Dynamic Question Ordering: Obtaining Useful Information while Reducing Burden

Abstract:

As data become more pervasive and computing power increases, the opportunity for transformative use of data grows. Collecting data from individuals can be useful to the individuals (by providing them with personalized predictions) and the data collectors (by providing them with information about populations). However, collecting these data is costly: answering survey items, collecting sensed data, and computing values of interest deplete finite resources of time, battery life, money, etc. Dynamically ordering the items to be collected, based on already known information (such as previously collected items or paradata), can lower the costs of data collection by tailoring the information-acquisition process to the individual. This thesis presents a framework for an iterative dynamic item ordering process that trades off item utility with item cost at data collection time. The exact metrics for utility and cost are application-dependent, and this framework can apply to many domains. The two main scenarios we consider are (1) data collection for personalized predictions and (2) data collection in surveys. We illustrate applications of this framework to multiple problems ranging from personalized prediction to questionnaire scoring to government survey collection. We compare data quality and acquisition costs of our method to fixed order approaches and show that our adaptive process obtains results of similar quality at lower cost.

For the personalized prediction setting, the goal of data collection is to make a prediction based on information provided by a respondent. Since it is possible to give a reasonable prediction with only a subset of items, we are not concerned with collecting all items. Instead, we want to order the items so that the user provides information that most increases the prediction quality, while not being too costly to provide. One metric for quality is prediction certainty, which reflects how likely the true value is to coincide with the estimated value. Depending whether the prediction problem is continuous or discrete, we use prediction interval width or predicted class probability to measure the certainty of a prediction. We illustrate the results of our dynamic item ordering framework on tasks of predicting energy costs, student stress levels, and device identification in photographs and show that our adaptive process achieves equivalent error rates as a fixed order baseline with cost savings up to 45%.

For the survey setting, the goal of data collection is often to gather information from a population, and it is desired to have complete responses from all samples. In this case, we want to maximize survey completion (and the quality of necessary imputations), and so we focus on ordering items to engage the respondent and collect hopefully all the information we seek, or at least the information that most characterizes the respondent so imputed values will be accurate. One item utility metric for this problem is information gain to get a “representative” set of answers from the respondent. Furthermore, paradata collected during the survey process can inform models of user engagement that can influence either the utility metric (e.g., likelihood the respondent will continue answering questions) or the cost metric (e.g., likelihood the respondent will break off from the survey). We illustrate the benefit of dynamic item ordering for surveys on two nationwide surveys conducted by the U.S. Census Bureau: the American Community Survey and the Survey of Income and Program Participation.