International Journal of Cheminformatics Research
Vol. 3: Issue 1
www.journalspub.com

Data Mining in Chemical Process Industry

*Y.S. Choudhary**
Indian Council of Agricultural Research, New Delhi, India

ABSTRACT

Data mining is the computing process of learning patterns in huge data sets involving methods at the intersection of statistics, and database systems. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining techniques are becoming increasingly important in chemistry as databases become very large to examine by hand. Data mining methods from the field of Inductive Logic Programming (ILP) have potential advantages for structural chemical data. Data mining was used to find all frequent substructures in the database, and knowledge of these frequent substructures is shown to add value to the database. Only by using a data mining algorithm, and by doing a complete search, is it possible to prove such a result. In this paper, study of data mining methods was presented which is the main novel tool for studying chemical databases.

Keywords: chemical databases, chemical industry, chemical modeling, data-based models, data mining

***Corresponding Author**
E-mail: yschoudharyicar@yahoo.co.in

INTRODUCTION

Development in database technologies has brought about the collection of tremendous amount of process data from chemical plants. Various process quantities, for example, temperature, weight, stream rates, level, arrangement, and pH can be just measured. Chemical processes are dynamic frameworks and are outfitted with hundreds or thousands of sensors that produce readings at regular intervals. Furthermore, derived quantities that are elements of the sensor estimations and in addition alarm and alarm are created frequently. Several commercial data warehouses, referred to as plant historians in chemical plants are in common use today around the world. These history students store immense amount of recorded process operation information. This information is accessible for mining, investigation and decision support both real-time and offline.

Process estimations can be characterized in light of their inclination as binary (on/off) or continuous. Estimations can likewise be characterized in light of their part amid operation as controlled, and non-control related factors. Controlled factors are specifically or in a roundabout way identified with the plant's quality, creation, or wellbeing targets and are kept up at indicated set focuses, even notwithstanding unsettling influences, by simple or advanced controllers. This control is accomplished by adjusting controlled factors, for example, stream rates. Chemical plants are typically well-integrated – an alteration in one variable would spread 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 [1].

In general, a plant can operate in a number of positions which can be broadly classified into steady-states and transitions. Large scale plants for example refineries regularly keep running for long periods in steady-states but undergo changes if there is a modification in feedstock or product grades. Changes also result due to large process troubles, repairs activities, and abnormal events. During steady-states, the process factors vary within a narrow range. In comparison, 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*. Several decisions are required 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 empower people to reveal data, learning, examples, patterns, and connections from the historical data are hence essential [2].

BACKGROUND CONTEXT

Numerous challenges of data mining are 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, indications from two instances could be significantly different due to difference in impurity profiles, initial conditions, and 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 reasons, special purpose methodologies to analyze chemical process data are essential. In this article, analysis of these data mining approaches was studied [3].

KEY APPLICATIONS OF DATA MINING

Given that large amounts of operational data are readily available from the plant historians, data mining can be used to extract knowledge and improve process understanding – both in an offline and online versions [4]. There are two chief areas where data mining techniques can enable knowledge extraction from plant historians, namely

(i) Process visualization and state-identification
(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 [5].

A main 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 [6].

For instance, an important application is to identify and segregate the various operating regimes of a plant. Visualization techniques and dimensionality reduction can be applied to detect oscillations. Other applications include process automation and control, inferential sensing, alarm management, control loop performance analysis and preventive maintenance.

Data-based models are also recurrently 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.
- 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.

Well-organized data mining algorithms are therefore 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 statistical, machine-learning, and signal processing approaches.

Provided that large amounts of operational data are eagerly available from the plant historian, data mining can be used to extract knowledge and improve process understanding – both in an offline and online versions. There are two main areas where data mining techniques can facilitate knowledge extraction from plant historians, namely:

- Process visualization and state-identification
- 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 postmortem analysis of operations towards improving process understanding or developing process models or online decision support systems [7].

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 condition as follows:

(i) Normal or abnormal process or its constituent equipment.
(ii) In case of abnormality, the root cause of the abnormal situation should be identified.

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. 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 [8].

Jayanthi Ranjan has analyzed the applications of data mining techniques in pharmaceutical industry. She explains the role of data mining in pharmaceutical industry. She shows how data mining on large sets of data uses tools like association, clustering, segmentation and classification for helping better manipulation of the data help the pharma firms compete on lower costs while improving the quality of drug discovery and delivery methods. Dan Braha and Armin Shmilovici have shown how data mining can be used for improving a Cleaning Process in the Semiconductor industry [9].

They have presented a comprehensive and successful application of data mining methodologies to the refinement of a new dry-cleaning technology that utilizes a laser beam for the removal of micro-contaminants. They suggest that data mining methodologies may be particularly useful when data is scarce, and the various physical and chemical parameters that affect the process exhibit highly complex interactions. Another implication is that on-line monitoring of the cleaning process using data mining may be highly effective [9].

Mohamed Azlan Hussain has reviewed the applications of neural networks in chemical process control using simulation and online implementation. He provides an extensive review of the various applications utilizing neural networks for chemical process control, both in simulation and online implementation. He has highlighted the broad, extensive and continuing increase in the application of neural network in many chemical process control applications, both online and in simulation.

CONCLUSION

In chemical process industries, data historians continuously collect process data from most of the process variables through the whole process chain for several months and sometimes for years. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that can be beneficially used for quality improvement in chemical process industries.

REFERENCES

[1] M.D. Giess, S.J. Culley, A. Shepherd. *Informing Design Using Data Mining Methods*. ASME DETC, Montreal, Canada, 2002, 98–106p.

[2] M.D. Giess, S.J. Culley. *Investigating Manufacturing Data for Use Within Design*. ICED 03, Stockholm, Sweden, 2003, 1408–13p.

[3] M. Perzyk, A. Kochanski, J. Kozlowsk. Data mining in manufacturing: significance analysis of process parameters, *Proc IMechE Part B: J Eng Manuf.* 2008; 222: 1503–16p.

[4] Chen-Fu Chien, Wen-Chih Wang, Jen-Chieh Cheng, Data mining for yield enhancement in semiconductor manufacturing and an empirical study, *Exp Syst Appl.* 2007; 33: 192–8p.

[5] B. Al-Salim, M. Abdoli. Data mining for decision support of the quality improvement process, In: *Proceedings of the Eleventh Americas Conference on Information Systems*. Omaha, NE, USA, August 11th–14th 2000, 1462–9p.

[6] S.-g. He, Z. He, G.A. Wang, L. Li. In: *Data Mining and Knowledge Discovery in Real Life Application*. J. Ponce, A. Karahoca (eds.), February 2009, I-Tech, Vienna, Austria, 438p.

[7] G. Koksa, I. Batmaz, M.C. Testik. A review of data mining applications for quality improvement in manufacturing industry, *Exp Syst Appl*. 2011; 38: 13448–67p.

[8] J. Ranjan. Applications of data mining techniques in pharmaceutical industry, *J Theor Appl Inform Technol*. 2005–2007; 61–7p.

[9] M.A. Hussain. Data mining for improving a cleaning process in the semiconductor industry, *IEEE Trans Semicon Manuf*. 2002; 15(1): 91–101p.