Novel algorithm for discovering anomalies in data outperforms current software

July 26 2024, by Tina Hilding

Anomaly detection poses several challenges that are not seen in traditional AI problems. Credit: Trevor Vannoy on Unsplash

An algorithm developed by Washington State University researchers can better find data anomalies than current anomaly-detection software, including in streaming data.

The work, reported in the *Journal of Artificial Intelligence Research*, makes fundamental contributions to artificial intelligence (AI) methods that could have applications in many domains that need to quickly find anomalies in large amounts of data, such as in cybersecurity, power grid management, misinformation, and medical diagnostics.

Being able to better find the anomalies would mean being able to more easily discover fraud, disease in a medical setting, or important unusual information, such as an asteroid whose signals overlap with the light from other stars.

"This work presents advances on how AI and humans can work together to synergistically solve anomaly discovery problems," said Jana Doppa, Huie-Rogers Endowed Chair Associate Professor of Computer Science who supervised the work.

"With all this generative AI technology, there is so much data which includes misinformation, and if you want humans to go over all of this, it's impossible as it's huge. If you have finite human resources, and you want to detect something like misinformation quickly, you want algorithms that prioritize which items should be labeled."

Anomaly detection poses several challenges that are not seen in traditional AI problems. The number of anomalies is very small compared to the normal data—typically less than 2%. Furthermore, there may not be a big difference between an anomaly and normal data.

"So, it's like finding needles in a big haystack kind of a problem," said Doppa. "And you don't even know in a lot of domains what needles to look for."

Another problem is that with large amounts of data, AI will oftentimes find too many candidate anomalies to pass along for people to check.

"Whenever you have these false positives, you are wasting a lot of humans' time, which we want to minimize," said postdoctoral researcher and lead author Shubhomoy Das. "How can we use minimal feedback from the human to adapt the anomaly detector so that the false positives go down over time, and we discover more and more diverse anomalies?"

As part of the work, the researchers provided new theoretical and empirical findings for why an ensemble of computer models worked well for anomaly discovery. They found that with just a small amount of step-by-step feedback, the AI algorithm can learn much better and discover many more diverse anomalies compared to a system where there was no feedback. The human needs an explanation regarding the candidate anomalies to understand why AI selected them for labeling.

"Some notion of interpretability or explainability is important," said Ph.D. student and co-author Rakibul Islam. "What we figured was that this was largely missing in the existing literature."

The researchers used their new findings to develop an algorithm that looks at anomalies in batches, which improved the ability to discover diverse types of anomalies. So, in the case of anomalous credit card data, the algorithm discovers different types of unusual behavior, such as a person's oddly expensive purchases and/or ones that are made in an odd location.

Unlike current AI models, the algorithm the researchers developed was able to handle streaming data, which is common in many real-world applications. Their algorithm can detect and quantify drift in the data distribution and then take corrective action.

"The problem of discovering anomalies when the data is coming in a stream was less studied," Doppa said.

The researchers' computer code and data are publicly available, and they now plan to deploy their algorithms in real-world systems to measure their accuracy and usability.

More information: Shubhomoy Das et al, Effectiveness of Tree-based Ensembles for Anomaly Discovery: Insights, Batch and Streaming Active Learning, *Journal of Artificial Intelligence Research* (2024). DOI: 10.1613/jair.1.14741

Provided by Washington State University