Generating realistic stock market data for deeper financial research

February 18 2020

Credit: CC0 Public Domain

Financial markets are among the most well-studied and closely watched

complex systems in existence. This rich literature on market modeling and analysis has led to many important innovations, such as automated tools for detecting market manipulation. But a large gap still exists between the current state-of-the-art and the powerful insights needed to fully understand the complex dimensions of market behavior.

Ultimately, these models need huge volumes of data—beyond even what's produced from real stock orders. Real-world stock order data offers researchers only a limited, historical view of the behavior the market can display. Models also require hypothetical scenarios and branching possibilities to inform deeper research.

A team at the University of Michigan has provided one answer to this need in the form of automatically generated, fake data. The team, led by Lynn A. Conway Professor of Computer Science and Engineering Michael Wellman, proposes an approach to generating realistic and high-fidelity stock market data based on a deep learning technique called generative adversarial networks (GANs). The resulting synthetic order streams open many doors for financial researchers in need of huge datasets to study the complex cause and effect relationships that play out every day in real markets.

In a nutshell, GANs work by placing two learning models against each other, one called the "generator" and the other the "discriminator." The two operate in a competitive relationship, where the generator learns how to spit out synthetic data based on what it's fed, while the discriminator learns how to tell the difference between the real and fake data streams.

As the discriminator gets better at catching fakes, the generator gets better at making its fakes more convincing. The end result is a generator capable of mimicking the target datasets very closely; in this case, stock order streams.

Called Stock-GAN, the instance used by the Michigan team was trained on two types of data sets composed of stock orders: one from an agent-based market simulator and another from a real stock market. They evaluated their generated data using a variety of statistics, such as the distribution of price and quantity of orders, inter-arrival times of orders, and the best bid and best ask evolution over time. The results showed that their generated data closely matched the corresponding statistics in real data, for both the simulated and real markets.

While this work is just a first step toward generating realistic order streams, says Xintong Wang, a Ph.D. student on the team, "acing this task may help to prepare datasets which can make other tasks possible."

In particular, new machine learning algorithms that specialize in automated trading can be trained and validated on the generated datasets, and automated anomaly detection could be made possible by comparing generated data with the actual market.

As Wang puts it, this system essentially allows finance researchers to undertake alt-history, or counterfactual, research—a technique that's not possible when restricted to real-world order streams.

"Real, historical market data can be viewed as one run out of many possible outcomes realized by nature," she explains, "and Stock-GAN can generate many more at low cost."

In addition to changing history, fully-realized synthetic stock data can also help finance researchers explore hypothetical scenarios, inserting specific data in order streams and observing the resulting permutations of future data.

"This allows us in principle to inject events into the system and observe a counterfactual evolution of the market," Wang says, "which is something

we can never get out of observational data directly."

Beyond detecting fraudulent or manipulative behavior, models trained on this data could offer researchers insight into the different kinds of legitimate trading practice exercised in markets and what results those yield.

"We would like to be able to more generally figure out what kinds of strategies traders are using," says Wellman. "With that knowledge, we could determine when an order stream contains certain strategies."

The researchers also note that running financial research on synthetic data overcomes the privacy and security issues associated with publicizing real trading data.

"Overall," the authors write, "our work provides fertile ground for future research at the intersection of deep learning and finance."

This research was published in the paper "Generating Realistic Stock Market Order Streams" at the 2020 Association for the Advancement of Artificial Intelligence (AAAI) Conference.

  More information: Generating Realistic Stock Market Order Streams: pdfs.semanticscholar.org/bfcf/ … 6ea28227f5908b51.pdf

Provided by University of Michigan