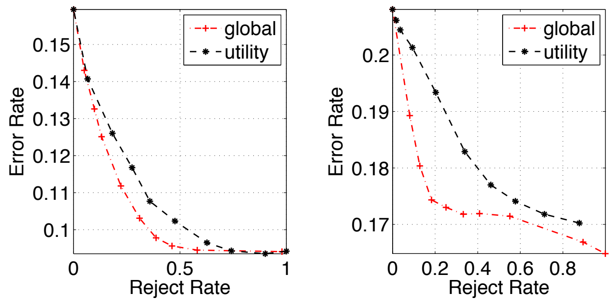
Name	Size	1st Stage	2nd Stage
Gassian Mixture	1000	1st dim	2nd dim
Mammogram Mass	830	3 CAD meas.	Radioligist Rating
Pima Diabetes	810	6 simple tests: BMI, sex, ..	2 blood tests
Polyps	310	12 freq. bins	126 freq. bins
Threat	1300	Images in IR, PMMW	Images in AMMW

Figure 5: Dataset Descriptions

Discrete Valued Data Experiments: To compare our method to the utility approach, we consider discrete data. The first dataset is a quantized (with 20 levels) Gaussian mixture synthetic data in two dimension. The 1st dimension is stage one; the 2nd dimension is stage two. The second dataset is Mammogram Mass from UCI Machine Learning Repository. It is used to predict the severity of a mammographic mass lesion (malicious or benign). It contains 3 attributes extracted from the CAD image and also an evaluation by a radiologist on a confidence scale in addition to the true biopsy results. The first stage are features extracted from the CAD image, and the second stage is the expert confidence rated on a discrete scale $1-5$. Automatic analysis of the CAD image is cheaper than employing an opinion of a radiologist.

Simulations in Fig. 6 demonstrate that utility performs worse when compared to our approach. This is possibly due to poor probability estimates in limited data setting.

4. While there are many different ways to estimate a probability likelihood we used a Gaussian mixture due to its computational efficiency
5. stump classifier is threshold on dth dimension: $h_{d,g,\{+1/-1\}}(x) = \{+1/-1\}sign(x(d) - g)$



(a) GaussMixQuant (b) mammogram

Figure 6: Comparison of Global to Utility on (a) quantized two gaussian clusters and (b) mammogram dataset. Reject Rate vs System Error. Reject Rate is the fraction of examples with measurements from both stages. Our approach outperforms Utility possibly because we do not need to estimate probability likelihoods

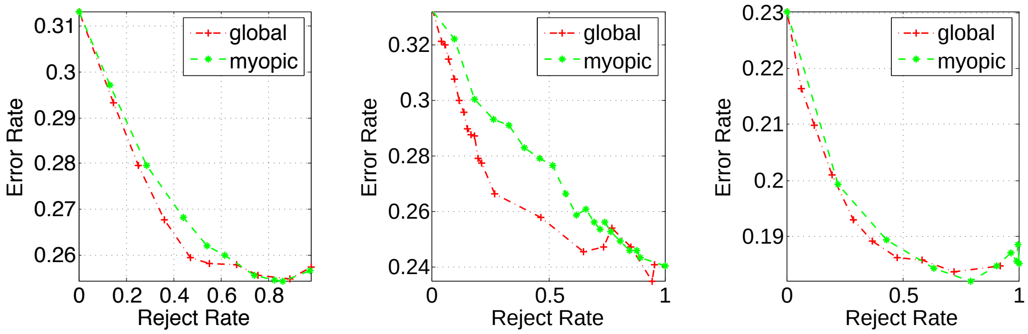
Continuous Valued Data Experiments We compare our global method to the myopic method on three datasets. The Pima Indians Diabetes Dataset (UCI MLR) consists of 8 measurements. 6 of the measurements are inexpensive to acquire and consist of simple tests such as body mass index, age, pedigree. These we designate as the first stage. The other two measurements constitute the second stage and require more expensive procedures.

The polyp dataset consists of hyper-spectral measurements of colon polyps collected during colonoscopies (Rodríguez-Díaz and Castañón (2009)). The attribute is a measured intensity at 126 equally spaced frequencies. Finer resolution requires higher photon count which is proportional to acquisition time. For a first stage, we use a coarse measurement downsampled to only 12 frequency bins. The second stage is the full resolution frequency response. Using the course measurements is cheaper than acquiring the full resolution.

The threat dataset contains images taken of people wearing various explosives devices. The imaging is done in three modalities: infrared (IR), passive millimeter wave (PMMW), and active millimeter (AMMW). All the images are registered. We extract many patches from the images and use them as our training data. A patch carries a binary label, it either contains a threat or is clean. IR and PMMW are the fastest modalities but also less informative. AMMW requires raster scanning a person and is slow but also the most useful.

In Fig. 7, global performs better than margin in most cases. On threat data, margin appears to be doing just marginally worse than global, however, we get only a few points on the curve with reject rates less than 50%. Due to the heuristic nature of margin, we cannot construct a multistage classifier with an arbitrary reject rate.

The goal is to reach the performance of a centralized classifier (100% reject rate) while utilizing the 2nd stage sensor only for a small fraction of examples. Overall, the results demonstrate the benefit of multi-stage classification: rejection rate can be set to less than 50% with only small sacrifices in performance. For the mammogram data, this implies that for half of the patients a diagnoses can be made solely by an automatic analysis of a CAD image without an expensive opinion of a radiologist. For the Pima data, similar error can be achieved without an expensive medical procedures. For the polyps dataset, a fast low resolution measurement is enough to classify a large fraction of patience. In the threat dataset, IR and PMMW are sufficient to decide whether or not a threat is present for the majority of instances without requiring a person to go through a slower AMMW scanner.



(c) pima (d) polyps (e) threat

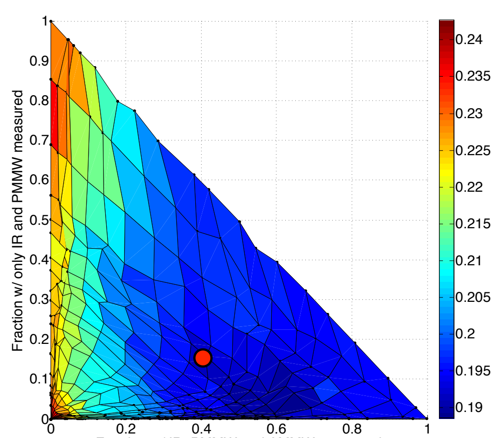
Figure 7: Three datasets are evaluated: pima, polyps and threat. Reject Rate vs Error Rate for a varying reject cost δ. Reject Rate is the fraction of examples with measurements from both stages. Global and Myopic are compared. Global (our approach) has a better performance over all while Myopic does better in some situations.

	1st Stage	Centralized	**Global**		Utility		Myopic	
Rej. Rate:	0%	100%	30%	50%	30%	50%	30%	50%
GaussMix	15.9	9.4	**10.3**	**9.5**	11.4	10.1		
Mammogram	20.1	16.5	**17.2**	**17.1**	18.6	17.5		
Pima	31.3	25.7	**27.5**	**25.9**			27.8	26.5
Polyps	33	24	**26.6**	**25.7**			29.1	27.9
Threat	23	18.5	**19**	**18.6**			19.3	18.8

Figure 8: Comparison of algorithms. Selected error for fixed reject rates of 30 and 50 % (quantative view of the curves)

Three Stages: Lastly, we demonstrate a three stage system, we apply our algorithm to three stages of threat dataset. Note for margin it is unclear how to generalize it to a multistage scenario and there is no way to define reject costs for different stages. We set the first stage to be IR, second PMMW and AMMW as third. There is no cost for acquiring IR. We vary the costs for the PMMW (2nd) stage, δ_1, and AMMW (3rd), δ_2, to generate an error map (color in Fig. 9). A point on the map corresponds to a performance of a particular multistage classification strategy. The vertical axis is the fraction of examples for

Figure 9: Three Stage System. The color maps error. A point on the map corresponds to a performance of a particular multistage classification strategy. The vertical axis is the fraction of examples for which only IR and PMMW measurements are used in making a decision. The horizontal axis is the fraction of examples for which all three modalities are used. An example red point in the figure, {.4, .15, .195}, correspond to a system where 40% of examples use IR and PMMW, 15% use only IR and the rest of data (45%) use all the modalities. And this strategy achieves a system error rate of 19.5%.



which only IR and PMMW measurements are used in making a decision. The horizontal axis is the fraction of examples for which all three modalities are used. For example, a red point in the figure, $\{.4, .15, .195\}$, correspond to a system where 40% of examples use IR and PMMW, 15% use only IR and the rest of data (45%) use all the modalities. And this strategy achieves a system error rate of 19.5%. Note that the support lies below the diagonal. This is because the sum or reject rates has to be less than one. Results demonstrate some interesting observations. While best performance (about 19%) is achieved when all the modalities are used for every example, we can move along the vertical lines and allow a fraction to be classified by IR and PMMW, avoiding AMMW all together. This strategy achieves performance comparable to a centralized system, (IR+PMMW+AMMW).

5. Conclusion

In this paper, we propose a general framework for a sequential decision system in a non-parametric setting. Starting from basic principles, we derive the bayesian optimal solution. Then, to simplify the problem, we parameterize a classifier at each stage in terms of two binary decisions. We formulate an ERM problem and optimize it by alternatively minimizing one stage at a time. Remarkably, all subproblems turn out to be weighed binary error minimizations. We introduce a practical boosting algorithm that minimizes a global surrogate of the empirical risk and test it on several datasets. Results show the advantage of our formulation to more heuristic approaches. Overall, our experiments demonstrate how multi-stage classifiers can achieve good performance by acquiring full measurements only for a fraction of samples.

Acknowledgments

This work is supported by the U.S. Department of Homeland Security 2008-ST-061-ED0001

References

P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 2008.

P. Bartlett, Y. Freund, W. S. Lee, and R. E. Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Annals of Stat.*, 1998.

M. Bilgic and L. Getoor. Voila: Efficient feature-value acquisition for classification. In *AAAI*, 2007.

M. Chen, Z. Xu, K. Q. Weinberger, O. Chapelle, and D. Kedem. Classifier cascade: Tradeoff between accuracy and feature evaluation cost. In *AISTATS*, 2012.

C. Chow. On optimum recognition error and reject tradeoff. *Inf. Theory, IEEE*, 1970.

L.P. Cordella and C. Sansone. A multi-stage classification system for detecting intrusions in computer networks. *Pattern Anal. Appl.*, 2007.

R. El-Yaniv and Y. Wiener. Agnostic selective classification. In *NIPS*, 2011.

W. Fan, W. Lee, S. J. Stolfo, and M. Miller. A multiple model cost-sensitive approach for intrusion detection. In *ECML*, 2000.

W. Fan, F. Chu, H. Wang, and P. S. Yu. Pruning and dynamic scheduling of cost-sensitive ensembles. In *AAAI*, 2002.

Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu. Support vector machines with a reject option. In *NIPS*, 2008.

S. Ji and L. Carin. Cost-sensitive feature acquisition and classification. In *Pattern Recognition*, 2007.

P. Kanani and P. Melville. Prediction-time active feature-value acquisition for cost-effective customer targeting. In *NIPS*, 2008.

A. Kapoor and E. Horvitz. Breaking boundaries: Active information acquisition across learning and diagnosis. In *NIPS*, 2009.

W. Lee, W. Fan, M. Miller, S. J. Stolfo, and E. Zadok. Toward cost-sensitive modeling for intrusion detection and response. *J. Comput. Secur.*, 2002.

Li-Ping Liu, Yang Yu, Yuan Jiang, and Zhi-Hua Zhou. Tefe: A time-efficient approach to feature extraction. In *ICDM*, 2008.

D. J.C. MacKay. Information-based objective functions for active data selection. *Neural Comp.*, 1992a.

H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *NIPS*, 2009.

E. Rodríguez-Díaz and D.A. Castañón. Support vector machine classifiers for sequential decision problems. In *IEEE CDC*, 2009.

V. S. Sheng and C. X. Ling. Feature value acquisition in testing: A sequential batch test algorithm. In *ICML*, pages 809–816, 2006.

K. Trapeznikov, V. Saligrama, and D. A. Castañon. Cost sensitive sequential classification. Technical report, Arxiv, http://arxiv.org/abs/1205.4377, 2012.

P. Viola and M. Jones. Robust real-time object detection. In *Int. J. of Comp. Vis*, 2001.

H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *KDD*, 2003.

C. Yuan and D. Casasent. A novel support vector classifier with better rejection performance. In *CVPR*, 2003.

C. Zhang and Z. Zhang. A survey of recent advances in face detection. In *Microsoft Research Technical Report*, 2010.

V. B. Zubek and T. G. Dietterich. Pruning improves heuristic search for cost-sensitive learning. In *ICML*, 2002.