Researchers teach 'machines' to detect Medicare fraud

October 30 2018

Taghi M. Khoshgoftaar, Ph.D., co-author and Motorola Professor in FAU's Department of Computer and Electrical Engineering and Computer Science. Credit: Florida Atlantic University

Using a highly sophisticated form of pattern matching, researchers from Florida Atlantic University's College of Engineering and Computer Science are teaching "machines" to detect Medicare fraud. Medicare, the primary health care coverage for Americans 65 and older, accounts for 20 percent of health care spending in the United States. About $19 billion to $65 billion is lost every year because of Medicare fraud, waste or abuse.

Like the proverbial "needle in a haystack," human auditors or investigators have the painstaking task of manually checking thousands of Medicare claims for specific patterns that could indicate foul play or fraudulent behaviors. Furthermore, according to the U.S. Department of Justice, right now fraud enforcement efforts rely heavily on health care professionals coming forward with information about Medicare fraud.

A study published in the journal *Health Information Science and Systems* is the first to use big data from Medicare Part B and employ advanced data analytics and machine learning to automate the fraud detection process. Programming computers to predict, classify and flag potential fraudulent events and providers could significantly improve fraud detection and lighten the workload for auditors and investigators.

Researchers from FAU's Department of Computer and Electrical Engineering and Computer Science examined Medicare Part B dataset from 2012 to 2015. They focused on detecting fraudulent provider claims within the dataset, which consisted of 37 million cases. Fraudulent activities include patient abuse or neglect as well as billing for services not rendered. Physicians and other providers who commit fraud are excluded from participating in federal health care programs like Medicare, and these cases are labeled as "fraud."

For the study, the researchers aggregated the 37 million cases down to a smaller dataset of 3.7 million and identified a unique process to map

fraud labels with known fraudulent providers.

Medicare Part B data included provider information, average payments and charges, procedure codes, the number of procedures performed as well as the medical specialty, which is referred to as provider type. In order to obtain exact matches, the researchers only used the National Provider Identifier (NPI) to match fraud labels to the Medicare Part B data. The NPI is a single identification number issued by the federal government to health care providers.

Researchers directly matched the NPI across the Medicare Part B data, flagging any provider in the "excluded" database as being "fraudulent." The research team classified a physician's NPI or specialty and specifically looked at whether the predicted specialty differed from the actual specialty, as indicated in the Medicare Part B data.

"If we can predict a physician's specialty accurately based on our statistical analyses, then we could potentially find unusual physician behaviors and flag these as possible fraud for further investigation," said Taghi M. Khoshgoftaar, Ph.D., co-author and Motorola Professor in FAU's Department of Computer and Electrical Engineering and Computer Science. "For example, if a dermatologist is accurately classified as a cardiologist, then this could indicate that this particular physician is acting in a fraudulent or wasteful way."

For the study, Khoshgoftaar, along with Richard A. Bauder, senior author, a Ph.D. student at FAU and a data scientist at FPL, and Matthew Herland, a Ph.D. student in FAU's Department of Computer and Electrical Engineering and Computer Science, had to address the fact that the original labeled big dataset was highly imbalanced. This imbalance occurred because fraudulent providers are much less common than non-fraudulent providers. This scenario can be likened to "where's Waldo," and is problematic for machine learning approaches because the

algorithms are trying to distinguish between the classes—and one dominates the other thereby fooling the learner.

To combat this imbalance, the researchers used random undersampling to reduce the dataset from the 3.7 million cases down to about 12,000 cases. They created seven class distributions and used six different learners across class distributions from severely imbalanced to balanced.

Results from the study show statistically significant differences between all of the learners as well as differences in class distributions for each learner. RF100 (Random Forest), a learning algorithm, was the best at detecting the positives of potential fraud events.

More interestingly, and contrary to popular belief that balanced datasets perform the best, this study found that was not the case for Medicare fraud detection. Keeping more of the non-fraud cases actually helped the learner/model better distinguish between the fraud and non-fraud cases. Specifically, the researchers found the "sweet spot" for identifying Medicare fraud to be a 90:10 distribution of normal vs. fraudulent data.

"There are so many intricacies involved in determining what is fraud and what is not fraud such as clerical error," said Bauder. "Our goal is to enable machine learners to cull through all of this data and flag anything suspicious. Then, we can alert investigators and auditors who will only have to focus on 50 cases instead of 500 cases or more."

This detection method also has applications for other types of fraud including insurance and banking and finance. The researchers are currently adding other Medicare-related data sources such as Medicare Part D, using more data sampling methods for class imbalance, and testing other feature selection and engineering approaches.

"Given the importance of Medicare, which insures more than 54 million

Americans over the age of 65, combating fraud is an essential part in providing them with the quality health care they deserve," said Stella Batalama, Ph.D., dean of FAU's College of Engineering and Computer Science. "The methodology being developed and tested in our college could be a game changer for how we detect Medicare fraud and other fraud in the United States as well as abroad."

More information: Richard A. Bauder et al, The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data, *Health Information Science and Systems* (2018). DOI: 10.1007/s13755-018-0051-3

Provided by Florida Atlantic University