Conformal Anomaly Detection on Spatio-Temporal Observations with Missing Data

Chen Xu, Yao Xie
Industrial and Systems Engineering (ISyE)
Georgia Institute of Technology
cxu310@gatech.edu, yao.xie@isye.gatech.edu

ICML International Conference On Machine Learning
Workshop on Distribution-Free Uncertainty Quantification

1. Introduction and Problem Setup

Goal and Challenges: We modify the recent EnbPI method in [2] to detect real-valued anomalous spatial-temporal observations that arrive in sequence, assuming missing values are present in training data. The task is challenging because:

- *Spatio-temporal Correlation* exists within data from different location.
- *Distribution-Free Detection* is desirable for wider applicability.
- *Quick SequentialDetection* requires computationally efficient methods.

Contribution: Traditional conformal prediction methods are distribution-free but likely have no Type-1 error control guarantee beyond exchangeable data. In contrast, our Ensemble Conformal Anomaly Detector (**ECAD**) can:

A. Computationally:

- Wrap around *most machine learning models*, from simple linear regression to complex neural networks.
- *Does not require data-splitting* of the training sample and *does not overfit* on training observations.
- Merely requires *the cost of fitting one ensemble predictor* but outputs "leave-one-out" (LOO) ensemble predictors, increasing statistical power.

B. Theoretically:

- The detector *approximately controls the Type-I error* under mild assumptions on estimator consistency and on dependency of unobservable errors. It is fully distributional-free.

C. Empirically:

- It outperforms a wide variety of supervised and unsupervised anomaly detectors on anomalous traffic flow detection in terms of F_1 score. We believe it is applicable to many other types of data, such as solar radiation data and financial crime data.

Setup: Let there be K sensors, each of which records a sequence of data $\{Y_{tk}\}_{t\geq 1}$. Assume that normal data are

$$Y_{tk} = f(X_{tk}) + \epsilon_{tk}, \tag{1}$$

where $Y_{tk} \in \mathbb{R}, X_{tk} \in \mathbb{R}^d$, and $\epsilon_{tk} \sim D$. Here, X_{tk} is an unknown feature vector (e.g., may be comprised of past and neighboring observations) and D is the distribution function for errors. Let the data matrix $\boldsymbol{Y} = \{Y_{tk}\}_{t=1,k=1}^{T,K}$ be available for training, which may have missing rows and/or columns entries. Therefore, our goals are to detect Y_{tk} that violate (1). In general, we can assume that $Y_{tk} = f_k(X_{tk}) + \epsilon_{tk}$, $X_{tk} \in \mathbb{R}^{d_k}$, and $\epsilon_{tk} \sim D_k$, but doing so is not scalable to large K. Therefore, the current assumptions on f, the dimension of features d, and D are only for computation.

2. Detection Procedure

In short, **ECAD** first fits ensemble predictors using a regression model $\mathscr{A}$ and then outputs p-values via locally ranking the test anomaly scores against LOO training scores. It has two primary phases, the training phase that trains the ensemble anomaly detector and the detection phase that detects anomalies sequentially. We only consider data with single subscripts below for notation simplicity but clarify how to apply **ECAD** to spatio-temporal data in experiments.

A. Training Phase in **ECAD**:

1. Fix a regression algorithm $\mathscr{A}$, which trains on N data points $\{(X_i, Y_i)\}_{i=1}^N$ and outputs a non-conformity mapping (NCM) $\mathscr{N} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. The input to $\mathscr{N}$ is any new data point (X_j, Y_j) and the output is the anomaly score $\hat{s}_j := |Y_j - \hat{f}(X_j)|$. Larger scores indicate more abnormal data.

2. Construct B bootstrap NCMs as follows: Assume T' out of T data are non-missing. For each $b = 1, \ldots, B$, compute $\mathscr{N}_b := \mathscr{A}(\{(X_j, Y_j)\}_{b_1}^{b_{T'}})$, which is trained on T' data sampled with replacement from the non-missing ones.

3. For the i-th non-missing training observation, we aggregate NCM from step 2 in a LOO fashion via a function ϕ (e.g. mean or median) to compute the aggregated anomaly score $\hat{s}_i^\phi$ for (X_i, Y_i). Formally, calculate $\hat{s}_i^\phi := |Y_i - \phi(\{\hat{f}_b(X_i) : i \notin \{b_1, \ldots, b_{T'}\}\}_{b=1}^B)|$, where $\hat{f}_b$ is the b-th bootstrap predictor in $\mathscr{N}_b$.

B. Detection Phase in **ECAD**:

1. Let $\hat{f}_{-i}^\phi := \phi(\{\hat{f}_b : i \notin \{b_1, \ldots, b_{T'}\}\}_{b=1}^B)$ be the i-th LOO ensemble predictor, where aggregation is applied pointwise. To detect whether the test datum $(X_t, Y_t), t > T$ is an anomaly, calculate

$$\hat{s}_t^\phi := |Y_t - (1-\alpha) \text{ quantile of } \{\hat{f}_{-i}^\phi\}_{i=1}^{T'}|,$$

where α is the user-defined significance level (or tolerated Type-I error). Then, compute the p-value $p_t := T'^{-1} \sum_{i=1}^{T'} \mathbf{1}(\hat{s}_i^\phi \geq \hat{s}_t^\phi)$ where $\mathbf{1}(\cdot)$ is the indicator function. Lastly, call x_t an anomaly if $p_t \leq \alpha$.

2. Slide the past sequence of T' anomaly scores ahead, so s_1 is dropped and s_t appended. Reset indices afterwards. Exist this phase if no more detection needs to be made. Otherwise, return to step 1 in the detection phase.

3. Experiment: Marginal Coverage

Data: We use hourly traffic flow observations from 20 sensors in 2020. Detailed data-preprocessing steps such as anomaly definitions and feature construction are not included here due to space limitation. We use data from the first six months for training and the rest for testing.

Result: The detector is called **ECAD** RNN, which lets $\mathscr{A}$ be the RNN with 2 LSTM layers, each with 100 hidden sensors and the Tanh activation function. The dense layer uses the ReLU activation function. Table 1 on the left shows the superior performance of our model comparing to other supervised and unsupervised techniques. **ECAD** Ridge uses the Ridge regression as $\mathscr{A}$. To better understand how **ECAD** RNN make decisions, we include in Figure 1 the trajectory of p-values over time, along with true/false positive/negatives. **ECAD** RNN behaves much worse at sensor 4 with only 0.66% anomalies likely because we assumed a common data-generating model (e.g. relatively similar amount of anomalies) for traffic flows at all sensors, but the model at this sensor differs very much from the rest. To improve the performance of ECAD RNN, one should fit it separately on groups of data with similar percentages of anomalies.

Sensor (Anomaly Fraction) 19 (32%) 14 (38%) 5 (31%) 3 (24%)

F_1 Score

Sensor \ Detector	RGness	ECAD RNN	ECAD Ridge	HBOS	IForest	OCSVM	PCA	GBoosting	MLP	Logistic	RF
19	0.48	**0.86**	*0.51*	0.03	0.11	0.48	0.31	0.26	0.20	0.21	0.24
14	0.55	**0.84**	*0.61*	0.02	0.00	0.57	0.00	0.32	0.30	0.43	0.28
5	0.48	**0.84**	*0.49*	0.05	0.04	0.47	0.02	0.29	0.33	0.29	0.33
3	0.39	**0.81**	*0.61*	0.07	0.00	0.40	0.02	0.04	0.05	0.02	0.04

Precision

Sensor \ Detector	RGness	ECAD RNN	ECAD Ridge	HBOS	IForest	OCSVM	PCA	GBoosting	MLP	Logistic	RF
19	0.32	**0.79**	*0.52*	0.10	0.61	0.31	0.87	0.24	0.28	0.23	0.23
14	0.38	**0.77**	*0.67*	0.10	0.05	0.40	0.06	0.37	0.45	0.40	0.33
5	0.31	**0.80**	*0.58*	0.06	0.06	0.31	0.03	0.26	0.43	0.27	0.33
3	0.24	**0.72**	*0.59*	0.17	0.03	0.25	0.16	0.04	0.06	0.03	0.05

Recall

Sensor \ Detector	RGness	ECAD RNN	ECAD Ridge	HBOS	IForest	OCSVM	PCA	GBoosting	MLP	Logistic	RF
19	**1.0**	*0.93*	0.50	0.01	0.06	1.0	0.19	0.29	0.15	0.20	0.25
14	**1.0**	*0.91*	0.56	0.01	0.00	1.0	0.00	0.28	0.22	0.46	0.24
5	**1.0**	*0.88*	0.43	0.04	0.03	1.0	0.01	0.32	0.27	0.30	0.33
3	**1.0**	*0.92*	0.62	0.04	0.00	1.0	0.01	0.03	0.04	0.02	0.04

Table 1: F_1 scores, Precision, and Recall by 11 anomaly detectors on selected sensors. The scores are sorted in descending order by the F_1 score of ECAD RNN, our desired method. Bold and italicized cells indicate the highest and second highest scores. In terms of F_1 scores, ECAD RNN yields superior performance on this task.

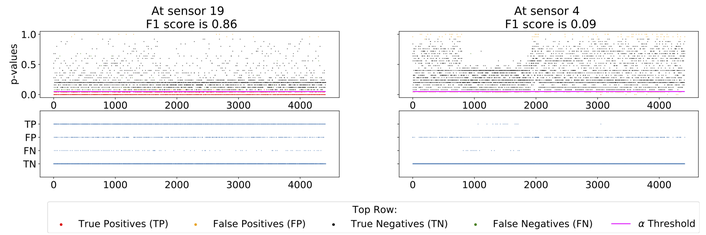

Figure 1: Performance of **ECAD** RNN on sensors at two extremes: sensor 19 on the left has roughly 32% anomalies, so that most predictions are TP and TN. In contrast, sensor 4 on the right has only 0.66% anomalies, so that FP and TN dominate predictions.

4. Extensions

Theoretically, we want to 1) Establish sequential FDR control so as to control the number of false positives. We will seek insights from works on online false discovery rate control and analyze the dependency of our p-values. 2) Study the performance guarantee when data have very few percentages of anomalies.

Experimentally, we will test ECAD on more spatial-temporal data under various regression models to examine its generality.

References

[1] Xu, Chen and Yao Xie. "Conformal Anomaly Detection on Spatio-Temporal Observations with Missing Data" In: ICML 2021 Workshop —Distribution-free Uncertainty Quantification.

[2] Xu, Chen and Yao Xie. "Conformal prediction interval for dynamic time-series" To appear as an oral paper/long talk in the Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021 (ICML 2021).

arXiv