Die Datenfülle zu Qualität verdichten

Moderne Data Mining-Techniken erlauben, versteckte Beziehungen zwischen Produktions-Qualitätsmerkmalen und Prozeßparametern zu identifizieren. Die oft mühsame Suche nach den qualitätsmindernden Variablen wird durch Regressionsbäume vereinfacht, wie sie etwa zur Analyse eines komplexen Trocknungsprozesses herangezogen wurden.

Data Mining
Methodenbericht

Riesige Datenmengen, die heute in der Produktion anfallen, bieten die Chance, den Ursachen von Produktionsabweichungen auf die Spur zu kommen und so schließlich Prozesse zu optimieren. Allzu oft liegen die Daten allerdings ungenutzt herum, teils aufgrund mangelnder Kenntnisse in Datenanalyse, teils wegen beschränkter Rechnerkapazität.

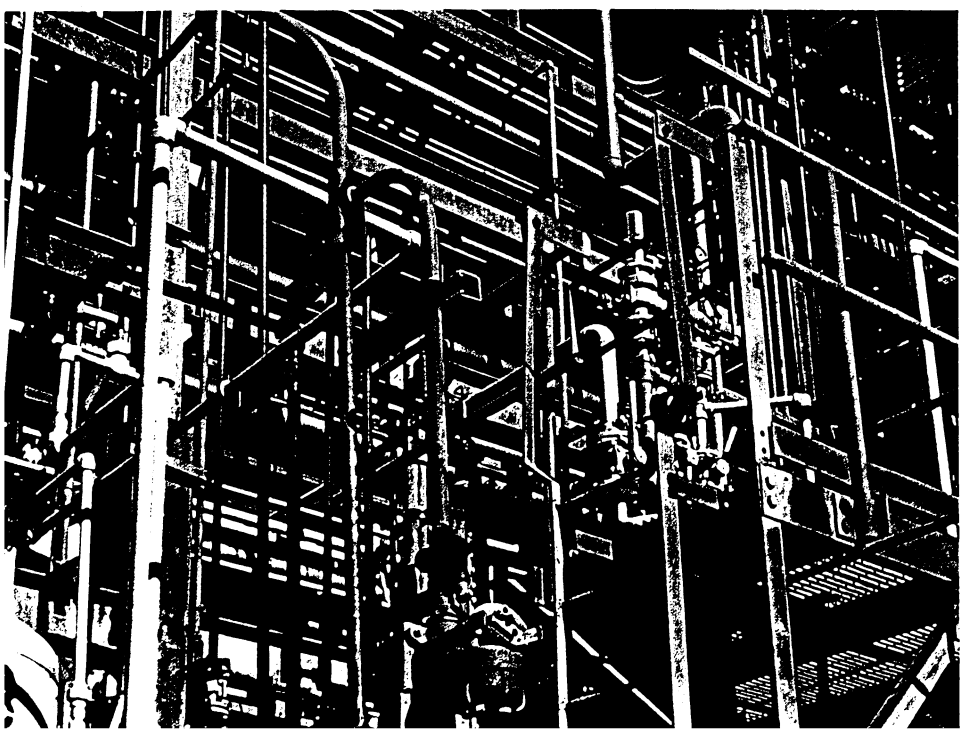
Gewiß stellen Menge und komplexe Struktur der Datensätze eine große Herausforderung an die statistischen Analysemethoden dar. Trotz fehlender Werte und Ausreißer in Produktionsdaten, trotz komplizierter Wechselwirkungen zwischen einzelnen Variablen möchte man relevante Zusammenhänge entdecken und die Abhängigkeit zwischen Prozeßparametern modellieren. Moderne Data Mining-Techniken erlauben, solche Problemstellungen zu lösen. Oft erweisen sich dabei sogenannte Klassifikations- und Regressionsbäume (CART) als besonders geeignete Hilfsmittel.

Unter Data Mining versteht man die Gewinnung verborgener Information aus üblicherweise sehr großen, komplexen Datensätzen. Das Ziel kann dabei die Entdeckung von Korrelationen, die Erkennung von Trends oder die Bildung von Modellen sein, die für Prognosen oder Optimierungen benutzt werden können. Im Vordergrund steht die Analyse von Daten. Die dazu verwendeten Techniken stammen aus der Statistik. Trotzdem gibt es zwischen der klassischen Statistik und dem Data Mining konzeptionelle Unterschiede. Im ersten Fall werden die Daten im Hinblick auf eine spezifische Fragestellung gesammelt, wie etwa für die Versuchsplanung für Prozeßoptimierungen.

Beim Data Mining dagegen geht es darum, in vorhandenen Daten unbekannte, interessante Zusammenhänge zu finden. Da in solchen Fällen oft mit einer schlechten Datenqualität (fehlende oder fehlerhafte Werte) zu rechnen ist, ist der Vorbereitungsphase (Data Screening) besondere Beachtung zu schenken.

Klassifikations- und Regressionsbäume sind flexible Hilfsmittel, um komplizierte, multivariante Datensätze zu beschreiben und einfache Vorhersageregeln zu formulieren. Ziel ist, die Wirkung von erklärenden Variablen auf eine interessierende Größe, die sogenannte Zielvariable, zu modellieren. Handelt es sich bei dieser um eine kontinuierliche Größe, spricht man von einem Regressionsbaum. Für eine kategorielle Zielvariable verwendet man die Bezeichnung Klassifikationsbaum.

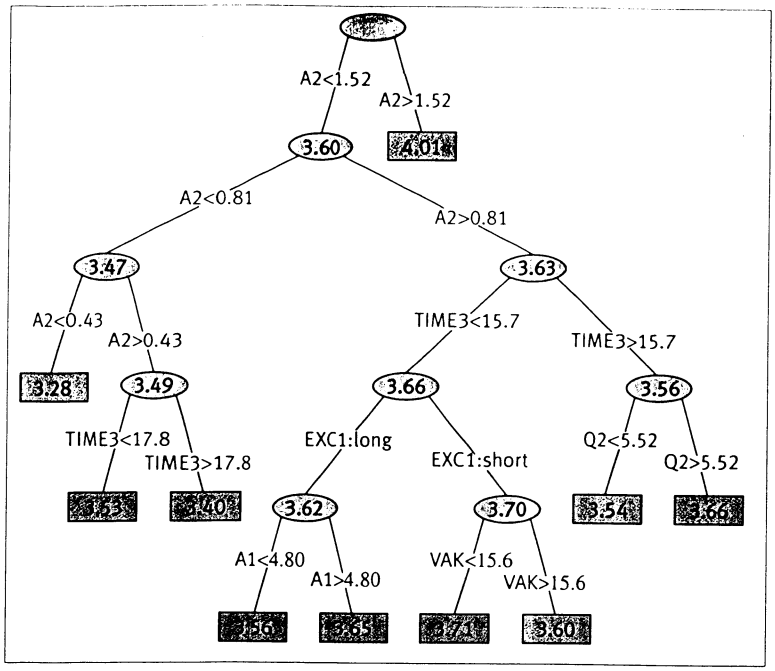
Für das folgende, auf Wunsch des Kunden hin anonymisierte Praxis-



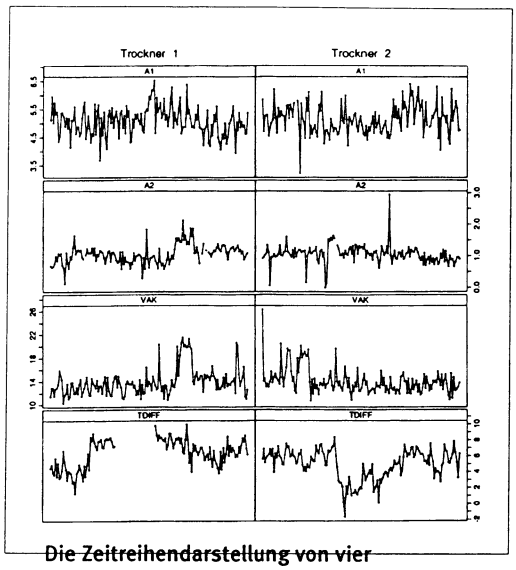
Die Datenfülle aus komplexen Anlagen läßt sich erst mit Methoden wie Data Mining zu verwertbaren Informationen für die Prozeßoptimierung verarbeiten.

Professionelle Beratung zahlt sich aus

Oft werden Auswertungsverfahren wie Klassifikations- und Regressionsbäume mißverstanden oder falsch angewendet. Durch nicht erfüllte Erwartungen werden wertvolle Techniken vernachlässigt und so wird die Chance auf Wettbewerbsvorteile verpaßt. Zudem ist es eine Illusion, zu glauben, durch ein paar Mausklicks könnte die Komplexität von Prozessen erklärt werden. Fachmännische Beratung durch professionelle Spezialisten spart viel Zeit, Ärger und Geld. Präzise Problemdefinition, sorgfältige Datenaufbereitung, Auswahl der richtigen Analysetechniken sowie kompetente Interpretation der Ergebnisse sind die Schlüssel zu effizienter Entdeckung versteckter Zusammenhänge.



Ein Regressionsbaum für den Endgehalt des Qualitätsmerkmals Q erlaubt die Analyse eines Trocknungsprozesses.



Die Zeitreihendarstellung von vier Prozeßvariablen liefert Hinweise auf Prozeßstörungen. Zudem sind Unterschiede zwischen den beiden Trocknern ersichtlich.

Info-Dienst

Der Baum der Erkenntnis

Vorteile von Klassifikations- und Regressionsbäumen :

- *Sowohl lineare wie auch nichtlineare Zusammenhänge zwischen erklärenden Variablen und Zielvariablen werden beschrieben;*
- *Wechselwirkungen zwischen Variablen werden automatisch berücksichtigt, müssen nicht im Voraus spezifiziert werden;*
- *fehlende Werte in den Variablen werden geeignet behandelt;*
- *nur jene Variablen, die für die Beschreibung der Zielvariable wichtig sind, werden ins Modell aufgenommen, quasi automatische Variablenselektion;*
- *Resultate sind auch ohne vertiefte statistische Kenntnisse einfach zu interpretieren.*

beispiel eines Trocknungsprozesses wurden alle Zahlen simuliert:

Bei einem chemischen Produkt wurde eine zu große Variabilität eines Qualitätsmerkmals Q festgestellt. Als kritische Phase wurde der Trocknungsprozeß in Betracht gezogen, durch welchen der Gehalt von Q und anderen Stoffen (A und B) im Produkt reduziert wird.

In zwei Trocknern werden jeweils drei Trocknungsphasen mit Temperaturen zwischen 30 und 70 °C durchgeführt, wobei die Dauer einer Phase vom Feuchtgewicht des Loses zu Beginn der Trocknung abhängt. Der Trocknungsprozeß dauert zwischen 30 und 50 Stunden. Nur durch die optimale Wahl von Temperatur und Dauer sowie anderer Parameter kann der Gehalt von Q auf einen akzeptablen Wert reduziert werden.

Es standen Daten von 306 Losen zur Verfügung. Neben dem Qualitätsmerkmal Q wurden ca. 30 Prozeßvariablen gemessen, wovon vier in der Grafik links oben dargestellt sind. Bei dieser geringen Größe des Datensatzes handelt es sich zwar nicht um eine typische Data Mining-Anwendung; der Nutzen der Methode wird damit dennoch illustriert.

Wie kann man nun mit Hilfe der Daten eines solch komplexen Prozesses die Ursachen der Qualitätsabweichungen identifizieren sowie Empfehlungen zur Prozeßverbesserung formulieren? Dazu ist die Baum-Methode besonders geeignet.

Einen Regressionsbaum liest man von oben nach unten wie einen Entscheidungsbaum. Dabei basieren die Abzweigungen auf einfachen Regeln wie: Ist eine erklärende Variable kleiner oder größer als eine vorgegebene Schranke? Man stellt also fest, daß für einen Gehalt des Stoffes A nach Phase 2 (Variable A2) größer als 1,52 Prozent ein hoher Endgehalt von 4,01 Prozent für Q resultiert. Ansonsten beträgt der mittlere Gehalt von Q nach Abschluß der Trocknung nur 3,60 Prozent. In diesem Fall müssen weitere Variablen herangezogen oder neue Einschränkungen der benutzten Variablen gemacht werden, um präzisere Voraussagen zu formulieren. Falls also der Gehalt von A nach Phase 2 zwischen 0,43 und 0,81 Prozent liegt und die Trocknungsphase 3 (TIME3) länger als 17,8 Stunden dauert, führt das zum relativ tiefen Endgehalt von Q von 3,40 Prozent.

Aus dem Baum geht hervor, daß hohe Werte des Stoffs A nach Phase 2 eine Ursache für die Variabilität des Qualitätsmerkmals Q sind. Die Reduktion des Gehalts von A muß deshalb bereits in Phase 2 erfolgen. Ferner erlaubt der Baum, die Anzahl potentiell einflußreicher Variablen von insgesamt 28 auf nur sechs zu reduzieren.

Dr. Yves L. Grize, Christian Keller,
Aicos Technologies