OSVJEŽIMO ZNANJE
Uređuje: Nenad Bolf

Ž. Kurtanjek *
Sveučilište u Zagrebu
Prehrambeno-biotehnološki fakultet
Pierrotijeva 6, 10 000 Zagreb

Kad zaključivanje matematičkim modelom može biti pogrešno: Primjer protočnog kemijskog reaktora PKR

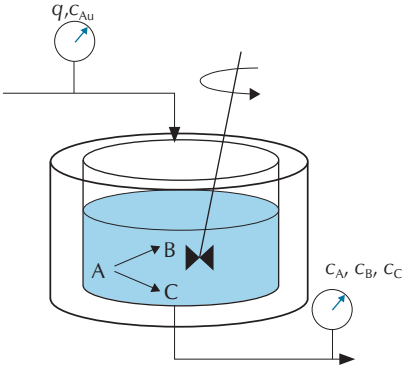
Cilj ovog priloga je pokazati kako strojno učenje i agnostički (empirijski) modeli ponekad mogu dovesti do pogrešnih zaključaka koji su protivni osnovnim načelima kemijskog inženjerstva. Prikazan je jednostavan primjer modeliranja zavisnosti koncentracije produkta o protoku u izotermom protočnom kemijskom reaktoru (PKR). Svrha modela je matematičkim modelom analizirati regulaciju koncentracije produkta manipulacijom protoka. U reaktoru se odvijaju dvije usporedne reakcije pretpostavljeno nepoznate kinetike. Predložen je model na načelima strojnog učenja, to jest, samo na osnovi mjerenih podataka protoka i koncentracija u stacionarnom stanju. Zbog jednostavnosti primjera umjesto neuronske mreže ili stabla odlučivanja, predložen je linearni model i metoda najmanjih kvadrata za procjenu parametara modela. Model je testiran na skupu s malim brojem uzoraka, $N = 10$, i sa skupom uzoraka $N = 30$. Validacija modela provedena je nezavisnim skupom od $N = 10$ uzoraka, koji nisu uključeni u skupove za razvoj (učenje) modela. Statistička ocjena modela i validacija pokazuju visku točnost, $R^2 = 0{,}98$ i vrijednost $p = 2{,}9 \cdot 10^{-5}$. Dobivenim modelom analizirana je kauzalnost protoka i koncentracije produkta te je dobiven pogrešan zaključak, suprotan osnovnim načelima kemijskog inženjerstva.

Uvod

Matematičko modeliranje procesa u kemijskom inženjerstvu najvažniji je digitalni alat za projektiranje, simulaciju, upravljanje i nadzor procesa (npr. "Aspen Plus" i "Chemcad"). Klasični modeli kemijske industrije osnivaju se na matematičkim jednadžbama bilanci tvari, energije, količine gibanja, termodinamičkim bazama i regresijskim kinetičkim modelima. Točnost tih modela za projektiranje procesa je na vrlo visokoj razini (1 – 2 %) i određena je točnošću ulaznih parametara modela. Danas su kemijski inženjeri, koji su zaduženi za upravljanje procesa modelima s velikim brojem podataka suvremenih industrijskih procesa, suočeni digitalizacijom (internet G5 generacije, internet stvari IoT), sustavima komunikacije i integracijom podataka u složenim sustavima za projektiranje i upravljanje. Upotreba velikih podataka (engl. *Big Data*) u modeliranju i umjetna inteligencija (engl. *Artificial Intelligence*, AI) od velike su važnosti za suvremeno kemijsko inženjerstvo. Zaključivanje i upravljanje procesa AI metodologijom bitno odstupa od egzaktnosti zaključivanja klasičnim kemijsko inženjerskim modelima i zahtijeva primjenu metodologije analize kauzalnosti i umjetne opće inteligencije (engl. *Artificial General Intelligence*, AGI).[1–2] Računalni programi za strojno učenje (npr. "Deep Neural Networks" i "Random Forest of Decision Trees") i analizu kauzalnosti ("DoWhy") vrlo su učinkoviti i dostupni u otvorenoj računalnoj platformi *Github*. Njihova primjena zahtijeva znanje kemijskog inženjera da bi se izbjegla vrlo česta situacija pogrešnog zaključivanja na osnovi agnostičkih (samo empirijskih) modela. Kao primjer pogrešnog zaključivanja regresijskim modelom prikazan je vrlo jednostavan zadatak analize regulacije produkta manipulacijom protoka u protočnom kemijskom reaktoru (PKR).

Model

Na slici 1 prikazan je izotermni protočni kotlasti reaktor (PKR) s pritokom reaktanta A i dvije usporedne reakcije kojima nastaju produkti B i C. Provedena su mjerenja (simulacijom) protoka i koncentracija u pritoku i reaktoru. Matematički model su tri stacionarne bilance tvari za A, B i C.[3]



Slika 1 – Shematski prikaz protočnog kemijskog reaktora (PKR) i mjerenih varijabli

Bilance za reaktant A i produkte dane su izrazima (1–3):

$$q\,(c_{Au} - c_A) = V\,(r_B + r_C), \quad (1)$$

$$qc_B = V\,r_B, \quad (2)$$

$$qc_C = V\,r_C. \quad (3)$$

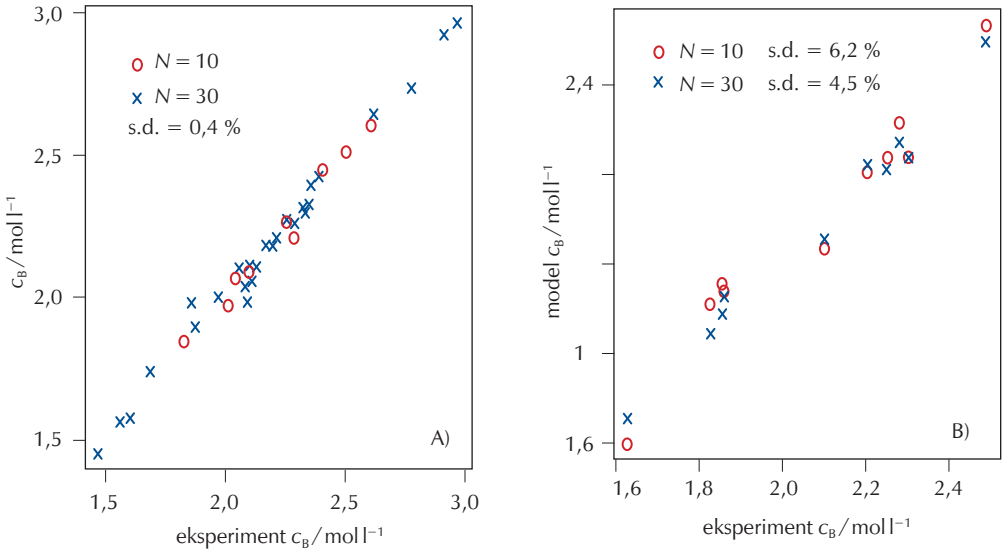
Protok q i ulazne koncentracije c_{Au} simulirani su kao slučajni uzorci iz normalne razdiobe n uz konstantni volumen reaktora V:

$$q \in n\,(\mu_1, \sigma_1), \quad (4)$$

$$c_{Au} \in n\,(\mu_2, \sigma_2). \quad (5)$$

Iako su koncentracije tvari u reaktoru zavisne o prosječnom vremenu zadržavanja $\tau = V/q$, zbog pretpostavljenog konstantnog volumena modelom, kao nezavisna varijabla analizira se utjecaj protoka.

* Želimir Kurtanjek, prof. u mirovini
e-pošta: zelimir.kurtanjek@gmail.com



Slika 2 – Prikazi točnosti predikcije koncentracije produkta c_B sa skupovima simuliranih eksperimentalnih podataka $N = 10$ i $N = 30$. Slika (A) su podatci iz skupa za modeliranje (učenje modela) i novih (nenaučenih) podataka (B).

Procjena parametara i validacija modela

Zbog jednostavnosti primjera, za potrebe simulacije odabrani su linearni kinetički modeli, $r_A = k_1 c_A$ i $r_B = k_2 c_A$, i simulirane vrijednosti koncentracija u reaktoru određene su iz linearnih jednadžbi (1–3). Izlazna veličina modela y koncentracija je produkta B. Za simulaciju utjecaja mjernih pogrešaka $\boldsymbol{y}_m$ vektoru izračunatih koncentracija B pridodane su slučajne vrijednosti iz normalne razdiobe:

$$\boldsymbol{y}_m = \boldsymbol{c}_B \text{ '+ } \boldsymbol{n}\ (0, \sigma). \tag{6}$$

Simulirana relativna standardna pogreška mjerenja upravljane (izlazne) veličine, koncentracije produkta B je 2,2 %. Ulazni podatci modela za skup s $N = 10$ uzoraka, danih u tablici 1, su elementi matrice $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{X}\,[q, c_{Au}, c_A, c_C]. \tag{7}$$

Parametri modela procijenjeni su metodom najmanjih kvadrata[4]

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}_m \tag{8}$$

i procjena izlazne veličine, koncentracije B, je model:

$$y = \beta_0 + \beta_1 q + \beta_2 c_{Au} + \beta_3 c_A + \beta_4 c_B \tag{9}$$

Tablica 1 – Eksperimentalni podatci simulirani modelom bilanci tvari

N	q / l min^{-1}	c_{Au} / mol l^{-1}	c_A / mol l^{-1}	c_B / mol l^{-1}	c_C / mol l^{-1}
1	1,860	12,948	4,955	2,611	5,350
2	1,942	11,220	4,409	2,259	4,525
3	2,390	11,302	5,011	2,046	4,239
4	2,018	10,721	4,311	2,100	4,317
5	2,032	9,388	3,792	1,834	3,772
6	2,429	14,074	6,296	2,508	5,219
7	2,115	11,496	4,754	2,289	4,522
8	1,684	6,567	2,361	1,410	2,801
9	1,828	11,903	4,507	2,408	4,915
10	1,889	9,554	3,691	2,017	3,890

Procijenjeni parametri i ocjena točnosti modela dani su u tablici 2.

Tablica 2 – Procjena parametara i statistička ocjena točnosti modela

β_0	β_1	β_2	β_3	β_4	R^2	P	F
−0,451	0,292	0,957	−1,077	−0,855	0,983	$2{,}9 \cdot 10^{-5}$	132,5

Utjecaj broja podataka N na točnost modela ispitana je s dvama skupovima podataka: s $N = 10$ i $N = 30$ uzoraka. Na slici 2A prikazane su usporedbe predikcije modela i simuliranih eksperimentalnih podataka koncentracije produkta B. Standardne pogreške u oba primjera su 0,4 % i jasno je da su predikcije modela "stabilne" i ne mijenjaju se brojem podataka. Validacija modela provedena je sa skupovima "novih" podataka koji nisu upotrijebljeni za razvoj modela. Rezultati točnosti predikcije prikazani su na slici 2B. Za skup s $N = 10$ standardna pogreška iznosi 6,2 %, a za skup s $N = 30$ iznosi 4,5 %. Iz dobivenih rezultata točnosti predikcije i validacije modela moglo bi se zaključiti da je model pouzdan za upravljanje (regulaciju) koncentracije produkta B u reaktoru u uvjetima poremećaja procesnih uvjeta.

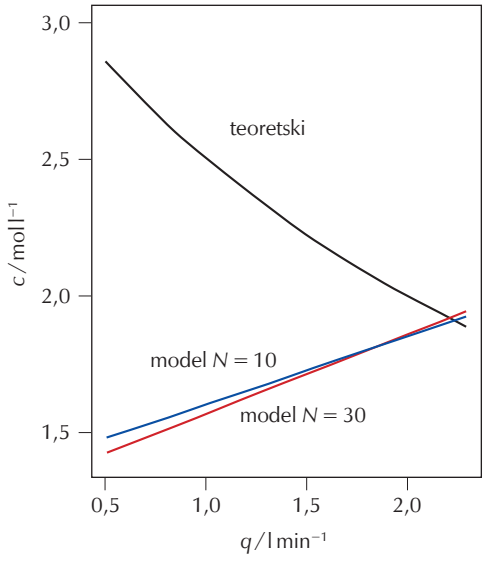
Upravljanje i kauzalnost

Za pouzdano upravljanje procesa nužna je nepristrana kauzalna veza između upravljačke (manipulativne, protok) i upravljane (regulirane, koncentracija produkta B) veličine. Da bi se matematičkim modelom mogla utvrditi kauzalna veza, potrebno je analizirati svojstvo modela na višoj odnosno drugoj razini (intervencije ili odlučivanja) Pearlove kauzalne hijerarhije PCH.[5] Zahvaljujući linearnosti modela bilanci tvari, teoretska kauzalna veza dana je izrazom gdje su k_1 i k_2 konstante brzine reakcija:

$$c_B = V k_1 \frac{c_{Au}}{q + V(k_1 + k_2)}. \tag{10}$$

Kad nemamo mogućnost analitičke analize određivanja kauzalnosti složenih modela, za kauzalnost modela potrebno je odrediti usmjereni aciklički graf (engl. *Directed Acyclic Graph*, DAG) i primijeniti razdvajanje ("d-separation") interferirajućih utjecaja kovarijantnih varijabli na odnos uzroka i posljedice. Analiza ka-

uzalnosti je statistički i numerički zahtjevna, ali se može provesti računalnom podrškom "DoWhy" u otvorenom pristupu na platformi Github.[6]



Slika 3 – Pogrešna procjena P(c_B | do(q)) modela o kauzalnosti koncentracije produkta B o volumnom protoku q

Zahvaljujući jednostavnosti izvedenog empirijskog, ali agnostičkog modela (jedn. (9)) moguće je neposredno odrediti predikciju zavisnosti koncentracije produkta B o protoku q. Kao primjer odabrano je stacionarno stanje $c_{Au} = 10$ mol l^{-1}, $c_A = 4{,}3$ mol l^{-1}, $c_C = 3{,}8$ mol l^{-1} i određen je učinak promjene protoka u intervalu $q \in [0{,}5 - 2{,}3$ l min$^{-1}]$. Rezultati su prikazani na slici 3, iz kojih je vidljivo da empirijski model koji daje točnu predikciju stanja koncentracija (prva razina PCH kauzalne hijerarhije) daje pogrešnu predikciju kauzalnosti (druga razina PCH). Neslaganje predikcije kauzalnosti temeljno je pogrešno, jer predviđa povećanje koncentracije produkta povećanjem protoka, dok u stvarnosti dolazi do smanjenja koncentracije produkta.

Zaključak

Ovim jednostavnim primjerom ukazano je na moguće pogrešno zaključivanje o donošenju odluka i upravljanja samo na osnovi empirijskih podataka i modela strojnog učenja. Do pogrešnog zaključivanja najčešće dolazi kad su empirijski modeli agnostički, odnosno bez poznavanja strukture funkcionalnih zavisnosti na osnovi temelja kemijskog inženjerstva. Točnost predikcije na prvoj razini modelima strojnog učenja ne garantiraju točnost predikcije kauzalne zavisnosti na višoj razini. Kemijski inženjeri primjenom modeliranja procesa strojnim učenjem mogu postići potrebnu točnost predikcije za vrlo složene sustave, ali za donošenje odluka (poslovnih, znanstveno istraživačkih, nadzor i upravljanje procesa proizvodnje) modele je potrebno istražiti na višoj (drugoj) razini PCH kauzalnosti.

Literatura

1. *Ž. Kurtanjek*, Važnost kauzalnosti za studije kemije i kemijskog inženjerstva, Kem. Ind. **70** (7-8) (2021) 467–471.
2. *N. Bolf*, Strojno učenje, Kem. Ind. **70** (9-10) (2021) 591–593.
3. *Z. Gomzi*, Kemijski reaktori, Hinus, Zagreb, 1998.
4. *Z. Gomzi, Ž. Kurtanjek*, Modeliranje u kemijskom inženjerstvu, HDKI, Zagreb, 2019.
5. *J. Pearl, D. Mackenzie*, The Book of Why, Penguin Books, Oxford, UK, 2018.
6. *A. Sharma, E. Kiciman*, DoWhy: A Python package for causal inference, 2021., URL: https://microsoft.github.io/dowhy.

SUMMARY

When Inference by a Mathematical Model can be Erroneous: Example of Chemical Continuous Stirred Tank Reactor (CSTR)

Želimir Kurtanjek

Application of machine learning for modelling chemical engineering processes by neural networks, decision trees or others, result in agnostic models which do not provide functional dependencies between variables. They belong to models of the first rung of Pearl's causal hierarchical (PCH) ladder and usually have high predictive power. However, for process control, optimisation, and monitoring chemical engineers need models which belong to the second rung of PCH which are able to answer the questions on causal effects do to considered process interventions. The problem is illustrated by a simple example of an isothermal continuous stirred tank reactor with two parallel reactions, A->B and A->C, with assumed unknown reaction rate kinetics. The process is modelled under steady state conditions with mass balance equations and Gaussian noise for measurement noise. The data matrix is defined by the measured volumetric flow through rate, feed concentration of the reactant, and outlet concentrations of the reactant and products. Developed is a linear multivariate model for prediction of product B from the input data given by the data matrix. The model parameters are estimated by ordinary least squares (OLS) for two training sets with $N = 10$ and $N = 30$ samples. The model is validated by a data set with $N = 10$ untrained samples. Prediction errors and the model performance are evaluated by standard statistical metrics. The results show the model high prediction accuracy, standard error for the training set is 0.4 %, and 4–6 % for the untrained data set. However, assumed application of the model for process control of the product B by manipulation of the input flow rate, P(c_B | do(q)), failed revealing fundamentally erroneous predictions, *i.e.*, opposite behaviour to the theoretical mass balance law. Hence, this simple example shows the importance of causal approach in modelling of chemical engineering processes for decision making in research and process control.