Data Mining for the Chemical Process Industry

Ng Yew Seng
National University of Singapore, Singapore

Rajagopalan Srinivasan
National University of Singapore and Institute of Chemical & Engineering Sciences, Singapore

INTRODUCTION

Advancements in sensors and database technologies have resulted in the collection of huge amounts of process data from chemical plants. A number of process quantities such as temperature, pressure, flow rates, level, composition, and pH can be easily measured. Chemical processes are dynamic systems and are equipped with hundreds or thousands of sensors that generate readings at regular intervals (typically seconds). In addition, derived quantities that are functions of the sensor measurements as well as alerts and alarms are generated regularly. Several commercial data warehouses, referred to as plant historians in chemical plants, such as the DeltaV Continuous Historian (from Emerson Process Management), InfoPlus.21™ (from AspenTech), Uniformance® PHD (from Honeywell), and Industrial SQL (from Wonderware) are in common use today around the world. These historians store large amount (weeks) of historical process operation data at their original resolution and an almost limitless amount (years) in compressed form. This data is available for mining, analysis and decision support – both real-time and offline.

Process measurements can be classified based on their nature as binary (on/off) or continuous. However, both are stored in discrete form in the historians. Measurements can also be classified based on their role during operation as controlled, manipulated, and non-control related variables. Controlled variables are directly or indirectly related to the plant's quality, production, or safety objectives and are maintained at specified setpoints, even in the face of disturbances, by analog or digital controllers. This regulation is achieved by altering manipulated variables such as flow-rates. Chemical plants are typically well-integrated – a change in one variable would propagate across many others. Non-control related variables do not have any role in plant control, but provide information to plant personnel regarding the state of the process.

In general, a plant can operate in a number of states which can be broadly classified into steady-states and transitions (Srinivasan *et al.*, 2005b). Large scale plants such as refineries typically run for long periods in steady-states but undergo transition if there is a change in feedstock or product grades. Transitions also result due to large process disturbances, maintenance activities, and abnormal events. During steady-states, the process variables vary within a narrow range. In contrast, transitions correspond to large changes / discontinuities in the plant operations; *i.e.*, change of set points, turning on or idling of equipments, valve manipulations, etc. A number of decisions are needed on the part of the plant personnel to keep the plant running safely and efficiently during steady states as well as transitions. Data mining and analysis tools that facilitate humans to uncover information, knowledge, patterns, trends, and relationships from the historical data are therefore crucial.

BACKGROUND

Numerous challenges bedevil the mining of data generated by chemical processes. These arise from the following general characteristics of the data:

1. *Temporal*: Since the chemical process is a dynamic system, all measurements vary with time.
2. *Noisy*: The sensors and therefore the resulting measurements can be significantly noisy.
3. *Non-stationarity*: Process dynamics can change significantly, especially across states because of structural changes to the process. Statistical properties of the data such as mean and variance can therefore change significantly between states.
4. *Multiple time-scales*: Many processes display multiple time scales with some variables varying quickly (order of seconds) while others respond over hours.

5. *Multi-rate sampling*: Different measurements are often sampled at different rates. For instance, online measurements are often sampled frequently (typically seconds) while lab measurements are sampled at a much lower frequency (a few times a day).
6. *Nonlinearity*: The data from chemical processes often display significant nonlinearity.
7. *Discontinuity:* Discontinuous behaviors occur typically during transitions when variables change status – for instance from inactive to active or no flow to flow.
8. *Run-to-run variations*: Multiple instances of the same action or operation carried out by different operators and at different times would not match. So, signals from two instances could be significantly different due to variation in initial conditions, impurity profiles, exogenous environmental or process factors. This could result in deviations in final product quality especially in batch operations (such as in pharmaceutical manufacturing).

Due to the above, special purpose approaches to analyze chemical process data are necessary. In this chapter, we review these data mining approaches.

MAIN FOCUS

Given that large amounts of operational data are readily available from the plant historian, data mining can be used to extract knowledge and improve process understanding – both in an offline and online sense. There are two key areas where data mining techniques can facilitate knowledge extraction from plant historians, namely (i) process visualization and state-identification, and (ii) modeling of chemical processes for process control and supervision.

Visualization techniques use graphical representation to improve human's understanding of the structure in the data. These techniques convert data from a numeric form into a graphic form that facilitates human understanding by means of the visual perception system. This enables post-mortem analysis of operations towards improving process understanding or developing process models or online decision support systems. A key element in data visualization is dimensionality reduction. Subspace approaches such as principal components analysis and self-organizing maps have been popular for visualizing large, multivariate process data. For instance, an important application is to identify and segregate the various operating regimes of a plant. Figure 1 shows the various periods over a course of two weeks when a refinery hydrocracker was operated in steady-states or underwent transitions. Such information is necessary to derive operational statistics of the operation. In this example, the states have been identified using two principal components that summarize the information in over 50 measurements. As another illustration, oscillation of process variables during steady-state operations is common but undesirable. Visualization techniques and dimensionality reduction can be applied to detect oscillations (Thornhill and Hagglund, 1997). Other applications include process automation and control (Amirthalingam *et al*., 2000; Samad *et al.*, 2007), inferential sensing (Fortuna *et al.*, 2005), alarm management (Srinivasan *et al.,* 2004a), control loop performance analysis (Huang, 2003) and preventive maintenance (Harding *et al.*, 2007).

Data-based models are also frequently used for process supervision – fault detection and identification (FDI). The objective of FDI is to decide in real-time the condition – normal or abnormal – of the process or its constituent equipment, and (ii) in case of abnormality, identify the root cause of the abnormal situation. It has been reported that approximately 20 billion dollars are lost on an annual basis by the US petrochemical industries due to inadequate management of abnormal situations (Nimmo, 1995). Efficient data mining algorithms are hence necessary to prevent abnormal events and accidents. In the chemical industry, pattern recognition and data classification techniques have been the popular approaches for FDI. When a fault occurs, process variables vary from their nominal ranges and exhibit patterns that are characteristic of the fault. If the patterns observed online can be matched with known abnormal patterns stored in a database, the root cause of a fault can generally be identified.

In the following, we review popular data mining techniques by grouping them into machine-learning, statistical, and signal processing approaches.

Machine Learning Methods

Neural-network (NN) is a popular machine learning technique that exhibits powerful classification and func-