The Journal of Open Source Software

nonconform: Conformal Anomaly Detection (Python)

Oliver Hennhöfer[1]

1 Intelligent Systems Research Group, Karlsruhe University of Applied Sciences, Karlsruhe, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Quantifying uncertainty is fundamental for AI systems in safety-critical, high-cost-of-error domains, as reliable decision-making depends on it. The Python package `nonconform` offers statistically principled uncertainty quantification for semi-supervised anomaly detection based on one-class classification (Tax, 2001). It implements methods from conformal anomaly detection (Bates et al., 2023; Jin & Candès, 2025; Laxhammar & Falkman, 2010), grounded in conformal inference (Lei & Wasserman, 2013; Papadopoulos et al., 2002; Vovk et al., 2005).

The package `nonconform` calibrates anomaly detection models to produce statistically valid p-values from raw anomaly scores. Conformal calibration uses a hold-out set $\mathcal{D}_{\text{calib}}$ of size n containing normal instances, while the model is trained on a separate normal dataset. For a new observation X_{n+1} with anomaly score $\hat{s}(X_{n+1})$, the p-value is computed by comparing this score to the empirical distribution of calibration scores $\hat{s}(X_i)$ for $i \in \mathcal{D}_{\text{calib}}$. The conformal p-value $\hat{u}(X_{n+1})$ is defined as the normalized rank of $\hat{s}(X_{n+1})$ among the calibration scores (Liang et al., 2024):

$$\hat{u}(X_{n+1}) = \frac{|\{i \in \mathcal{D}_{\text{calib}} : \hat{s}(X_i) \leq \hat{s}(X_{n+1})\}|}{n}.$$

By framing anomaly detection as a sequence of statistical hypothesis tests, these p-values enable systematic control of the *marginal* (average) false discovery rate (FDR) (Bates et al., 2023; Benjamini & Hochberg, 1995) at a predefined significance level via appropriate statistical procedures. The library integrates seamlessly with the widely used `pyod` library (Chen et al., 2025; Zhao et al., 2019), extending conformal techniques to a broad range of anomaly detection models.

Statement of Need

A major challenge in anomaly detection lies in setting an appropriate anomaly threshold, as it directly influences the false positive rate. In high-stakes domains such as fraud detection, medical diagnostics, and industrial quality control, excessive false alarms can lead to *alert fatigue* and render systems impractical.
The package `nonconform` mitigates this issue by replacing raw anomaly scores with p-values, enabling formal control of the FDR. Consequently, conformal methods become effectively *threshold-free*, since anomaly thresholds are implicitly determined by underlying statistical procedures.

$$FDR = \frac{\text{Efforts Wasted on False Alarms}}{\text{Total Efforts}}$$

(Benjamini et al., 2009)

Conformal methods are *nonparametric* and *model-agnostic*, applying to any model that produces consistent anomaly scores on arbitrarily distributed data. Their key requirement is the assumption of *exchangeability* between calibration and test data, ensuring the validity of resulting conformal p-values.
Exchangeability only requires that the joint data distribution is invariant under permutations, making it more general—and less restrictive—than the independent and identically distributed (*i.i.d.*) assumption common in classical machine learning.

To operationalize this assumption, `nonconform` constructs calibration sets from training data using several strategies, including approaches for low-data regimes (Hennhofer & Preisach, 2024) that do not require a dedicated hold-out set. Based on these calibration sets, the package computes *standard* or *weighted* conformal p-values (Jin & Candès, 2025), which are particularly useful under covariate shift, when exchangeability is only approximate.
These tools enable practitioners to build anomaly detectors whose outputs are statistically controlled to maintain the FDR at a chosen nominal level.

Overall, reliance on exchangeability makes these methods well-suited to cross-sectional data but less appropriate for time series applications, where temporal ordering conveys essential information.

Acknowledgements

This work was conducted in part under the research projects *Biflex Industrie* (Grant no. 01MV23020A) funded by the *German Federal Ministry of Economic Affairs and Climate Action*.

References

Bates, S., Candès, E., Lei, L., Romano, Y., & Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, *51*(1). https://doi.org/10.1214/22-aos2244

Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4255–4271. https://doi.org/10.1098/rsta.2009.0127

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Chen, S., Qian, Z., Siu, W., Hu, X., Li, J., Li, S., Qin, Y., Yang, T., Xiao, Z., Ye, W., Zhang, Y., Dong, Y., & Zhao, Y. (2025). PyOD 2: A python library for outlier detection with LLM-powered model selection. *Companion Proceedings of the ACM on Web Conference 2025*, 2807–2810. https://doi.org/10.1145/3701716.3715196

Hennhofer, O., & Preisach, C. (2024).Leave-One-Out-, Bootstrap- and Cross-Conformal Anomaly Detectors . *2024 IEEE International Conference on Knowledge Graph (ICKG)*, 110–119. https://doi.org/10.1109/ICKG63256.2024.00022

Jin, Y., & Candès, E. J. (2025). Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika*, asaf066. https://doi.org/10.1093/biomet/asaf066

Laxhammar, R., & Falkman, G. (2010). Conformal prediction for distribution-independent anomaly detection in streaming vessel data. *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, 47–55. https://doi.org/10.1145/1833280.1833287

Lei, J., & Wasserman, L. (2013). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 71–96. https://doi.org/10.1111/rssb.12021

Liang, Z., Sesia, M., & Sun, W. (2024). Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(3), 671–693. https://doi.org/10.1093/jrsssb/qkad138

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine learning: ECML 2002* (pp. 345–356). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-36755-1_29

Tax, D. (2001). *One-class classification; concept-learning in the absence of counter-examples* [Dissertation (TU Delft)]. Delft University of Technology. ISBN: 90-75691-05-X

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer. https://doi.org/10.1007/b106715

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, *20*(96), 1–7. http://jmlr.org/papers/v20/19-011.html