Learning for Sequential Information Fusion: Wald Kernel Density Ratio Fitting

Diyan Teng and Emre Ertin
The Ohio State University

The Ohio State University

Problem Statement

Given samples $\{\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_M^{(0)}\}$ and $\{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_N^{(1)}\}$ from hypothesis H_0 and H_1 respectively.

Goal: Learn a mechanism $\{\delta, \gamma, \eta\}$ to sequentially classify testing sequence, where $\delta : \mathbb{R} \to \{0,1\}$ is the stopping rule, $\gamma : \mathbb{R} \to \{0,1\}$ is the final decision rule, $\eta : \mathbb{R}^d \to \mathbb{R}$ is the information accumulation rule using a single real number.

Criterion: Minimize $\pi_0 N_0 + \pi_1 N_1$ given P_M and P_F constraints.

Assumption: $\{\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_M^{(0)}\}$ and $\{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_N^{(1)}\}$ are conditionally iid, same for testing sequence.

Background

- Optimality of Sequential Probability Ratio Test(SPRT)
- SPRT requires knowing class conditional density
- Standard machine learning techniques are designed and optimized for single sample scenario

Our Approach

- ***Kernel Functional Estimation:*** Construct kernel based estimate for the log-likelihood ratio function
- ***Log-likelihood Ratio Accumulation:*** Use cumulative log-likelihood ratio as aggregation rule
- ***Cost Approximation:*** Derive error and sampling cost using Martingale theory
- ***Kernel Weights Optimization:*** Optimize kernel weights to minimize sampling cost with constraints

Upper bound of sampling costs

Applying method similar to [1, 2], obtain upper bound of the sampling costs using convex conjugate:

$$C(r) \le \inf_f \frac{\omega_0}{\int \{f \cdot p_1 + p_0 + \log(-f) p_0\}} + \frac{\omega_1}{\int \{\frac{1}{f} \cdot p_0 + p_1 + \log(-\frac{1}{f}) p_1\}}$$

Final objective:

$$\hat{r}^\star = \arg\min_{\hat{r}} \frac{\omega_0}{\int -\log(\hat{r}) p_0} + \frac{\omega_1}{\int \log(\hat{r}) p_1}$$
$$\text{s.t.} \int \hat{r} p_0 = 1$$
$$\text{and} \int \hat{r}^{-1} p_1 = 1$$

Kernel Density Ratio Estimation

Impose Reproducing Kernel Hilbert Space(RKHS) structure on $\log(\hat{r})$:

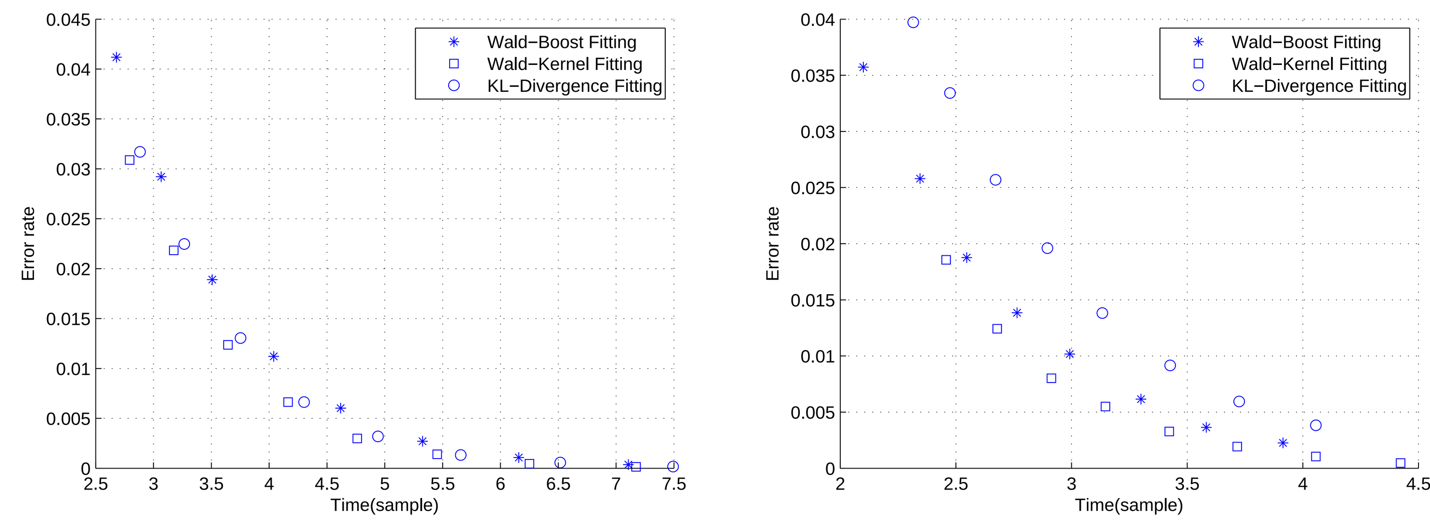
$$\log(\hat{r}(\mathbf{x})) = -\sum_{i=1}^{l} \alpha_i \cdot K_i(\mathbf{x}, \mathbf{x}_i)$$

Kernel based objective:

$$\alpha^\star = \arg\min_\alpha \frac{\omega_0}{\frac{1}{M}\sum_{j=1}^{M} \alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_j^{(0)})} - \frac{\omega_1}{\frac{1}{N}\sum_{i=1}^{N} \alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_i^{(1)})}$$
$$\text{s.t.} \frac{1}{M}\sum_{j=1}^{M} \exp(-\alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_j^{(0)})) \le 1$$
$$\text{and} \frac{1}{N}\sum_{i=1}^{N} \exp(\alpha^{\mathrm{T}} \mathbf{K}(\mathbf{x}_i^{(1)})) \le 1$$

Experimental Results

Human activity recognition using smartphones



(a) Moving *vs.* static (b) Moving *vs.* static

Human activity recognition using smartphones

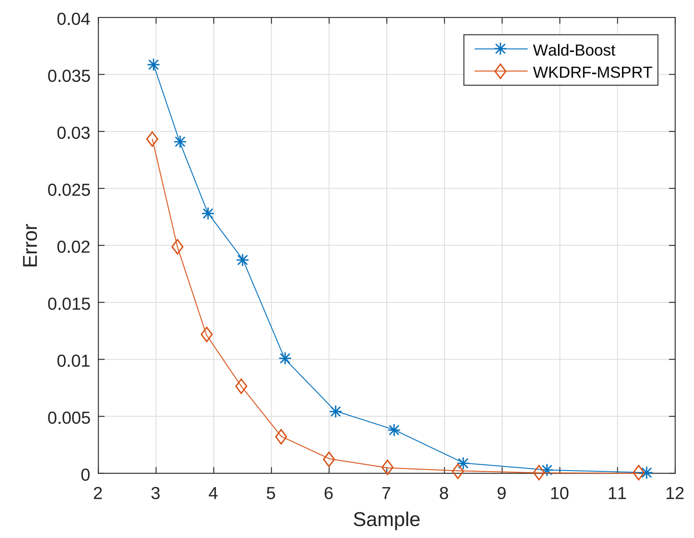
Extensions of The Basic Model

Multi-class Sequential Classification

Testing rule for MSPRT [3]:

$$\hat{H} = H_k \text{ if: } N_A \text{ is the smallest n such that} \frac{\sum_{j \ne k} \pi_j \prod_{i=1}^{n} p_j(\mathbf{x}_i)}{\pi_k \prod_{i=1}^{n} p_k(\mathbf{x}_i)} < A_k \text{ for some k}$$

Experimental result



Moving forward *vs.* moving upstairs *vs.* moving downstairs

Maximizing Multi-view Information Efficiency through View Dependent Training

Motivation

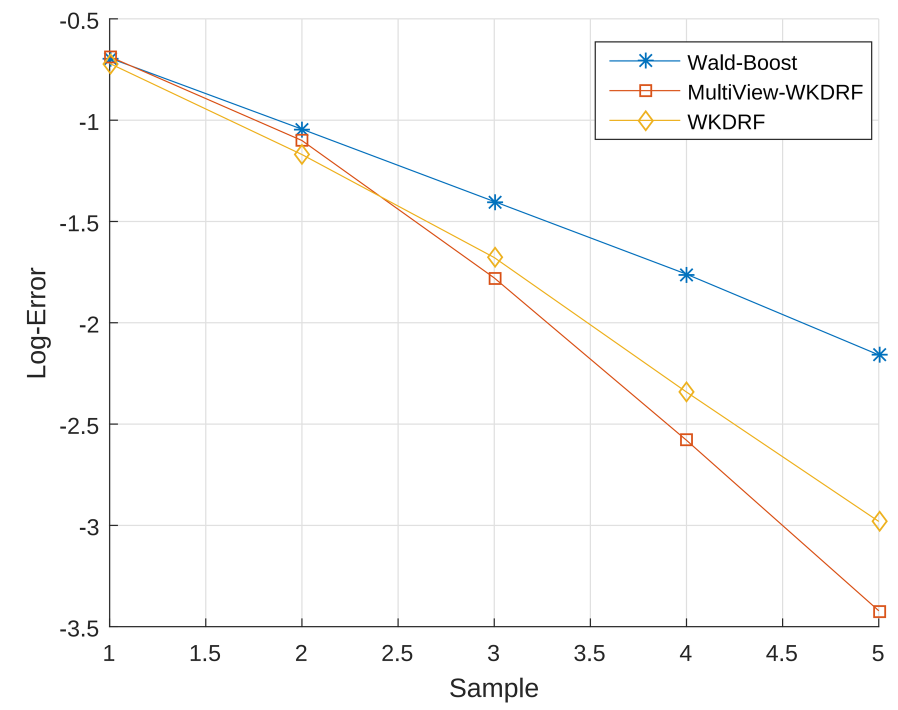
1. High dimensional objects' data usually lie on a low dimensional manifold when the measurement is proceeded in a controlled way smoothly
2. The measurement procedure is typically identical across targets' class, resulting in different targets' data being isometric to a few common underlying parameters
3. Recovering the underlying parameters will enable the algorithm to aggregate information more efficiently

MultiView-WKDRF

Log-likelihood ratio function is redesigned to capture the dependence of the underlying parameters:

$$\log(\hat{r}(\mathbf{x}, \theta)) = -\sum_{i=1}^{L}\sum_{j=1}^{T} \beta_{i,j} \cdot K_i(\mathbf{x}, \mathbf{x}_i) \cdot K_j(\theta, \theta_j)$$

Experimental Result: 2 Class Problem (BTR70 vs T72) from MSTAR Public Dataset



References

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

[2] T. Kanamori, T. Suzuki, and M. Sugiyama, "Statistical analysis of kernel-based least-squares density-ratio estimation," *Machine Learning*, vol. 86, no. 3, pp. 335–367, 2012.

[3] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *Information Theory, IEEE Transactions on*, vol. 40, no. 6, 1994.