Baum der Erkenntnis

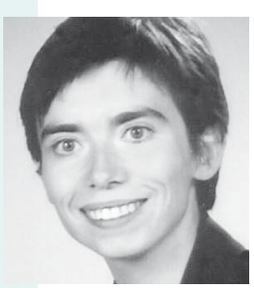
Data Mining macht aus einem Dickicht aus Pharma-Produktionsdaten wertvolle Basisinformationen für die Prozessoptimierung

Betreiber in der Prozessindustrie sitzen häufig auf einem Schatz. Im Lauf der Zeit hat sich eine gewaltige Datenmenge angesammelt, die größtenteils nicht verwertet wird. Diese Rohdaten verbergen Informationen, Beziehungen, ursächliche Zusammenhänge, die – einmal entdeckt – erlauben, bestimmte Phänomene genauer zu verstehen und im weiteren Verlauf die Prozesse und die Qualität der Produkte zu optimieren. Denn Qualitätsprobleme sind Ursache von Verspätungen, unbefriedigten Kunden und Glaubwürdigkeitsverlust und stellen letztendlich einen großen finanziellen Verlust für das Unternehmen dar.

Isabelle Giraud Zindy, Philippe Solot

Quelle: Boehringer Ingelheim

Die Qualitätssicherung, in der Pharmaindustrie von besonders hoher Bedeutung, wird durch Data Mining wirkungsvoll unterstützt.



Dr. Isabelle Giraud Zindy
ist für Marketing bei Aicos Technologies in Basel zuständig
T +41/61/68698-82
igiraud@aicos.com

Dr. Philippe Solot
ist Geschäftsführer von Aicos Technologies in Basel
T +41/61/68698-76
psolot@aicos.com

Den in Produktionsanlagen der Prozessindustrie bestehenden wertvollen Datenbestand zu vernachlässigen, wäre im derzeitigen angespannten Wettbewerbsumfeld inkonsequent. Die zur Datenverwertung nutzbaren Methoden des Data Mining („Daten schürfen“) existieren seit langem und haben den Nachweis ihrer Effektivität erbracht. Ihr Ziel besteht darin, aufgrund großer Datenmengen ein erklärendes oder Vorhersagemodell zu bauen, und dies mit automatischen oder halbautomatischen Methoden. Dennoch erschrecken sie manchmal noch den Nicht-Statistiker, der nicht weiß, wie er sein Problem angehen soll, und befürchtet, die Ergebnisse nicht richtig interpretieren zu können. Tatsächlich ist man in der Praxis mit Datenmengen konfrontiert, die oft mehr als fünfzig, wenn nicht gar hundert Parameter umfassen. Jedoch fegt die Nutzung moderner und benutzerfreundlicher Software diese Hindernisse beiseite. Derartige Software übernimmt den ganzen Teil der wissenschaftlichen Berechnung und langweiligen Statistik und erlaubt, sich der eigentlichen Problematik des Prozesses zu widmen.

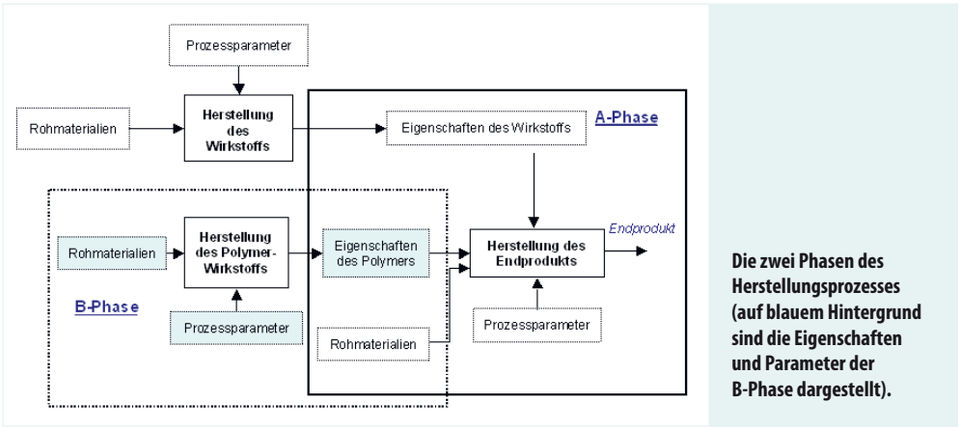
Qualität mikrogekapselter Pharmaprodukte gesteigert

Ein pharmazeutisches Produktionsunternehmen, das auf Qualitätsprobleme in der Herstellung eines seiner Produkte (eines Hormon-Analogons) gestoßen war, hat erfolgreich auf eine solche Software zurückgegriffen. Dieser Hormonersatz kann nur als Injektion verab-

reicht werden. Um den Komfort der Patienten zu verbessern und ihnen zu ersparen, für eine kontinuierliche Zufuhr der Substanz ständig mit einem Infusionsgerät verbunden bleiben zu müssen, ist das aktive Molekül in einem anderen, neutralen und löslichen Molekül verkapselt. Dadurch wird seine Bioverfügbarkeit optimiert. Somit können sich die Patienten auf tägliche oder halbtägliche Injektionen beschränken. Aber das Endprodukt zeigte in der Herstellung eine leicht höhere Auflösungsgeschwindigkeit, als es die Entwicklungsphase vorhersehen ließ.

Der Herstellungsprozess besteht aus zwei Phasen: einerseits die Synthese des eigentlichen aktiven Moleküls und andererseits die Herstellung des Polymers, das die Mikro-Kapsel bildet. Man vermutete, dass die zu schnelle Auflösungsgeschwindigkeit einem Problem in dieser letzten Phase angelastet werden kann (Phase B, siehe Abbildung unten). Diese Phase besteht aus vier verschiedenen Produktionsschritten, in welchen über verschiedene Zwischenprodukte das Endpolymer entsteht.

Die Datenmenge, die im Lauf mehrerer Herstellungsjahre gesammelt wurde, ist bedeutend genug, um Data-Mining-Methoden darauf anwenden zu können, wie etwa die Methode der CART-Regressionsbäume. Das Ziel dieser a posteriori-Datenuntersuchung besteht darin, diejenigen Parameter zu identifizieren, die die Zusammensetzung oder Qualität des Polymers am deutlichsten beeinflussen. Auf der einen Seite sollten hierbei die Prozessfaktoren untersucht werden (wie zum Beispiel die Reaktionstemperatur) und auf der anderen Seite der Einfluss der Ausgangsmaterialien ausgewertet werden. Die Charakterisierung des Polymers geschieht durch die Messung seiner molaren Masse.

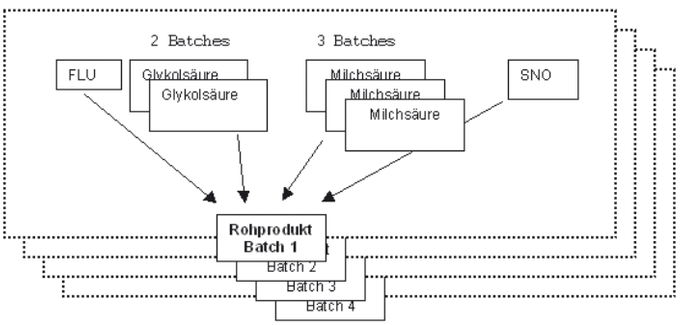


Die zwei Phasen des Herstellungsprozesses (auf blauem Hintergrund sind die Eigenschaften und Parameter der B-Phase dargestellt).

Die Auswertung basierte auf der Methode der CART-Entscheidungsbäume. CART ist das Akronym für „Classification And Regression Trees“; dies ist ein Algorithmus für die Erstellung binärer Entscheidungsbäume, der 1984 von weltbekannten Statistikern der Universität von Kalifornien entwickelt wurde. Die Auswertung wurde in hohem Maße durch die Benutzung der auf diesem Algorithmus aufbauenden Software „CART“ vereinfacht.

Kritische Variablen identifiziert

Hinsichtlich der klassischen statistischen Methoden haben Entscheidungsbäume folgende Vorteile: ›



Produktionsablauf mit Zusammenführung mehrerer Rohmaterial-Batches.

- Sie können sowohl lineare als auch nichtlineare Beziehungen modellieren;
- Wechselwirkungen zwischen Faktoren werden automatisch berücksichtigt und müssen nicht im voraus spezifiziert sein;
- fehlende Werte werden angemessen behandelt;
- sie sind Ausreißern gegenüber robust;
- es ist unnötig, die Vorhersagevariablen zu transformieren;
- das Endmodell umfasst nur die wichtigsten Faktoren zur Vorhersage des Ergebnisses;
- und schließlich ist das Ergebnis leicht interpretierbar, ohne besondere statistische Kenntnisse zu erfordern.

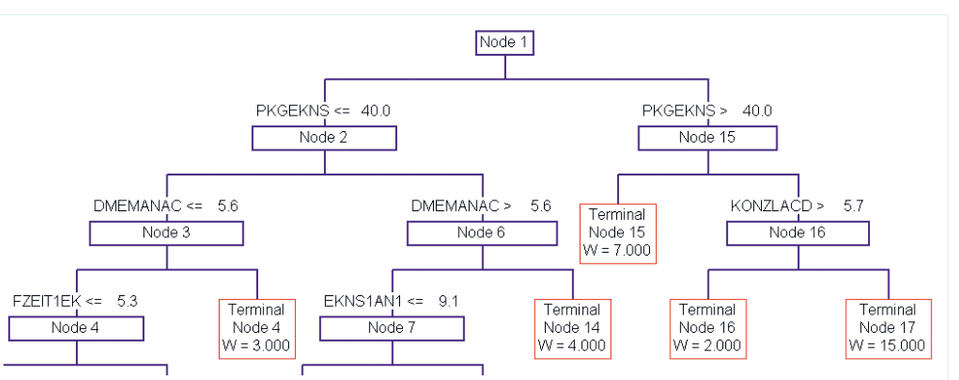
Die einflussreichsten Faktoren werden automatisch durch den Entscheidungsbaum bestimmt. Diese Siebeigenschaft ist besonders wertvoll, wenn die potenziell wichtigen Variablen zahlreich und ungenügend bekannt sind – wie es bei Produktionsdaten oft der Fall ist.

Bruchstückhafte Daten

Der Datensatz litt unter vielen fehlenden Werten, was daran liegt, dass die Datenerfassung über mehrere Jahre hinweg durchgeführt wurde, und dass das Messprotokoll im Lauf der Zeit leichten Änderungen unterlag. Nicht alle Messungen wurden immer durchgeführt, und folglich sind die Daten manchmal lückenhaft.

Die CART-Software umgeht dieses Problem, indem sie Aufspaltungsregeln mit Surrogat-Substitution einführt. Wenn eine Variablenangabe fehlt, ersetzt sie diese Variable durch eine andere mit vergleichbarer Wirkung. Wenn eine Regel des Baums die Variable „AbsolutAusbeute“ enthalten würde und mehrere Werte dieser Variable fehlen würden, würde CART zum Beispiel die Variable „RelativAusbeute“ als Ersatz nehmen, deren Wirkung ähnlich ist. Diese Methode gibt den erhaltenen Modellen eine größere Robustheit, selbst wenn große Datensätze mit etwa hundert Parametern und vielen fehlenden Werten vorliegen, wie dies in Produktionsanwendungen typischerweise der Fall ist.

Da in die Herstellung eines einzigen Endproduktbatches mehrere Rohstoffbatches eingingen, musste man zudem zunächst die Daten konsolidieren, indem man Durchschnittswerte unterschiedlicher Batches berechnete (siehe Abbildung oben). Die hier zu untersuchende Zielgröße ist die Mol-Masse des Polymers, und somit eine quantitative Antwort. Der Entscheidungsbaum wird also ein so genannter Regressionsbaum sein. CART erlaubt aber auch die Untersuchung von qualitativen Antworten, wie beispielsweise eine Qualitätsbeurteilung (schlecht, erträglich, gut, hervorragend); in diesem Fall wird der Baum als Klassifikationsbaum bezeichnet. Die Einflussfaktoren können natürlich ebenfalls entweder zum quantitativen oder qualitativen Typ gehören; die Software findet automatisch diese zwei Typen heraus und behandelt ihre Werte angemessen.



Eine eindeutige Darstellung des Ergebnisses.

Kinderleichte Benutzung

Unter dem Nutzungsaspekt ist die Verwendung von CART extrem einfach: Daten, die in Form von Excel-Tab-Dateien gespeichert sind, können direkt in die Software importiert werden. Es reicht danach aus, die zu untersuchenden Einflussfaktoren sowie die zu charakterisierende Zielgröße in einem Dialogfenster mit der Maus auszuwählen.

Im vorliegenden Fall bezogen wir 46 Faktoren in die Auswertung ein. Diese beinhalten sowohl Parameter, die die Verfahrensweise betreffen (wie Latenzzeiten oder Reaktionstemperaturen), als auch Einflussgrößen, die die Rohmaterialien beschreiben (zum Beispiel die Farbe des Produkts, seine Konzentration, sein Gewicht…).

Das Programm eliminiert die Eingaben, bei denen eine Angabe über die Zielgröße fehlt und erzeugt danach automatisch den Entscheidungsbaum (Abbildung unten). Dieser Baum kann mit mehr oder weniger Details dargestellt werden (Abbildung Seite 45). So kann jeder Knoten mit seiner Aufspaltungsvariable, seinem Schwellenwert, der Aufspaltungsvariable des Elternknotens, der Zielgrößenvarianz usw. dargestellt werden.

Der „Report Writer“ der Data-Mining-Software ist besonders leistungsstark und benutzerfreundlich. Er erlaubt, den Auswertungsbericht der Studie nach Maß zu erzeugen.

Berichte „à la carte“

Jede von CART erstellte Graphik kann durch einen einfachen Mausklick in den Bericht aufgenommen werden, indem man den „Add to Report“-Befehl ausführt. Es ist auch möglich, zunächst einen Bericht, der die Hauptergebnisse umfasst, automatisch zu erzeugen und diesen danach zu personalisieren. Das Dokument kann im rtf-Format gespeichert werden, was eine gute Kompatibilität mit der Mehrzahl der Textverarbeitungsprogramme gewährleistet.

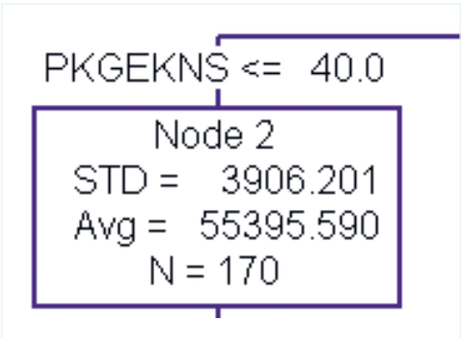
Das Programm liefert zudem eine weitere Größe, die „Wichtigkeit“ der Variablen. Die Berechnung dieser Wichtigkeit berücksichtigt nicht nur ihr Auftreten als Aufspaltungsvariable der Knoten, sondern auch ihre Rolle als Ersatzvariable – eine Rolle, die nicht sichtbar ist, wenn man sich darauf beschränkt, den Baum zu betrachten.

In unserem Beispiel sind die zwei wichtigsten Variablen (siehe Tabelle 1) also PPREKNS und PKGEKNS. Das Aufspaltungskriterium für den ersten Knoten des Baums ist die Variable PKGEKNS. Wenn man sich auf den detaillierten Bericht bezieht, der durch CART erstellt wurde, sieht man, dass die Ersatzvariable (das „Surrogat“) von PKGEKNS die Variable PPREKNS ist. Diese zwei Variablen stellen das absolute und das relative Gewicht des Produkts EKNS dar.

Man stellt sofort fest, dass die Etappe 2 deutlich die kritischste im Fertigungsverfahren des Polymers ist. Die Massen (relativ und absolut) des Endprodukts dieser Etappe, EKNS, werden als die zwei wichtigsten Faktoren identifiziert. Die vier nächstwichtigen Variablen betreffen alle die EKN- und CEF-Zusatzmengen, die ihrerseits einflussreiche Faktoren für die Masse des Endprodukts der Etappe 2 darstellen. Man kann daraus folgern, dass die Menge des letzteren Produkts der wichtigste Einflussfaktor für die Endzielgröße, also die molare Masse des Polymers, ist. Das Zwischenprodukt EKNS ist selbst von der Masse der Zusätze EKN und CEF abhängig.

Fazit

In dieser Anwendung in der Pharmaproduktion liegt das Hauptinteresse der Benutzung von CART deutlich in der unglaublichen Geschwindigkeit, mit der die potenziellen Kausalitätsbeziehungen erläutert werden konnten. Dies hat ermöglicht, eine schnelle Sortierung der zahlreichen Ausgangsvariablen durchzuführen. Es ist von nun an möglich, diejenigen Parameter, die als die einflussreichsten identifiziert wurden, gezielt zu überwachen. Außerdem kann man sich auf diese wenigen, als die am wichtigsten identifizierten Faktoren stützen, um eine optimale Einstellung der Parameter zu erhalten; hier bietet es sich an, eine statistische Versuchsplanung durchzuführen, was zuvor mit 46 Produktionsfaktoren undenkbar gewesen wäre. ■



Detailsicht des zweiten Knotens mit Mittelwert, Standardabweichung und Anzahl Batches im Teilbaum.

Tabelle 1: Bedeutung und Wichtigkeit der wichtigsten Variablen.

Bedeutung	Kürzel	Wichtigkeit	
Masse des Endprodukts relativ (Etappe 2)	PPREKNS	100.00	
Masse des Endprodukts absolut (Etappe 2)	PKGEKNS	89.59	
Masse des Zusatzstoffs EKN (Etappe 2)	EKNS1AN1	83.03	
Masse des Zusatzstoffs CEF (Etappe 2)	CEFOKAN1	73.94	
Messung am Zusatzstoff EKN (Etappe 2)	IMPSULEK	72.27	
Messung am Zusatzstoff CEF (Etappe 2)	EXWATEKN	72.27	

Weiterführende Infos auf www.PuA24.net

more @ click PA076054