Prozessoptimierung durch Datenanalyse und -prognose mit R in der Praxis

Simon Hofinger
Robotron Datenbank-Software GmbH
Berlin

Schlüsselworte
Business Intelligence, Datenanalyse, Prozessdatenanalyse, Prozessoptimierung, R, ROracle

Einleitung
Dieser Vortrag erläutert und demonstriert anhand eines Beispiels aus der Automobilzulieferindustrie, wie mithilfe der Analyse von Prozessdaten eine Prozessoptimierung in der Herstellung von Sensoren erreicht werden konnte. Für die Datenanalyse wurde die Open-Source-Statistiksoftware „R“ und „ROracle“ eingesetzt.

Ausgangslage
Das Ziel des Projektes war es die Anzahl an Messpunkten zu reduzieren die dazu notwendig sind, die hergestellten Sensoren zu kalibrieren und damit voll funktionstüchtig zu machen. Bei einem Messpunkt handelt es sich hierbei um eine vorgegebene Position der Kalibrierungseinrichtung. Diese Positionen werden für jeden Sensor durchlaufen und dabei jeweils die dortigen Messwerte aufgezeichnet. Insgesamt gab es über fünfhundert verschiedene Messpunkte an denen jeweils Daten zum dortigen magnetischen Fluss erhoben wurden. Diese wurden anschließend an eine Software übergeben, welche mit Hilfe eines komplizierten Rechenverfahrens die Werte zur korrekten Programmierung der Mikrochips auf den Sensoren ermittelte.

Vorgehen
Der Weg um eine Reduzierung der Messpunkte zu erreichen verlief über die Analyse historischer Messdaten und den daraus errechneten Werten zur Programmierung der Mikrochips. Eine vorangestellte Analyse der Daten ergab, dass man die Anzahl der betrachteten Werte, welche für jeden Sensor variabel und folglich von den Messungen abhängig waren, auf 10 beschränken konnte. Das konkrete Ziel war damit, diese zehn Werte auf Basis der Messdaten mit Hilfe von Data Mining prognostizieren zu können. Vom Unternehmen wurden dafür Datensätze zu mehreren zehntausend bereits produzierten und kalibrierten Sensoren zur Verfügung gestellt. Diese enthielten jeweils die Messungen an allen Messpunkten und einige aus der Produktion der Sensoren stammenden Werte. Insgesamt waren somit über 50 Millionen Messung vorhanden. Über die Eingabe aller Daten in die bisher verwendete Software wurden die Sollwerte für die zehn Zielvariablen pro Sensor ermittelt. Letztlich sollten für die Prognosen dieser Werte die Daten aus möglichst wenigen Messpunkten ausreichen.

Zur Bewältigung dieser Aufgabe wurde in der Modellierungsphase des Data-Mining-Prozesses anhand der statistischen Programmiersprache R ein Algorithmus implementiert, welcher diese Daten selbstständig auswertet. Er soll zuerst den aussagekräftigsten Messpunkt zur Prognose der Zielwerte bestimmen und anschließend den Messpunkt finden der, in Kombination mit dem ersten Messpunkt, nun am meisten Aussagekraft mit sich bringt. So sollen iterativ Messpunkte hinzugefügt werden, bis mit Hilfe der daraus stammenden Messungen die Zielwerte ausreichend gut prognostiziert werden können. Im Folgenden wird die Funktionsweise des Algorithmus im Detail erläutert.

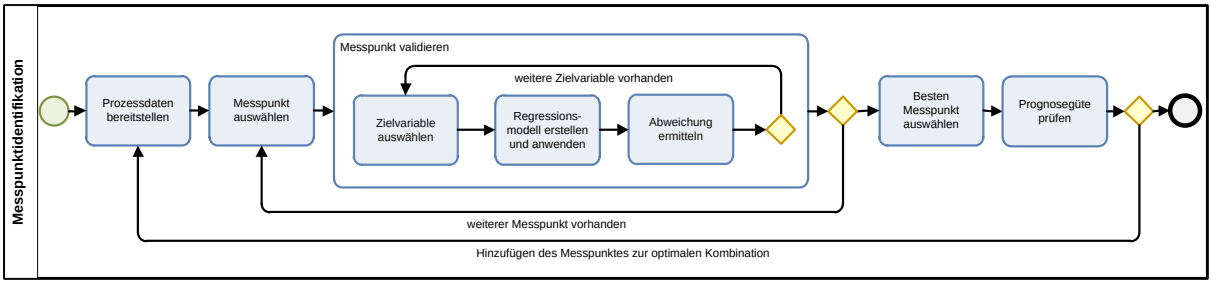

Abb. 1: Prozessdiagramm zur Messpunktidentifikation

Zuerst erhält der Evaluierungsalgorithmus Zugriff auf die vorbereiteten Prozessdaten. Dazu wird das ROracle-Paket geladen und eine Verbindung zur Datenbank hergestellt. Nach einer Synchronisierung können die Tabellen referenziert oder, wie im folgenden Beispiel, in R kopiert werden.

```
library(ORE)

ore.connect(user="user1",sid="sid1",host="localhost",
password="***", port=1521)
ore.sync()

messpunkte_ore <- ore.get("MESSPUNKTE")
messungen_ore <- ore.get("MESSUNGEN")
zielwerte_ore <- ore.get("ZIELWERTE")
```

Daraufhin wird ein Messpunkt mitsamt den daraus stammenden Messwerten zur Evaluierung ausgewählt. Da insgesamt zehn Zielwerte zu prognostizieren sind, wird die Analyse vorerst auf den Ersten eingeschränkt. Im nächsten Schritt wird auf Grundlage der Messungen und der zugehörigen Ausprägungen des gewählten Zielwertes ein Regressionsmodell geschätzt. Dieses Modell hat das Ziel anhand der hier übergebenen Messungen die gewählten Zielwerte möglichst genau prognostizieren zu können. Der folgende R-Codeausschnitt verdeutlicht die Verschachtelung der verschiedenen Schleifen.

```
for(durchlauf in 1:anzahl_durchlaeufe){

 for(messpunkt in 1:anzahl_messpunkte){

  for(zielwert in 1:zielwerte){

   for(i in 1:10){

     glm_modell <- glm(formel, family=gaussian(link="identity"),
                                          data=trainingsdaten)
```

Um die vorhandenen Daten optimal zu nutzen und die Ergebnisse zuverlässig validieren zu können wird dieser Vorgang im Rahmen einer 10-fachen Kreuzvalidierung zehnmal wiederholt. Dabei werden jeweils 90% der Daten als Trainingsdaten dafür verwendet das Modell zu schätzen. Auf die restlichen 10% der Daten, welche sich bei keinem der zehn Durchgänge überschneiden, wird das Modell angewendet und die dabei prognostizierten Zielwerte mit den Sollwerten verglichen.

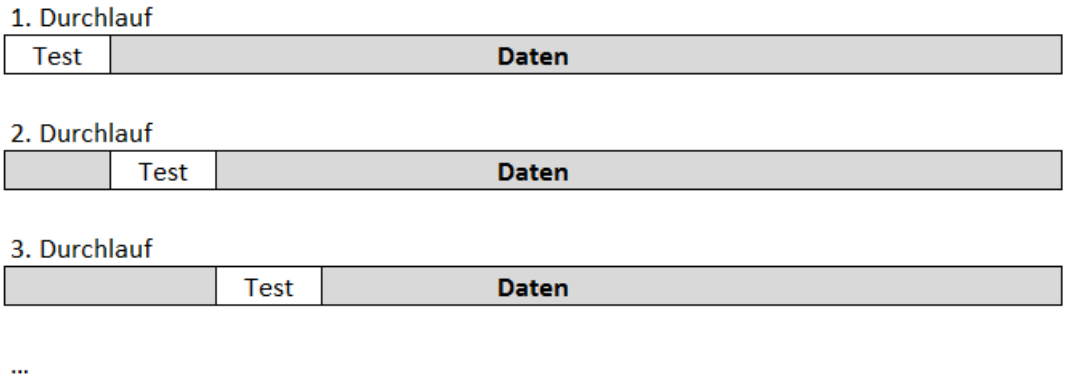

Abb. 2: 10-fache Kreuzvalidierung

Über dieses Validierungsverfahren werden die Abweichungen der Prognosen von den Sollwerten und damit anschließend die Güte des Modells berechnet. Im Detail werden hierbei die Quadratfehlersummen durch die Standardabweichung und die Anzahl der Sensoren geteilt. Danach werden alle übrigen Zielwerte auf die gleiche Art und Weise behandelt und abschließend daraus mit folgender Formel ein für den Messpunkt aggregiertes Gütemaß berechnet.

$$Gütemaß = \sum_{i=1}^{10} \left(\frac{Fehlerquadratsummen_i}{n * Standardabweichung_i}\right)^2$$

mit i = Zielvariable, n = Anzahl der Sensoren

Abb. 3: Formel zum Gütemaß

An dieser Stelle ist der aktuelle Messpunkt vollständig evaluiert und es kann zum Nächsten übergegangen werden. Nachdem für alle Messpunkte die aggregierten Gütemaße berechnet wurden kann aus Ihnen derjenige mit dem niedrigsten Gütemaß und somit der Beste bestimmt werden. Dieser Messpunkt gehört nun zur optimalen Kombination von Messpunkten und die daraus stammenden Messungen werden bei den folgenden Durchläufen des Prozesses jedem zu evaluierenden Messpunkt hinzugefügt. Im Rahmen dieser Forward-Selection wird nun nacheinander der jeweils beste Messpunkt gewählt bis das aggregierte Gütemaß die gewünschte Prognosegüte erreicht hat. Während

anfangs der Informationsbeitrag jedes neuen Messpunkts die Prognosequalität stark verbessert ist schon beim fünften Messpunkt kaum noch eine Verbesserung zu beobachten. Das Ergebnis des Analyseverfahrens ist folglich die zuletzt verwendete optimale Kombination von Messpunkten.

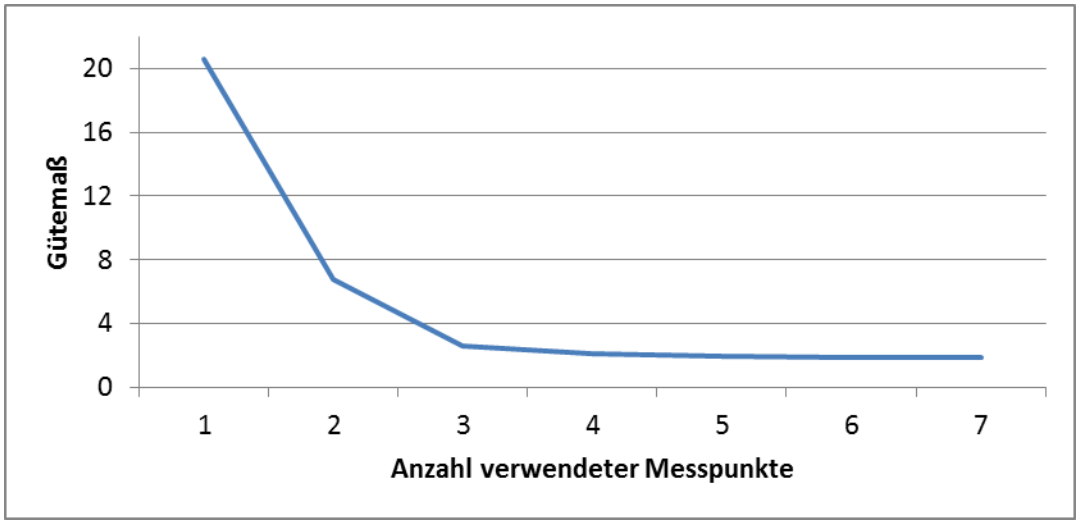

Abb. 4: Verbesserung des Gütemaßes

Zur Modellbildung wurden unter anderem folgende statistische Verfahren getestet: Generalisierte Lineare Modelle, Neuronale Netze, Support Vector Machine und Random Forest. In diesem Projekt haben sich, je nach Zielvariable, verschiedene Variationen und Kombinationen der generalisierten linearen Modelle und Random Forest als optimal erwiesen. Einen wichtigen Beitrag hierbei lieferte die Erzeugung von zusätzlichen Merkmalen durch die Quadrierung der Messwerte und das Bilden von Verhältnissen der Messwerte zueinander. Um ein Overfitting durch die große Anzahl an erzeugten Merkmalen zu verhindern wurden anschließend durch eine Variablenselektion diejenigen ohne Einfluss herausgefiltert.

Um die Anfälligkeit der Prognosemodelle für Fehler durch Änderungen im Produktionsprozess zu reduzieren, wurden die Daten von drei verschiedenen Losen verwendet. Die Modelle wurden mit den beiden älteren Losen erstellt und anschließend mit den beiden neueren evaluiert. Trotz dieser Erschwerung der Aufgabe konnten mit nur sieben der ursprünglich über fünfhundert Messpunkte in 98,7% der Fälle die Sensoren genauso programmiert werden wie es das bisher verwendete Computerprogramm vorgab. Nach abgeschlossener Evaluierung konnte ein für den praktischen Einsatz bestimmtes Prognosemodell erstellt werden, welches alle verfügbaren Daten zum Training verwendete.

Ergebnis

Der Datenanalyse-Prozess bringt schlussendlich zwei evaluierte Ergebnisse hervor. Einerseits konnte ein Erklärungsmodell für den Prozessschritt der Sensorkalibrierung unter Verwendung von nur sieben der einst 500 Messpunkte erzeugt werden. Andererseits entstand ein vollständig anwendbares Prognosemodell zur direkten Integration in die Prozesssteuerung. Das erzeugte Erklärungsmodell

beinhaltet alle relevanten Einflussgrößen (sieben Messpunkte) und deren Wirkungsbeziehungen (Prognosemodell) zur qualitätsgerechten Bestimmung der zehn Zielvariablen.

Kontaktadresse:
Simon Hofinger
Robotron Datenbank-Software GmbH
Albert-Einstein-Straße 16
D-12489 Berlin

Telefon: +49 (30) 26-39292 437
Fax: +49 (30) 26-39292 955
E-Mail simon.hofinger@robotron.de
Internet: www.robotron.de