CONFORMAL PREDICTORS FOR ONLINE TRACK CLASSIFICATION

Michael J. Pekala, I-Jeng Wang, Ashley J. Llorens

Johns Hopkins University Applied Physics Laboratory
firstname.lastname@jhuapl.edu

ABSTRACT

This paper considers online classification problems where each object to be classified consists of a sequence of measurements, termed here a *track*. We present an approach that combines ideas from sequential hypothesis testing with those from conformal prediction to address track level outliers - entire measurement sequences that are novel relative to the statistical model. We show with analysis and empirical results that this approach preserves the optimal performance of the underlying sequential hypothesis testing when outliers are absent and provides an error rate guarantee in the presence of contamination by novel tracks.

Index Terms— pattern classification, conformal prediction, statistics, robustness.

1. INTRODUCTION

Automated screeners play a key role in many signal processing applications where a noisy sequence of observations taken from a single object forms the basis for deciding whether the object belongs to a category of interest (target) or not (clutter). In the following, we refer to the sequence of measurements associated with a single object as a *track*.

There are a number of desiderata and assumptions that accompany practical screening applications. Often, target predictions trigger a costly response by a system operator or more sophisticated automated system; hence, adherence to a pre-defined error rate is an important requirement. Experimental conditions often constrain the number of measurements available in each track (e.g. in remote sensing problems, mobile objects move in and out of sensor detection range), and therefore track lengths are finite and not known in advance. Thus, while tracks may be statistically independent, they are not identically distributed due to length variations. We also assume the possible presence of outlier tracks that are drawn from distributions not anticipated at design time. Finally, we consider an online setting where classifiers are occasionally provided with the object's true label following prediction. The availability of label feedback is a reasonable assumption in real-time applications where the classifier acts as a decision aid in conjunction with an operator. Our goal is to design an online classification framework that will provide a theoretical guarantee on pre-defined error rate despite uncertainty in track length and existence of outlier tracks.

In the case of a single track with two possible class labels, existing theory in sequential hypothesis testing (such as Wald's *sequential probability ratio test* (SPRT) [1]) solves the track classification problem subject to various constraints on the decision time and/or Type I and Type II error rates. Robust variants of these tests address the issue of occasional measurement outliers by clipping extreme test values in accordance with an assumed model for outlier contamination [2]. These approaches, however, are not robust to the presence of outlier tracks as the realized error rates may deviate from designed limits when entire measurement sequences are mismatched with the likelihood functions.

Various schemes for detecting and screening track-level outliers are possible; our approach is motivated by the need to meet predetermined performance requirements. Conformal Prediction (CP) is an online framework for classification that provides control over long-term error rates [3, 4]. CPs generate sets of possible labels for each object, akin to a confidence interval. These predictions are created from the outputs of *nonconformity measures*, real-valued functions that quantify how different a new example is from previously seen examples. Assuming the objects to be classified adhere to a suitable model of randomness, such as exchangeability, and feedback is provided sufficiently often, CP theory ensures predictions contain the true class label for the object $(1 - \epsilon)$ percent of the time in the long run, where ϵ is a user-specified error rate. Such a CP is said to be *valid*. Using prediction sets (versus point predictions) permits "hedging" for more difficult instances; however, predictions containing more than one label are less informative. Therefore, an important CP metric is *efficiency*, the rate of unambiguous predictions (predictions with fewer than two labels).

In this paper, we design a CP with a nonconformity measure determined by statistics derived from the SPRT. A key feature of our approach is that it induces an exchangeable sequence of objects when tracks are statistically independent, despite variations in track length (thus ensuring a valid CP). When the likelihood functions are known, the resulting CP will be maximally efficient in that it never produces multi-

The authors would like to thank Kelly Bennett and the US Army Research Laboratory for providing the ACIDS data set.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

label predictions. Furthermore, the CP continues to achieve the desired error rate in presence of outlier tracks, albeit at the cost of reduced efficiency.

2. PRIOR WORK

Prior work in robust SPRT includes [2, 5, 6]; more recently, [7] considered situations where the likelihood functions are not perfectly known. The authors propose an augmented likelihood ratio test that generates confidence intervals to decide between two Bernoulli hypotheses. The resulting statistical test terminates only after the entire confidence interval crosses one of the two decision thresholds. Our approach also utilizes confidence intervals via the conformal prediction framework but relies upon exchangeable examples and label feedback instead assuming a specific distributions for the hypotheses. In [8], the authors consider robust track classification in the case of dynamic objects whose state is estimated by a Kalman filter. Prior work in the conformal prediction literature for sequence classification includes [9, 10, 11]. Our approach differs in the use of SPRT (and the associated efficiency properties); as far as the authors are aware, this choice of CP statistic has not previously been considered.

3. APPROACH

This section presents our approach to online track classification that embeds the SPRT within a CP framework. Consider an online binary classification problem where a sequence of objects $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ are incrementally provided to a classifier which must predict the corresponding class labels $y_i \in \{0, 1\}$ based on prior examples $z_j := (\mathbf{x}_j, y_j), j = 1, \ldots, i-1$. Each track consists of a finite sequence of measurements, $\mathbf{x}_i = \{x_k^i : k = 1, \ldots, L(\mathbf{x}_i)\}$, where $L(\mathbf{x}_i)$ denotes the finite length of the track $\mathbf{x}_i$. With screening applications in mind, we refer to tracks with label $y = 1$ as target and $y = 0$ as clutter. We also assume (*i*) the sequence of track objects are statistically independent but not identically distributed (due to varying track lengths); (*ii*) the $L(\mathbf{x}_i)$ are not known in advance; and (*iii*) given the class of a track object, the sequence of measurements comprising the track are i.i.d. We denote the class-dependent density functions for the measurements of track $\mathbf{x}_i$ by $p_i(x|y_i = 0)$ and $p_i(x|y_i = 1)$.

In the ideal case, all tracks have the same class-dependent measurement-level distributions. However, there may exist *outlier tracks* following different measurement-level distributions. Specifically, for class labels $l \in \{0, 1\}$ and all i:

$$p_i(x|y_i = l) = \begin{cases} f_l(x) & \text{with probability } 1 - \pi, \\ \hat{f}_l(x) & \text{with probability } \pi, \end{cases}$$

where f_0, f_1 are clutter and target likelihood density functions, the $\hat{f}$'s are outlier density functions, and π is a small outlier probability.

3.1. Sequential Detection Review

Sequential detection problems involve deciding between two hypotheses: the null hypothesis H_0 that a sequence of i.i.d. observations is drawn from distribution Q_0 and the alternate hypothesis H_1 that the sequence is instead drawn from a distinct distribution Q_1. Sequential detection problems are often formulated as *optimal stopping problems*, where the goal is to reach a decision after some finite number of observations; once a decision is made, it is final and no further observations are considered. Wald's SPRT formulates the sequential decision problem as the optimal stopping problem to minimize the number of measurements required to decide between the hypotheses subject to an upper bound on the probability of Type I and Type II errors. SPRT operates by incrementally computing a statistic S_k and comparing to a pair of thresholds $a < b$. The SPRT statistic S_k is defined by the cumulative sum of the log likelihood ratio

$$S_k = S_{k-1} + \log\left(\frac{f_1(x_k)}{f_0(x_k)}\right). \tag{1}$$

The thresholds a, b are designed based on Wald's approximation as functions of α, β, the desired Type I and Type II probabilities (see [12] for full details). As long as $a < S_k < b$, no decision is made and the procedure continues. SPRT terminates the first time $S_k <= a$ (whereupon it decides H_0) or the first time $b <= S_k$ (decide H_1). Let τ be the time/index when the SPRT terminates and make its decision, that is,

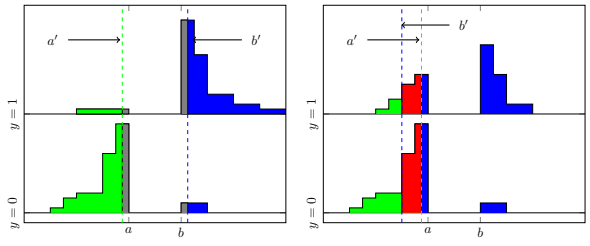
$$\tau = \min\{k \in \mathbb{N} \colon S_k \leq a \text{ or } S_k \geq b\}. \tag{2}$$

When the observations are i.i.d. and the class-conditional probability density functions are known, SPRT is provably optimal in the sense that no other sequential decision rule that similarly constrains the Type I and Type II errors can do so using fewer measurements.

3.2. Using SPRT Within Conformal Predictors

We propose a CP nonconformity measure based on the sequence of cumulative log likelihood ratios up to the stopping time τ. That is, given an object $\mathbf{x} \in \mathcal{X}$, a nonconformity measure is defined as a function of $[S_1, \ldots, S_\tau]^T$. There are two immediate consequences of this approach:

- For a track object with length less than the stopping time τ, the nonconformity measure is not defined. In this case, we will simply not include the track object in the CP. The interpretation is that the track is too short to provide sufficient information for the CP to maintain its validity while achieving high efficiency.
- For the (sub-)sequence of track objects that resulted in a well-defined nonconformity measure, the exchangeability condition required for the validity of CP is satisfied since $[S_1, \ldots, S_\tau]^T$'s are i.i.d. across these tracks.



(a) $a' < a < b < b'$ (b) $a' < a,\ b' < a$

Fig. 1: Example distributions of $S_\tau(\mathbf{x})$ by class label. The CP accepts hypothesis $y = 1$ when $S_\tau(\mathbf{x}) \geq b'$ and hypothesis $y = 0$ when $S_\tau(\mathbf{x}) \leq a'$. Panel (a) depicts the ideal (no outlier) case where $P_{fa} < \epsilon$ and $(1 - P_d) < \epsilon$ while panel (b) depicts a case when $(1 - P_d) > \epsilon$. Also shown are the CP decision regions for accepting only the clutter hypothesis (green), only the target hypothesis (blue) both hypotheses (red) and neither (gray).

Throughout the rest of this paper we assume that a nonconformity measure is always defined for the sequence of track objects presented to a CP with the understanding that tracks shorter than the stopping time of the SPRT have been excluded and are associated with multi-label predictions.

Given a finite sequence of track objects and their class labels $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{n-1}, y_{n-1})\}$, we consider a class-dependent nonconformity score defined by

$$\alpha_i = -(2 \cdot y_i - 1) \cdot S_\tau(\mathbf{x}_i), \tag{3}$$

where $S_\tau(\mathbf{x}_i)$ denotes the cumulative log likelihood ratio of the track object $\mathbf{x}_i$ at the stopping time of the SPRT. The class-conditional Mondrian CP framework [3, 13] includes hypothesis y in the prediction for track $\mathbf{x}_n$ when

$$p_n^y \triangleq \frac{|\{i = 1, \ldots, n-1 \colon y_i = y \ \&\ \alpha_i \geq \alpha_n\}|}{|\{i = 1, \ldots, n-1 \colon y_i = y\}|} > \epsilon, \tag{4}$$

i.e., the associated p-value p_n^y is larger than ϵ, the desired CP error rate for class y. Since the stochastic process $S_\tau(\mathbf{x}_1), S_\tau(\mathbf{x}_2), \ldots$ is i.i.d. (and therefore exchangeable), a CP defined using (4) is valid. Note that this validity result is independent of whether SPRT is able to achieve the desired Type I and Type II error rates.

To analyze the efficiency of the Mondrian CP defined in (4), it is convenient to define two scalar thresholds a' and b' such that the CP will include $y = 0$ in the prediction if and only if $S_\tau(\mathbf{x}_i) \leq a'$, and $y = 1$ in the prediction if and only if $S_\tau(\mathbf{x}_i) \geq b'$. Thus, a', b' play a role in the CP analogous to that of a, b in SPRT. Specifically,

$$b' \triangleq \inf\{|\{i = 1, \ldots, n-1 \colon y_i = 1 \ \&\ S_\tau(\mathbf{x}_i) \leq S\}| > \epsilon n_1\}, \tag{5}$$

$$a' \triangleq \sup\{|\{i = 1, \ldots, n-1 \colon y_i = 0 \ \&\ S_\tau(\mathbf{x}_i) \geq S\}| > \epsilon n_0\}, \tag{6}$$

where n_1 and n_0 denote the numbers of examples labeled $y = 1$ and $y = 0$ made available to the CP via feedback. Note that $b' = S_\tau(\mathbf{x}_i)$ and $a' = S_\tau(\mathbf{x}_j)$ for some i and j, and their values will change over time as feedback is accumulated. We first prove a simple lemma that relates a' and b' to the SPRT thresholds a and b:

Lemma. *Choose α and β for the SPRT such that*

$$\max\left\{\frac{\beta}{1-\alpha}, \frac{\alpha}{1-\beta}\right\} < \epsilon. \tag{7}$$

Then a' and b' defined by (5) *and* (6) *satisfy $a' \leq a$ and $b' \geq b$ for large n, where a and b are the thresholds for the SPRT designed according to the desired α and β.*

Proof. By class-dependent validity of the CP, for large n

$$P\{S_\tau(\mathbf{x}_n) < b' \mid y_n = 1\} \approx \epsilon, \tag{8}$$

$$P\{S_\tau(\mathbf{x}_n) > a' \mid y_n = 0\} \approx \epsilon. \tag{9}$$

Furthermore, by SPRT theory

$$P\{S_\tau(\mathbf{x}_n) \leq a \mid y_n = 1\} = 1 - P_d \leq \beta/(1-\alpha), \tag{10}$$

$$P\{S_\tau(\mathbf{x}_n) \geq b \mid y_n = 0\} = P_{fa} \leq \alpha/(1-\beta). \tag{11}$$

Together equations (7), (10) and (8) give that

$$P\{S_\tau(\mathbf{x}_n) \leq a \mid y_n = 1\} < P\{S_\tau(\mathbf{x}_n) < b' \mid y_n = 1\}$$

which in turn implies $a < b'$. By construction, $b' = S_\tau(\mathbf{x}_i)$ for some i; since no S_τ reside within the SPRT "gap" (a, b), it must be that b' is either $\geq b$ or $\leq a$. Since $a < b$ (by the definition of the SPRT) and $a < b'$ (above), it must be that $b \leq b'$. Similar reasoning shows that $a' \leq a$. □

Figure 1 illustrates the basic intuition behind the proof by depicting the distribution of SPRT scores for tracks of each class in an ideal (figure 1a) and a non-ideal (figure 1b) case. In the ideal case, since $a < b$ the Lemma implies that, for any choice of parameters satisfying (7), the CP thresholds satisfy $a' < b'$. This means the regions for accepting the hypotheses $y = 1$ and $y = 0$ are disjoint and the SPRT-based CP will never generate multi-label predictions.

When outlier tracks arrive, the SPRT may not achieve the desired Type I or Type II error rates due to possible mismatch between the actual measurement-level likelihood functions $\hat{f}_l$ and the estimated likelihood functions (based on f_l) used in the SPRT. In this case, equations (10), (11) no longer hold and the proportion of SPRT scores on the "wrong" side of either threshold may exceed ϵ (i.e., $P_{fa} > \epsilon$ or $(1 - P_d) > \epsilon$). In this case, it is possible that $b' \leq a'$, introducing a region of nonconformity scores for which neither hypothesis will be rejected (figure 1b, red region). Note that this mismatch may be difficult to address by online estimation of the likelihood functions since outlier tracks occur infrequently.

π	SPRT Type I	SPRT Type II	CP Err. $(y=1)$	CP Err. $(y=0)$	CP Mult. Pred. $(y=1)$	CP Mult. Pred. $(y=0)$	Total Mult. Pred.
CP Using S_τ							
.05	3.26 % (.39)	.07 % (.06)	5.03% (.1)	4.97 % (.1)	0	0	0
.15	9.27 % (.54)	.06 & (.05)	5.04 % (.1)	4.98 % (.1)	42.40 % (5.23)	3.70 % (.5)	23.09 % (2.84)
.20	12.44 % (.99)	.07 % (.05)	5.03 % (.1)	4.97 % (.1)	58.35 % (5.26)	6.78 % (.1)	32.60 % (2.89)
CP Using S as defined in (12)							
.05	3.26 % (.39)	.07 % (.07)	5.03 % (.1)	4.96 % (.1)	0	0	0
.15	9.27 % (.54)	.06 % (.05)	5.04 % (.1)	4.97 % (.1)	4.43 % (5.79)	.36 % (.4)	2.4 % (3.14)
.20	12.44 % (.99)	.07 % (.05)	5.03 % (.1)	4.96 % (.1)	20.4 % (8.64)	2.1 % (.5)	11.3 % (4.59)

Table 1: Results for experiments using two different CP statistics.

3.3. Extending to Address Outlier Tracks

Outlier tracks could negatively impact the efficiency of the CP with our particular choice of nonconformity measure since the values of $S_\tau(\mathbf{x}_i)$ tend to cluster right outside the SPRT thresholds. To address this potential lack of robustness, we propose an extension using statistics based on the cumulative log-likelihood ratio just beyond the stopping time τ. Specifically, we use the following statistic to replace $S_\tau(\mathbf{x}_i)$ in the definition of nonconformity measure:

$$S(\mathbf{x}_i) = \begin{cases} S_{\tau+d}(\mathbf{x}_i) + c & \text{if } S_\tau(\mathbf{x}_i) \geq b, \\ S_{\tau+d}(\mathbf{x}_i) - c & \text{if } S_\tau(\mathbf{x}_i) \leq a, \end{cases} \tag{12}$$

where $d \in \mathbb{N}$ and $c > 0$ are chosen such that $S(\mathbf{x}_i)$ is outside the SPRT gap. Extending the period of observation by d will "spread out" the cumulative log-likelihood ratio values and help reduce the impact of outlier tracks upon CP efficiency. Since d is fixed, the stochastic process $S_{\tau+d}(\mathbf{x}_1), S_{\tau+d}(\mathbf{x}_2), \ldots$ is also i.i.d., and the resulting CP remains valid. Note that in the idealized case, this modified nonconformity measure will achieve the same performance as the SPRT and exhibits the same maximal efficiency as the original approach discussed in section 3.2.

4. EXPERIMENTAL RESULTS

We now consider empirical results using the Acoustic-seismic Classification Identification Data Set (ACIDS), created by the US Army Research Laboratory for developing signal classification algorithms. ACIDS contains acoustic and seismic measurements obtained from nine different vehicle types under different environmental conditions. Tracks consist of a single vehicle making a single pass of the sensor system at varying speeds and closest points of approach [14].

The ACIDS raw acoustic measurements were first transformed into a spectrogram by means of a short-time Fourier transform and the resulting scans (snapshots of the spectrum in time) were passed through a spectral normalization filter to remove the broadband energy trend, focusing the feature extraction on the narrowband characteristics of the signature. The resulting spectrum was then normalized. Scans along each track were reduced to a single dimension by means of a principal component analysis followed by applying a nonlinear support vector machine (SVM). The SVM scores (one per scan) comprise the track.

Table 1 presents results for an experiment where vehicle type 1 represents the target class, type 6 the clutter class and 4 the clutter outlier class. To ensure i.i.d. measurements within tracks and to increase the number of test tracks, 5000 resampled tracks (split 50/50 between target and clutter) were each generated by uniformly sampling 30 measurements from the test data pool. In practice, probability of detection is a function of signal to noise ratio (SNR) which usually exhibits dependencies along tracks [15, 16]; the within-track i.i.d. assumption made here is an idealization that roughly corresponds to a far-field assumption in which SNR is more or less consistent. These results shown are for $\epsilon = .05$, SPRT parameters $\alpha = \beta = .03$ and various values of π, the negative class outlier probability. Values reported are averages over 10 experiments (standard deviations in parentheses). For this choice of α, β, SPRT makes decisions very quickly; SPRT made decisions for all tracks in the experiment.

For $\pi = .05$ the CP does not generate multiple predictions. Columns 4 and 5 show that the CP remains valid with respect to both classes even when outlier tracks preclude SPRT from achieving the desired Type I error rate. The price paid is a decrease in efficiency (columns 6-8). In this case, using $S(\mathbf{x}_i)$ in (12) as the CP statistic leads to a substantial reduction in the multiple prediction rate compared to $S_\tau(\mathbf{x}_i)$. This is consistent with the discussion in section 3.3 suggesting cumulative log likelihood ratio values clustering near the decision threshold makes for a less desirable CP statistic.

5. SUMMARY AND FUTURE WORK

This paper presents a novel combination of techniques from sequential hypothesis testing and conformal prediction to provide robustness to occasional track-level outliers in an online classification setting. We showed that this approach retains the ideal performance of SPRT when outliers are absent while providing valid predictions when track-level outliers are present. Future work includes considering adaptations that admit non-i.i.d. observations along each track and additional ways to improve CP efficiency as the proportion of outlier tracks increases (e.g. alternative nonconformity measures or distance metrics [17]).

6. REFERENCES

[1] Abraham Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[2] Peter J Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[3] Vladimir Vovk, Alex Gammerman, and Glenn Shafer, *Algorithmic learning in a random world*, Springer, 2005.

[4] Glenn Shafer and Vladimir Vovk, "A tutorial on conformal prediction," *The Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.

[5] Peter J Huber and Volker Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *The Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.

[6] Peter J Huber, *Robust statistics*, Springer, 2011.

[7] Ricardo Santiago-Mozos, R Ferndndez-Lorenzana, Fernando Perez-Cruz, and Antonio Artes-Rodriguez, "On the uncertainty in sequential hypothesis testing," in *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*. IEEE, 2008, pp. 1223–1226.

[8] Nathan Parrish, Hyrum Anderson, and Maya R Gupta, "Robust sequential classification of tracks," in *Information Fusion (FUSION), 2010 13th Conference on*. IEEE, 2010, pp. 1–8.

[9] Rikard Laxhammar and Göran Falkman, "Conformal prediction for distribution-independent anomaly detection in streaming vessel data," in *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*. ACM, 2010, pp. 47–55.

[10] Rikard Laxhammar and Göran Falkman, "Sequential conformal anomaly detection in trajectories based on Hausdorff distance," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.

[11] Rikard Laxhammar and Göran Falkman, "Online detection of anomalous sub-trajectories: A sliding window approach based on conformal anomaly detection and local outlier factor," in *Artificial Intelligence Applications and Innovations*, pp. 192–202. Springer, 2012.

[12] H. Vincent Poor and Olympia Hadjiliadis, *Quickest detection*, vol. 40, Cambridge University Press Cambridge, 2009.

[13] Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman, "Mondrian confidence machine," *Technical Report*, 2003.

[14] Tien Pham and Nassy Srour, "Acoustic-seismic Classification and Identification Data Set (ACIDS)," US Army Research Laboratory data CDs, 1999.

[15] Biao Chen and P Willett, "Detection of hidden Markov model transient signals," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 36, no. 4, pp. 1253–1268, 2000.

[16] Wayne R Blanding, Peter K Willett, Yaakov Bar-Shalom, and Stefano Coraluppi, "Multisensor track management for targets with fluctuating SNR," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 45, no. 4, pp. 1275–1292, 2009.

[17] Michael J Pekala, Ashley J Llorens, and I-Jeng Wang, "Local distance metric learning for efficient conformal predictors," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 2012, pp. 1–6.