

White Paper

Finding the Right Solution in the Process Industries

for

Data Mining – Based

Process Intelligence

Applications

Stel Kentritas



Abstract

The data mining “buzzword” has been floating around the process industries offices and control rooms over the past decade. Yet, from “buzzword” to strategic initiative and to necessity, data mining will find a place in what is called the Level 3 domain of process systems. Data mining has already established a presence in Level 4 Business Systems as means to perform a range of business analytics, and in Level 2 Process Control Systems - in the form of one of its most popular algorithm, Neural Networks.

Looking into a wider use of data mining in the process industries invites also use of a very appropriate term, Process Intelligence, which like Business Intelligence, implies extracting knowledge from plant data. This is done through exploratory and multi-dimensional data structures for drilling down on data sets and by “slicing and dicing” the data residing in “data cubes”, but a step further is through data mining. Thus, true Process Intelligence goes beyond exploratory analysis to predictive models developed through data mining. Subsequently, data mining in the process industries serves needs, requirements and purposes that can be based on data models that have structures specific to process units and unit operations. This is an approach similar to what is already implemented by such other business sectors as financial services and telecoms.

This White Paper attempts to address some of the main considerations for introducing data mining in the process industries in a way that extends beyond the classical applications addressed with basic and advanced statistics. Thus, data mining is not just about advanced statistics, but, much like other “crafty” applications, it becomes a bit of a science and an art when it comes to knowledge discovery, which is always associated with innovation about what and where to find, extract and distill the needed knowledge.

Introduction

An assessment of data mining approaches and solutions are presented in order to address their suitability and applicability when it comes to the process industries and for such requirements and applications as for Root Cause Analyses, Advanced Process Control, Exploratory Analysis and Predictive Modeling, Quality Control and Statistical Process Control. The term Process Intelligence is therefore used in a wider sense and it goes beyond these applications. It is used more in the sense of Knowledge Discovery for sustaining a competitive advantage, which can be in the form of answering previously unanswered questions ranging from why certain behaviors occur in a process plant to what I should do in order to prevent a trip or sustain operations within certain safety or profitable ranges.

There are many factors that contribute to a successful or failed data mining initiative. The most critical being the experience in specific domains and how the data mining solution supplier’s culture, in terms of this experience, aligns and matches with the end user experience.

Take for example the domain of telecoms; there is no doubt that the experience culture is totally different than the one in the process industries, where the process of analyses of plant data must be backed up by deep insight and knowledge of plant characteristics, and with the primary, secondary and sometimes tertiary effects on plant variables following a transient or incident. Or in the case of applying data mining for exploratory analyses, development of a data model to be used for data analysis in line with production and business management, this must be backed up by knowledge of how such plants behave and should be operated.

As such, although there is no doubt that each industry is unique, the process industries are unique when it comes to the impact of thermodynamics, chemical science, unit operations design and process control, which, as it is well known, all of these contribute to first principles models used in steady state and dynamic process simulation.

Data Mining in General

Data mining has been primarily used in business applications, as for example in the analytical CRM (Customer Relationship Management) domain. Since then, it has been extended to other specialized applications as for example risk management, development of credit scoring models, fraud detection, etc.

The majority of data mining professionals are experienced with such applications, while, in the process industries, core statistical functions for Quality Control (QC) and Statistical Process Control (SPC) have been traditionally applied.

It was not until recently that data mining has been gradually introduced for more advanced statistical applications in the process industries. As such, the use of data mining in the process industries is quite new and this can be easily concluded by researching the public internet domain and realizing that only limited resources are published for the process industries, which indicates the “newness” of the topic.

Even in the case of data mining professionals, “googling” for relative job postings, it seems that very few, i.e ExxonMobil, BP, Dow Chemical and Shell seem to have been searching for professionals for data mining analyst positions. As such, the combination of data mining knowledge and process knowledge is a rather rare case and therefore an attractive career path for those looking for specialization areas.

The Retail Industry Paradigm

Some characteristics associated with the ability of a data mining solution to address the process industries challenges are embedded capabilities, such as the one of Sequence Analysis, which is of great importance to the achievement of Root Cause Analysis (RCA) requirements.

Sequence, Association and Link Analysis (SLA) is an implementation of several techniques specifically designed for extracting rules from datasets, which has been generally referred to metaphorically as "market-baskets". It is interesting that such capability (SLA) has its roots and terminology from a non-process industry domain. In fact, the term "market-basket" comes from the use of data mining applications in the retail industry. Together with the "Market-Basket" metaphor or SLA, additional techniques applied for RCA in the process industries include Clustering, Association Rules, Decision Trees, etc. It is useful for potential data mining users in the process industries to look into this paradigm since as all people are consumers, they can understand the examples of what and how data mining deals with challenges in the retail industry and how some of these approaches can be extended to the process industries.

The "**Market-Basket**" metaphor deals with what it implies, that is, the market-basket problem in the retail industry. It assumes that there are a large number of products that can be purchased by a customer, either in a single transaction, or over time in a sequence of transactions. Such products can be goods displayed in a supermarket, spanning a wide range of items from groceries to electrical appliances, or they can be insurance packages which customers might be willing to purchase, etc. Customers fill their basket with only a fraction of what is on display or on offer.

Association rules can be extracted from a database of transactions, to determine which products are frequently purchased together. For example, one might find that purchases of flashlights also typically coincide with purchases of batteries in the same basket. Likewise, when transactions are time-stamped, allowing the analysts to track purchases.

Sequence analysis is concerned with a subsequent purchase of a product or products given a previous buy. For instance, buying an extended warranty is more likely to follow (in that specific sequential order) the purchase of a TV or other electric appliances. Sequence rules, however, are not always that obvious and sequence analysis helps extract such rules no matter how hidden they may be in a market-basket dataset. There is a wide range of applications for sequence analysis in many areas of industry and since including customer shopping patterns, phone call patterns, fluctuation of the stock market, DNA sequence and web-log streams.

A notably useful capability that encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories, that is, in order to develop taxonomies is known as clustering. **Cluster analysis**, as it is known, is an exploratory data analysis approach and capability, which aims at sorting different objects into groups in a way that the degree of association between two objects is maxima.

Just to limit the reference to data mining techniques, either for exploratory or predictive purposes, we can mention the most common perhaps and general computational approach, that of **Decision trees**, which over the past few years, has emerged as one of the most powerful methods for predictive data mining. Some implementations of these powerful algorithms allow them to be used for regression as well as classification problems, with continuous and/or categorical predictors.

Obviously, in retailing or marketing, knowledge of purchase "patterns" can help with the direct marketing of special offers to the "right" or "ready" customers (i.e., those that, according to the rules, are most likely to purchase some specific items given their observed past consumption patterns). However, transaction databases occur in many areas of business, such as banking, as well as general customer "intelligence." In fact, the term "link analysis" is often used when these techniques -- for extracting sequential or non-sequential association rules -- are applied to organize complex "evidence." It is easy to see how the "transactions" or "market-basket" metaphor can be applied to situations where individuals engage in certain actions, open accounts, contact other specific individuals, and so on. Applying these approaches to transactional databases may quickly extract patterns and associations between individuals and actions, and hence, reveal the patterns and structure in datasets.

Data Mining in the Process Industries

Over the years, the use and application of Neural Networks (NN) has found a "home" in the domain of industrial process control. It is also well known that NN is practically a core function in most popular data mining solutions. It is interesting though to note that NN algorithms have been embedded in process control solutions, yet sometimes seen or even projected as a bit of a "black box" or "magic box". Obviously, because of the complexity involved for most process control engineers to rationalize the output of an NN algorithm, except the performance of the controller.

RCA has traditionally been conducted by core statistical applications. RCA is classified based on the use or objectives as:

1. Safety-based RCA, which descends from the fields of accident analysis and occupational safety and health
2. Production-based RCA, which has its origins in the field of quality control for industrial manufacturing.
3. Process-based RCA, which is an "add-on" to production-based RCA, but with a scope that has been expanded to include business processes.
4. Failure-based RCA is rooted in the practice of failure analysis as employed in engineering and maintenance.
5. Systems-based RCA emerged as an amalgamation of the preceding uses, along with ideas taken from fields such as change management, risk management, and systems analysis.

In the course of an RCA initiative, the need to deal with substantial volumes of process data is well known. The combined dependent and independent variables can be in the range of hundreds for a single knowledge discovery problem addressed with data mining, and it is not uncommon that analyses without deep process knowledge can fail because of insufficient understanding of the process characteristics and behavior in question.

As such, data collection and the ability of a data mining solution to interface, sometimes in near real time, with plant data bases residing in a control systems (i.e. Integrated Control and Safety System) or a Historian data base, is key to the development of an integrated solution that can be deployed for use in a dynamic environment versus performing data mining analytics offline.

The approaches, principles or algorithms referred for the retail industry in the previous section can be extended in the process industries by considering sequence patterns of similar nature, except that we are dealing with much larger data sets, considering that the time resolutions of such sequences can be seconds or even milliseconds.

Take the scenario of a huge data set with a long sequence of events and alarms, with thousands of triggered flags or events, logged operator actions and also changes in battery limit conditions, and then add changes that are being tracked due to heat exchange fouling or catalyst characteristics, etc. , and then try to rationalize the outcome and impact on the behavior of continuous or discrete variable or a number of variables, be them trip events or an alarmed deviation of a safety or quality variable. Try also to visualize the endless number of dimensions (variables) involved as those related to the plant or process areas, the time dimension itself, the operators involved, the process unit operations or physical assets associated with the plant areas and units, etc., and then it becomes obvious that a data mining scenario of high complexity evolves. Yet , this is where data mining brings value and makes its money, because data can be both explored and analyzed in so many ways, but also used for predictive purposes by using the same techniques as in the retail example. It is by far a more complex situation that any other industry can offer.

In the case of the RCA example, data, once extracted, transformed and loaded for mining, rules about associations or the sequences of items as they occur in a transactional database can be established and make them useful not only for addressing the RCA problem in concern, but for many other applications, including exploratory and predictive data mining, as for example predicting runaway conditions for a catalytic reactor in a plant or preventing off spec production.

Development of data mining models in the process industries must take into consideration data model schemas, which are specific to process units and unit operations.

Data Models for Data Mining

The integration of data mining with an ICSS and Historian System is of paramount importance when it comes to deployment of a dynamic RCA solution.

Data for predictive modeling purposes can be utilized as flat files, which means that data is structured as an XY time-based matrix. Historians typically provide such flat files and as such they can be considered as Operational Data Stores (ODS). Results of exploratory analysis can be best utilized when the transformation of variables from an ODS is done based on a data model which is well thought out and designed in such a way that it meets the needs of various users. This means that a data model may feed another data model, typically known or called data mart.

The data model is practically the logical and physical structure of how data is mapped from existing data bases in real time systems (i.e. ICSS, Historian) onto flat files and ODS systems and then into multi-dimensional structures or schemas that are specific to a type of process unit or unit operation. It is important here to stress that the intellectual property of data

mining users is key here, since such data models should be considered as ownership of the data model developer, much like the ownership of a first principles process simulation models, which may be developed by using a specific tool, but the model is the ownership of the end user. It is true that a supplier of data mining services or tools may provide a customer with a generic data mining model, as for example a data model that represents the variables, data tables and schemas of a hydrocracker process unit, but this will have to be customized and fit into the specific process unit concerned by relating the actual variables and tags from the ICSS and Historian to the ones defined in the data model developed by the end user or by the organization employed by the end user. This is extremely important for end users of data mining.

Top 10 Data Mining Initiative Checks

The development and realization of a data mining initiative for Process Intelligence requires that end users take into consideration, and in advance, some issues listed below.

Issue	Comment
1. Clear definition of data mining objectives and benefits	It could be an RCA requirement, as in many cases known, or a combination of unexplained plant behaviors that need to be rationalized. In any event, a list of goals and objectives for any data mining initiative is where everything should start from. Ensure also engagement of a corporate sponsor who is stakeholder in the whole initiative.
2. In house knowledge	Even if there is some in house capability, getting a data mining subject matter expert (SME) aboard can save a company huge amounts. An SME may also help define goals and objectives of the initiative and also supervise project execution and knowledge transfer. Process knowledge is certainly a differentiator for any SME.
3. Investigating the market for similar data mining examples and experience	If someone has done it before, then it can be done again. If not, then risk of failure becomes higher. Supplier experience and culture in the process industries are critical.
4. Investigating the market for suppliers	This is typically a straight forward exercise as long as some principles are applied: <ul style="list-style-type: none"> • Supplier match means that supplier references and portfolio match the requirement of the specific initiative. • Not every data mining supplier is suitable for a data mining initiative in the process industries. • The biggest supplier does not means the best • The best does not mean the most suitable • The most suitable is not always the most cost effective Conclusion: experience, value and total cost of ownership are key.
5. Do a pilot project	Proof of concept is always useful, but do not expect to get all answers. And a free of charge pilot does not necessarily mean that the supplier willing to do it free of charge is the most suitable. It may be of strategic importance to the supplier, which is ok, but the long term commitment of the supplier in the process industry domain should be explored.
6. Avoid custom development	Some suppliers say that everything is possible to do and can be coded. Bespoke development should be avoided.
7. Productization	Use of off the shelf software products, as for example OPC compliant interfaces to historians like OSIsoft, which have been used in the past should be preferred.
8. Technology Features & Methodology	Most data mining solutions offer the most commonly required features. Focus on the results versus the features. Regarding methodology, most suppliers have adopted standard practices for the data mining workflow (i.e. SEMMA, SixSigma, etc.)
9. Licensing terms	The type of licensing and price ranges can vary dramatically. Consider also long term licensing schemes for multiple sites. Typical licensing schemes are as follows: <ul style="list-style-type: none"> • Perpetual license: allows the customer to install and use the software indefinitely. Technical support is included for a limited term, usually 90 days. • Subscription license or non – perpetual license: Allows the user to use the software for a specified time period (i.e. 1 year). This license usually includes technical support and access to upgrades and patches released during the term of the subscription. At the end of the term the user has several options: (1) renew the subscription; or (2) purchase a perpetual license at a discounted cost; or (3) remove the software from the computer. • Volume licenses: allows the Licensee to install the software on a certain number of computers. The licensee usually has to satisfy a minimum purchase requirement and obtains reduced prices in exchange. When purchasing the licenses, the licensee usually receives one copy of the media and documentation with the option of purchasing more. • Site/Enterprise: This license provides access to software at a single location. Typically, these licenses are individually negotiated with the publisher and vary widely in their provisions. • Server (Network): Licensed per server – This license type requires that you have a single copy of the software residing on the file server. With Per Server licensing, a

	<p>specified number of CALs are associated with a particular server. The number of devices that can legally access that server simultaneously is limited in Per Server licensing to the number of CALs purchased for that particular server.</p> <ul style="list-style-type: none"> • Per Seat (Machine): Licensed per machine/seat – This license requires that you purchase a license for each client computer and/or device where access to services is needed. This license is typically used in conjunction with a network license. • Per Processor: Under the Per Processor model, you acquire a Processor License for each processor in the server on which the software is running. A Processor License usually includes access for an unlimited number of users to connect. You do not need to purchase additional server licenses, CALs, or Internet Connector Licenses. • Concurrent Use: This license type requires that you purchase licenses for the maximum number of people who will be running the software simultaneously. However, you can usually install the software on more computers than
<p>10. System Architecture for deployment</p>	<p>Data mining tools are just tools. But deployment of data mining solutions, in a real plant environment, including interfaces to real time systems, web (intranet) enablement, etc. is a different story and needs to be dealt with caution, otherwise you may end up with a very limited solution.</p>

Conclusions

Process Intelligence is a term that has wide interpretation in the process industries, but for many visionaries and innovators it relates to the ability to exploit plant data for competitive advantage. Thus, process intelligence can be empowered by data mining.

Data mining should be approached with specific objectives and take into consideration of how the outcome of such initiative can bring value to as many stakeholders and of course how. Use of data models is rather new but critical for the successful establishment of standards versus ad hoc deployments.

Establishing strategic relationships with a data mining solution provider and subject matter experts is of paramount importance for the long terms sustainability of such initiatives, including lifecycle management and total cost of ownership of these solutions.

About the Author:



Founder of ziconNET and executive consultant with >30 year career mainly in the process and IT industries and in the development and implementation of process systems, dynamic simulation, advanced process control, information management, manufacturing execution and supply chain management solutions, data warehousing and data mining systems, business intelligence and decision support systems, geographical information management, CAD/CAM/CAE, facilities management and photogrammetry systems for land information and cadastre applications.

Author of a number of articles, white papers and extensive participation in forums as a speaker and visionary for information technology solutions and applications both in the process and business applications domains.