Prozessentwicklung in der KI-Cloud

Datengetriebene Modellierung in der Bioprozesstechnik

Jonathan Sturm[1] und Frank Eiden[1]

Warum nicht die Prozessentwicklung in einem kollaborativen Netzwerk durchführen? Warum nicht PAT bereits in der Prozessentwicklung nutzen? Warum nicht in Echtzeit aus Prozessdaten Informationen generieren?

Zur Bewältigung dieser komplexen Fragenstellungen im Kontext biotechnologischer Prozesse müssen drei Voraussetzungen erfüllt werden:

- 1. Die Zusammenführung von fragmentierten In- und Outputs (Mess- und Regelgrößen) von Sensoren und Aktoren aus unterschiedlichen Plattformen in ein organisiertes, durchsuchbares Datenformat.
- 2. Die Vernetzung aller Prozessdaten (auch von unterschiedlichen Standorten) in einer kollaborativen Cloud.
- 3. Die Anwendung von Methoden der künstlichen Intelligenz (insbes. des maschinellen Lernens), der multivariaten und statistischen Datenanalyse auf die zusammengeführten Daten.

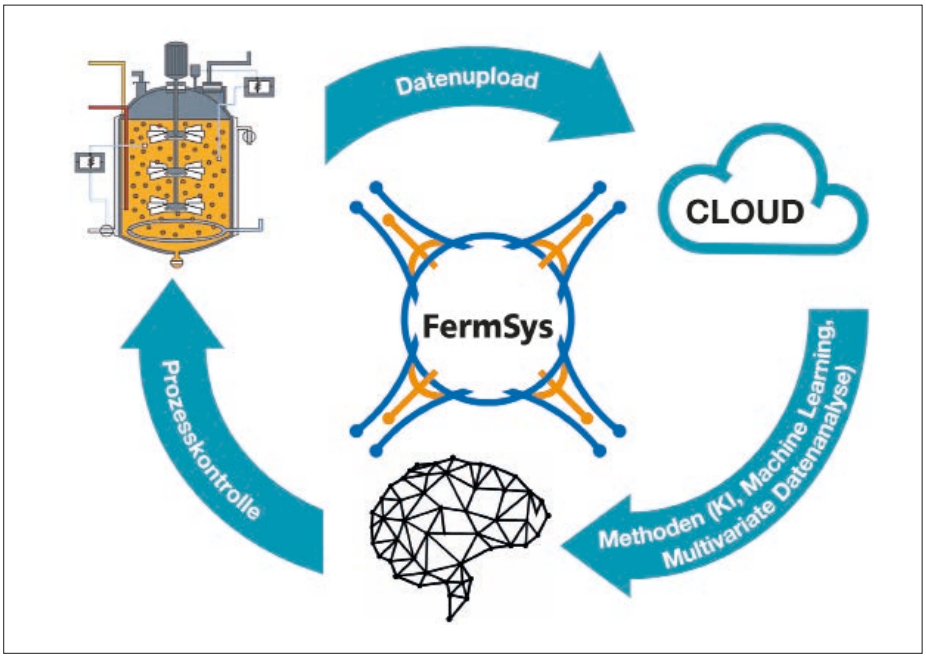

Abb. 1: Die cloudbasierte Plattform Fermsys ermöglicht durch Zusammenführung von Prozessdaten aus unterschiedlichen Quellen eine einfache Anwendung von KI-Methoden.

Immer mehr und immer komplexere Daten müssen in immer kürzerer Zeit verarbeitet werden. Smarte Sensoren und Verfahren generieren in Wissenschaft und angewandter Forschung kaum bewältigbare Datenmengen. Dadurch steigen auch die Anforderungen für Datenmanagement, -Visualisierung und -Analyse, die zukünftig nur auf Cloud-basierten Plattformen skalierbar erfüllt werden können. Um diese Herausforderung zu bewältigen und relevante Ergebnisse aus vernetzten wissenschaftlichen und industriellen Forschungsverbünden schnell wertvolles Prozessverständnis zu generieren, dürfen Systeme nicht fragmentiert arbeiten. Alle Phasen der Forschungs- und Entwicklungsabläufe müssen in einem Datenerfassungs- und -Analysesystem integriert werden und ein vernetztes, digitales Ökosystem darstellen. Um den Prozess der Datenerfassung und -Analyse zu gewährleisten, sind die unterschiedlichen Datensätze in ein organisiertes, einheitliches Format umzuwandeln, um sie dann in Ihrer Gesamtheit den entsprechenden Analysen zuzuführen. Die so zu einem Datenformat zusammengeführten Daten erlauben es anschließend KI-Programmen, wie maschinellen Lernmethoden, komplexe unstrukturierte biologische Daten zu analysieren und ermöglichen, dass wichtige Forschungsfragen schneller bearbeitet werden können als durch herkömmliche Ansätze. Auf dieser gemeinschaftlichen Grundlage können spezifische Auswerte- und Visualisierungsapps den jeweiligen besonderen Herausforderungen des Projektes angepasst und einer Community zur Verfügung gestellt werden. Darüber hinaus sind Cloud-basierte Plattformen für kollaborative Arbeitsansätze ideal geeignet und ermöglichen einen einfachen Datenaustausch zwischen verschiedenen Projektpartnern auch an unterschiedlichen Standorten und können Forschungs- und Entwicklungsprozesse effizient beschleunigen [1,2].

Die Kombination aus Zukunftstechnologien mit Cloud-basierten digitalen Plattformen stellt somit ein großes Potenzial für die Bewältigung komplexer und Datenintensiver F&E-Vorhaben dar [3].

Zur Realisierung dieser Forderungen wurde eine entsprechende Cloud-basierte Plattform (Fermsys) mit Tools zu Datenstandardisierung und -analyse entwickelt und in einem Projektverbund eingesetzt (Abb.1) [12].

Im Folgenden soll insbesondere die Anwendung von Methoden aus dem Bereich des maschinellen Lernens aufgezeigt werden.

Maschinelles Lernen

Das typische Vorgehen bei der Modellbildung mittels maschinellem Lernen (ML) besteht aus folgenden fünf Schritten (Abb. 2): Daten importieren, aufbereiten, ein Modell trainieren, dieses validieren und schließlich implementieren – fertig ist die „künstliche Intelligenz“. In der praktischen Umsetzung trifft man dabei allerdings auf diverse Herausforderungen. Allein schon die Auswahl des besten Trainings-Algorithmus ist nicht trivial, da von diesen bereits ein ganzer Baukasten zur Verfügung steht, welcher stetig erweitert wird. Und zu allem Überfluss gibt es bis dato nicht einmal eine Möglichkeit, um im Vorfeld zu bestimmen, welche Methode sich am besten für welches Problem eignet. Daher ist eine gründliche Validierung vor dem tatsächlichen Einsatz eines Modells nach wie vor unerlässlich. Kommt ein Modell nach gründlichen Tests dann im realen Prozess in den Einsatz, sollte es auch

im Nachhinein weiter überwacht und optimiert werden, um das Potenzial der künstlichen Intelligenz vollständig ausschöpfen zu können.

Der große Vorteil von ML-Methoden besteht darin, dass – im Gegensatz zur „klassischen" Modellierung - kein Vorwissen über einen Vorgang mehr benötigt wird, um ihn zu beschreiben. Es ist nur noch erforderlich, dass genügend Informationen in Form von Beobachtungen, also z.B. Messdaten, verfügbar sind. Ob es sich dabei um Daten von Abgas-, Temperatur- oder sonstigen anderen Sensoren handelt ist im Grunde unerheblich, solange überhaupt ein Zusammenhang zwischen Messwert und Zielgröße vorliegt. Ist diese Bedingung erfüllt, sind die Methoden der datengetriebenen Modellierung dazu in der Lage, beliebig komplexe Funktionen zu approximieren [4].

Doch wozu kann ein so trainiertes Modell überhaupt nützlich sein? Zum einen führen Prozessmodelle zu besserem Verständnis über einen Vorgang, und auf der anderen Seite lassen sich dank ihnen sogar Abläufe automatisieren. Aufgrund dieser Entwicklung gibt es auch für die Modellierung biologischer Systeme moderne Ansätze, die perspektivisch helfen könnten, die Anforderungen der PAT-Initiative zu erfüllen. So können sog. Softsensoren geschaffen werden, also eine Mischung aus Software und Sensor (Abb. 3). Diese setzen beliebig viele Eingabegrößen miteinander in Beziehung und errechnen so eine Zielgröße [5]. Der berechnete Wert kann dann z.B. bei der Prozesskontrolle helfen.

Beispiel: Modellierung der Biomasse

Ein Beispiel für so einen Softsensor soll im Folgenden dargestellt werden. Der Softsensor soll für die Modellierung der Zielgröße Biomasse genutzt werden, damit diese in Echtzeit zur Prozesskontrolle genutzt werden kann. Als Datengrundlage dienen Messwerte aus einer Reihe von Fed-Batch-Fermentationen mit dem Organismus *Hansenula polymorpha*.

Zunächst wurden die verfügbaren Daten gesichtet und vorverarbeitet. Dieser auch Preprocessing genannte Schritt umfasst unter anderem die Auswahl geeigneter Eingabewerte, die Entfernung von Ausreißern oder Normierung der Daten. Weiterhin wurden die Daten in ein Test- und Validierungsset aufgeteilt (85% und 15% der Daten). In unserem Beispiel wurden die Parameter Zeit, Rührergeschwindigkeit, pH, dO, sowie Kontrollsignale der Säure-, Base und Fütterungspumpe als Inputs ausgewählt.

Danach kommt es zum eigentlichen Training des Systems. In diesem Schritt „lernt"

Abb. 2: Workflow der datengetriebenen Modellierung.

Tab. 1: Liste der getesteten ML-Algorithmen

Name	Methode
Modell 1	Multiple lineare Regression (MLR) [6]
Modell 2	Support vector machine (SVM) fine gaussian [7]
Modell 3	Gauß-Prozess Regression (GPR) – exponential [8]
Modell 4	Multiple lineare Regression (MLR) - robust
Modell 5	Decision Tree (Entscheidungsbaum) – fine [9]
Modell 6	Decision Tree (Entscheidungsbaum) - coarse
Modell 7	Support vector machine (SVM) - quadratic
Modell 8	Support vector machine (SVM) – cubic
Modell 9	Support vector machine (SVM) – coarse gaussian

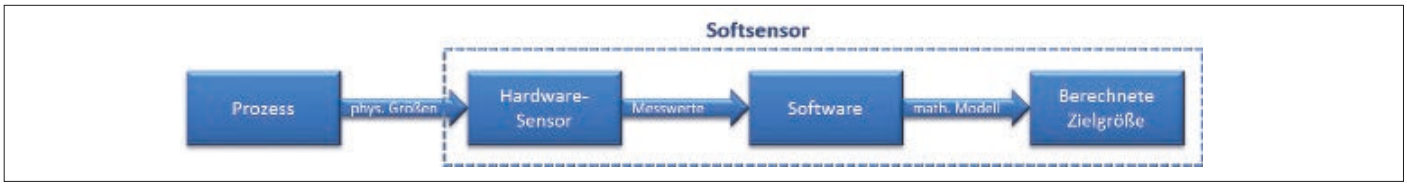

Abb. 3: Funktionsweise von Softsensoren (gestrichelte Linien).

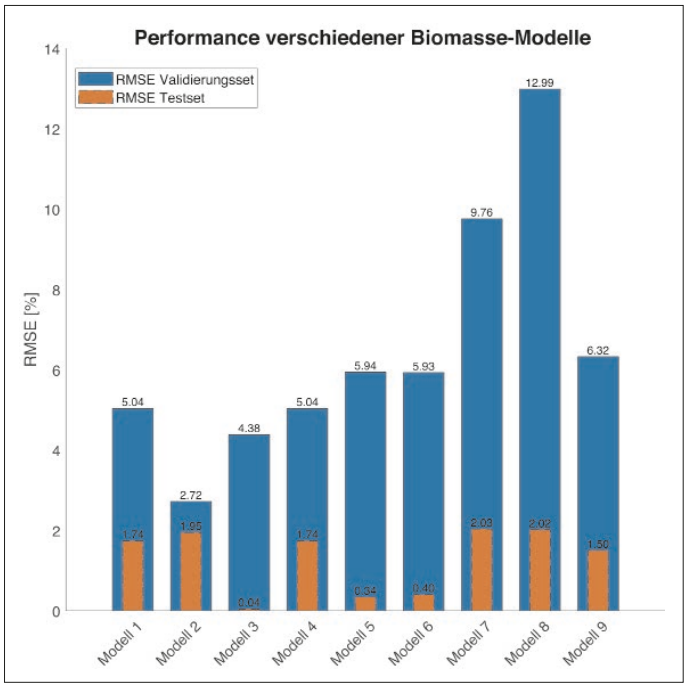

Abb. 4: Ergebnis von Training und Validierung der 9 getesteten Modelle. Blau: Fehler (RMSE) auf dem Validierungsset; Orange: Fehler auf dem Testset.

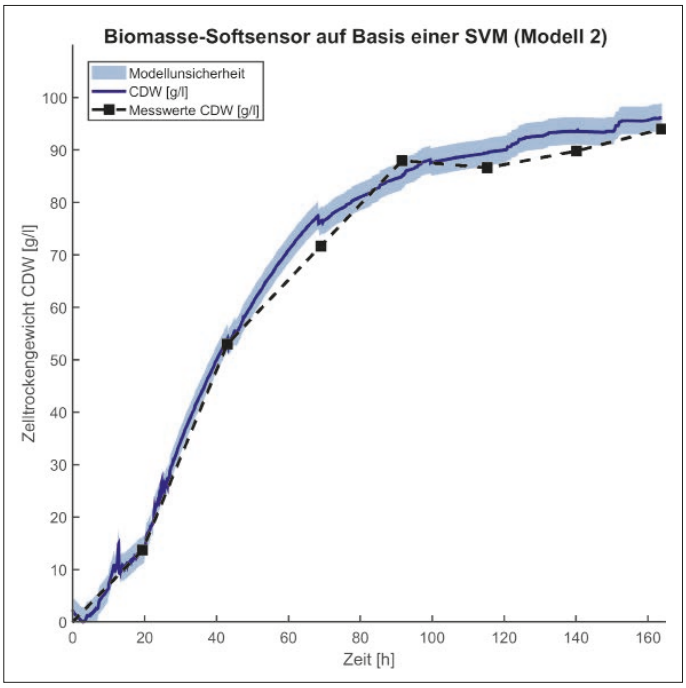

Abb. 5: Ergebnis des fertigen Softsensors. Dunkelblau ist die modellierte Biomasse dargestellt, hellblau hinterlegt der Unsicherheitsbereich. Schwarz dargestellt ist die tatsächlich gemessene Biomasse.

der gewählte Trainingsalgorithmus die komplexen mathematischen Beziehungen zwischen Ein- und Ausgabewerten und minimiert iterativ den Fehler auf dem Testset.

Ergebnis des Trainings ist ein fertiges Modell. Dieses wird nun mittels des Validierungssets überprüft, um ein sog. Overfitting zu vermeiden. In diesem Schritt überprüft man also, ob das Modell nur auf den Trainingsdaten funktioniert oder allgemein verwendbar ist. In dem hier aufgeführten Beispiel wurden mehrere Trainingsalgorithmen getestet, bis ein Optimum gefunden war (Tab.1).

Von den 9 getesteten Modellen wies Modell 2 den kleinsten RMSE (root mean squared error) auf dem Validierungsset auf (Abb. 4). Auch der Fehler auf dem Testset war hinreichend klein. Somit konnte das Modell im realen Prozess implementiert werden.

Abbildung 5 zeigt die modellierte und die tatsächlich gemessene Biomasse einer der Fermentationen. Mit einem Fehler von 2,7 % ist der Softsensor bereits geeignet, um für die Prozesskontrolle genutzt zu werden. Durch Einspeisen weiterer Versuchsdaten kann die Präzision zukünftig sogar noch gesteigert werden.

Fazit

Bereits mit einfachen Methoden aus dem Bereich der KI können wertvolle Informationen über Prozesse gewonnen werden, wodurch Prozessteuerung und -Verständnis verbessert werden. Cloud-basierte Technologien bilden hierfür einen skalierbaren Grundstein.

Bei geeignetem Modelldesign könnte prinzipiell sogar eine in silico Optimierung des Prozesses erreicht werden. Die virtuelle Manipulation von Prozessparametern, bevor ein Prozess überhaupt gestartet wird, führt so zu optimaler Produktivität. Führt man diesen Gedanken konsequent zu Ende, könnte die datengetriebene Modellierung zukünftig tatsächlich der Entwicklung zu selbstlernenden smart Factories den Weg bereiten.

Zugehörigkeiten
[1]AG Bioprozesstechnik, Westfälische Hochschule, Recklinghausen, Deutschland

KONTAKT |

Prof. Dr.-Ing. Frank Eiden
AG Bioprozesstechnik
Westfälische Hochschule
Recklinghausen, Deutschland
frank.eiden@w-hs.de

Mehr zu Bioprozesstechnik: http://bit.ly/GIT-BPT

Mehr zu KI: http://bit.ly/GIT-KI

[1] Literatur: http://bit.ly/GIT-Eiden2