Implementacija procesno analitske tehnologije (PAT) s pomočjo tehnik strojnega učenja

Matjaž Gams, Jana Krivec, Matej Ožek, Tea Tušar

[1] *Institut »Jožef stefan«, Jamova cesta 39, 1000 Ljubljana*
E-pošta: jana.krivec@ijs.si

Povzetek

V prispevku je prestavljena implementacija umetne inteligence v procesno analitske tehnologije (PAT) pri procesu izdelave tablet. S pomočjo algoritmov strojnega učenja smo zgradili modele, ki na podlagi parametrov izdelave preteklih serij in aktualnih podatkov napovedujejo pričakovano kakovost novih serij tablet. Modele smo vgradili v orodje Orange, kjer smo razvili dodatno komponento za prikazovanje načrtovalskega prostora in interaktivno napovedovanje kakovosti novih serij s pomočjo izbranega modela.

1 Uvod

Administracija Združenih držav za hrano in zdravila (United states Food and Drug Administration, FDA) je leta 2004 [1] uvedla mehanizem Procesno analitskih tehnologij (PAT), danes sprejetih širom po svetu [1]. PAT je definirala kot mehanizem za oblikovanje, analizo in kontrolo farmacevtskih proizvodnih procesov, kjer se s sprotnim merjenjem kritičnih procesnih parametrov zagotavlja visoka kvaliteta končnega izdelka. Z namenom, da motivira farmacevtsko industrijo k izboljšanju produkcijskega postopka, je uvedla tudi dokaj stroge regulativne standarde [1].
Mnoge farmacevtske firme se soočajo s problemom, kako PAT vpeljati v svoje že utečene produkcijske postopke. Pogosti so problemi pri izbiri kompleksnega procesa ter pridobivanju in obdelavi podatkov. Odkrivajo se vedno novi načini implementacije PAT-a, ki spreminjajo poti znanstvenega in inženirskega razvoja produktov in procesov. V našem primeru smo s pomočjo tehnik strojnega učenja poskušali PAT implementirali v proces izdelovanja tablet.

Implementacija PAT v proizvodnjo tablet s pomočjo tehnik strojnega učenja.

V postopku izdelave tablet nastopajo številni parametri, ki vplivajo tako na kakovost izdelka kot na učinkovitost izdelave. Cilj našega projekta je bil definirati ***načrtovalski prostor*** (angl. *design space*) parametrov določenega tipa tablet, znotraj katerega je kvaliteta končnega izdelka še sprejemljiva za prodajo. V ta namen smo se posluževali metod strojnega učenja [2], s katerimi lahko na osnovi podatkov iz preteklih serij zgradimo model, ki na podlagi aktualnih vrednosti posameznih parametrov napove kakovost predvidenega končnega izdelka. S tem smo želeli doseči večjo fleksibilnost pri proizvodnji, poglobljeno razumevanje tehnološkega procesa in učinkovito nadzorovanje postopka izdelave posamezne vrste tablet.
V ta namen smo razvili tudi prototipni računalniški program, ki nam na osnovi predlaganega modela s pregledno in interaktivno vizualizacijo pomaga pri nastavljanju primernih vrednosti procesnih parametrov, pri katerih dobimo kvaliteten končni izdelek.

2 Metoda

Metode strojnega učenja in izkopavanja znanja so med najuspešnejšimi sodobnimi aplikacijami umetne inteligence. Ko imamo opravka z veliko količino podatkov in učnih primerov, se ti sistemi lahko naučijo, katere so zakonitosti posamezne domene in predvidevajo, kaj se bo zgodilo s prihodnjimi primeri. Zgrajeno znanje oz. modeli so pogosto v človeku razumljivi predstavni obliki.
Uspešnost metode strojnega učenja bazira na dobro zastavljenih in kvalitetnih vhodnih podatkih, uporabi različnih algoritmov, parametrov in kombinacij z namenom, da se najdejo najbolj smiselne in pomembne relacije med vhodnimi podatki in preiskovanim rezultatom.

2.1 Opis vhodnih podatkov

Vsaka tableta gre skozi predpisan proces izdelave. Posamezen proces izdelave tablet se imenuje *serija*. Če želimo s pomočjo strojnega učenja zgraditi model izdelave določenih tablet, potrebujemo vhodne podatke o preteklih serijah izdelave takšnih tablet, pri čemer vsaka pretekla serija predstavlja en učni primer.
Vhodni podatki, ki jih je v našem primeru pridobila farmacevtska firma, imajo glede na različne odmerke naslednjo sestavo: najmanjše, male, srednje, velike. Za učenje modela smo imeli skupaj na voljo manj kot 40 serij, kar je malo glede na običajne postopke strojnega učenja. Vsaka serija je določena s parametri (lastnostmi) uporabljenih vhodnih surovin in parametri procesa, ki v matriki vhodnih podatkov predstavljajo atribute strojnega učenja ter kakovostjo končnega izdelka, ki predstavlja razred.

2.1.1 Parametri surovin

Priprava atributov surovin je zahtevala dodatno obdelavo podatkov, saj se lahko v eni seriji uporabljajo

surovine različnih lastnosti. V takšnih primerih smo atribute surovin za neko serijo izračunali kot uteženo vsoto parametrov surovin, ki so bile uporabljene v tej seriji, kjer je bila utež sorazmerna s količino uporabljene surovine. Skupaj smo imeli okoli 30 parametrov, ki so bili dobljeni z analitskimi izvidi vhodnih surovin. Vsi parametri so numerični.

2.1.2 Parametri procesa izdelave

Proces izdelave je opisovalo okoli 40 t.i. »procesnih parametrov«, ki jih lahko podrobneje združimo v 5 faz: granulacijo, sušenje, hlajenje, mletje in tabletiranje. Pri tem je bilo v celotnem procesu le nekaj nad 10 takšnih parametrov, ki so direktno nastavljivi, medtem ko večina parametrov predstavlja le rezultat procesa. Nas so najbolj zanimali nastavljivi parametri v fazi granulacije, saj je to faza kjer je manevrski prostor še najbolj odprt in kjer se oblikujejo osnove za rezultat končne kakovosti. Od vseh parametrov so bili 4 nominalni (dvojiški), vsi ostali pa numerični. Atributi strojnega učenja so večinoma enaki podanim parametrom proizvodnega procesa. Dva nova atributa, ki smo jih izračunali iz obstoječih parametrov procesa smo dodali le pri postopku granulacije. Tri atribute procesa smo normirali glede na jakost tablete.

2.1.3 Kakovost izdelka

Vsaka serija tablet je na koncu podvržena analizi kakovosti, kar predstavlja razred, ki smo ga s strojnim učenjem želeli napovedovati. Kakovost je bila v osnovi izražena numerično, za potrebe našega projekta pa smo jo s pomočjo mnenj strokovnjakov razdelili v 3 razrede: zelo dobra kvaliteta - 1, srednje dobra kvaliteta - 2 in slaba kvaliteta - 3. Kasneje smo dodali še kategorijo, kjer smo združili srednje dobro in slabo kvaliteto - 2+3, saj smo želeli napovedovati le najbolj kvalitetne tablete.

Tabela 1. Načrtovalski prostor različnih podproblemov.

		DELITEV GLEDE NA ATRIBUTE							
			vsi	S	P	S+P1	S+P1+P2	S+ P1+P2+P3	S+P1+P2+P3+P4
DELITEV GLEDE NA LIČNE PRIMERE	Nova+stara tehnologija/ nova tehnologija	vse	vsi / vse	S/ vse	P/vse	S+P1/ vse	S+P1+P2/ vse	S+P1+P2+P3/ vse	S+P1+P2+P3+P4/ vse
		najmanjše	vsi / najmanjše	S/ najmanjše	P/ najmanjše	S+P1/ najmanjše	S+P1+P2/ najmanjše	S+P1+P2+P3/ najmanjše	S+P1+P2+P3+P4/ najmanjše
		male	vsi / male	S/ male	P/ male	S+P1/ male	S+P1+P2/ male	S+P1+P2+P3/ male	S+P1+P2+P3+P4/ male
		srednje	vsi / srednje	S/ srednje	P/ srednje	S+P1/ srednje	S+P1+P2/ srednje	S+P1+P2+P3/ srednje	S+P1+P2+P3+P4/ srednje
		velike	vsi / velike	S/ velike	P/ velike	S+P1/ velike	S+P1+P2/ velike	S+P1+P2+P3/ velike	S+P1+P2+P3+P4/ velike

Legenda: S…atributi surovin, P…atributi procesa, P1…atributi granulacije, P2…atributi sušenja, P3…atributi hlajenja, P4…atributi mletja

Ker je vseh parametrov kar skoraj 100, učnih podatkov pa manj kot 40, lahko že iz vhodnih podatkov sklepamo, da je bilo učenje modela zahtevna naloga. Poleg tega je bilo v podatkih mnogo vrednosti manjkajočih (predvsem med parametri surovin) in malo neprimernih serij (le dve), kar nalogo dodatno otežuje.

2.2 Načrtovalski prostor reševanja problema

Reševanja problema smo se lotili na različne načine oz. ga razdelili na več podproblemov, pri čemer smo modificirali učne primere in/ali atribute učenja.

2.2.1 Modifikacija učnih primerov

Problem smo razdelali najprej glede na izbrano učno množico. Ločili smo 5 skupin učnih primerov: vsi primeri, le najmanjši primeri, le mali primeri, le srednji primeri in le veliki primeri.
Poleg tega smo tablete razdelili še v 2 skupini glede na to, ali so bile narejene s staro ali z novo tehnologijo. Ker obe tehnologiji nista povsem različni, smo najprej delali odločitvena drevesa na podlagi vseh podatkov. Ker pa bodo v prihodnje vse serije narejene po novi tehnologiji, smo odstranili podatke iz serij, narejenih po stari tehnologiji, in preverili, če se drevesa kakorkoli spremenijo.

2.2.2 Modifikacija atributov

Ker se proces izvaja po korakih, smo tudi mi združevali atribute po skupinah, ki so vključene v posamezne dele izdelave in na njih gradili pravila. Tako so pravila razdeljena v naslednje skupine: samo atributi surovin, samo atributi procesa, atributi surovin + granulacije, atributi surovin + granulacije + sušenja, atributi surovin + granulacije + sušenja + hlajenja, atributi surovin + granulacije + sušenja + hlajenja + mletja in vsi atributi.

Na ta način smo dejansko dobili vrsto podproblemov (glej Tabelo 1), ki smo jih reševali posebej. Najprej smo na različne načine iskali odločitvena drevesa, nato pa smo na njihovi podlagi zgradili še odločitvena pravila.

2.3 Izgradnja modelov

Zaradi zahteve po razumljivosti modela smo se omejili na odločitvena drevesa in odločitvena pravila. Pri

gradnji modelov smo si pomagali z dvema zbirkama orodij strojnega učenja: Orange [3] in Weka [2]. Oba sistema ponujata na desetine metod strojnega učenja ter tehnik dodatne obdelave podatkov in vizualizacije. Poskusili smo več algoritmov za gradnjo odločitvenih dreves in se na koncu odločili za algoritem C4.5 [4], ki je implementiran v obeh zbirkah orodij. Algoritem je primeren, predvsem ko se pojavi zahteva po transparentnosti znanja, kar v našem primeru vsekakor drži, saj je potrebno iz velike količine zgrajenih odločitvenih modelov oz. dreves in pravil izluščiti najbolj smiselne relacije. Najbolj smiselne relacije so tiste, ki se zdijo pomembne ekspertom in ki imajo hkrati najboljšo klasifikacijsko točnost. Slednja nam pove, kolikšna je verjetnost, da bo nov primer v domeni pravilno uvrščen. Točnost klasifikacijskih dreves smo preverjali z vgrajeno metodo 10-kratnega prečnega preverjanja.

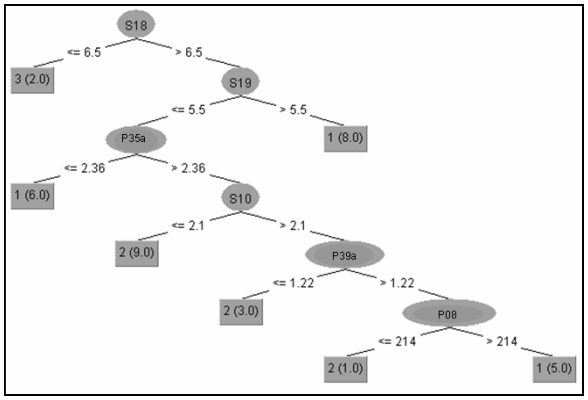
Z algoritmom C4.5 in različnimi podmnožicami atributov smo na več načinov zgradili množico odločitvenih dreves oz. pravil, ki na čim bolj smotrn način opozorijo, kadarkoli bi se približali nevarnemu območju. Ta pristop je skladen s principom mnogoterega znanja [5], originalno razvitem na Odseku za inteligentne sisteme Instituta »Jožef Stefan«. Gre za to, da ne zgradimo samo enega preprostega modela, ki opozori na nevarno območje, ampak množico modelov oz. opisov, ki skozi svoj zorni kot opozorijo na nevarnost. Vsak tak model mora biti 100% točen (100% loči negativne učne primere od pozitivnih), hkrati pa mora biti čimbolj preprost. Zato smo algoritem C4.5 poganjali brez naknadnega rezanja poddreves in z minimalnim številom primerov v listih enakem 1. Glede na princip Ockhamove britve je znano, da so preveč podrobni samostojni modeli zavajajoči. Znano pa je tudi, da je smiselna kombinacija več modelov (mnogotero znanje) optimalni model v smislu točnosti napovedovanja in razlage. Zato smo vedno gradili 100% drevesa (oz. pravila iz dreves, kar je le bolj kompaktna oblika opisa), ki so hkrati čim krajša.

3 Rezultati

Rezultati obdelave podatkov s tehnikami strojnega učenja predstavljajo odločitvena drevesa (Slika 1) in pravila. Ker je bilo vseh upoštevanih dreves za vse podprobleme preveč, da bi jih v prispevku predstavili, prikazujemo le en primer drevesa in iz njega izpeljanih pravil (Slika 1). Za potrebe računalniškega modela smo namreč najboljše veje odločitvenih dreves pretvorili v odločitvena pravila.

Iz drevesa na Sliki 1 lahko izpeljemo naslednja pravila:

IF S18≤6.5 THEN CLASS=3 (2)
IF S18>6.5 AND S19>5.5 THEN CLASS=1 (8)
IF S18>6.5 AND S19≤5.5 AND P35A≤2.36 THEN CLASS=1 (6)
IF S18>6.5 AND S19≤5.5 AND P35A>2.36 AND S10≤2.1 THEN CLASS=2 (9)
IF S18>6.5 AND S19≤5.5 AND P35A>2.36 AND S10>2.1 AND P39A>1.22 AND P08>214 THEN CLASS=1 (5)



Slika 1. Odločitveno drevo za vse atribute. Prečno preverjanje = 65%.

Primer branja odločitvenega pravila

Iz prvega pravila lahko razberemo naslednje: v primeru da je atribut S18≤6.5 potem je razred (kvaliteta izdelka) 3 (slaba), kar velja za dva učna primera (seriji tablet v učni množici).
Iz drugega pravila vidimo, da je kvaliteta dobra (razred 1), če je atribut S18>6.5 in je hkrati atribut S19>5.5, kar velja za 8 učnih primerov.

Iz množice izgrajenih dreves smo na zgoraj predstavljen način za vse kategorije podproblemov izločili okoli 600 odločitvenih pravil (glej Tabelo 2).

Tabela 2. Število pravil, ki smo jih za posamezne podprobleme vključili v računalniški program.

		DELITEV PO UČNIH PRIMERIH IN TEHNOLOGIJI					
		najmanjše	male	srednje	velike	vse	vse
		NT+ST					NT
RAZRED	1	30	5	26	15	74	52
	2	44	5	32	26	60	60
	2+3	40	/	/	/	15	21
	3	33	/	/	/	4	6

Legenda: NT…nova tehnologija, ST...stara tehnologija

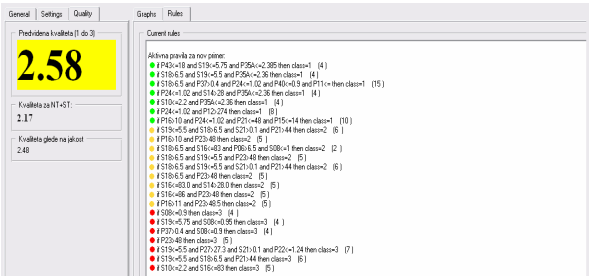
4 Izdelava računalniškega programa

Ko ekspert dobi nove surovine za tablete, mora najprej ugotoviti, ali so te primerne za kvaliteten končni izdelek in nato še ustrezno nastaviti procesne parametre. Zato smo v pomoč temu v okvirju orodja Orange izdelali orodje Multiscatter za vizualizacijo podatkov, razumljiv prikaz pravil in napoved kvalitete novega izdelka. Glede na interaktivnost dela, tj. človeški interakciji s programom, mora biti model uporabniku prijazen, množico informacij pa je potrebno prikazati na razumljiv in pregleden način. *MultiScatter* omogoča tako prikaz načrtovalskega prostora kot tudi

napovedovanje kakovosti novih serij s pomočjo izbranega modela, ki je v ozadju programa.

4.1 Prikaz načrtovalskega prostora in napovedovanje kakovosti novih serij

Uporabnik v računalnik vnese podatke o atributih surovin in procesa. Računalnik sprotno izpisuje pravila, ki so na podlagi podatkov lahko določila kvaliteto. Pravila so ustrezno obarvana glede na razred, ki ga določajo (zeleno za dobro kvaliteto, rumeno za srednje dobro kvaliteto in rdeče za slabo kvaliteto) in različno obtežena. Uteženost pravila se izračuna iz kombinacije števila parametrov v pravilu in števila primerov, ki ga pravilo pokrije.
Na podlagi rezultatov pravil se izračuna predvidena kvaliteta novega izdelka (glej Sliko 2).



Slika 2: Prikaz pravil in napovedovanje kakovosti nove serije

Na sliki 2 je prikazana napovedana kvaliteta novega izdelka. Izračun 2.58 je uteženo povprečje kakovosti, ki ga napovedujejo pravila. Pod kvaliteto 2.58 še dve kvaliteti. Druga vrednost je izračunana iz pravil, ki smo jih dobili iz podatkov o serijah, narejenih po novi in stari tehnologiji. Tretja vrednost pa je izračunana iz pravil, ki so prirejena posebej za določeno težo (jakost) tablete. Zato bo tretja vrednost manjkala, dokler ne vnesemo podatka o želeni teži tablet. Najbolj zanesljiva je prva vrednost. Drugi dve vrednosti pa služita le za primerjavo in sta manj zanesljivi.

Atribute, ki se uporabljajo v pravilih (t.i. ključne atribute), računalnik zapiše v posebno okno. Iz njih lahko izrišemo množico grafov, ki imajo za osi izbrane atribute in sestavljajo načrtovalski prostor. Na grafu se izrišejo zelene točke, ki za kombinacijo izbranih atributov predstavljajo kakovostne primere iz učne množice, rumene točke (primeri srednje kvalitete) in rdeče točke (primeri slabe kvalitete). Poleg tega se z uporabo algoritma najbližjih sosedov [6] glede na pričakovano verjetnost razredov obarva tudi ploskev grafa načrtovalskega prostora. Nov primer, ki ga testiramo se na grafu pokaže kot moder križec. V primeru, da kombinacija teh dveh lastnosti oz. nastavitev ne predvideva kvalitetnega končnega izdelka, lahko poskušamo le-to primerno spremeniti. Vrednost atributov lahko spremenimo enostavno tako, da kliknemo na modri križec in ga potegnemo na območje, ki nam napoveduje ugoden končni izid. Večkrat nam zelena in rdeča področja ne dajo prave informacije. Dejstvo, da smo blizu slabemu primeru, je lahko opozorilo, vendar pa to samo po sebi ni zanesljiv kriterij, da bo naš izdelek slab. Zato so s črtkanimi vodoravnimi in navpičnimi črtami na grafu prikazane tudi meje, ki jih uporabljajo pravila. Ali je bolje na levi ali desni strani črte, pa mora ugotoviti uporabnik sam.
Po popravku vrednosti določenih atributov pogledamo, če se napovedana kakovost končnega izdelka dovolj izboljša. Če želimo parametre še bolj optimalno nastaviti celoten postopek ponovimo.

Zaključek

Za farmacevtsko podjetje smo razvili sistem strojnega učenja za vodenje in sprotno diagnosticiranje proizvodnje tablet. Tehnične podrobnosti v referatu niso na voljo, opisan pa je algoritem, ki vsebuje nekatere izvirne rešitve.

Literatura

[1] *"Guidance for Industry PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance"*. U.S. Department of Health and Human Services, Food and Drug Administration, 2004.

[2] Ian H. Witten and Eibe Frank. *"Data Mining: Practical machine learning tools and techniques"*. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[3] Janez Demšar, Blaž Zupan and Gregor Leban. *"Orange: From experimental machine learning to interactive data mining"*. White Paper, Faculty of Computer and Information Science, University of Ljubljana, 2004.

[4] J. Ross Quinlan. *"C4.5: Programs for Machine Learning"*. Morgan Kaufmann, 1993.

[5] Matjaž Gams. *"Weak Intelligence: Through the Principle and Paradox of Multiple Knowledge"*. Advances in Computation, vol. 6., Huntington, New York, Nova Science, 2001.

[6] Thomas M. Cover and Peter E. Hart. "Nearest neighbor pattern classification". *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.