Introduction to the HICSS-53 Collaboration for Data Science Minitrack

Lina Zhou
The University of North Carolina at Charlotte
Charlotte, NC 28223
lzhou8@uncc.edu

Souren Paul
Nova Southeastern University
Fort Lauderdale, FL 33314
souren.paul@gmail.com

Data science is an interdisciplinary field that uses statistics, data analysis, machine learning, and their related methods to extract knowledge and insights from both structured and unstructured data. Collaboration enables data scientists to be more productive and efficient in identifying relevant questions or problems, collecting data from multitude of different sources, organizing and making sense of the vast majority of data and information, and communicating their findings in such a way that can be used easily across different roles in support of business decision making. The new actionable knowledge and insights gained is in turn expected to support achieving collaborative goals such as innovation, idea generation, decision making, negotiation, and problem solving. Therefore, collaboration is a critical success factor for data science.

As the data is being generated at an exponential pace, there is a growing interest in reaping the value of data science to address increasingly complex business problems. Encouraging and facilitating collaboration in data science between members of a data science team, groups and organizations is one promising way for businesses and organizations to enhance their operational excellence or competitive advantages. In addition, the collaboration is not limited between humans and organizations but include that between humans and computers as well. For example, human knowledge and expertise can provide guidance in building computational models or in search of effective and/or efficient solutions to business problems.

This minitrack includes one paper session, consisting of four papers covering the following areas of interest: social media driven collaborative data science, humans in the loop data science, crowdsourcing analytics, data science for collaborative work, and case studies on collaborative data science. The first paper, "WeSAL: Applying Active Supervision to Find High-quality Labels at Industrial Scale", proposes a new method that incorporates weak supervision and active learning in order to create labelled data for supervised machine learning. Obtaining accurately-labelled training data is a very important yet difficult task in machine learning. Weakly supervised learning is an effective approach to obtaining labels for training data in building classification models, which helps overcome the issue associated with label scarcity. The proposed method is evaluated on six different datasets across binary and multi-class classification tasks. The results indicate that WeSAL can generate high-quality labels at a large scale while reducing the labeling cost.

The second paper, "Crowdsourcing Data Science: A Qualitative Analysis of Organizations' Usage of Kaggle Competitions", attempts to address the combination of crowdsourcing and data science through conducting semi-structured interviews with data science experts and exploring Kaggle services. Content analysis of the interview data uncovers three categories of factors that influence an organization's perceived success when hosting a data science competition, which include platform-related, organization-related, and outcome-related factors.

The third paper, "Metrics for Analyzing Social Documents to Understand Joint Work", proposes an approach to measure and interpret collaboration based on the structure of social documents. It develops seven metrics for measuring group collaborations from the document instead of user perspective and evaluates them through an analysis of large-scale enterprise collaboration systems that integrate a range of different functional modules. The findings suggest that the purpose of a group workspace has an influence on the richness of its documents.

The last paper, "Dissecting Moneyball: Improving Machine Learning Model Interpretability in Baseball Pitch Prediction", addresses the interpretability problem of machine learning models to facilitate the collaboration between data science team members. It extends advanced methods for drawing explanations from the results of classification models in two main aspects: aggregating explanations from the instance level to the user-defined level, and providing explanations using the original input features. Using the prediction of baseball pitch outcome as a test case, the proposed methods demonstrate its improved interpretability while preserving prediction performance.

URI: https://hdl.handle.net/10125/63766
978-0-9981331-3-3