Actionable Data Monitoring in Modern Data Streams

Ananya Joshi
aajoshi@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Bryan Wilder
bwilder@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Roni Rosenfeld
rrosenfeld@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

Traditional outlier detection methods are inadequate for modern data monitoring, where humans must identify actionable data in real time across vast volumes of noisy, nonstationary, and heterogeneous data streams. This work introduces a new system tailored for actionable modern data monitoring, featuring: a) a framework and evaluation strategy for deployed monitoring systems, and b) a novel outlier detection approach and new machine learning task designed for modern streaming data. Data experts using a real-world implementation of this system in a public health context currently identify approximately 200 actionable data points per week related to outbreaks or data quality issues. By effectively addressing the critical challenges traditional methods face in practical contexts like this, our system demonstrates significant potential for improving actionable data monitoring in modern settings.

CCS CONCEPTS

• **Information systems** → **Data stream mining**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Data Monitoring, Heterogeneous Time Series, Big Data Analytics, Actionable Machine Learning

ACM Reference Format:
Ananya Joshi, Bryan Wilder, and Roni Rosenfeld. 2024. Actionable Data Monitoring in Modern Data Streams. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Data reviewers are humans who continuously inspect individual data points from large volumes of streaming time series data to identify actionable insights necessary for maintaining complex systems or careful downstream processing (a process called data monitoring). As data volumes increase and include heterogeneous, short, and nonstationary data streams (modern setting), existing approaches for data monitoring, such as online anomaly or outlier detection, struggle to adapt because they cannot: (a) quickly process large volumes of data to maintain a real-time setting, (b) identify insights from modern data streams, and (c) meet human-centric requirements like accounting for limited data reviewer attention. Our experiments [9, 10] show that modern outlier detection methods (e.g., using deep learning) struggle to identify insightful data from short, nonstationary streams [13], and simpler statistical methods struggle with noisy and seasonal data [1]. These methods often only identify global outliers that rarely require human verification. Moreover, they usually handle large data volumes via dimension reduction [5, 8], which may assume static stream correlation structures that are violated in modern settings and may not provide the requisite individual data-stream level insights for data monitoring by design. Finally, these methods can be opaque [7] and result in reviewing loads so large [4] that trained reviewers frequently miss nuanced insights due to being overwhelmed [9].

Despite these challenges, data monitoring is increasingly relevant in industrial[1] and practical applications. We work with a group that publishes large volumes of updated public health data [2] and needs data monitoring to quickly identify insights related to changes in disease dynamics or data quality for their stakeholders [11, 12]. As technological advancements continue to increase the demand for real-time monitoring across more variables (streams) [15], that are updated more quickly [14], and are more complex [2, 6], like that in the public health setting, new approaches are needed.

2 APPROACH

In our approach, a data reviewer inspects only the top entries from a computer-generated ranked list of insights to prioritize their attention. This list is created using a novel real-time multiple-univariate outlier detection framework, comprising of: (**M1**) a method to identify insights per data stream and (**M2**) a method to rank these insights across all streams.

M1: Per Stream We use a modified hypothesis-driven prediction-based univariate outlier detection approach [3] that identifies insightful data requiring human review (e.g. excluding obvious global outliers corresponding to evident failures). We assume that each data stream is actually produced by a complex and unknown model. We then test the hypothesis that the most recent point in the stream is drawn from a simple and quick approximation of that model. If the recent data has a low probability under this hypothesis, it indicates that this most recent point had a larger deviation from the simple model than it has had historically and can thus potentially correspond to valuable insights. Specifically, our approach accounts for seasonality and global outliers in different steps of the framework, and we build test statistics that nonparametrically capture the fit of simple data-generating models across heterogeneous streams with short training histories (due to nonstationarity) by combining information across similar streams.

M2: Scaling Across All Streams Applying univariate outlier detection methods like **M1** independently to a large number of

[1]For example, companies like MetaPlan, CrowdStrike, and Splunk have already designed products that address the continuous monitoring setting.
[2]Delphi Research Group

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

data streams often produces many tied scores, leading to numerous false positives. This is because outputs from methods like **M1** are contextualized within individual or small sets of similar streams and should not be directly ranked across all streams. To address this, we introduce a new machine learning task, called **multi-stream outlier ranking**, to rank the highest-priority outliers across all data streams using the test statistics from univariate outlier detection methods. Our method prioritizes data deviance that is extreme relative to the recent historical behavior of the univariate method outputs across all streams. Specifically, given a list of similar streams and their test statistics from **M1**, our approach constructs a reference test statistic distribution using a modified extreme value analysis technique that only requires a limited amount of recent history to produce granular and interpretable scores for ranking. Because **M2** works with any univariate outlier detection method and on different types of streams, it can adapt to fluctuations in modern data that result in changes to **M1**.

3 EVALUATION AND RESULTS

Accurately evaluating algorithms for unsupervised time series outlier detection is challenging [17]. Our approach utilizes IRB-approved and preregistered surveys with custom evaluation interfaces to obtain high-quality labeled data from experts who curate, publish, and use public health data (n=13-17)[3]. Unlike binary outlier detection labels, which may be insufficient as our survey results show that different experts have varying thresholds for outlier determination, asking experts to rank data points corresponding to insights provides a more informative comparison. In these evaluations, we also perform ablation studies, and our approaches match or outperform 13+ previous outlier detection methods used for data monitoring.

M1: In performance experiments and expert evaluations, we found that: (a) **M1** was computationally feasible, with a few algorithms (mainly deep learning algorithms) which did not finish training within one day and needed a tradeoff between addressing nonstationarity or timeliness. (b) **M1** performs as well as the unusable deep learning approaches and better than other methods on standard binary and ranking outlier detection metrics. (c) **M1** identifies useful points that were unlikely to have been inspected without computational assistance (via an algorithm identifying the point) [4], which is a result of **M1**'s emphasis the types of insightful data points that our prototyping showed are difficult for humans to recognize in noisy streams. Methods designed with this approach can leverage the statistical properties of particular domains and innately enhance trust in method outputs that may also translate to performance gains.

M2: Because multi-stream outlier ranking is a new task, we compared our approach to three other approaches currently used to adapt streaming outlier detection methods to monitoring settings across 6+ relevant univariate outlier detection methods. Our evalutions showed that **M2** (a) is **4.5x** faster than the M1 approach alone, (b) outperforms all other approaches across different combinations

[3]This is a significant increase over previous work (e.g., n=3 in [16]).

[4]This metric was the set of points that (a) the majority of humans rated as warranting investigation after a full examination, and (b) at least 40% of such respondents said that they were "unlikely" or "somewhat unlikely" to have identified the point without algorithmic assistance

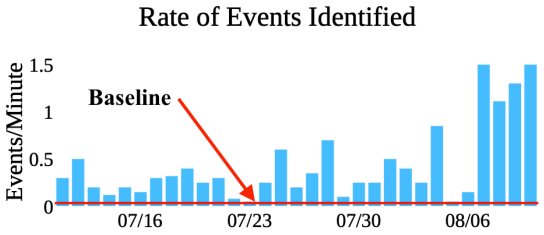

Figure 1: Deployed results of the monitoring methods, originally from [10], show that reviewers can identify actionable insights more quickly with M1/M2 output than without.

of univariate outlier detection methods (as shown in [10]) , and (c) increases the rate at which data reviewers parse insights from the data monitoring process in a deployed setting by **9.1x** compared to their prior approach, as shown in Fig 1.

4 DISCUSSION

Recent changes to the framework and user interface of our monitoring system have led to a 53x efficiency increase during a 3-month IRB-approved evaluation. Additionally, over the past year, users have identified up to 200 data insights per week, split between data quality issues and disease dynamics. Experts have also found various patterns within these insights ("meta-insights"). Based on their feedback, we propose three new functionalities: (1) methods to identify and handle overlapping anomalies of different lengths, (2) a system using reviewer feedback to improve efficiency, and (3) using foundational models to summarize insights from outliers over time. Preliminary progress for (1) using synthetic datasets shows a 95% chance of a reviewer detecting an insight each time they review up to 50 streams based on 20 simulations with 2000 streams.

Overall, we use computational approaches to tackle the core challenges of monitoring modern data for actionable insights. Our contributions include two novel methods, novel evaluations in deployed settings, and results from deployment demonstrating the actionability of these methods. Future work will address challenges in making these methods accessible to stakeholders by reducing the overhead required for this approach's maintenance and updates.

Acknowledgements The larger project this work is part of includes contributions from Nolan Gormley, Richa Gadgil, Tina Townes, Luke Neureiter, and Katie Mazaitis. This material was pulled together from [9, 10], my thesis drafts, other application materials, and additional preliminary results.

REFERENCES

[1] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33, 2021.

[2] A. P. Brady, B. Allen, J. Chong, E. Kotter, N. Kottler, J. Mongan, L. Oakden-Rayner, D. P. Dos Santos, A. Tang, C. Wald, et al. Developing, purchasing, implementing and monitoring ai tools in radiology: practical considerations. a multi-society statement from the acr, car, esr, ranzcr & rsna. *Insights into Imaging*, 15(1):16, 2024.

[3] H. S. Burkom, S. Murphy, J. Coberly, and K. Hurt-Mullen. Public health monitoring tools for multiple data streams. *Morbidity and Mortality Weekly Report*, 54(Supplement on Syndromic Surveillance):55–62, 2005.

[4] M. Coletta and H. Zhou. What can you really do with 35,000 statistical alerts a week anyways? *Online Journal of Public Health Informatics*, 11(1), 2019.

[5] F. Gottwalt, E. Chang, and T. Dillon. Corrcorr: A feature selection method for multivariate correlation network anomaly detection techniques. *Computers & Security*, 83:234–245, 2019.

[6] J. Howard. Artificial intelligence: Implications for the future of work. *American journal of industrial medicine*, 62(11):917–926, 2019.

[7] K. J. Hurt-Mullen and J. Coberly. Syndromic surveillance on the epidemiologist's desktop: making sense of much data. *MMWR Morb Mortal Wkly Rep*, 54(Suppl):141–6, 2005.

[8] Z. Jadidi, S. Pal, M. Hussain, and K. Nguyen Thanh. Correlation-based anomaly detection in industrial control systems. *Sensors*, 23(3):1561, 2023.

[9] A. Joshi, K. Mazaitis, R. Rosenfeld, and B. Wilder. Computationally assisted quality control for public health data streams. *arXiv preprint arXiv:2306.16914*, 2023.

[10] A. Joshi, T. Townes, N. Gormley, L. Neureiter, R. Rosenfeld, and B. Wilder. Outlier ranking for large-scale public health data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22176–22184, 2024.

[11] M. D. of Health. Data: Quality, analysis, and interpretation, 10 2022.

[12] U. S. G. A. Office. Covid-19 data quality and considerations for modeling and analysis. https://www.gao.gov/assets/gao-20-635sp.pdf, 07 2020.

[13] A. Paleyes, R.-G. Urma, and N. D. Lawrence. Challenges in deploying machine learning: a survey of case studies. *ACM Computing Surveys*, 55(6):1–29, 2022.

[14] A. Reinhart, L. Brooks, M. Jahja, A. Rumack, J. Tang, S. Agrawal, W. Al Saeed, T. Arnold, A. Basu, J. Bien, et al. An open repository of real-time covid-19 indicators. *Proceedings of the National Academy of Sciences*, 118(51):e2111452118, 2021.

[15] G. Shurkhovetskyy, N. Andrienko, G. Andrienko, and G. Fuchs. Data abstraction for visualizing large time series. In *Computer Graphics Forum*, volume 37, pages 125–144. Wiley Online Library, 2018.

[16] W.-K. Wong. *Data mining for early disease outbreak detection*. Carnegie Mellon University, 2004.

[17] R. Wu and E. J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE transactions on knowledge and data engineering*, 35(3):2421–2429, 2021.