IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9323615

A MACHINE LEARNING APPROACH FOR DATA QUALITY CONTROL OF EARTH OBSERVATION DATA MANAGEMENT SYSTEM

Weiguo Han[1, 2], *Matthew Jochum*[2]

[1] Cooperative Programs for the Advancement of Earth System Science, University Corporation for Atmospheric Research, 3090 Center Green Drive, Boulder, CO 80301, USA
[2]Center for Satellite Applications and Research (STAR), National Oceanic and Atmospheric Administration
5830 University Research Court, College Park, MD 20740, USA

ABSTRACT

In the big data era, innovative technologies like cloud computing, artificial intelligence, and machine learning are increasingly utilized in the large-scale data management systems of many industry sectors to make them more scalable and intelligent. Applying them to automate and optimize earth observation data management is a hot topic. To improve data quality control mechanisms, a machine learning method in combination with built-in quality rules is presented in this paper to evolve processes around data quality and enhance management of earth observation data. The rules of quality check are set up to detect the common issues, including data completeness, data latency, bad data, and data duplication, and the machine learning model is trained, tested, and deployed to address these quality issues automatically and reduce manual efforts.

Index Terms— Big Data, Machine Learning, Earth Observation Data, Data management, Data Quality, Random Forest

1. INTRODUCTION

Data warehouse modernization, machine learning, and modern data hubs are listed as the top three trends of today's development of data management system[1]. Machine learning helps optimize system performances of data management in data cataloging, data provenance, metadata management, data quality, data security, data exploration, data dissemination, and so on [1].

A robust data quality process is a key component of the modern data management system. Machine learning techniques bring the significant impacts on the traditional data quality control mechanism based on user experiences and predefined rules [2, 3]. They were used to identify and remediate data quality issues [4], check the quality of Volunteered Geographic Information (VGI) data like OpenStreetMap (OSM) [5], automate verification of large-scale data quality like constraint suggestions and anomaly detection [6], and recommend geospatial data check and correctness [7].

As a centralized data management system within a research organization, STAR Central Data Repository (SCDR) integrates various types of earth observation data from multiple sources [8]. Collecting over than 800,000 files (~ 5.25 TB) for more than 600 earth observation data products every day, SCDR becomes a reliable data source of our scientists and developers. During the past years, we kept on enhancing and evolving this system by adoption of innovative technologies to facilitate STAR's research and development activities [9 - 11]. In this paper, we will study how to utilize a machine learning approach for data quality monitoring and control of SCDR.

2. METHODS

Machine learning is increasingly leveraged to solve various real world problems from image recognition to self-driving cars and recommender systems. It can be employed in the tasks of data management systems, including data creation, data maintenance, data quality, data discovery, etc. [1] By automating repetitive manual task, machine learning saves the costs for data correction, data transform, data recommendation, data linking, duplication removal, capacity management, and so on. It offers an intelligent approach to enhance data management functions, and has been adopted in data management of many industrial sectors like banking, health care, e-learning, stock market, retail, manufacturing, and among others.

Data quality monitoring and control is one of most important parts in the modern data management system. Considering the characteristics of volume, variety, and velocity of big earth observation data, data quality is one of main concerns for data managers, data consumers and IT technicians. New machine learning methods like supervised,

[1] Profisee Group, Inc, https://profisee.com/blog/dba-data-quality-trends-making-waves-in-data-management/.

semi-supervised, unsupervised, reinforcement are utilized to monitor, manage, and improve data quality. They can help check data completeness, validate data, identify latency, standardize data, remove duplicates, and offer suggestions of system configuration [12], and have been developed in commercial data quality software, like IBM InfoSphere® QualityStage®, and Syncsort Trillium.

SCDR includes many processes to check quality of the earth observation data in various formats, structures, and sizes derived from multiple sources [9]. The limitations of current method are time-consuming, human intervention, and inefficient. For example, when an exception of missing data occurs, the system administrator needs to identify the root causes, contact data provider, fix the issues, backlog the missing files, and notify data users. Sometimes, the issue is not reported and handled timely; it will cause longer data latency, even losing the missing data permanently. Therefore, a machine learning method is put forward to automate and improve the set of defined processes for data quality control.

As seen in Figure 1, the Random Forest algorithm is leveraged to build a machine learning model. The model captures the outputs of data quality rules check as the input of a set of features, and is trained and tested against known results before deployment. In addition, the model also generates quality report, raises alarms of new quality issue, and activates the defined processes.

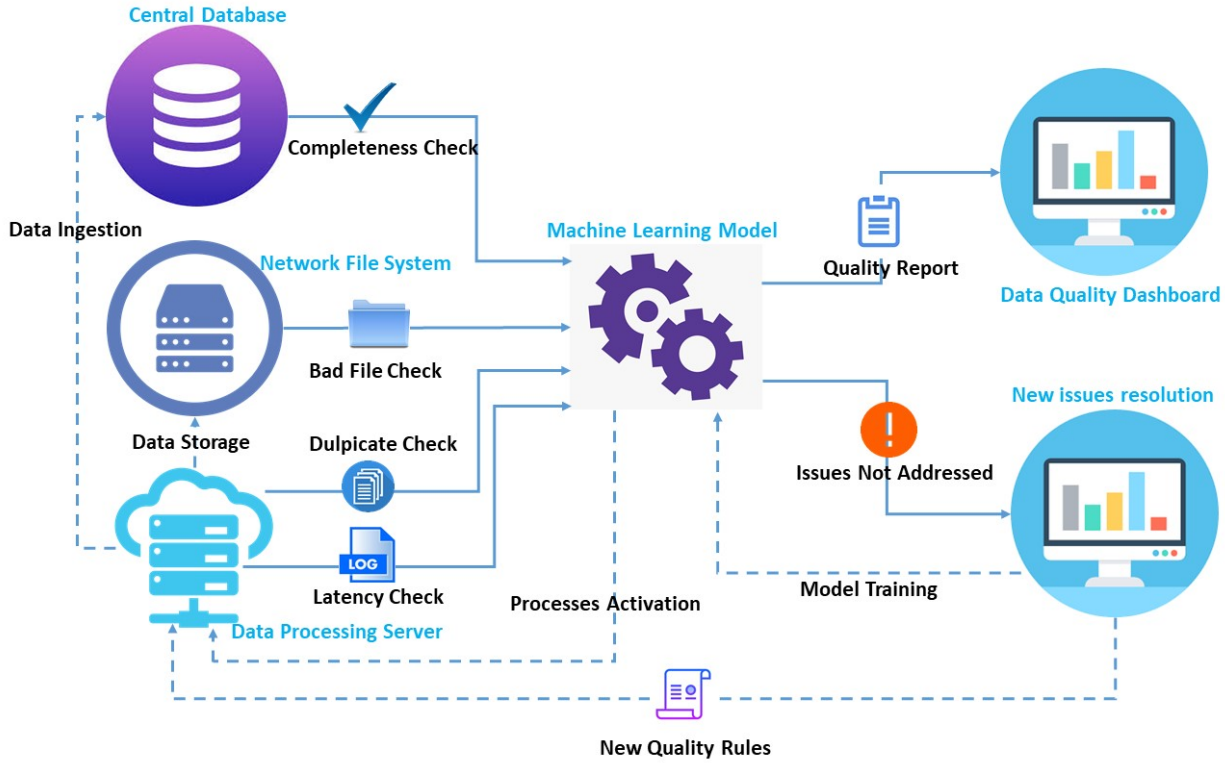

Figure 1 Machine learning based data quality control

The domain knowledge on each earth observation data product is required to create data quality rules. The commonly encountered data quality issues and the built-in quality rules include:

1) Data completeness
   The completeness check is built to check whether any files are missing according to the observation continuity, channels, or other information. The data gaps caused by satellite maintenance or anomaly are ignored.
2) Bad data
   In the download process, the files failed validation by checksum or with zero-byte in size are labelled as bad and quarantined in the specified location. Moreover, so do the ones failed retrieval of metadata information or with wrong metadata information when ingesting it.
3) Data latency
   The system outages or maintenances of both data providers and SCDR cause delay of data availability. The latency threshold of each collected product is set based on its average latency.
4) Data duplication
   The checksum value embedded in the database is used to check if the data is existing in the system. Regarding different versions of same data, a version control is configured to determine which ones are kept.

When data quality issues occur, the machine learning model will report them to data quality dashboard, notify them to the right person, and activate the related processes to handle them. The defined processes are given in order of importance:

1) Correcting data: correct the bad files like wrong date and time in the file name;
2) De-duplicating data: remove the duplicated data from data repository;
3) Re-pulling data: obtain data from original data providers or other sources to fill gaps caused by outage, bad data, or other reasons;
4) Restoring ingestion: restore data ingestion interrupted by outages;
5) Reconfiguring system: use system backup configurations to balance load of data pulling and ingestion among hosts and disks when the system outage causes long data latency;
6) Restarting system: restart database and routine jobs of data collection and ingestion if necessary.

It is impossible that the built-in rules can cover all quality issues. When the issue is marked as new, a new quality rule need to be built manually and the model will be trained to learn how to handle new issue. The machine learning method greatly reduces complexity and manual efforts in data quality control of SCDR system.

3. CONCLUSIONS

Applying machine learning in management system of earth observation data is an emerging topic. The challenges are constantly evolving. This paper presents applying method of machine learning in complement to built-in quality rules to automate data quality control with less manual and time-consuming tasks. Implementation of an automated and intelligent management system for big earth observation data still requires a long-term plan. We will investigate the applications of machine learning to automate and optimize functions of data cataloging, metadata management, and data dissemination.

4. ACKNOWLEDGEMENTS

The authors would like to thank STAR's IT team lead (Joseph Brust), Data Management Working Group members, and researchers for their valuable comments and suggestions. The contents are solely the opinions of the authors and do not constitute a statement of policy, decision, or position on behalf of NOAA or the U.S. government.

5. REFERENCES

[1] G. Nelson, "Data Management Meets Machine Learning", SAS Conference Proceedings: SAS Global Forum, 2018.

[2] R. Dhana, N. G. Venkat, and R.V. Vijay, "Data Quality Issues in Big Data", IEEE International Conference on Big Data (Big Data), 2015.

[3] L. Franchina, F. Sergiani, "High Quality Dataset for Machine Learning in the Business Intelligence Domain", In Proceedings of SAI Intelligent Systems Conference, Cham, 2019, pp. 391-401.

[4] T. Cagala, D. Bundesbank, "Improving data quality and closing data gaps with machine learning", IFC Bulletins chapters, 46, 2017.

[5] A. L. Ali, F. Schmid, "Data quality assurance for volunteered geographic information", In International Conference on Geographic Information Science, Cham, 2014, pp. 126-141), Cham.

[6] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification", Proceedings of the VLDB Endowment, 2018 pp.1781-1794.

[7] A. Parthy, A., L. Silberstein, E. Kowalczyk, J. P. High, A. Nagarajan, and A. Memon, "Using machine learning to recommend correctness checks for geographic map data", In Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice, 2019, pp. 223-232.

[8] W. Han, J. Brust, "Central satellite data repository supporting research and development", AGU Fall Meeting, 2015.

[9] W. Han, M. Jochum, "Near real-time satellite data quality monitoring and control", 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, 2016, pp. 206-209.

[10] W. Han, M. Jochum, "Assessing a central satellite data repository and its usage statistics", 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, 2018, pp. 6528-6531.

[11] W. Han, M. Jochum, "Latency analysis of large volume satellite data transmissions", 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, 2017, pp. 383-387.

[12] W. Dai, K. Yoshigoe, and W. Parsley, "Improving data quality through deep learning and statistical models", In Information Technology-New Generations, 2018, pp. 515-522.