Machine Learning Applications in Chemical and Biological Engineering

Kristen A. Severson

Chemical and biological systems are increasingly implemented with advanced sensor systems that collect large amounts of data. For example, a single microarray can measure thousands of genes and a typical offshore oil platform generates 1 to 2 TB of data per day. New algorithms are needed to efficiently and effectively use these datasets to increase predictive capability and improve system understanding. In this thesis, algorithmic advances to bridge the gap between data and system insights are addressed in a series of case studies.

In the first case study, the problem of predicting critical quality attributes for a monoclonal antibody using data from the manufacturing process is addressed. In this setting, the main challenge is that there is only a limited dataset available for modeling. To tackle this issue, Monte Carlo sampling was used in conjunction with an elastic net approach to subset selection.

The second case study is also within the biological domain but considers a discrete outcome. The proposed algorithm addresses two common issues when building classification models for biological studies: learning a sparse model, where only a subset of a large number of possible predictors is used, and training in the presence of missing data. The resulting algorithm leverages expectation-maximization to tackle both issues simultaneously.

In the third case study, the goal was to identify anomalous operating periods using production data from an oil and gas well without access to historical examples of such periods. The proposed approach recasts the problem as a semi-supervised problem and leverages approaches from the positive and unlabeled literature.

The final case study considers the task of prediction lithium-ion battery cycle life. Cycle life is defined as the number of charge and discharge cycles the battery undergoes before 80% capacity fade. Several, difficult to identify factors can contribute to capacity fade. Even in batteries with the same chemistry, operated using the same conditions, there is considerable cycle life variability. Therefore, the challenge was to build a model to capture individual capacity trajectories.

Each case study is benchmarked using state-of-the-art approaches. In all settings, the value of data-driven methods is demonstrated.

Thesis Supervisor: Richard D. Braatz
Title: Edwin R. Gilliland Professor