The Journal of Open Source Software

rrcf: Implementation of the Robust Random Cut Forest algorithm for anomaly detection on streams

Matthew D. Bartos[1], **Abhiram Mullapudi**[1], **and Sara C. Troutman**[1]

1 Department of Civil and Environmental Engineering, University of Michigan

DOI: 10.21105/joss.01336

Software

- Review
- Repository
- Archive

Submitted: 04 March 2019
Published: 29 March 2019

Summary

In this paper, we present the first open-source implementation of the *robust random cut forest* (RRCF) algorithm—an unsupervised ensemble method for anomaly detection on streaming data (Guha, Mishra, Roy, & Schrijvers, 2016). RRCF offers a number of features that many competing anomaly detection algorithms lack. Specifically, RRCF:

- Is designed to handle large volumes of streaming data.
- Is well-suited to data of high dimension.
- Reduces the influence of irrelevant dimensions in the input data.
- Gracefully handles duplicates and near-duplicates that could otherwise mask the presence of outliers.
- Features an anomaly-scoring metric with a clear underlying statistical meaning.

The RRCF algorithm is currently used for anomaly detection in the *Amazon Kinesis* real-time analytics engine. The goal of our repository is to provide an open-source implementation of the RRCF algorithm and its core data structures for the purposes of facilitating experimentation and enabling future extensions.

Background

Anomaly detection is an important unsupervised learning problem with economic and social implications in a variety of fields. In finance, online anomaly detection algorithms are used to alert customers to potentially fraudulent credit card transactions (Aleskerov, Freisleben, & Rao, 1997). In web infrastructure, anomaly detection algorithms facilitate improved intrusion detection (Lazarevic, Ertoz, Kumar, Ozgur, & Srivastava, 2003), and can be used to flag and deflect malicious IPs during distributed denial of service (DDoS) attacks (Mirkovic & Reiher, 2004). With respect to monitoring of industrial and infrastructure control systems, outlier monitoring can be used to identify malfunctioning industrial equipment, flag quality assurance problems, and alert supervisors to hazardous conditions (Filev, Chinnam, Tseng, & Baruah, 2010). These application areas demand powerful but flexible approaches that can process large volumes of streaming data while at the same time adapting to non-stationary conditions.

Existing anomaly detection approaches typically suffer from a few key limitations that hinder their usefulness in real-time applications. First, many conventional methods are not suited for streaming data (Guha et al., 2016). Methods such as isolation forest (Liu, Ting, & Zhou, 2012) and local outlier factor detection (Breunig, Kriegel, Ng, & Sander,

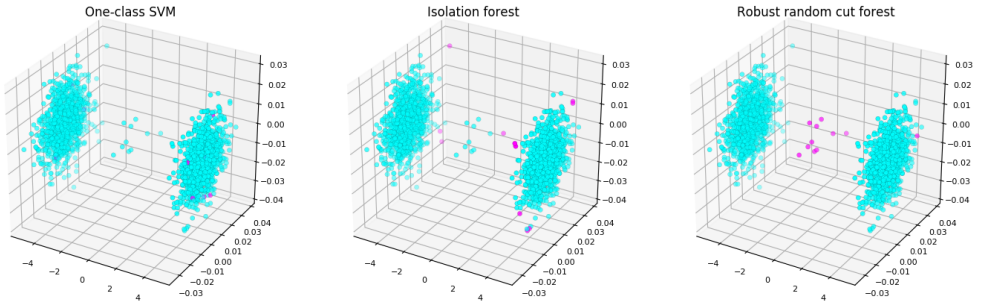

Figure 1: Figure 1: Detecting collusive outliers using One-Class SVM, Isolation Forest and RRCF (left to right).

2000) must reconstruct the entire model when a new data point is added, while methods such as replicator neural networks (Williams, Baxter, He, Hawkins, & Gu, 2002) and one-class support vector machines (Tax & Duin, 2004) must be retrained if the distribution of the input data changes over time. Second, many existing algorithms fail to detect anomalies in the presence of duplicates or near-duplicates—a phenomenon known as "outlier masking" (Guha et al., 2016). Finally, existing algorithms often struggle to detect outliers in high-dimensional data, especially in the presence of "irrelevant dimensions" that offer little relevant information to the outlier detection problem (Guha et al., 2016). The *robust random cut forest* algorithm addresses these problems by using a novel sketching algorithm to construct a real-time summary of the data (Guha et al., 2016). This sketching algorithm works by (i) constructing an ensemble of space-partitioning binary trees on the point set, and then (ii) generating an anomaly score based on the conditional change in model complexity imposed by the insertion or deletion of each point. This method efficiently detects anomalies on streaming data while at the same time adapting to changing input signals and handling "collusive" outliers.

Example applications

Test 1: Detection in the presence of collusive outliers and irrelevant dimensions

Here, we validate a test case from the original RRCF paper involving two dense clusters of 3-dimensional points in which almost all the variation occurs in the first dimension (Guha et al., 2016). Consider two clusters consisting of 1000 points each distributed according to $x_i \sim \mathcal{N}([5\ 0\ 0]^T, 0.01 \cdot I)$ for the first class and $x_i \sim \mathcal{N}([-5\ 0\ 0]^T, 0.01 \cdot I)$ for the second class. Add to this point set 10 collusive outlier points distributed according to $x_i \sim \mathcal{N}([0\ 0\ 0]^T, 0.01 \cdot I)$. The goal of this test case is to detect the 10 collusive outlier points in the center. From Figure 1, it can be seen that RRCF successfully detects the outliers, while One-Class SVM and Isolation Forest do not.

Test 2: Detection of event onset

We validate another result from the original paper by showing that RRCF can effectively detect both the onset and offset of anomalies on streaming time series data (Guha et al., 2016). As in the original paper, we generate a sine wave from time $t \in [0, 730]$, with period $T = \pi/50$ and a phase offset of $\phi = 30$. An anomaly is then injected into the sine wave along the interval $t \in [235, 255]$. Using a technique recommended by the original

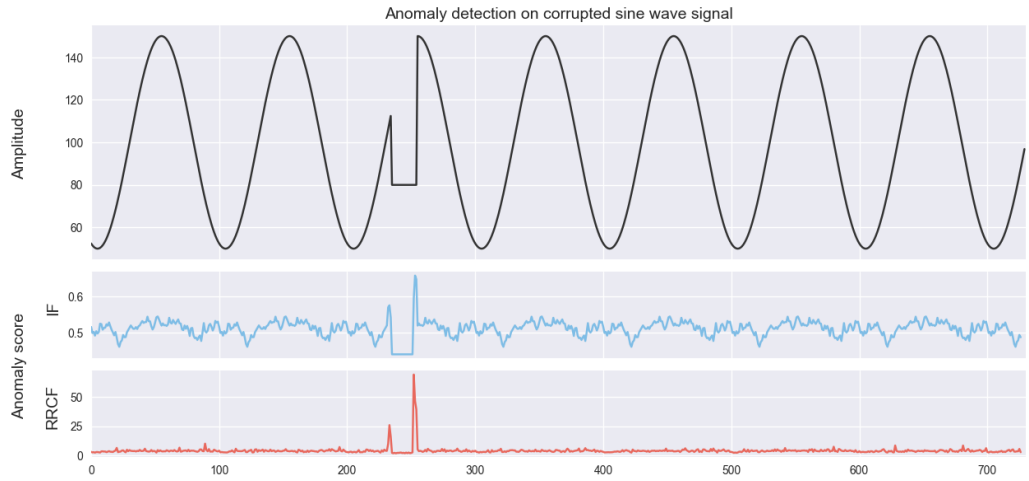

Figure 2: Figure 2: Detecting event onset using Isolation Forest (center) and RRCF (bottom).

paper, we sample the data using a "shingling" approach. In this approach, a window of length ℓ is passed over the data, and each consecutive windowed sequence is treated as an $\ell-$dimensional point (e.g. for a shingle of length 4, the sampled point at time t will be $[x_{t-3},\ x_{t-2},\ x_{t-1},\ x_t]^T$). We apply both Isolation Forest and RRCF with a forest size of 100 trees, and a tree size of 256 points. From Figure 2, it can be seen that RRCF performs better than Isolation Forest at detecting the onset of the anomaly.

Acknowledgements

We thank Dr. Alfred Hero for bringing the RRCF algorithm to our attention, and Dr. Branko Kerkez for his invaluable support.

References

Aleskerov, E., Freisleben, B., & Rao, B. (1997). CARDWATCH: A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr)*. IEEE. doi:10.1109/cifer.1997.618940

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *SIGMOD Rec.*, *29*(2), 93–104. doi:10.1145/335191.335388

Filev, D. P., Chinnam, R. B., Tseng, F., & Baruah, P. (2010). An industrial strength novelty detection framework for autonomous equipment monitoring and diagnostics. *IEEE Transactions on Industrial Informatics*. doi:10.1109/TII.2010.2060732

Guha, S., Mishra, N., Roy, G., & Schrijvers, O. (2016). Robust random cut forest based anomaly detection on streams. In *International conference on machine learning* (pp. 2712–2721).

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., & Srivastava, J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM international conference on data mining*. Society for Industrial; Applied Mathematics. doi:10.1137/1.9781611972733.3

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, *6*(1), 1–39. doi:10.1145/2133360.2133363

Mirkovic, J., & Reiher, P. (2004). A taxonomy of DDoS attack and DDoS defense mechanisms. *ACM SIGCOMM Computer Communications Review*. doi:10.1145/997150.997156

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*(1), 45–66. doi:10.1023/B:MACH.0000008084.60811.49

Williams, G., Baxter, R., He, H., Hawkins, S., & Gu, L. (2002). A comparative study of RNN for outlier detection in data mining. In *Data mining, 2002. ICDM 2003. Proceedings. 2002 ieee international conference on* (pp. 709–712). IEEE. doi:10.1109/ICDM.2002.1184035