Using artificial intelligence to find anomalies hiding in massive datasets

February 24 2022, by Adam Zewe

Credit: CC0 Public Domain

Identifying a malfunction in the nation's power grid can be like trying to find a needle in an enormous haystack. Hundreds of thousands of interrelated sensors spread across the U.S. capture data on electric

current, voltage, and other critical information in real time, often taking multiple recordings per second.

Researchers at the MIT-IBM Watson AI Lab have devised a computationally efficient method that can automatically pinpoint anomalies in those data streams in real time. They demonstrated that their artificial intelligence method, which learns to model the interconnectedness of the power grid, is much better at detecting these glitches than some other popular techniques.

Because the machine-learning model they developed does not require annotated data on power grid anomalies for training, it would be easier to apply in real-world situations where high-quality labeled datasets are often hard to come by. The model is also flexible and can be applied to other situations where a vast number of interconnected sensors collect and report data, like traffic monitoring systems. It could, for example, identify traffic bottlenecks or reveal how traffic jams cascade.

"In the case of a power grid, people have tried to capture the data using statistics and then define detection rules with domain knowledge to say that, for example, if the voltage surges by a certain percentage, then the grid operator should be alerted. Such rule-based systems, even empowered by statistical data analysis, require a lot of labor and expertise. We show that we can automate this process and also learn patterns from the data using advanced machine-learning techniques," says senior author Jie Chen, a research staff member and manager of the MIT-IBM Watson AI Lab.

The co-author is Enyan Dai, an MIT-IBM Watson AI Lab intern and graduate student at the Pennsylvania State University. This research will be presented at the International Conference on Learning Representations.

Probing probabilities

The researchers began by defining an anomaly as an event that has a low probability of occurring, like a sudden spike in voltage. They treat the power grid data as a probability distribution, so if they can estimate the probability densities, they can identify the low-density values in the dataset. Those data points which are least likely to occur correspond to anomalies.

Estimating those probabilities is no easy task, especially since each sample captures multiple time series, and each time series is a set of multidimensional data points recorded over time. Plus, the sensors that capture all that data are conditional on one another, meaning they are connected in a certain configuration and one sensor can sometimes impact others.

To learn the complex conditional probability distribution of the data, the researchers used a special type of deep-learning model called a normalizing flow, which is particularly effective at estimating the probability density of a sample.

They augmented that normalizing flow model using a type of graph, known as a Bayesian network, which can learn the complex, causal relationship structure between different sensors. This graph structure enables the researchers to see patterns in the data and estimate anomalies more accurately, Chen explains.

"The sensors are interacting with each other, and they have causal relationships and depend on each other. So, we have to be able to inject this dependency information into the way that we compute the probabilities," he says.

This Bayesian network factorizes, or breaks down, the joint probability

of the multiple time series data into less complex, conditional probabilities that are much easier to parameterize, learn, and evaluate. This allows the researchers to estimate the likelihood of observing certain sensor readings, and to identify those readings that have a low probability of occurring, meaning they are anomalies.

Their method is especially powerful because this complex graph structure does not need to be defined in advance—the model can learn the graph on its own, in an unsupervised manner.

A powerful technique

They tested this framework by seeing how well it could identify anomalies in power grid data, traffic data, and water system data. The datasets they used for testing contained anomalies that had been identified by humans, so the researchers were able to compare the anomalies their model identified with real glitches in each system.

Their model outperformed all the baselines by detecting a higher percentage of true anomalies in each dataset.

"For the baselines, a lot of them don't incorporate graph structure. That perfectly corroborates our hypothesis. Figuring out the dependency relationships between the different nodes in the graph is definitely helping us," Chen says.

Their methodology is also flexible. Armed with a large, unlabeled dataset, they can tune the model to make effective anomaly predictions in other situations, like traffic patterns.

Once the model is deployed, it would continue to learn from a steady stream of new sensor data, adapting to possible drift of the data distribution and maintaining accuracy over time, says Chen.

Though this particular project is close to its end, he looks forward to applying the lessons he learned to other areas of deep-learning research, particularly on graphs.

Chen and his colleagues could use this approach to develop models that map other complex, conditional relationships. They also want to explore how they can efficiently learn these models when the graphs become enormous, perhaps with millions or billions of interconnected nodes. And rather than finding anomalies, they could also use this approach to improve the accuracy of forecasts based on datasets or streamline other classification techniques.

More information: Paper: Graph-augmented Normalizing Flows for Anomaly Detection of Multiple Time Series, openreview.net/forum?id=45L_dgP48Vd

Provided by Massachusetts Institute of Technology