Leveraging Common Sense

Knowledges for Accurate Regression Analysis of Social Media Data

Specialist Meeting on “Human Dynamics and Big Data”

hosted by the Center for Human Dynamics in the Mobile Age (HDMA)

San Diego State University (SDSU), San Diego, CA

August 1-2, 2016

Author: Xiaobai Liu, Assistant Professor, Department of Computer Science, SDSU

Regression of high-dimensional data (texts, videos, images, or their combinations) aims to map the original data to a low-dimensional continuous space while preserving the original data structure. It has been playing a critical role in social media analysis, and attracting a lot of attentions in the fields of computer vision, image processing, and Artificial Intelligences (AI) in general. Although impressive results achieved, existing studies are limited to extracting low-level perceptual features, e.g., color, texture, gradients, etc., which are not necessarily powerful in discovering the underlying data structures of social media data.

This position paper presents my on-going research on leveraging common sense knowledges from multi-modality social media data towards effective and accurate regression analysis of human dynamics. This line of research will directly address the workshop’s research question: How to reduce the dimensionality of big datasets in ways that help humans conduct analysis or create knowledge? The developed models, theories, and algorithms will enable the other questions, e.g., how to detect communities predictive of social policy information.

Project Description

High-dimensional regression is one of the fundamental tasks in statistics and machine learning, and has been widely used to predict real-world outcomes from social media data by creating a form of collective wisdom. These quantitative predictions, for instance, including, stock price, return rate of advertising, movie revenue, etc. Analyzing the correlations between social media data and these predictions of interests can be readily addressed with a wide variety of regression techniques, e.g., support vector machine [1], linear regression [2], boosting [3], random forest [4] and neural network [5].

However, parsing social media data imposes three major challenges. First, how to deal with high-dimensional data, known as the curse of dimensionality. Second, how to mitigate the effect of data noises that is ubiquitous in social media data. Last but most importantly, how to take advantage of multi-modality data, e.g., textual data, images, which usually appear simultaneously in a single item of social media messages (e.g., in TWEET).

To address the above challenges, the investigation team introduces a novel perspective for solving quantitative predictions over social media data. The key idea is to utilize a small set of common sense knowledges to help understand the social data, which leads to a well confined regression solution space. Taking movie revenue for examples, a movie with more reviews/comments is likely to earn more revenue; a successful production team will be likely to succeed again. The above knowledges are referred to *common sense* that is frequently used by our human beings in daily life to interpret observations/data. While these knowledges do not necessarily hold true for every particular case, they can be used to confine the solution space in particularly when the space is complicated and highly non-convex. The knowledge can also be extracted from various modalities, serving as an effective way to integrate multi-modality representations of the same data.

In the workshop we will show some preliminary studies of the above research direction on predicting house values from both textual and imagery data. We will demonstrate that house photos, if properly modeled, can be used to boost the accuracies of house appraisal in multiple settings. In particular, we found that a small set of common sense knowledge about community, school-rating, etc., can make significantly improvements in terms of system accuracies and robustness.

Reference

1. Basak, D., Pal, S., Patranabis, D.: Support vector regression. In: NIPS. (2007)

2. Mangasarian, O., Wild, E.: Generalized linear models. JASA 95(452) (2000)

3. Viola, P.A., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS. (2006)

4. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A uni_ed framework for classi_cation, regression, density estimation, manifold learning and semi-supervised learning. Foundations and Trends in Computer Graphics and Vision 7(2) (2012)

5. Specht, D.: A general regression neural network. IEEE Transactions on Neural Network 2(6) (1991)