AICOS Technologies AG
ADVANCED INDUSTRIAL CONSULTING IN OPERATIONS RESEARCH AND STATISTICS

Christian Keller, Dr. Yves-L. Grize
Efringerstrasse 32
CH-4057 Basel, Switzerland
Tel.: +41-61-686 98 77
Fax: +41-61-686 98 88
Web: `http://www.aicos.com`

Prozessdatenanalyse mit Regressionsbäumen

Ein Projekt von AICOS Technologies AG

Einführung

Bei vielen Produktionsprozessen werden heute Daten erhoben, wobei die Messungen vermehrt automatisch erfolgen. Die Verfügbarkeit von Daten ermöglicht, Zusammenhänge zwischen Prozessparametern zu studieren, um dadurch die Ursachen von Produktionsabweichungen zu identifizieren und um Prozesse zu optimieren.

Die wachsende Menge und die komplexe Struktur der Datensätze stellen aber eine grosse Herausforderung an die statistischen Analysemethoden dar. So ist mit fehlenden Werten und Ausreissern in Produktionsdaten häufig zu rechnen. Zudem treten oft komplizierte Wechselwirkungen zwischen einzelnen Prozessvariablen auf. Trotzdem möchte man relevante Zusammenhänge in den Daten entdecken und die Abhängigkeit zwischen Prozessparametern modellieren. Moderne Data Mining Techniken erlauben, solche Problemstellungen zu lösen. Als besonders geeignet erweisen sich dabei sogenannte Klassifikations- und Regressionsbäume (CART).

Was sind Klassifikations- und Regressionsbäume?

Klassifikations- und Regressionsbäume stellen eine flexible Methode dar, die verwendet wird, um komplizierte multivariate Datensätze zu beschreiben und einfache Vorhersageregeln zu formulieren. Das Ziel ist, die Wirkung von erklärenden Variablen auf eine interessierende Grösse, die sogenannte Zielvariable, zu modellieren. Handelt es sich bei der Zielvariable um eine kontinuierliche Grösse, spricht man von einem Regressionsbaum. Für eine kategorielle Zielvariable verwendet man die Bezeichnung Klassifikationsbaum.

Im Vergleich zu anderen Methoden sind folgende Vorteile von Klassifikations- und Regressionsbäumen zu erwähnen:

- Sie können sowohl lineare wie auch nichtlineare Zusammenhänge zwischen den erklärenden Variablen und der Zielvariable beschreiben.
- Allenfalls vorhandene Wechselwirkungen zwischen einzelnen Variablen werden automatisch berücksichtigt und müssen nicht im Voraus spezifiziert werden.
- Fehlende Werte in den Variablen werden geeignet behandelt.
- Es werden nur jene Variablen ins Modell aufgenommen, die für die Beschreibung der Zielvariable auch wichtig sind. Es wird also in einem gewissen Sinne eine automatische Variablenselektion durchgeführt.
- Die Resultate sind auch ohne vertiefte statistische Kenntnisse einfach zu interpretieren.

Klassifikations- und Regressionsbäume können im Statistik-Programm S-PLUS mit der Funktion `tree` angepasst werden. Die Modelle lassen sich dabei mit der flexiblen Formelnotation von S-PLUS spezifizieren. Zahlreiche weitere Funktionen dienen der Modellwahl und Modellüberprüfung sowie der grafischen Darstellung der Resultate (vgl. z.B. Abbildung 2).

Eine praktische Anwendung: Analyse eines Trocknungsprozesses

Bei einem chemischen Produkt wurde eine zu grosse Variabilität des Ethylester-Gehaltes in Bezug auf die Spezifikationen festgestellt. Als kritische Phase in der Produktion wurde der Trocknungsprozess in Betracht gezogen. Die Trocknung dient unter anderem dazu, den Gehalt von Ethylester und anderen Stoffen im Produkt zu reduzieren. Drei Trocknungsphasen mit verschiedenen Temperaturen werden durchgeführt, wobei die Dauer einer Phase von zahlreichen Merkmalen abhängt, wie beispielsweise dem Feuchtgewicht des Loses zu Beginn der Trocknung. Nur mit einer optimalen Wahl von Temperatur und Dauer sowie anderer Parameter ist es möglich, den Ethylester-Gehalt auf einen akzeptablen Wert zu reduzieren.

Das Ziel der statistischen Analyse bestand einerseits in der Identifikation der Ursachen für die Qualitätsabweichungen und andererseits in der Bildung von Modellen, die ermöglichen, Empfehlungen zu formulieren, wie die Variabilität reduziert und die Prozessqualität verbessert werden kann. Es standen Daten von insgesamt 275 Losen zur Verfügung. Neben dem Ethylester-Gehalt wurden verschiedene Prozessvariablen gemessen.

Im ersten Teil der Analyse wurden die Daten mit einigen ausgewählten grafischen Techniken untersucht (vgl. Abbildung 1). Diese explorative Datenanalyse hilft, die

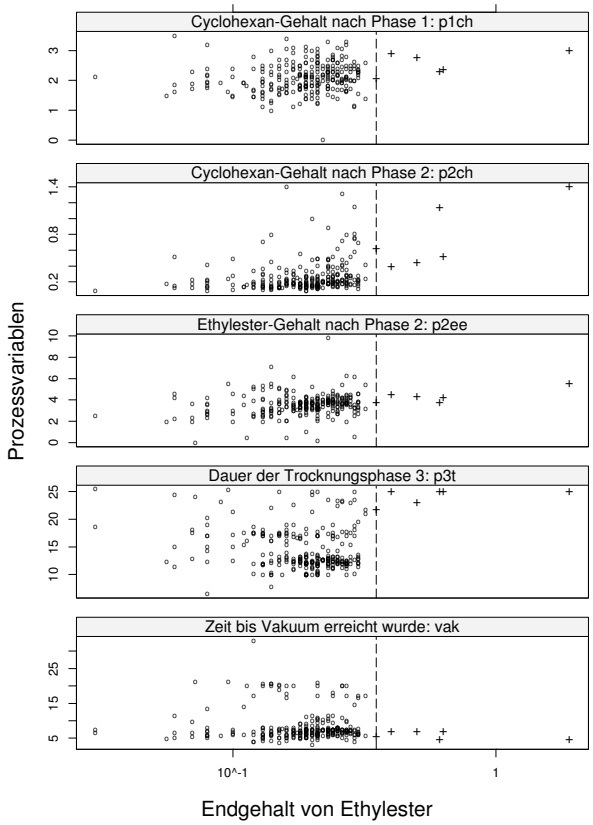

ABBILDUNG 1: *Streudiagramme des Endgehaltes von Ethylester gegen verschiedene Prozessvariablen. Die vertikale Referenzlinie gibt die obere Spezifikationsgrenze an.*

Prozessstörungen zu identifizieren und erleichtert die Interpretation der gefundenen statistischen Modelle.

Der angepasste Regressionsbaum für den Ethylester-Gehalt ist in Abbildung 2 gezeichnet. Er lässt sich wie folgt interpretieren: Bei jedem Knoten, dargestellt als Oval, wird für die Zielvariable eine Vorhersageregel aufgrund einer erklärenden Variable formuliert. Im Baum aus Abbildung 2 basiert die erste Regel auf dem Cyclohexan-Gehalt nach Trocknungsphase 2 (Variable p2ch). Für p2ch $>$ 1.35% wird ein hoher Endgehalt von 1% für Ethylester vorhergesagt. Im anderen Fall (p2ch $<$ 1.35%) beträgt der mittlere Ethylester-Gehalt nach der Trocknung hingegen nur 0.2%. Die Zahlenwerte in den Rechtecken, den Blättern des Baumes, sind die Vorhersagewerte für die Zielvariable. Falls also beispielsweise die Trocknungsphase 3 (Variable p3t) länger als 14.3 Stunden dauert und der Cyclohexan-Gehalt nach Phase 2 unter 0.17% liegt, führt das zu einem tiefen Ethylester-Gehalt von 0.1%.

Aus dem Baum geht hervor, dass hohe Werte des Cyclohexan-Gehaltes nach Phase 2 ebenfalls hohe Endwerte des Ethylester-Gehaltes zur Folge haben. Falls der gefundene Zusammenhang zwischen diesen Variablen kausaler Natur ist, muss deshalb die Reduktion des Cyclohexan-Gehaltes bereits in Phase 2 erfolgen.

Weitere Variablen im Modell sind der Cyclohexan-Gehalt nach Phase 1 (p1ch), der Ethylester-Gehalt nach Phase 2 (p2ee), die Zeit bis ein bestimmtes Vakuum erreicht wurde (vak) sowie die kategorielle Variable f1d, die angibt, ob die Überschreitung der Normtemperatur in Phase 1 kurz oder lange gedauert hat.

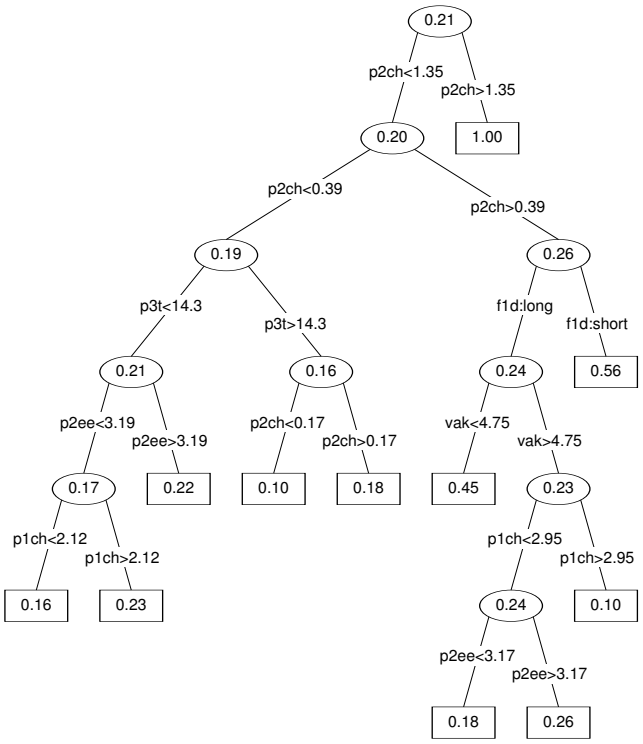

ABBILDUNG 2: *Angepasster Regressionsbaum für den Endgehalt von Ethylester.*

Verfeinerungen der Methodologie

Da die Grösse des Regressionsbaumes, d.h. die Anzahl der Blätter, bei der Modellbildung nicht limitiert ist, kann es vorkommen, dass ein zu kompliziertes Modell angepasst wird. Solche Modelle bergen die Gefahr, dass sie zwar die für die Modellschätzung verwendeten Daten gut beschreiben, aber für neue Beobachtungen schlechte Vorhersagen liefern. Die Grösse des Baumes muss deshalb sorgfältig gewählt werden. Dazu bieten sich die Techniken *cost-complexity pruning* und Kreuzvalidierung an, welche in S-PLUS einfach mit den Funktionen `prune.tree` und `cv.tree` durchgeführt werden können.

Bei der Kreuzvalidierung wird der Datensatz zuerst in ungefähr zehn gleich grosse Stücke geteilt, wobei man die ersten neun Teile als Trainingsdaten und den letzten Teil als Testdatensatz bezeichnet. Danach wird der Regressionsbaum aufgrund der Trainingsdaten in seiner maximalen Grösse angepasst und anschliessend jener zurückgestutzte Teilbaum gewählt (*pruning*), der für die Testdaten die beste Vorhersage liefert.

Zusammenfassung

Mithilfe eines Regressionsbaumes konnte der Zusammenhang zwischen Prozessparametern und dem Endgehalt von Ethylester geeignet beschrieben werden. Die Resultate liefern Hinweise, wie der Trocknungsprozess geändert werden muss, damit die Variabilität verkleinert und die Spezifikationen eingehalten werden.

Da offenbar zwischen einzelnen Prozessvariablen starke Wechselwirkungen auftreten, hat sich die Verwendung von Regressionsbäumen für die Modellierung als besonders geeignet erwiesen.