Computational Intelligence und Data Mining

Datenanalyse und Prozessoptimierung am Beispiel Kläranlagen

Prof. Dr. Thomas Bartz-Beielstein
thomas.bartz-beielstein@fh-koeln.de, Tel. 02261/8196-6391

Prof. Dr. Wolfgang Konen
wolfgang.konen@fh-koeln.de, Tel. 02261/8196-6275

Prof. Dr. Hartmut Westenberger
hartmut.westenberger@fh-koeln.de, Tel.02261/8196-6385

Zielsetzung

Die zunehmende Vernetzung industrieller und wirtschaftlicher Anlagen sowie vermehrt auftretende automatische Datenerhebungen ergeben die Möglichkeiten, aber auch die Last, immer detailliertere Datenmengen zu analysieren. Dies geschieht oftmals vor dem Hintergrund, Prozesse optimal zu steuern oder Prognosen über den zukünftigen Verlauf anzustellen.

Ziel des Projektes ist es, sowohl aktuelle bis aktuellste Methoden zur Modellierung, Simulation und Optimierung komplexer Prozesse einzusetzen. Hierzu werden praxisbewährte Methoden der Computational Intelligence (CI) und des Data Mining am Institut für Informatik der FH Köln gebündelt zum Einsatz gebracht. In Kooperationsprojekten mit Partnern aus Industrie und Wirtschaft werden die Methoden auf Einsetzbarkeit und Leistungsfähigkeit geprüft. Der Einsatz in diesen konkreten Anwendungsfällen ermöglicht es, die Reichweite und die Grenzen verschiedener, oftmals komplexer CI und Data Mining Methoden auch für Praktiker aus Industrie und Wirtschaft gut fassbar darzustellen. Das Institut für Informatik der FH Köln unterstützt Unternehmen beim Einsatz dieser Methoden.

Zum gegenwärtigen Themen- und Methodenspektrum gehören

- Evolutionäre Algorithmen
- Neuronale Netze
- Fuzzy Logic
- Entscheidungsbäume und Random Forests
- Modellierung dynamischer Systeme, Echo State Networks
- Experimentelles Design, Versuchsplanung
- Standardverfahren des Data Minings
- Hybride Kombinationen der o.g. Methoden

Inwieweit diese Verfahren sinnvoll kombiniert werden können, soll anhand eines aktuellen Forschungsprojektes im Folgenden skizziert werden. Weitere aktuelle Forschungsprojekte, die mit diesen Methoden bearbeitet werden, beschreiben

- wie durch eine systematische Auswertung der Daten des Hochschulinformationssystems der FH Köln Anhaltspunkte für ein wirksames Hochschulmarketing gewonnen werden können [3] und
- wie Portfoliooptimierungen unter Berücksichtigung von Nebenbedingen effizient durchgeführt werden können [4].

Anwendungsbeispiel: Modellierung Regenüberlaufbecken

Im Rahmen einer Kooperation mit unserem Kollegen Prof. Michael Bongards zum NRW-geförderten Forschungsprojekt KANNST (**KAN**al**N**etz-**St**euerung) [1] haben wir als ein erstes Beispiel untersucht, ob es möglich ist, die Füllstandshöhe in einem Regenüberlaufbecken des Aggerverbandes auf Basis einzelner Regenmessungen zu modellieren und damit schließlich im Kurzzeitbereich auch zu prognostizieren. Hierbei besteht die Schwierigkeit darin, dass der den Regen aufnehmende Boden mit seinen vielen verschiedenen Schichten und der großen Menge an unterschiedlichen Zuständen ein komplexes dynamisches System darstellt. Es macht einen großen und quantitativ schwer zu erfassenden Unterschied, ob eine bestimmte Menge Regen auf einen trockenen oder bereits mit Nässe gesättigten Boden fällt.

Wir versuchten daher eine Modellierung, die von empirisch erhobenen Daten getrieben wird. Als solche dienten minütliche Messdaten aus einem 108 Tage umfassenden Zeitraum, insgesamt über 150.000 Datensätze.

Ein erster Ansatz startete mit ESN (Echo State Network) [2], die prinzipiell in der Lage sind, in einer Zeitreihe mit komplexer Dynamik die richtigen Zusammenhänge zu finden. Bei dieser Aufgabe aus der Praxis stellte sich aber heraus, dass sie nicht einsetzbar waren, da die Daten, gekennzeichnet von plötzlichen Regenphasen und langen Pausen, zu unstetig und evtl. zu verrauscht waren. Dies heißt nicht, dass ESN prinzipiell für die Aufgabe ungeeignet sind, sondern nur, dass die ungefilterten Input-Daten erstmal ein Problem für ESN darstellen. Wir planen, nach geeigneter Aufbereitung von Daten durchaus noch einmal auf ESN zurückzukommen.

Ein weiterer Versuch der Modellierung mit ODEs (gewöhnlichen Differentialgleichungen) brachte Teillösungen, war aber ebenfalls von starken numerischen Schwierigkeiten gekennzeichnet, die in der unstetigen, "burst-artigen" Struktur der Daten begründet liegt.

Als wesentlich besser geeignet erwies sich eine Modellierung mit Integralgleichungen, welche naturgemäß mit unstetigen Inputdaten besser umgehen können. Eine Möglichkeit, die Füllhöhe y(t) als Funktion des aktuellen Regens r(t) und des Bodenzustandes K(t), der wiederum eine Funktion des vorher gefallenen Regens ist, zu beschreiben, ist die Integralgleichung

$$y(t) = \int_{-\infty}^{t} r(\tau - \tau_{rain}) g(t-\tau) d\tau + \int_{-\infty}^{t} K(\tau - \tau_{rain}) h(t-\tau) d\tau$$

Hierin sind $g(\tau)$ und $h(\tau)$ aus den empirischen Daten zu modellierende Integralkerne, die die Ablaufcharakteristika beschreiben.

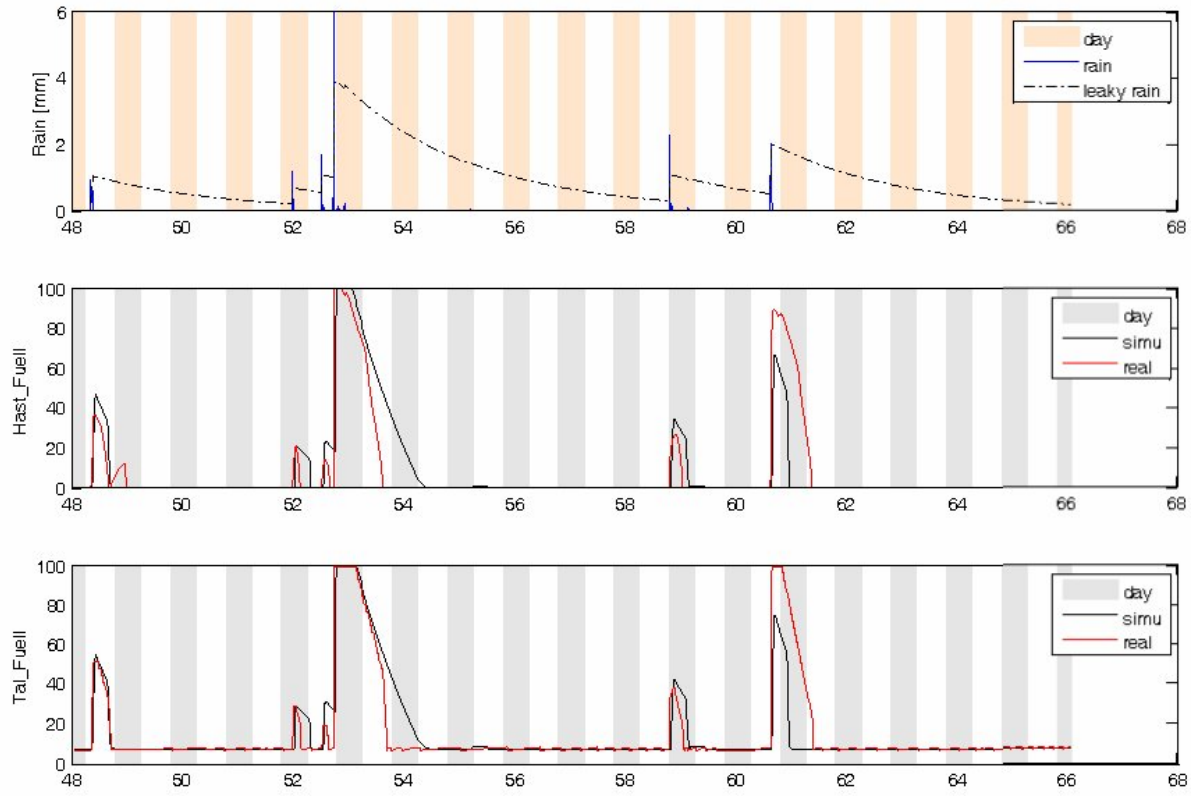

Abbildung 1: Regenmenge (oben, blau), Füllstand Becken 1 (Mitte), Becken 2 (unten) für den Zeitraum Tag 48 bis Tag 66. Die Balken kennzeichnen die Nachtphasen.

Im Ergebnis erhalten wir in Abbildung 1 eine schon recht gute Modellierung der Füllhöhe. Allerdings betrachten wir hier erst einen kleinen Ausschnitt der Daten. Ein wichtiges Ziel ist die Entwicklung eines für den gesamten Jahresverlauf gültigen Modells. Hierzu ist geplant, auch weitere meteorologische Kenngrößen wie Temperatur, Sonnenscheindauer einzubeziehen und ferner eine methodische Parameteroptimierung, z. B. mit EA (Evolutionären Algorithmen) vorzunehmen.

Das Ergebnis unserer Simulationen und den darauf basierenden Vorhersagen hängt von einer Vielzahl unterschiedlicher Einflussgrößen (Parameter) ab. Grob können diese in problemspezifische Parameter (z. B. Bodenbeschaffenheit) und algorithmenspezifische Parameter (z. B. Art des zur Simulation herangezogenen neuronalen Netzes) unterschieden werden [5]. Mittels Methoden der statistischen Versuchsplanung sollen in den weiterführenden Untersuchungen modellrelevante Einflussgrößen bestimmt werden.

Zur Untersuchung der problemspezifischen Parameter wurde bereits ein vereinfachtes Simulationsmodell basierend auf dem EPA Storm Water Management Model (SWMM) entwickelt. SWMM wird seit 1971 für unterschiedliche Simulationen benutzt und stellt ein verlässliches Werkzeug zur Generierung von Simulationsdaten dar. Wir sind somit in der Lage, z. B. den Einfluss unterschiedlicher Bodenbeschaffenheiten auf den Ausgang der Simulation zu untersuchen und wichtige Einflussgrößen zu bestimmen.

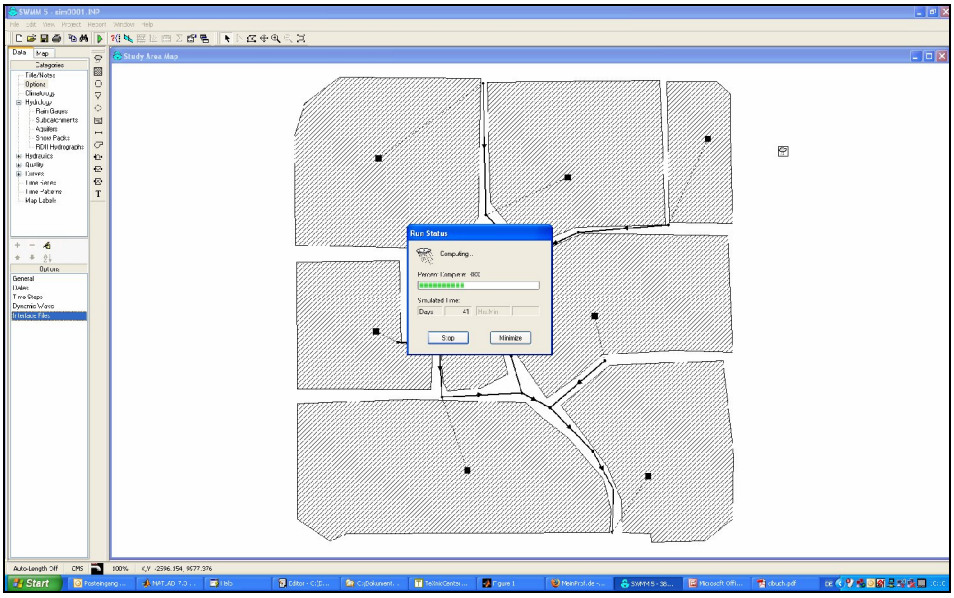

Abbildung 2: Kanalnetzoptimierung mit SWMM, einem von der amerikanischen EPA (Environmental Protection Agency) entwickelten Simulator.

Zudem ermöglichen die Verfahren der statistischen Versuchsplanung, relevante Parameter der einzelnen Algorithmen, die zur Modellierung herangezogen werden, zu bestimmen. So können neuronale Netze entwickelt werden, die optimal für die Prognose der Füllstandshöhe der Regenüberlaufbecken geeignet sind.

Der Anwender erhält somit statistisch abgesicherte Informationen, dass das empfohlene Prognosemodell besser geeignet ist als die anderen betrachteten Modelle und zusätzlich die Sicherheit, dass diese Aussagen nicht von zufälligen Messdaten abhängig sind, sondern dass diese Ergebnisse für ein großes Spektrum an unterschiedlichen problemspezifischen Parametern (wie z.B. die Bodenbeschaffenheit, Niederschlagsmengen) Gültigkeit behält.

Danksagung: Für viele interessante Diskussionen und die Bereitstellung anwendungsbezogener Daten danken wir unserem Kollegen Prof. Dr. Michael Bongards und seinem Team, besonders Dipl.-Ing. Tanja Hilmer und Dipl.-Ing. Andreas Stockmann.

Literatur

[1] M. Bongards. Online-Konzentrationsmessung in Kanalnetzen - Technik und Betriebsergebnisse, Forschungsbericht FH Köln, 2007.

[2] H. Jaeger, H. Haas: *Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication.* Science, April 2, S. 78-80, 2004.http://www.faculty.iu-bremen.de/hjaeger/pubs/ESNScience04.pdf

[3] H. Westenberger, W. Konen, T. Bartz-Beielstein, Business Intelligence an Hochschulen, Forschungsbericht FH Köln, 2007.

[4] T. Bartz-Beielstein, W. Konen, H. Westenberger. Portfoliooptimierung unter Nebenbedingungen. Forschungsbericht FH Köln, 2007.

[5] T. Bartz-Beielstein, W. Konen, H. Westenberger. Moderne statistische Verfahren zur experimentellen Versuchsplanung. Forschungsbericht FH Köln, 2007.