Thesis Defense

Institute for Software Research
Societal Computing

Multi-view Clustering of Social-based Data

Iain J. Cruickshank

Thursday, July 16th, 2020
11:00 AM - 2:00 PM
Zoom Meeting: https://cmu.zoom.us/j/6246193856

Real-world, social phenomena produce various types of data, like explicit networks or user-emitted text. When different sets of data describe the same entities, the data is termed multi-view or multi-modal. A distinct advantage of multi-view data is that different views may better capture different aspects of the latent structure of the data. However, there are difficulties in combining that data to produce, something like one set of cluster labels. Multi-view clustering techniques, primarily developed for image or biological use cases or network only use cases, have typically not been used for clustering social-based use cases. I investigate the use of multi-view clustering on various social-based multi-view data sets, and propose new techniques for multi-view clustering of social-based data.

In the first part of this thesis I discuss the use of multi-view clustering for social-based data, and propose a new paradigm and new techniques for multi-view clustering. In chapter two I propose a new hybrid paradigm of multi-view clustering, which combines elements of late paradigm and intermediate paradigm integration. I test the various intermediate, late, and hybrid paradigm algorithms on a wide range of benchmark data sets from social-based data scenarios. The results of the empirical testing demonstrate a wide variance in the performance of multi-view clustering techniques. This is likely because social-based data often have high inter- and intra-view variances that are not present in other data scenarios, which presents difficulties for existing techniques. Only two techniques proposed in the chapter have good performance across all of the data sets and are robust to inter- and intra-view differences. From the results in chapter two, I devise a new algorithm based in network modularity and graph learning to cluster multi-view social data in chapter three. I present the results of a series of empirical tests of the new technique, as well as possible variations on the technique. The results demonstrate that the presented technique often performs well across a wide range of social-scenarios that give rise to multi-view data, is scalable to large data sets, and is robust to inter- and intra-view variance.

In the second part of this thesis I use the new techniques to do clustering analysis of real-world data. In chapter four I use multi-view clustering on Twitter data collected during the initial stages of the COVID-19 pandemic. This analysis is the first ever use of multi-view clustering to cluster hashtags from large, social-media data sets. The results display that hashtags form topical clusters and that these topical clusters have changed over the course of the pandemic. In chapter five I use multi-view clustering to cluster malware samples. The results demonstrate that a multi-view clustering of malware samples provides insight into communities of malware use, and confirm the techniques developed can be applied to a wide range of social-based data scenarios.

In sum, I demonstrate the suitability of, and create techniques for, multi-view clustering of complex, multi-view, social-based data. This thesis advances practical clustering analyses of large-scale, noisy, social-based data and contributes to the field of multi-view clustering in general.

Committee: Kathleen M. Carley (Chair), L. Rick Carley,
J. Zico Kolter, Tanya Berger-Wolf (Ohio State University)

institute for SOFTWARE RESEARCH | Carnegie Mellon University School of Computer Science