RELIABLE EARLY CLASSIFICATION OF TIME SERIES

Hyrum S. Anderson

Sandia National Laboratories *
Albuquerque, NM 87123

Nathan Parrish, Kristi Tsukida, and Maya R. Gupta†

University of Washington
Department of Electrical Engineering
Seattle, WA

ABSTRACT

Early classification of time series is important in time-sensitive applications. An approach is presented for early classification using generative classifiers with the dual objectives of providing a class label as early as possible while guaranteeing with high probability that the early class matches the class that would be assigned to a longer time series. We give a specific algorithm for early quadratic discriminant analysis (QDA), and demonstrate that this classifier meets the requirement of reliable early classification.

***Index Terms*—** classification, minorization, Pareto optimal

1. INTRODUCTION

The ability to confidently classify time-series data as soon as possible is critical in military, medical, and commercial applications. For example, matching internet users to advertisements as soon as possible increases the chance of being able to serve them a profitable ad before they go offline. Making such classification decisions from less data generally carries increased risk of error, thus it is desirable that one be able to judge whether the classification would change if one waited for more data. We formalize this as the two goals:

Timeliness: classify the time series as early as possible;

Reliability: guarantee that with probability greater than or equal to some threshold, the class label assigned early matches the classification decision given a longer signal.

Recently, Xing et al. developed *early classification on time-series* (ECTS) based on the nearest-neighbor (NN) classifier [1]. In this paper we develop an approach for early classification of signals using a generative classifier, with a focus on the quadratic discriminant analysis (QDA) classifier. We prove that our early classifier decision will meet a desired reliability. Equivalently, we provide a reliability bound on the classifier's decision for every point in time. Experiments show that our approach is both early and reliable, that it performs well compared to the ECTS algorithm, and provides the user with a parameter to choose the trade-off between reliability and timeliness.

*Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporatoin, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

†Funding for this research was provided by Sandia National Labs, the United States Office of Naval Research, and a PECASE Award.

2. EARLY GENERATIVE CLASSIFICATION

We assume that we are given iid training pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i^{th} sampled time-series vector with corresponding class label $y_i \in \mathcal{G}$ for some discrete set of class labels $\mathcal{G}$. A generative classifier uses the labeled training data to estimate the parameters of the generating distribution for each class: $p(\mathbf{x}|y)$. At test time, the generative classifier classifies an unlabeled test example $\mathbf{x}$ according to the class which maximizes the a-posteriori probability given estimates of the generating distribution and class prior:

$$\begin{aligned}\hat{y}(\mathbf{x}) &= \max_{y\in\mathcal{G}} \hat{p}(y|\mathbf{x}) \\ &\equiv \min_{y\in\mathcal{G}} q_y(\mathbf{x}) \text{ and } q_y(\mathbf{x}) = -2\log(\hat{p}(\mathbf{x}|y)\hat{p}(y)). \quad (1)\end{aligned}$$

For the early classification problem, instead of the "full" time-series $\mathbf{x} \in \mathbb{R}^d$, we take as given the partial time series $\mathbf{x}_{t,k} \in \mathbb{R}^t$ for some $0 \le t \le d$, where $\mathbf{x} = \begin{bmatrix}\mathbf{x}_{t,k}^T & \mathbf{x}_{t,u}^T\end{bmatrix}^T$, and $\mathbf{x}_{t,u} \in \mathbb{R}^{d-t}$ is not known. We treat the unknown portion of the time series $\mathbf{x}_{t,u}$ as a random variable, $\mathbf{X}_{t,u}$, whose distribution we estimate from the training data. For each class $y \in \mathcal{G}$, we bound the possible classifier score $q_y(\mathbf{x})$ for the probable values of the unknown part of the signal:

$$q_{y,t}^{\max} = \max_{\mathbf{x}_{t,u}\in A} q_y(\mathbf{x}) \quad (2)$$

$$q_{y,t}^{\min} = \min_{\mathbf{x}_{t,u}\in A} q_y(\mathbf{x}), \quad (3)$$

where the set A is defined such that $\Pr(\mathbf{X}_{t,u} \in A) \ge \tau$; see Section 3 for details on A. Then, the following lemma gives conditions for making reliable early decisions.

Lemma: Let $\mathbf{X} = \left[\mathbf{x}_{t,k}^T \, \mathbf{X}_{t,u}^T\right]^T$. If $q_{g,t}^{\max} \leq q_{h,t}^{\min}$ for some g and all $h \neq g$, then $\Pr(\hat{y}(\mathbf{X}) = g) \geq \tau$.

Proof: Let B be the event $\hat{y}(\mathbf{X}) = g$ and C be the event $\mathbf{X}_{t,u} \in A$, and by the lemma's condition there is a g for which $q_{g,t}^{\max} \leq q_{h,t}^{\min}$ for all $h \neq g$. Then $\Pr(B|C) = 1$, as there is no realization of $\mathbf{X}_{t,u}$ in A that results in class g not having the minimum in (1). Therefore,

$$\Pr(B) = \underbrace{\Pr(B|C)\Pr(C)}_{\geq \tau} + \underbrace{\Pr(B|\bar{C})\Pr(\bar{C})}_{\geq 0} \geq \tau. \; \square$$

3. EARLY QDA

Next, we choose the constraint set A in (2) and (3) for the case of a quadratic discriminant analysis (QDA) classifier. QDA generalizes linear discriminant analysis [2], and models the generating distribution as Gaussian, $\hat{p}(\mathbf{x}|y) = \mathcal{N}(\mathbf{x}; \hat{\mu}_y, \hat{\Sigma}_y)$, so that (1) becomes

$$q_y(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_y)^T \hat{\Sigma}_y^{-1} (\mathbf{x} - \hat{\mu}_y) + \ln(|\hat{\Sigma}_y|) - 2\ln(\hat{p}(y)). \quad (4)$$

3.1. Chebyshev Constraint

We first construct the constraint set A using the multidimensional Chebyshev inequality, which states that for a random variable $\mathbf{X}_{t,u} \in \mathbb{R}^{d-t}$ with mean $\mathbf{m}_{t,u}$ and covariance $R_{t,u}$:

$$\Pr\left((\mathbf{X}_{t,u} - \mathbf{m}_{t,u})^T R_{t,u}^{-1} (\mathbf{X}_{t,u} - \mathbf{m}_{t,u}) \leq \alpha^2\right) \geq 1 - \frac{d-t}{\alpha^2}. \quad (5)$$

Thus $\Pr(\mathbf{X}_{t,u} \in A) \geq \tau$ implies

$$A = \left\{ \mathbf{x}_{t,u} \;\middle|\; (\mathbf{x}_{t,u} - \mathbf{m}_{t,u})^T R_{t,u}^{-1} (\mathbf{x}_{t,u} - \mathbf{m}_{t,u}) \leq \frac{d-t}{1-\tau} \right\}. \quad (6)$$

Note that A in (6) is non-empty for $\tau \in (-\infty,\ 1]$, although $\tau \leq 0$ provides an uninformative lower bound on the reliability of the classifier. Nevertheless, smaller values of τ reduce the size of A, and will result in earlier classification. Next, let

$$\hat{\mu}_y = \begin{bmatrix} \hat{\mu}_{y,k} \\ \hat{\mu}_{y,u} \end{bmatrix} \text{ and } \hat{\Sigma}_y^{-1} = \begin{bmatrix} S_{kk} & S_{ku} \\ S_{uk} & S_{uu} \end{bmatrix}$$

be the sample mean and covariance, partitioned into the known and unknown subsets. Substituting (4) and (6) into optimization problems (2) and (3) produces

$$q_{y,t}^{\max} = \max_{\mathbf{x}_{t,u} \in A} (\mathbf{x}_{t,u} - \mathbf{b})^T S_{uu} (\mathbf{x}_{t,u} - \mathbf{b}) + c \quad (7)$$

$$q_{y,t}^{\min} = \min_{\mathbf{x}_{t,u} \in A} (\mathbf{x}_{t,u} - \mathbf{b})^T S_{uu} (\mathbf{x}_{t,u} - \mathbf{b}) + c, \quad (8)$$

where A is given by (6) with

$$\begin{aligned} \mathbf{b} &= \hat{\mu}_{y,u} - S_{uu}^{-1} S_{uk} (\mathbf{x}_{t,k} - \mu_{y,k}) \\ c &= \log(|\Sigma_y|) + (\mathbf{x}_{t,k} - \hat{\mu}_{y,k})^T S_{kk} (\mathbf{x}_{t,k} - \hat{\mu}_{y,k}) \\ &\quad - 2\log(\hat{p}(y)) + 2(\mathbf{x}_{t,k} - \hat{\mu}_{y,k})^T S_{ku} \hat{\mu}_{y,u} \\ &\quad - (\mathbf{x}_{t,k} - \hat{\mu}_{y,k})^T S_{ku} S_{uu}^{-1} S_{uk} (\mathbf{x}_{t,k} - \hat{\mu}_{y,k}). \end{aligned}$$

Since the matrix S_{uu} is positive semi-definite, the objective function is convex and the $\min$ problem in (8) can be solved using standard convex optimization techniques.

Strong duality holds for any problem with a quadratic objective and quadratic constraints [3], so although the $\max$ problem in (7) is non-convex, $q_{y,t}^{max}$ can be found solving the dual problem which is a convex semidefinite program (SDP) [3, Appendix B]. However, we found that solving the $\max$ problem by the dual SDP was computationally prohibitive. Instead, we use a minorization approach [4] to reach a local solution of the $\max$ problem iteratively. A function $g(x|x^{(m)})$ is said to minorize function $f(x)$ if $f(x^{(m)}) = g(x^{(m)}|x^{(m)})$ and $f(x) \geq g(x|x^{(m)}) \forall x$. Since the objective function in (7) is convex, by Jensen's inequality

$$\begin{aligned} f(\mathbf{x}_{t,u}) &\geq f(\mathbf{x}_{t,u}^{(m)}) + (\mathbf{x}_{t,u} - \mathbf{x}_{t,u}^{(m)})^T \nabla f(\mathbf{x}_{t,u}^{(m)}) \\ &= (\mathbf{x}_{t,u}^{(m)} - \mathbf{b})^T S_{uu} (\mathbf{x}_{t,u}^{(m)} - \mathbf{b}) + c \\ &\quad + (\mathbf{x}_{t,u} - \mathbf{x}_{t,u}^{(m)})^T (2 S_{uu} \mathbf{x}_{t,u}^{(m)} - 2 S_{uu} \mathbf{b}). \end{aligned} \quad (9)$$

Therefore, the function

$$\begin{aligned} g(\mathbf{x}_{t,u}|\mathbf{x}_{t,u}^{(m)}) = 2\mathbf{x}_{t,u}^T S_{uu} (\mathbf{x}_{t,u}^{(m)} - \mathbf{b}) - \mathbf{x}_{t,u}^{(m)^T} S_{uu} \mathbf{x}_{t,u}^{(m)} \\ + \mathbf{b}^T S_{uu} \mathbf{b} + c \end{aligned}$$

is a linear function that minorizes the objective function in (7). We can solve for the $\mathbf{x}_{t,u}$ which gives a local maximum for (7) by iteratively solving the convex optimization problem:

$$\begin{aligned} \mathbf{x}_{t,u}^m = &\operatorname*{argmax}_{\mathbf{x}_{t,u}} g(\mathbf{x}_{t,u}|\mathbf{x}_{t,u}^m) \\ &s.t. \, (\mathbf{x}_{t,u} - \mathbf{m}_{t,u})^T R_{t,u}^{-1} (\mathbf{x}_{t,u} - \mathbf{m}_{t,u}) \leq \frac{d-t}{1-\tau}. \end{aligned} \quad (10)$$

3.2. Näive Bayes Constraints

Recall our goal of classifying $\mathbf{x}$ as early as possible with reliability $\geq \tau$. From our general problem formulation, given in Section 2, it is clear that the constraint set has a great impact on the earliness of the classifier. Since the Chebyshev constraint set (6) guarantees reliability $\geq \tau$ for any distribution of the unknown data, it may be overly conservative. Therefore, we develop two constraint sets based on a Gaussian assumption for the unknown data. Because these constraint sets rely on the Gaussian assumption for the unknown data, they can result in earlier decisions than the Chebyshev constraint.

Näive Bayes assumes that the covariates of a random variable are independent [2], so that $p(\mathbf{X}_{t,u})$ is given by

$$p\left(\mathbf{X}_{t,u}(1), \ldots, \mathbf{X}_{t,u}(d-t)\right) = \prod_{\ell=1}^{d-t} p\left(\mathbf{X}_{t,u}(\ell)\right), \quad (11)$$

where $\mathbf{X}_{t,u}(\ell)$ is the ℓth element of $\mathbf{X}_{t,u}(\ell)$. Further applying a Gaussian assumption, we have $\mathbf{X}_{t,u} \sim \mathcal{N}(\mathbf{m}_{t,u}, R)$, where

R is a diagonal matrix. Then, the smallest set A such that $\Pr(\mathbf{X}_{t,u} \in A) \geq \tau$ is given by

$$\begin{aligned}
&\Pr\left((\mathbf{X}_{t,u} - \mathbf{m}_{t,u})^T R_{t,u}^{-1} (\mathbf{X}_{t,u} - \mathbf{m}_{t,u}) \leq \beta^2\right) = \tau \\
\equiv &\Pr\left(\sum_{\ell=1}^{d-t} \left(\frac{\mathbf{X}_{t,u}(\ell) - \mathbf{m}_{t,u}(\ell)}{\sqrt{R_{t,u}(l,l)}}\right)^2 \leq \beta^2\right) = \tau \\
\equiv &\Pr\left(\mathbf{Z}_{t,u} \leq \beta^2\right) = \tau, \qquad (12)
\end{aligned}$$

where $\mathbf{Z}_{t,u} = \sum_{\ell=1}^{d-t} \left(\frac{\mathbf{X}_{t,u}(\ell) - \mathbf{m}_{t,u}(\ell)}{\sqrt{R_{t,u}(l,l)}}\right)^2$ is a chi-squared random variable with $d-t$ degrees of freedom [5]. Given a desired reliability rate τ, we solve for β^2 that satisfies (12) using the chi-squared inverse cdf, and denote that value as $\beta^2(\tau)$. The resulting constraint set is given by

$$A = \left\{\mathbf{x}_{t,u} | \left(\mathbf{x}_{t,u} - \mathbf{m}_{t,u}\right)^T R_{t,u}^{-1} \left(\mathbf{x}_{t,u} - \mathbf{m}_{t,u}\right) \leq \beta^2(\tau)\right\}. \qquad (13)$$

A second constraint set that stems from the näive Bayes Gaussian assumption is a box constraint. We define the box constraint set to be

$$A = \{\mathbf{x}_{t,u} | \mathbf{x}_{t,u}(\ell) \in [\mathbf{m}_{t,u}(\ell) - \mathbf{s}(\ell),\ \mathbf{m}_{t,u}(\ell) + \mathbf{s}(\ell)], \forall \ell\}, \qquad (14)$$

By näive Bayes, the constraint boundaries are set independently for each covariate, $\mathbf{s}(\ell)$, by solving for the $\mathbf{s}(\ell)$ that satisfies $\Pr(\mathbf{X}_{t,u}(\ell) \in [\mathbf{m}_{t,u}(\ell) - \mathbf{s}(\ell),\ \mathbf{m}_{t,u}(\ell) + \mathbf{s}(\ell)]) = \tau^{\frac{1}{d-t}}$ that results in $\Pr(\mathbf{X}_{t,u} \in A) = \tau$. Substituting the box constraint set in (14) into the $\min$ and $\max$ problems yields

$$\begin{aligned}
q_{y,t}^{\max} = &\max_{\mathbf{x}_{t,u}} (\mathbf{x}_{t,u} - \mathbf{b})^T S_{uu} (\mathbf{x}_{t,u} - \mathbf{b}) + c \qquad (15) \\
&s.t.\ \mathbf{x}_{t,u}(\ell) \leq \mathbf{m}_{t,u}(\ell) + \mathbf{s}(\ell),\ \ell = 1, ..., d-t \\
&\quad \mathbf{x}_{t,u}(\ell) \geq \mathbf{m}_{t,u}(\ell) - \mathbf{s}(\ell),\ \ell = 1, ..., d-t,
\end{aligned}$$

$$\begin{aligned}
q_{y,t}^{\min} = &\min_{\mathbf{x}_{t,u}} (\mathbf{x}_{t,u} - \mathbf{b})^T S_{uu} (\mathbf{x}_{t,u} - \mathbf{b}) + c \qquad (16) \\
&s.t.\ \mathbf{x}_{t,u}(\ell) \leq \mathbf{m}_{t,u}(\ell) + \mathbf{s}(\ell),\ \ell = 1, ..., d-t \\
&\quad \mathbf{x}_{t,u}(\ell) \geq \mathbf{m}_{t,u}(\ell) - \mathbf{s}(\ell),\ \ell = 1, ..., d-t.
\end{aligned}$$

The optimal $\mathbf{x}_{t,u}$ for (15) and (16) can be solved algebraically. For the $\max$ problem in (15), each $\mathbf{x}_{t,u}(\ell)$ lies at the edge of the box that maximizes the distance from $\mathbf{b}(\ell)$. Similarly, for the $\min$ problem in (16), $\mathbf{x}_{t,u}(\ell) = \mathbf{b}(\ell)$ if $\mathbf{b}(\ell) \in [\mathbf{m}_{t,u}(\ell) - \mathbf{s}(\ell),\ \mathbf{m}_{t,u}(\ell) + \mathbf{s}(\ell)]$. Otherwise, $\mathbf{x}_{t,u}(\ell)$ lies at the edge of the box that minimizes the distance to $\mathbf{b}(\ell)$.

3.3. Estimation of the mean and variance parameters

For each method, we estimate the mean $\mathbf{m}_{t,u}$ and covariance $R_{t,u}$ of $\mathbf{X}_{t,u}$ using the training data under a joint Gaussian assumption, as follows. We first estimate the class independent maximum likelihood mean, $\hat{\bar{\mathbf{x}}}$, and regularized maximum likelihood covariance, $\hat{\Sigma}$, from the training data. Assuming that the the complete time series $\mathbf{X}$ is Gaussian distributed,

$$\begin{bmatrix}\mathbf{X}_{t,k} \\ \mathbf{X}_{t,u}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\bar{\mathbf{x}}}_{t,k} \\ \hat{\bar{\mathbf{x}}}_{t,u}\end{bmatrix}, \begin{bmatrix}\hat{\Sigma}_{k,k} & \hat{\Sigma}_{k,u} \\ \hat{\Sigma}_{u,k} & \hat{\Sigma}_{u,u}\end{bmatrix}\right),$$

then the mean and covariance of $\mathbf{X}_{t,u}$ given $\mathbf{X}_{t,k} = \mathbf{x}_{t,k}$ is

$$\begin{aligned}
\mathbf{m}_{t,u} &= \hat{\bar{\mathbf{x}}}_{t,u} + \hat{\Sigma}_{u,k} \hat{\Sigma}_{k,k}^{-1} (\mathbf{x}_{t,k} - \hat{\bar{\mathbf{x}}}_{t,k}) \\
R_{t,u} &= \hat{\Sigma}_{u,u} + \hat{\Sigma}_{u,k} \hat{\Sigma}_{k,k}^{-1} \hat{\Sigma}_{k,u}.
\end{aligned}$$

Although the time series vector is assumed to be jointly Gaussian when estimating the mean and covariance, the maximum and minimum QDA scores using the Chebyshev bounds in (7) and (8) do not require Gaussian assumption, but hold for any distribution with mean $\mathbf{m}_{t,u}$ and covariance $R_{t,u}$.

4. EXPERIMENTS

In this section we perform experiments using one synthetic and four real datasets from the UCR Time Series Page [6]. In all experiments, we implement a local version of QDA [7] that fits the mean and covariance for class y by choosing the k nearest class y neighbors to the test sample. Additionally, we use diagonal class covariance matrices, $\hat{\Sigma}_y$, in (4).

In all figures, we plot Pareto curves for the early QDA classifier by varying the value of τ. Varying τ provides a tradeoff between reliability and earliness, with smaller values resulting in earlier classification but lower reliability, and vice versa for larger values of τ. In all figures we plot reliability, the percentage of early labels that match the final labels, vs. the average early classification time over the test samples.

The tradeoff between reliability and earliness is shown explicitly in Fig. 1 using the *Synthetic Control* dataset and early QDA with the Chebyshev constraint set (6). A result of note in this figure is that, although the value of τ given by the Chebyshev inequality meets the desired reliability, we can achieve the target reliability with earlier classification by reducing τ. For instance, suppose that we want reliability of $\geq 95\%$. By setting $\tau = 0.95$, we achieve reliability of 100% with average early classification time of 57.8. At $\tau = -15$, we still achieve reliability of 95.6% and an average early classification time of 22.88. This indicates that in practice we can set τ by cross-validation given enough training data.

Due to space constraints, we chose four diverse real datasets from the UCR repository according to the following criteria: longest time-series length (*Lightning-2*), shortest time-series length (*ECG*), most training data (*Two Patterns*), and fewest training data (*Face Four*). We show the results for ECTS [1] and for early QDA with the three constraint sets: the Chebyshev constraint set for values of τ between -400 and 0.95, and the näive Bayes constraint sets for values of τ between 10^{-80} and 0.95. We also plot two baselines, 'Fixed t QDA' and 'Fixed t 1-NN', that classify a test sample at time t with a classifier trained only on training data up to time t.

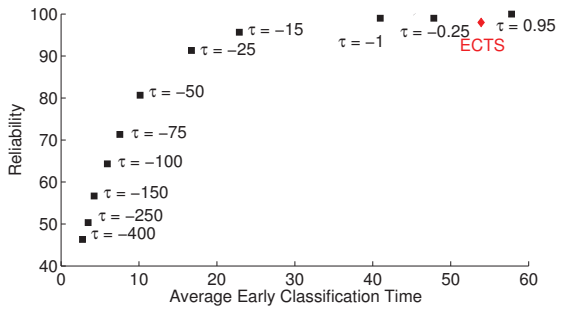


Fig. 1. Pareto optimal curve for the *Synthetic Control* dataset. The black boxes show the results for the indicated value of τ using early QDA with the Chebyshev constraint set (6).

We plot the results in Fig. 2. We can see that in all plots the Chebyshev constraint is more conservative than the constraints based on the näive Bayes Gaussian assumption, as the average classification time for $\tau = 0.95$ (the rightmost plot point) is the greatest under the Chebyshev constraint. We can also see that the reliability of early QDA and ECTS dominates the respective 'fixed t /methods. Finally, comparing early QDA directly to ECTS, we can see that early QDA dominates ECTS in reliability in all experiments.

5. CONCLUSIONS

We have presented an early classification framework for generative classifiers that guarantees high reliability, and have provided an implementation for early quadratic discriminant analysis (early QDA). Experimental results show that early QDA performs well in practice when compared to baseline methods that classify at a fixed time t and compared to ECTS.

6. REFERENCES

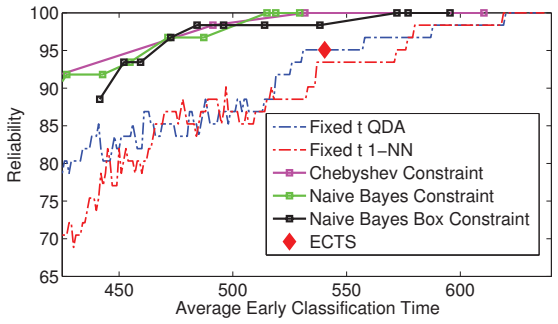
[1] Z. Xing, J. Pei, and P. S. Yu, "Early prediction on time series: a nearest neighbor approach," pp. 1297–1302, 2009.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.

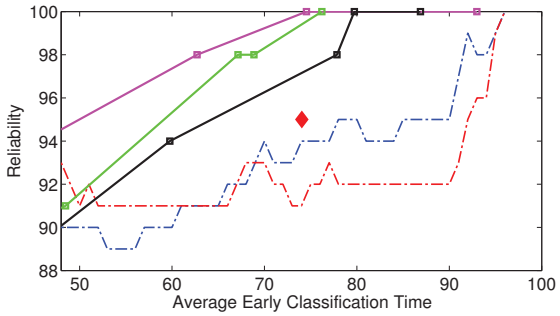
[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[4] K. Lange, D. R. Hunter, and Y. Ilsoon, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

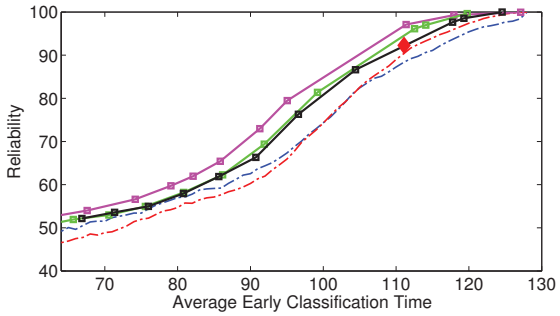
[5] M. K. Simon, *Probability Distributions Involving Gaussian Random Variables*, Springer, New York, 2002.



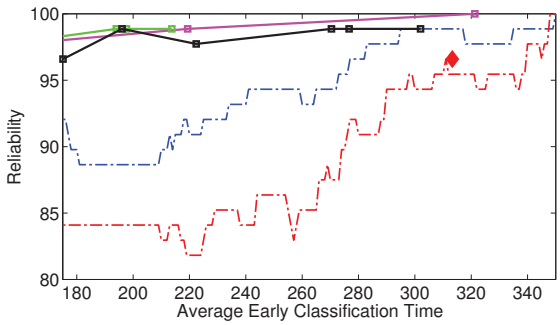
(a) Lightning 2



(b) ECG



(c) Two Patterns



(d) Face Four

[6] E. Keogh, X. Xi, L. Wei, and C. Ratanamahatana, "UCR time series classification and clustering page," `http://www.cs.ucr.edu/~eamonn/time_series_data/`.

[7] E. Garcia, S. Feldman, M. R. Gupta, and S. Srivastava, "Completely lazy learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1274–1285, 2010.