Learning to Aggregate Information for Sequential Inferences

Diyan Teng and Emre Ertin

May 4, 2016

Abstract

We consider the problem of training a binary sequential classifier under an error rate constraint. It is well known that for known densities, accumulating the likelihood ratio statistics is time optimal under a fixed error rate constraint. For the case of unknown densities, we formulate the learning for sequential detection problem as a constrained density ratio estimation problem. Specifically, we show that the problem can be posed as a convex optimization problem using a Reproducing Kernel Hilbert Space representation for the log-density ratio function. The proposed binary sequential classifier is tested on synthetic data set and UC Irvine human activity recognition data set, together with previous approaches for density ratio estimation. Our empirical results show that the classifier trained through the proposed technique achieves smaller average sampling cost than previous classifiers proposed in the literature for the same error rate.

1 Introduction

Sequential decision strategies outperform their fixed sample size counterparts in achieving same decision risk using less number of samples on the average. Initially, developed by Wald [1] to reduce the number of inspections in industrial quality control, it becomes widely used in clinical studies to reduce the average number of patients that are undergoing potentially risky treatments. Even when the cost of samples are not a major concern, sequential techniques can be used to reduce the computational cost of obtaining relevant information from a data sample. Thus sequential test is still a method of great potential in any time sensitive scenario. For example, in many computer vision problems, more sophisticated feature is usually expensive and slow to obtain even though they provide higher accuracy. Therefore cascading classifier such as Viola-Jones[2] is widely used due to their sequential nature.

For the case of known class conditional densities accumulating likelihood statistics and comparing with fixed thresholds minimizes the average stopping time under fixed error constraints. In this paper, we consider the case where

the class conditional densities generating the data is unknown and sequential decision rule has to be learned directly from labeled data samples. While there exists plethora of supervised learning algorithms to learn fixed sample test rules using parametric and non-parametric forms, there exist relatively few algorithms designed to learn to perform sequential classification. Unlike the single sample classification problems where only the decision boundary is critical, sequential decision rules require a mapping from sample space to a state space for aggregation of evidence and making stopping rules. To be concrete we focus on the specific problem of learning a binary sequential classifier. The class conditional distribution is assumed to be unknown, but identical and conditionally independent over time resulting in a stationary decision/aggregation rule. For temporal aggregation of information across samples constructing an estimate of the likelihood or posterior probability emerges as an obvious framework for constructing sequential rule.

The information summarizing problem itself has been discussed in [3] and the reference therein without considering sequential testing scenario. In the same framework of this paper, Sochman and Matas [4] constructed a likelihood ratio function estimator using Adaboost [5, 6] to perform binary sequential classification based on accumulation and thresholding of the likelihood ratio estimate, resulting in an algorithm called Wald-Boost algorithm. Similarly, other methods of constructing density ratio estimates based on maximizing information theoretic functionals [7, 8] can be employed to perform sequential decisions.

However, the optimization criteria used by these methods for constructing likelihood ratio functions estimates are not directly related to the performance in sequential detection. We note that errors in the likelihood estimate effect the average stopping time and error probabilities in a non-trivial way due to accumulation of errors across samples. Kuh *et al.* [9, 10] used reinforcement learning methods to propagate errors in terminal decisions to adjust weights in a parametric likelihood ratio function estimate to learn binary sequential classifiers. However, again stopping time is not considered as a direct optimization criteria. In this paper we derive a variational bound on the sampling cost of SPRT and associated non-parametric log-density ratio estimate which minimizes this bound. Our empirical results show that the sequntial classifier trained through the proposed technique achieves smaller average sampling cost than learned sequential tests employing likelihood function estimates proposed in the literature.

2 Problem Statement

In this paper, the problem of learning a binary sequential detector from training data is studied. The training data consists of M samples $\left\{\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_M^{(0)}\right\}$ from class 0, and N samples $\left\{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_N^{(1)}\right\}$ from class 1, sampled i.i.d. with unknown densities $p_0(x)$ and $p_1(x)$ respectively. Each sample $\mathbf{x}_n^{(c)} \in \mathbb{R}^d$ is a d dimensional feature vector from class c. The learning problem is to

design a sequential decision making mechanism which consists of an information aggregating rule, a stopping criterion and a decision rule to make a terminal decisions between the two hypotheses $\{\mathrm{H}_0, \mathrm{H}_1\}$ regarding the density used to generate series of samples in a test set. Here, the information aggregating rule is assumed to be stationary and only use a one dimensional state variable summarizing information received up to current sample regarding prevailing class label. Recall that in the classic setting for sequential detection with known class conditional density, Sequential Probability Ratio Test (SPRT) minimizes stopping time for both classes under constraints on miss detection and false alarm probability [1]. SPRT compares the cumulative the log-likelihood ratio with fixed thresholds to choose between terminal decision or continue to sample:

$$\begin{aligned}
&\text{Stop and Declare } \mathrm{H}_0 \text{ if:} \\
&\qquad \sum_i \log \frac{p_1(x_i)}{p_0(x_i)} \leq a(\mathrm{P_F}, \mathrm{P_M}) \\
&\text{Stop and Declare } \mathrm{H}_1 \text{ if:} \\
&\qquad \sum_i \log \frac{p_1(x_i)}{p_0(x_i)} \geq b(\mathrm{P_F}, \mathrm{P_M}) \\
&\text{continue sampling if:} \\
&\qquad a(\mathrm{P_F}, \mathrm{P_M}) < \sum_i \log \frac{p_1(x_i)}{p_0(x_i)} < b(\mathrm{P_F}, \mathrm{P_M})
\end{aligned}$$

where a and b are respectively the lower and upper terminating boundaries. Under zero-overshoot assumption on the accumulated likelihood at stopping time , the expected sampling cost for standard binary SPRT is given in [11] as:

$$\begin{aligned}
\mathrm{N}_0 &= \frac{1}{\mathrm{D}_{01}}[\mathrm{P_F} \log \frac{\mathrm{P_F}}{1-\mathrm{P_M}} + (1-\mathrm{P}_F) \log \frac{1-\mathrm{P_F}}{\mathrm{P_F}}] \\
\mathrm{N}_1 &= \frac{1}{\mathrm{D}_{10}}[\mathrm{P_M} \log \frac{\mathrm{P_M}}{1-\mathrm{P_F}} + (1-\mathrm{P_M}) \log \frac{1-\mathrm{P_M}}{\mathrm{P_F}}]
\end{aligned} \tag{1}$$

Inspired by the structure of the SPRT, an appealing choice for learning a binary sequential detector is to construct a function estimate for the likelihood ratio function from training samples and design termination and decision rule as threshold comparisons as in SPRT. Directly estimating class conditional density function independently and computing the ratio results in poor performance [8] since fit errors in different regions of the sample space is emphasized when dividing the two densities. A number of techniques have been suggested in the literature for directly estimating density ratio functions which can be characterized into three classes: parametric approaches [12] that assume a parametric form and use regression methods to fit through maximization of binomial likelihood on the training data, boosting based methods [4] that rely on asymptotic properties of weighted sum of weak learners and non-parametric techniques[7, 8]

that construct likelihood ratio function estimates through maximization of information theoretic divergence metrics on the training data.

An estimate of log-posterior ratio can be converted to density ratio estimate through canceling the term produced by the prior ratio. For example, a common parametric form for log-posterior ratio is the additive functions of training samples [12]:

$$\log \frac{p(\mathrm{H}_1|\mathbf{x})}{p(\mathrm{H}_0|\mathbf{x})} = \sum_m f_m(\mathbf{x};\gamma_m) = F(\mathbf{x};\gamma) \tag{2}$$

then the estimate for the density ratio can be formed as:

$$\hat{r}(\mathbf{x}) = e^{F(\mathbf{x};\gamma)} \frac{p(\mathrm{H}_0)}{p(\mathrm{H}_1)} \tag{3}$$

the parameter vectors $\{\gamma_m\}_m$ are typically through maximization of the binomial log-likelihood function. When the model was specified correctly, the solution has the asymptotic optimality of a maximum likelihood estimator. Interestingly, as shown by Friedman *et al.* [13] boosting approaches [6, 5] that combine binary decision of weak classifiers to train classifiers with improved performance can be analyzed under the same framework of fitting an additive model through maximization of likelihood. Specifically consider weighted sum of binary decisions from weak classifier outputs $\{f_i(\mathbf{x})\}$ with associated weights $\{c_i\}$:

$$F_A(\mathbf{x}) = \sum_i c_i f_i(\mathbf{x}) \tag{4}$$

The function $F_A(\mathbf{x})$ represents the aggregate decision of the ensemble of weak classifier. The design process is to iteratively add new classifier to the existing ones while optimizing the weights associated with them. In boosting, each new weak classifier is tasked to minimized a weighted classification error for the training set, in which higher weights are assigned to incorrectly classified samples using current classifier. Friedman *et al.* [13] have shown that the iterative weighted minimization procedure is equivalent to minimizing the expected exponential error $\mathrm{E}(e^{-yF(\mathbf{x})})$, which is a second order approximation to the binomial log-likelihood function. And they pointed out that the density ratio can be retrieved from the final classifier through:

$$\hat{r}(\mathbf{x}) = e^{2F_A(\mathbf{x})} \frac{p(\mathrm{H}_0)}{p(\mathrm{H}_1)} \tag{5}$$

The boosting approach is fast with good empirical performance and is resistant to over-fitting and provides a fast approach for constructing density ratio estimates.

In principle all these methods provide density ratio function estimates that can be used to form a binary sequential classifier incorporating into the SPRT structure. However, the optimization criteria employed in these techniques is decoupled from the performance of these function estimates in a sequential

decision task. This mismatch results in sub-optimal performance as illustrated by our empirical experiments in Section 4.

In a novel direction, Nguyen *et al.* [7] derived variational characterizations of f-divergences which enabled estimation of divergence functionals and likelihood ratios through convex risk minimization. Following this work, we derive a variational bound on the expected sampling cost of SPRT (with known densities) and obtain an associated density ratio estimate $\hat{r}(\mathbf{x})$. Next, using a Reproducing Kernel Hilbert Space representation for the log-density ratio function we obtain a convex optimization approach for fitting density ratio estimate $\hat{r}(\mathbf{x})$ to training data dubbed as Wald Kernel Density Ratio Fit (WKDRF).

3 Wald Kernel Density Ratio Fit (WKDRF)

In this section, we formulate a new algorithm for learning log-density ratio function estimates that are tailored for performing sequential binary detection in the SPRT decision structure of accumulation and thresholding. Our goal is to form a log-density ratio estimate such that the resulting sequential decision structure minimizes the average stopping time (or equivalently expected number of samples) for a desired level of probability of error. Towards that end, we first extend known results on SPRT error probabilities [11] to the case of learned sequential test based on a given density ratio estimate $\hat{r}(\mathbf{x})$.

Theorem 1. *In a learned SPRT, if the estimated density ratio function $\hat{r}(\cdot)$ is not constant and normalized as:*

$$\mathrm{E}[\hat{r}|\mathrm{H}_0] = 1 \quad \textit{and} \quad \mathrm{E}[\hat{r}^{-1}|\mathrm{H}_1] = 1 \tag{6}$$

then for fixed lower and upper thresholds a and b on log-likelihood ratio, probability of false alarm and miss detection of terminal decisions is given by :

$$\mathrm{P_F} = \frac{1-e^a}{e^b - e^a} \quad \textit{and} \quad \mathrm{P_M} = \frac{e^a(e^b-1)}{e^b-e^a} \quad . \tag{7}$$

We note that learned SPRT performance with normalized density ratio estimates matches SPRT performance, albeit with a potentially longer average stopping time. To prove Theorem 1, the following Lemma is required.

Lemma 1. *Let $\hat{z}_i = \log \hat{r}(\mathbf{x}_i)$, $\hat{\Lambda}_n = \sum_{k=1}^{n} \hat{z}_k$ and $\hat{G}(u) = \mathrm{E}[e^{u\hat{z}}]$, under both hypotheses the following process is a Martingale:*

$$\hat{M}_n = \frac{e^{u\hat{\Lambda}_n}}{\hat{G}(u)^n} \tag{8}$$

Proof. First of all, it is easy to check $\hat{M}_0 = 1$. Next, one can verify that:

$$\begin{aligned}
\mathrm{E}[\hat{M}_{n+1}|\hat{M}_k, 0 \le k \le n] &= \mathrm{E}[\frac{e^{u\hat{\Lambda}_{n+1}}}{\hat{G}(u)^{n+1}}|\hat{M}_k, 0 \le k \le n] \\
&= \mathrm{E}[\frac{e^{u\hat{z}_{n+1}}}{\hat{G}(u)} \cdot \hat{M}_n|\hat{M}_k, 0 \le k \le n] \\
&= \hat{M}_n \cdot \frac{\mathrm{E}[e^{u\hat{z}_{n+1}}]}{\hat{G}(u)} \\
&= \hat{M}_n
\end{aligned}$$

And since $\hat{M}_n$ are all positive valued, we have:

$$\mathrm{E}[|\hat{M}_n|] = \mathrm{E}[\hat{M}_n] = \mathrm{E}[\hat{M}_0] = 1$$

Thus we proved the process $\hat{M}_n$ satisfies the two properties to be a Martingale. □

Next we prove Theorem 1.

Proof. Define $\hat{G}_0(u) = \mathrm{E}[e^{u\hat{z}}|\mathrm{H}_0]$ and $\hat{G}_1(u) = \mathrm{E}[e^{u\hat{z}}|\mathrm{H}_1]$. The special case of $\hat{r} = 1$ satisfies both constraints, but when $\hat{r} = 1$ the test never stops. Other than that, there is no constantly valued $\hat{r}(\cdot)$ satisfies both constraints. When $\hat{r}$ is not constantly 1, the test will stop at finite time. Let N be the random stopping time, then we have the two types of error when the test stops:

$$\hat{\mathrm{P}}_{\mathrm{F}} = \Pr\{\hat{\Lambda}_N \ge b|\mathrm{H}_0\} \quad \text{and} \quad \hat{\mathrm{P}}_{\mathrm{M}} = \Pr\{\hat{\Lambda}_N \le a|\mathrm{H}_1\}$$

For the special case of $u = 0$, $u = -1$ and $u = 1$:

$$\hat{G}_0(0) = 1, \quad \hat{G}_1(0) = 1$$

and

$$\hat{G}_0(1) = \int \hat{r}(\mathbf{x})p_0(\mathbf{x})\mathrm{d}\mathbf{x}, \quad \hat{G}_1(-1) = \int \hat{r}(\mathbf{x})^{-1}p_1(\mathbf{x})\mathrm{d}\mathbf{x}$$

Now one can evaluate the expected value of the Martingale $\hat{M}_N$ at the stopping time N under H_0 with $u = 1$, which gives:

$$\begin{aligned}
& \mathrm{E}[\frac{e^{\hat{\Lambda}_N}}{\hat{G}_0(1)^N}|\mathrm{H}_0] = 1 \\
\overset{\text{when } \mathrm{E}[\hat{r}|\mathrm{H}_0]=1}{\Leftrightarrow} \quad & \mathrm{E}[e^{\hat{\Lambda}_N}|\mathrm{H}_0] = 1 \\
\overset{\text{applying zero-overshooting assumption}}{\Leftrightarrow} \quad & \hat{\mathrm{P}}_{\mathrm{F}}B + (1 - \hat{\mathrm{P}}_{\mathrm{F}})A = 1 \\
\Leftrightarrow \quad & \hat{\mathrm{P}}_{\mathrm{F}} = \frac{1 - A}{B - A}
\end{aligned}$$

where $A = e^a$ and $B = e^b$. Similarly, by constraining $\mathrm{E}[\hat{r}^{-1}|\mathrm{H}_1] = 1$, one can get:

$$\hat{\mathrm{P}}_{\mathrm{M}} = \frac{A(B-1)}{B-A}$$

□

Next, following the approach used in [7] to develop estimate of divergence functionals and likelihood ratio functionals, we derive a variational upper bound on the sampling cost of SPRT, which reveals a density ratio function estimate linked to the sequential test performance.

Theorem 2. *The average stopping time for SPRT can be upper bounded by the solution of the following problem:*

$$\begin{aligned} &\min_{\hat{r}} \frac{\omega_0}{\int p_0 \log(\hat{r})} - \frac{\omega_1}{\int p_1 \log(\hat{r})} \\ &s.t. \int \hat{r} p_0 = 1, \quad \int \hat{r}^{-1} p_1 = 1 \end{aligned} \tag{9}$$

Proof. Recall that the expected number of sample in the standard SPRT is given in (1). Since the terms inside the bracket are constant after fixing the error rate, we define the following two constants for simplicity:

$$\omega_0 = \pi_0[\mathrm{P}_{\mathrm{F}} \log \frac{\mathrm{P}_{\mathrm{F}}}{1-\mathrm{P}_{\mathrm{M}}} + (1-\mathrm{P}_F) \log \frac{1-\mathrm{P}_{\mathrm{F}}}{\mathrm{P}_{\mathrm{F}}}]$$

and

$$\omega_1 = \pi_1[\mathrm{P}_{\mathrm{M}} \log \frac{\mathrm{P}_{\mathrm{M}}}{1-\mathrm{P}_{\mathrm{F}}} + (1-\mathrm{P}_{\mathrm{M}}) \log \frac{1-\mathrm{P}_{\mathrm{M}}}{\mathrm{P}_{\mathrm{F}}}]$$

where π_0 and π_1 are the prior probability of H_0 and π_1. The standard SPRT sampling cost then can be written as:

$$\begin{aligned} \mathrm{C} &= \frac{\omega_0}{\mathrm{D}_{01}} + \frac{\omega_1}{\mathrm{D}_{10}} \\ &= \frac{\omega_0}{\int p_0 \log r} + \frac{\omega_1}{\int p_1 \log r^{-1}} \end{aligned} \tag{10}$$

Applying similar method as [7], the cost objective can be upper bounded using the convex conjugate formula for $-\log(\cdot)$ function which is:

$$-\log(x) \Leftrightarrow -(1 + \log(-x^*)) \tag{11}$$

as:

$$\mathrm{C} = \frac{\omega_0}{\int p_0 \log r} + \frac{\omega_1}{\int p_1 \log r^{-1}} \tag{12}$$

$$\leq \frac{\omega_0}{\sup_g \int p_0 (g \cdot r + \log(-g) + 1)} + \frac{\omega_1}{\sup_f \int p_1 (f \cdot r^{-1} + \log(-f) + 1)} \tag{13}$$

$$\leq \inf_f \frac{\omega_0}{\int \frac{1}{f} p_1 - p_0 \log(-f) + p_0} + \frac{\omega_1}{\int f p_0 + p_1 \log(-f) + p_1} \tag{14}$$

Using the two constraints (6) and defining the density ratio estimate as $\hat{r} = \frac{1}{-f}$ we obtain the variational bound given in (9).

□

3.1 Kernel Based Density Ratio Fitting

If we adopt a Reproducing Kernel Hilbert Space representation for the log-density ratio function and replace the class conditional densities with empirical distributions defined by the training data, the variational problem given in (9) is equivalent to:

$$\begin{aligned} &\min_{\hat{r}} \frac{\omega_0}{\sum_{j=1}^{M} \log(\hat{r}(\mathbf{x}_j^{(0)}))} - \frac{\omega_1}{\sum_{i=1}^{N} \log(\hat{r}(\mathbf{x}_i^{(1)}))} \\ &\text{s.t.} \sum_{j=1}^{M} \hat{r}(\mathbf{x}_j^{(0)}) = 1, \quad \sum_{i=1}^{N} \hat{r}^{-1}(\mathbf{x}_i^{(1)}) = 1 \end{aligned} \tag{15}$$

Next, we impose the Reproducing Kernel Hilbert Space(RKHS) structure to the log-density ratio function. Any function in RKHS can be written as an inner product form of:

$$f(\cdot) =< \omega, \Phi(\cdot, \mathbf{u}) >= \sum_{c=1}^{C} \alpha_c K(\cdot, \mathbf{u}_c)$$

where $K(\cdot, \cdot)$ is the kernel function. In this paper, we choose Gaussian kernel with randomly sampled centers as suggested in [8]. Since the objective function is a pointwise cost whose minimizer is not unique and could even be infinite dimensional, we add a regularization term to penalize the l_2 norm of the estimated log-likelihood ratio function which gives:

$$\begin{aligned} &\min_{\alpha} \frac{\omega_0}{\sum_{j=1}^{M} \sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_j^{(0)} - \mathbf{x}_c\|^2}{\sigma^2})} - \frac{\omega_1}{\sum_{i=1}^{N} \sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_i^{(1)} - \mathbf{x}_c\|^2}{\sigma^2})} + \frac{\lambda}{2} \alpha^{\mathrm{T}} \mathbf{K} \alpha \\ &\text{s.t.} \sum_{j=1}^{M} e^{\sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_j^{(0)} - \mathbf{x}_c\|^2}{\sigma^2})} = 1, \quad \sum_{i=1}^{N} e^{-\sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_i^{(1)} - \mathbf{x}_c\|^2}{\sigma^2})} = 1 \end{aligned} \tag{16}$$

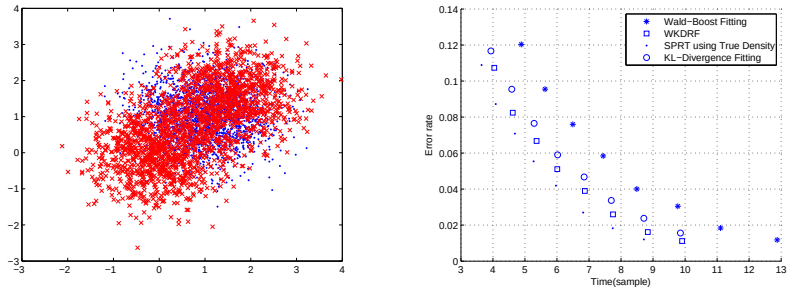
The equality constraints can be relaxed to inequality constraints to obtain a convex optimization problem:

$$\begin{aligned} &\min_{\alpha} \frac{\omega_0}{\sum_{j=1}^{M} \sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_j^{(0)} - \mathbf{x}_c\|^2}{\sigma^2})} - \frac{\omega_1}{\sum_{i=1}^{N} \sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_i^{(1)} - \mathbf{x}_c\|^2}{\sigma^2})} + \frac{\lambda}{2} \alpha^{\mathrm{T}} \mathbf{K} \alpha \\ &\text{s.t.} \sum_{j=1}^{M} e^{\sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_j^{(0)} - \mathbf{x}_c\|^2}{\sigma^2})} \leq 1, \quad \sum_{i=1}^{N} e^{-\sum_{c=1}^{C} \alpha_c \exp(-\frac{\|\mathbf{x}_i^{(1)} - \mathbf{x}_c\|^2}{\sigma^2})} \leq 1 \end{aligned} \tag{17}$$

where $\mathbf{K}$ is the kernel matrix with $\mathbf{K}(i,j) = \exp(-\frac{\|\mathbf{x}_{c_i} - \mathbf{x}_{c_j}\|^2}{\sigma^2})$. One may observe that since the two denominator terms are both linear functions of α, convexity preserving rule for composition of functions guarantees that the objective being convex in α as long as α is properly initialized. Specifically, $\frac{1}{\cdot}$ is a convex non-increasing function for positive valued denominator and the linear function is concave, resulting in the composite function being convex when the denominator is positive. In addition, we need to guarantee that the first term in (17) has a positive denominator while the second term has a negative denominator. This can be easily done by performing a proper initialization. Since those exponential function coefficients can be viewed as the normal of the hyperplane in terms of α, in (17) we need to choose the α vector such that it lies in the region that gives proper inner product value for both term. A natural yet simple choice of initial α could be the normalized equipartitioning vector of the two normal vectors. The resulting parameter vector α defines the estimator of the log-density ratio function, which summarizes each observation into a log-likelihood to be used in a learned SPRT. The testing phase is exactly the same as standard SPRT with known density, except that in the learned test the estimated density ratio function is used as the information aggregation mapping. The resulting learned SPRT automatically satisfies the error constraints with appropriately chosen thresholds as shown in Theorem 1.

4 Experimental Results

We compare the performance of the learned SPRT using WKDR fitting with the performance of Wald-Boost [4], which is based on AdaBoost training of the density ratio function, and the learned SPRT employing KL-divergence density ratio fit [7] which fits the density ratio by maximizing the lower bound to the one sided KL-divergence. The kernel width in our method and KL-divergence fitting method is chosen using cross validation.



(a) Training data (b) Performance curve

Figure 1: Synthetic example

We first tested the algorithm in a synthetic data set. In this example, H_0

samples are single component Gaussian random vector $\mathcal{N}(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix})$ and H_1 samples are Gaussian mixture as $\frac{1}{2}\mathcal{N}(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix})+\frac{1}{2}\mathcal{N}(\begin{bmatrix}1.5\\1.5\end{bmatrix}, \begin{bmatrix}0.5 & 0\\0 & 0.5\end{bmatrix})$. We used 2,000 samples from each class in training and sequentially draw samples in testing. For simplicity, we choose the termination boundary to be symmetric and the prior probability of the two hypotheses being equal. We use 25 randomly picked samples as kernel centers in both WKDRF and KL-Divergence Fitting methods. We use 200 stumps as weak classifier in Wald-Boost. In testing phase, the same sample is feed into all methods until they terminate, and the termination time is recorded. A scatter plot for the dataset is given in Figure 1a, and the empirical performance result is plotted in Figure 1b. The proposed method outperforms both the KL-divergence based method and outperform Wald-Boost in this example achieving lower sampling cost for a given probability of error.

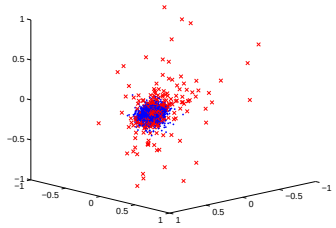
The second example we present is human activity recognition. The data set we used is the smartphone recorded human activity data from the UC Irvine Machine Learning Repository website [14]. Two feature sets are used in our evaluation: Features 1-3 which is the mean accelerometer value and Features 294-296 which is the mean frequency domain accelerometer value. We consider two classification tasks: 1) Classification task to determine whether a subject is moving or static, 2) Classification of subject that are moving on staircase to walking upstairs or downstairs. The size of training data is 3285 and 4067 respectively for moving v.s. static test, and 1073 and 986 for up v.s. down test. We picked 50 randomly chosen samples as kernel centers. Also the number of stumps used in Wald-Boost is 200. The data set is plotted in Figure 2a-d, and the results are in Figure 2e-h. Again in both classification tasks, our method outperforms the other methods. Notably, WaldBoost outperforms KL divergence based method in this learning task.

We note that, if the true density ratio and its inverse are indeed in the RKHS function class, then KL-divergence density ratio fitting would result identical log-density ratio estimates as our proposed method. Under model mismatch for the log-density ratio function, our optimization criteria balances the two errors in the SPRT expression to choose the density estimate, arguably resulting in better performance in sequential tasks.

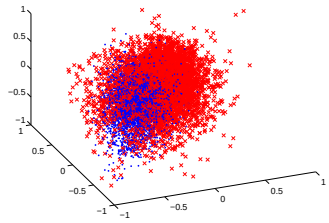
5 Conclusion

In this work, we proposed a method for learning binary sequential tests based on a optimizing a variational bound on sampling cost of SPRT. The proposed algorithm results in an convex program that can be solved efficiently. Experimental results show that the proposed algorithm outperforms previously proposed techniques achieving smaller stopping time for a given error rate. A potential direction for future work is characterization of the distance metric between the true and estimated log-density ratio metrics when the optimization criteria in 9 is utilized and use this metric to study convergence of the proposed method as

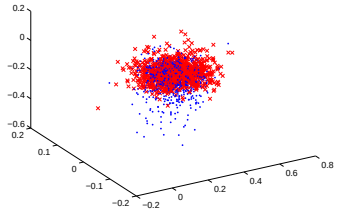
the number of training samples increase.



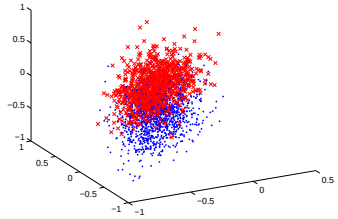
(a) Moving v.s. static training feature 1-3



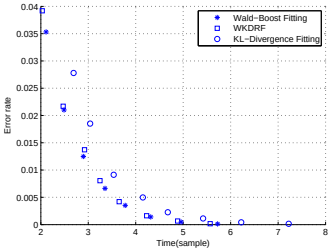
(b) Moving v.s. static training feature 294-296



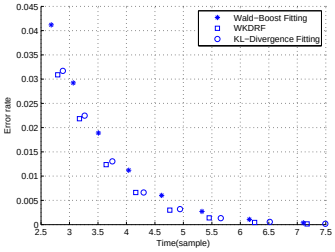
(c) Walking upstairs v.s. downstairs training feature 1-3



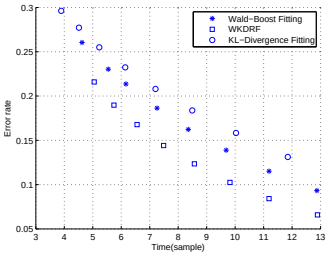
(d) Walking upstairs v.s. downstairs training feature 294-296



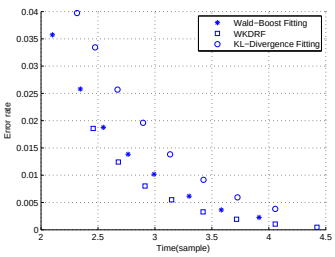
(e) Performance curve, moving v.s. static using feature 1-3



(f) Performance curve, moving v.s. static using feature 294-296



(g) Performance curve, walking upstairs v.s. downstairs using feature 1-3



(h) Performance curve, walking upstairs v.s. downstairs using feature 294-296

Figure 2: Human activity classification

References

[1] A. Wald, *Sequential analysis*. John Wiley and Sons, New York, 1947.

[2] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. I–511–I–518.

[3] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[4] J. Sochman and J. Matas, "Waldboost-learning for time constrained sequential detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 150–156.

[5] Y. Freund, "Boosting a weak learning algorithm by majority," *Information and computation*, vol. 121, no. 2, pp. 256–285, 1995.

[6] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, no. 2, pp. 197–227, 1990.

[7] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[8] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori, "Approximating mutual information by maximum likelihood density ratio estimation," in *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, vol. 4, 2008, pp. 5–20.

[9] C. Guo and A. Kuh, "Temporal difference learning applied to sequential detection," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 278–287, 1997.

[10] A. Kuh and D. Mandic, "Sequential detection using least squares temporal difference methods," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, 2006, pp. V–701–V–704.

[11] B. C. Levy, *Principles of signal detection and parameter estimation*. Springer, 2008.

[12] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*. CRC Press, 1990, vol. 43.

[13] J. Friedman, T. Hastie, R. Tibshirani *et al.*, "Additive logistic regression: a statistical view of boosting," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[14] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013.