The Machine Learning approach to data quality

wipro

The world of data is expanding at an unimaginable pace. Researchers estimate that by 2020, every human would create 1.7MB of information each second. The true power of data can be unlocked when it is refined and transformed into a high quality state where we can realize its true potential.

Many businesses and researchers believe that data quality is one of the primary concerns for data-driven enterprises and associated processes considering the pace of data growth. Repeatedly, bad data is sighted as one of the root causes for failure of data-dependent initiatives. Most of the operational processes and analytics rely on good quality data for being efficient and consistent in output. Data preparation tasks take up more than half of the data managers' and data scientists' time.

Data quality is not new, but has undergone constant makeover with time. Manual data quality assessment, cleansing and deduplication processes have gradually passed on the baton to rule-based automation, which uses data quality tools. This has relieved, to some extent, data preparation tasks but a huge scope of improvement still exists. One question that stands tall is "What saves cost emanating from poor data quality, is scalable to match up with unprecedented data growth and solves the time conundrum too?" The answer is "Multifaceted, insights-driven and new age solutions like Artificial Intelligence (AI)/Machine Learning(ML) that learn from situations and condense the efforts spent towards data quality management and also reduce the process time by leveraging distributed enhanced computations."

Data quality - The Machine Learning way

Data quality process has evolved in its capacity but the demand for pace and efficiency has been proliferating extensively. Data management experts believe that data quality remains a bottleneck that creeps repeatedly to bother the data management and business fraternity due to proliferating data volumes and the complexity involved to derive quality insights. Innovative technologies such as Big Data, AI, ML etc. have made the application of large scale advanced data analytics more tangible than before.

A transition in data quality process is noticeable from static rule-based approach to a dynamic, self-adapting, learning-based ML approach in various domains. ML has the potential to assess the quality of data assets, predict missing values, and provide cleansing recommendations, thereby reducing the complexity and efforts spent by data quality experts and scientists. Some businesses already use ML models to identify and eliminate fake customer records to target their marketing at genuine customers through reliable data.

Also, ML partially substitutes the role of data stewards by flagging the data points based on probabilistic ratings as per learning from training set of past data steward decisions and categorizing duplicate, vacant, incorrect or suspicious entries. This reduces manual effort and governance activities.

It is always beneficial for the organization to have a good hold of what data is being procured and consumed, and the business purpose it would solve. ML provides assistance in deriving a data quality index score to assess data sets' quality and reliability in real time based on deviation from predicted parameter values. It also has a marked ability to predict trends and identify outliers if trained properly and can make suggestions or take actions on the go.

ML algorithms can learn from human decision labels in the training datasets and replicate the scenarios in real-time. However, ML algorithms are also prone to biases that may reflect in these data sets and are learnt through fresh data sets. These biases could lead to erosion of data quality. External validity testing and audits on a regular basis will help in avoiding such situations.

The road ahead

Most businesses look for fast analytics with high quality insights to deliver real time benefits based on quick decisions. They consider this a top priority and means of competitive advantage. To enable this, there is an opportunity for the organizations to fine-tune and enhance current data quality approach using ML techniques.

Most leading data quality tools and solution providers have ventured out into ML territory in anticipation of increasing the effectiveness of their solutions. Thus, it has the propensity of being a game changer for the businesses in pursuit of improved data quality.
ML can complement the contemporary rule-based solutions, and synergistically and gradually lead to process evolution around data quality. Although the current maturity level of the use of ML for data quality assessment and enhancement is low, it has promising future prospects to churn large data sets and enhance data quality.

Endnote

1 - https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#5fc22a3617b1

About the author

Mohan Mahankali, Practice Leader and Principal Architect - Information Management, Data, Analytics & Artificial Intelligence, Wipro

Mohan has 20+ years of business and IT experience in the areas of information management and analytics solutions for global organizations. In his current role, he is responsible for practice vision and strategy, solution definition, customer advisory, consulting, competency development, and nurturing of emerging trends and partner ecosystem in the areas of data and information management.

Mohan is the co-owner of a patent in data management and governance awarded by USPTO (United States Patent and Trademark Office).

Wipro Limited
Doddakannelli, Sarjapur Road,
Bangalore-560 035, India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading global information technology, consulting and business process services company. We harness the power of cognitive computing, hyper-automation, robotics, cloud, analytics and emerging technologies to help our clients adapt to the digital world and make them successful. A company recognized globally for its comprehensive portfolio of services, strong commitment to sustainability and good corporate citizenship, we have over 160,000 dedicated employees serving clients across six continents. Together, we discover ideas and connect the dots to build a better and a bold new future.

For more information, please write to us at **info@wipro.com**

IND/BRD/SEP 2018-AUG 2019