Introduction to Data, Text, and Web Mining for Business Analytics Mini-track

Dursun Delen
Oklahoma State University

Behrooz Davazdahemami
University of Wisconsin-Whitewater

Hamed Zolbanin
University of Dayton

This mini-track has a total of six papers that are about developing analytics systems for decision support by means of data, text, or web mining. Three of the six papers focus on a variety of interesting methodologies to improve text mining and natural language processing. The remaining three papers deal with feature selection and dimensionality reduction, methods for linkage of data sets obtained from various resources, and application of data stream based regression for mass appraisal of properties.

The paper by Kiefer et al. presents a framework to help domain experts, with little knowledge and expertise on IT or data analytics, in automatic selection of the best-fitting training data set in supervised text analytics problems. They discuss that using default training data sets with the out-of-the-box analytics tools by such domain experts can lead to low-quality results. Specifically, they indicate that latent semantic analysis (LSA) and cosine similarity can be used as basis for automatic selection of the best-fitting training data for a part-of-speech tagging task.

The paper by Zaman and Goldberg addresses the need for surveillance and monitoring of consumer product reviews for the early identification of product safety hazards, particularly injuries. Given the scarcity of reporting safety hazards in consumer reviews, they employed a transfer learning technique using a data set with higher-than-normal representation of safety hazards (obtained from a government maintained data set) to develop a detection technique and then applied it to a set of consumer product reviews from Amazon. Their results indicate improved surveillance for monitoring hazards across multiple industries.

Lastly, yet equally importantly, the paper by Gu and Leroy proposes an approach to automatically optimize hyper parameters of a deep learning network using conventional machine learning methods in the specific context of NLP labeling task for identifying autism spectrum disorder (ASD). Particularly they show that using SVM to tune the class weights in a multinomial classification task with a bidirectional long short-term memory (Bi-LSTM) network, where the data set is highly unbalanced, can lead to a considerable increase in the performance of the network (i.e., increasing the micro-average F1 score from 45.9% to 47.1%).

The paper by Aghanavesi and Memedi uses a partial least square (PLS) regression model for dimensionality reduction of a data set collected using smartphones, to predict motor states of Parkinson's disease patients. The results outperformed those resulted from a combination of principal component analysis (PCA) and support vector machines (SVM), suggesting PLS as an efficient methodology for data-driven analysis.

Kruse et al. in their paper provide a theory-guided qualitative literature review on techniques for record linkage and entity linking. They mention that while majority of the extant literature deals with record linkage and structured data, processing unstructured data through entity linkage is receiving more interest with the trend Big Data. In addition, deep learning methodologies are being explored to be used for both purposes.

In an effort to tackle the shortcomings of standard statistical and machine learning approaches for mass evaluation of properties, Levitan et al. propose a prequential regression approach based on three data stream based regression methods, namely adaptive model rules (AMR), perceptron learning (PL), and random rules (RR) for mass appraisal. They show the high performance and efficiency of their proposed approach, compared to standard linear regression, using a data set of 110,525 sales transaction records from a municipality in the Midwest US.

URI: https://hdl.handle.net/10125/63865
978-0-9981331-3-3