World Academy of Science, Engineering and Technology
International Journal of Electrical and Information Engineering
Vol:16, No:12, 2022

Use of Machine Learning in Data Quality Assessment

Authors : Bruno Pinto Vieira, Armando Sérgio de Aguiar Filho

Abstract : Nowadays, a massive amount of information has been produced by different data sources, including mobile devices and transactional systems. In this scenario, concerns arise on how to maintain or establish data quality, which is now treated as a product to be defined, measured, analyzed, and improved to meet consumers' needs, which is the one who uses these data in decision making and companies strategies. Information that reaches low levels of quality can lead to issues that can consume time and money, such as missed business opportunities, inadequate decisions, and bad risk management actions. The step of selecting, identifying, evaluating, and selecting data sources with significant quality according to the need has become a costly task for users since the sources do not provide information about their quality. Traditional data quality control methods are based on user experience or business rules limiting performance and slowing down the process with less than desirable accuracy. Using advanced machine learning algorithms, it is possible to take advantage of computational resources to overcome challenges and add value to companies and users. In this study, machine learning is applied to data quality analysis on different datasets, seeking to compare the performance of the techniques according to the dimensions of quality assessment. As a result, we could create a ranking of approaches used, besides a system that is able to carry out automatically, data quality assessment.

Keywords : machine learning, data quality, quality dimension, quality assessment
Conference Title : ICDQCM 2022 : International Conference on Data Quality Control and Management
Conference Location : Paris, France
Conference Dates : December 29-30, 2022