Chapter 17

Use of Data Mining Techniques for Process Analysis on Small Databases

Matjaz Gams
Jozef Stefan Institute, Ljubljana, Slovenia

Matej Ozek
Jozef Stefan Institute, Ljubljana, Slovenia

ABSTRACT

The pharmaceutical industry was for a long time founded on rigid rules. With the new PAT initiative, control is becoming significantly more flexible. The Food and Drug Administration is even encouraging the industry to use methods like machine learning. The authors designed a new data mining method based on inducing ensemble decision trees from which rules are generated. The first improvement is specialization for process analysis with only a few examples and many attributes. The second innovation is a graphical module interface enabling process operators to test the influence of parameters on the process itself. The first task is creating accurate knowledge on small datasets. The authors start by building many decision trees on the dataset. Next, they subtract only the best subparts of the constructed trees and create rules from those parts. A best tree subpart is in general a tree branch that covers most examples, is as short as possible and has no misclassified examples. Further on, the rules are weighed, regarding the number of examples and parameters included. The class value of the new case is calculated as a weighted average of all relevant rule predictions. With this procedure the authors retain clarity of the model and the ability to efficiently explain the classification result. In this way, overfitting of decision trees and overpruning of the basic rule learners are diminished to a great extent. From the rules, an expert system is designed that helps process operators. Regarding the second task of graphical interface, the authors modified the Orange explanation module so that an operator at each step takes a look at several space planes, defined by two chosen attributes (Demšar et al., 2004). The displayed

DOI: 10.4018/978-1-60566-908-3.ch017
attributes are the ones that appeared in the classification rules triggered by the new case. The operator can interactively change the current set of process parameters in order to check the improvement of the class value. The task of seeing the influence of combining all the attributes leading to a high quality end product (called design space) is now becoming human comprehensible, it does not demand a high-dimensional space vision any more. The method was successfully implemented on data provided by a pharmaceutical company. High classification accuracy was achieved in a readable form thus introducing new comprehensions.

INTRODUCTION

Donald E. Knuth said almost twenty years ago (D. Knuth, interview, 1993): “I think the most exciting computer research now is partly in robotics, and partly in applications to biochemistry. Biology is so digital, and incredibly complicated, but incredibly useful.”

In 2004, the United States Food and Drug Administration (FDA) issued a document “PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance.” This document was written as guidance for a broad industry audience in different organizational units and scientific disciplines (“Guidance for industry PAT”, 2004). To a large extent, the guidance discusses principles with the goal of developing regulatory processes of drug production that encourage innovation.

As the FDA states, the conventional pharmaceutical manufacturing is generally accomplished using batch processing with laboratory testing conducted on collected samples to evaluate quality (“Guidance for industry PAT”, 2004). This conventional approach has been successful in providing quality pharmaceuticals to the public. However, today significant opportunities exist for improving pharmaceutical development, manufacturing, and quality assurance through innovation in product and process development, process analysis, and process control. This is where machine learning might help.

Unfortunately, the pharmaceutical industry has been generally hesitant to introduce innovative systems into the manufacturing sector for a number of reasons. One often cited reason is regulatory uncertainty, which may result from the perception that the existing regulatory system is rigid and unfavorable for the introduction of innovative systems. For example, many manufacturing procedures are treated as being frozen and many process changes are managed through regulatory submissions.

Because of the hesitancy of the pharmaceutical industry, the document encourages new production techniques with common name Process Analytical Technology (PAT). Its focus is innovation in development, manufacturing and quality assurance by removing “regulatory fear/uncertainty”, utilizing science and risk-based approach to regulatory requirements and oversight. This will provide a flexible and less burdensome regulatory approach for well understood processes, creating an environment that facilitates rationale science, risk, and business decisions.

Therefore, the pharmaceutical industry needs a system for designing, analyzing and controlling manufacturing process (Schneidir, 2006). The goal of PAT is to understand and control the manufacturing process in real time. The system must be able to make the recommendation during the process to achieve higher quality of the end product. The system should follow the performance attributes, raw and in-process materials and processes. A PAT system should use multiple tools for understanding and controlling the manufacturing process: multivariate tools for design, data acquisition and analysis, process analyzers, process control tools, continuous improvement and knowledge management tools.
Related Content

Discovery of Anomalous Windows through a Robust Nonparametric Multivariate Scan Statistic (RMSS)
[www.igi-global.com/article/discovery-anomalous-windows-through-robust/75614?camid=4v1a](www.igi-global.com/article/discovery-anomalous-windows-through-robust/75614?camid=4v1a)

Transmission Control Protocol for Mobile Ad Hoc Network
[www.igi-global.com/chapter/transmission-control-protocol-for-mobile-ad-hoc-network/139394?camid=4v1a](www.igi-global.com/chapter/transmission-control-protocol-for-mobile-ad-hoc-network/139394?camid=4v1a)

Collaborative Filtering Based Data Mining for Large Data
[www.igi-global.com/chapter/collaborative-filtering-based-data-mining-for-large-data/159498?camid=4v1a](www.igi-global.com/chapter/collaborative-filtering-based-data-mining-for-large-data/159498?camid=4v1a)

Applied Analytical “Distant Reading” using NVivo 11 Plus™
Shalin Hai-Jew (2017). *Social Media Data Extraction and Content Analysis* (pp. 159-201).
[www.igi-global.com/chapter/applied-analytical-distant-reading-using-nvivo-11-plus/161964?camid=4v1a](www.igi-global.com/chapter/applied-analytical-distant-reading-using-nvivo-11-plus/161964?camid=4v1a)