KLASIFIKACIJA IN VIZUALIZACIJA PROCESOV Z METODAMI STROJNEGA UČENJA

Matej Ožek, Matjaž Gams, Jana Krivec, Tea Tušar

Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
E-pošta: matej.ozek@ijs.si

POVZETEK

Razvili smo novo metodo za klasifikacijo procesnih podatkov. Metoda temelji na pridobivanju klasifikacijskih pravil iz odločitvenih dreves in je še posebej primerna za procese z veliko atributi in malo učnimi primeri, kjer je potrebno rezultat razložiti. V ta namen smo razvili tudi program za vizualizacijo in modeliranje procesa. Metodo smo uporabili na podatkih iz farmacevtske industrije.

1 UVOD

Metode za rudarjenje podatkov delujejo najbolje na podatkih, ki imajo malo atributov in veliko primerov. Kadar pa imamo na voljo malo primerov in veliko atributov, so standardne metode precej manj uspešne. Če poleg tega podatke dobimo pri procesu, kjer je določen atribut odvisen od vseh prejšnjih, potem je jasno, da je potrebno uporabiti nov pristop. Naročnik dela je farmacevtska družba, v kateri želijo poleg klasifikacije izdelka že med proizvodnim procesom dobiti tudi razlago, zakaj je bil določen polizdelek klasificiran za slabega.

1.1 Implementacija PAT

Administracija Združenih držav za hrano in zdravila (United states Food and Drug Administration, FDA) je leta 2004 uvedla mehanizem Procesno analitskih tehnologij (PAT) [1], ki je danes sprejet širom sveta. PAT je definirala kot ogrodje za razvoj novih pristopov k ohranjanju visoke kvalitete farmacevtskih proizvodov.

To je spodbudilo farmacevtske družbe, da se povežejo z raziskovalnimi ustanovami. Farmacevtske družbe se soočajo s problemom, kako PAT vpeljati v svoje že utečene produkcijske postopke. Težave so z nezanesljivimi meritvami podatkov, ter z velikimi stroški, povezanimi z izvajanjem eksperimentov. Zato se soočajo z velikim šumom v maloštevilnih podatkih.

Cilj projekta je razvoj programa, ki bi znal napovedovati kvaliteto izdelka že med postopkom izdelave. Pri tem je treba napovedi utemeljiti, vse znanje pa mora biti prikazano pregledno in razumljivo tudi nestrokovnjakom.

2 OPIS POSTOPKA

2.1 Gradnja dreves

Glede na to, da potrebujemo klasifikacijsko metodo, ki nudi tudi razlago, smo izbrali odločitvena drevesa kot preizkušeno in uspešno klasifikacijsko metodo. Uporabljali smo algoritem C4.5 [4].

Pri klasičnem načinu gradnje dreves, bi poskušali zgraditi le eno, čim boljše drevo. Zaradi specifičnosti problema želimo v drevesa vključiti čim več atributov, saj verjamemo, da prav vsi atributi prispevajo h kvaliteti izdelka. Zato ne zgradimo le eno, najboljše odločitveno drevo, ampak zgradimo množico dreves. Množica dreves je tudi manj občutljiva na napačno klasifikacijo nekaterih dreves, ki so nastala zaradi nezanesljivih podatkov. Tak pristop je skladen s principom mnogoterega znanja [5] in je podoben mnogim modernim ansambelskim metodam.

Pri gradnji odločitvenih dreves smo morali upoštevati, da podatke med proizvodnim procesom dobivamo postopoma. Zdravila se izdelujejo v šestih stopnjah, zato tudi drevesa gradimo po stopnjah. Najprej gradimo drevesa le z atributi iz prve stopnje. Nato le z atributi prve in druge stopnje, itn. vse dokler ne zgradimo drevesa z atributi vseh šestih stopenj procesa. Na ta način so atributi začetnih stopenj privilegirani, saj nastopajo v vseh drevesih. To se ujema z mnenjem farmacevtskih strokovnjakov, ki pravijo, da se o kvaliteti izdelka odloča že v začetnih stopnjah. Obstaja tudi praktični vidik privilegiranja začetnih stopenj procesa. Če model predvidi slabo kvaliteto končnega izdelka že na začetku, lahko poskušamo v naslednjih stopnjah izdelku zvišati kvaliteto. Če pa ugotovimo slabo kvaliteto šele proti koncu, nam odločitvena drevesa ne koristijo več mnogo.

2.2 Gradnja pravil

Postopek gradnje dreves končamo z množico dreves, od katerih so mnoga, zaradi izjemno majhnega števila primerov in mnogih atributov, preveč podrobna. Natančnost klasifikacije popravimo tako, da izberemo le najboljše veje v drevesih in jih napišemo v obliki klasifikacijskih pravil. Najboljše veje so tiste, ki so kratke, vsebujejo čim več primerov in nobenega primera ne klasificirajo narobe. Tako iz manj natančnih dreves dobimo natančna pravila. Pri gradnji pravil se ravnamo po načelu Ockhamove britve, ki pravi, da je v primeru dveh ekvivalentnih pravil boljše tisto, ki je enostavnejše.

Izdelava drevesa samo zato, da ga večino zavržemo, ni tako vprašljiva, kot se zdi na prvi pogled. Na ta način se izognemo slabostim, ki jih imajo algoritmi za generiranje pravil. Ti so večinoma nagnjeni k prevelikem krnjenju [2]. Pravila niso enakovredna. Vsako pravilo utežimo glede na število primerov učne množice, ki jih pravilo pokrije. Vsa pravila morajo pregledati farmacevtski strokovnjaki in izločiti tista, ki so po njihovem mnenju nesmiselna. Tako dobimo prečiščeno množico uteženih pravil, ki se dobro obnesejo na šumnih podatkih.

2.3 Klasifikacija

Za klasifikacijo primerov v razrede obstaja več metod. Mi smo se zaradi narave problema odločili za kombinacijo več metod. Ker lahko že ena slaba sestavina pokvari končni izdelek, vsako pravilo, ki polizdelek klasificira kot slab, sproži alarm. Po drugi strani pa lahko s pravilno izbiro atributov izboljšamo kvaliteto izdelka. Zato je končna kvaliteta utežena aritmetična sredina rezultatov pravil.

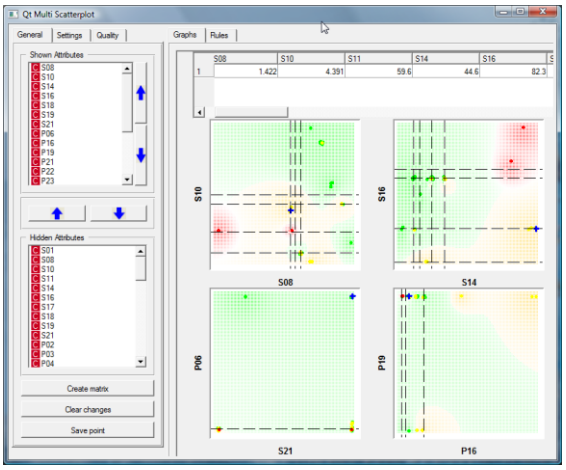
3 VIZUALIZACIJA

Glavna prednost dreves in iz njih narejenih pravil je enostavna razlaga rezultatov. Še posebej je razlaga pomembna med procesom izdelave, ko bi želeli vedeti, katere atribute bi bilo potrebno popraviti v nadaljevanju, da bi dobili boljšo kvaliteto izdelka.

Zato smo izdelali program za vizualizacijo in klasifikacijo, ki temelji na programu Orange [3]. Kljub temu, da so odločitvena drevesa pregledna in enostavna za razlago, je težko pregledovati več dreves hkrati. Zato je potrebno prikazati rezultate pravil na pregleden način. Zato se najprej uporablja utežena aritmetična sredina klasifikacij, ki nam da okvirno oceno kvalitete izdelka. Posebno pozornost je potrebno nameniti pravilom, ki napovejo, da bo izdelek neuporaben.

Ko bi radi med procesom spreminjali vrednost atributov, je vprašanje, kako spremeniti atribute, da bo izdelek kvalitetnejši. Popolno informacijo bi dobili, če bi znali prikazati meje varnega področja (načrtovalskega prostora) v visokodimenzionalnem prostoru, kjer bi vsak atribut tvoril novo dimenzijo. Če bi imeli dovolj učnih primerov, bi lahko aproksimirali načrtovalski prostor. Tega bi predstavljal hiperkvader, v notranjosti katerega bi bilo področje visoke kvalitete, zunaj pa področje slabše kvalitete.

Ker je zaslon računalnika dvodimenzionalen (2D), mora biti vsak graf, ki je tri ali večdimenzionalen, projekcija na dve dimenziji. Pri tem se del informacij izgubi. Zato program na zaslonu prikaže nabor 2D grafov, ki imajo za koordinatni osi dva izbrana atributa (primer na sliki 1). Program sam predlaga atribute, ki so v določeni točki procesa najbolj kritični. Na takšnem grafu so meje med enim in drugim razredom, ki jih določajo pravila, navpične ali vodoravne premice. Območja visoke kvalitete so omejeni ali neomejeni pravokotniki.



Slika 1: Prikaz načrtovalskega prostora z množico 2D grafov.

Prednost množice 2D grafov pred 1D nomogrami je v prikazu interakcije dveh atributov na kvaliteto izdelka, saj za noben atribut ne velja, da bi bil linearno povezan s kvaliteto izdelka.
Program dopušča enostavno modeliranje procesa. Tako lahko z uporabo miške spreminjamo atribute, program pa nam v realnem času izračunava predvideno kvaliteto izdelka in opozarja na kritične atribute.
Na sliki 1 je prikazan prikaz na zaslonu. Skušali smo hkrati prikazati vse podatke, ki so trenutno zanimivi za uporabnika.

4 IMPLEMENTACIJA METODE NA FARMACEVTSKIH PODATKIH

Uporabljali smo podatke, pridobljene iz procesa izdelave tablet. Podatke so beležili ročno med postopkom izdelave, zato so ti pomanjkljivi in večkrat netočni.

Posamezen proces izdelave tablet imenujemo serija. Za učenje modela smo imeli na voljo 30 serij in 70 atributov. Atribute razdelimo na parametre surovin in na parametre procesa izdelave. Slednje razdelimo na 5 faz. Tako smo atribute razdelili na 6 skupin.

Podatke je bilo potrebno pred procesom učenja še precej obdelati. Tako smo izločili nepomembne atribute, izračunali nove atribute ter popravili netočne podatke. Podrobneje je proces opisan v nadaljevanju.

4.1 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.

4.2 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.

4.3 Kakovost izdelka

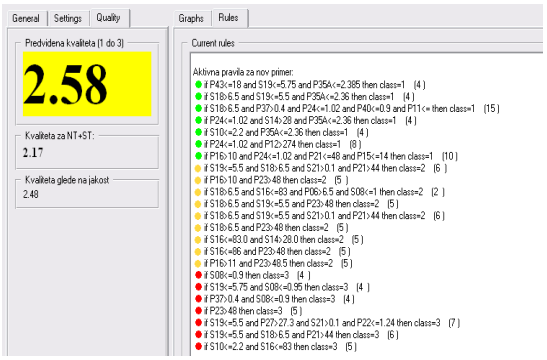
Vsaka serija tablet je na koncu podvržena analizi kakovosti, kar predstavlja razred, ki smo ga s strojnim učenjem želeli napovedati. Kakovost je bila v osnovi izražena numerično, za potrebe našega projekta pa smo jo s pomočjo mnenj strokovnjakov razdelili v 3 razrede: zelo dobra kvaliteta - 1, srednje dobra kvaliteta - 2 in slaba kvaliteta - 3.

4.4 Izgradnja modelov

Pri gradnji modelov smo si pomagali z dvema zbirkama orodij strojnega učenja: Orange in Weka [2].
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. Delali smo po postopku, ki je opisan v poglavju 2.

5 UPORABA RAČUNALNIŠKEGA PROGRAMA

Želimo si, da bi uporabnik čim bolj izkoristil kvalitete modela in množico informacij, ki jih računalnik izračuna. Uporabnika ni smiselno obremenjevati s tehničnimi podrobnostmi. Zanj je najpomembnejši del vizualizacija procesa in pregledan predstavitev podatkov in izračunov.
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, ki je prikazan na sliki 2. *MultiScatter* omogoča tako prikaz načrtovalskega prostora kot tudi napovedovanje kakovosti novih serij s pomočjo izbranega modela, ki je v ozadju programa.



Slika 2: Primer uporabe MultiScatter. Prikazana so pravila in ocena kakovosti izdelka.

Uporabnik v računalnik vnese podatke o atributih surovin in procesa. Računalnik sproti izpisuje pravila, ki so pomembna za trenutne podatke (slika 2). 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 utežena. Uteženost pravila se izračuna iz kombinacije števila parametrov v pravilu in števila primerov, ki ga pravilo pokrije.
Poleg tega se za pomembne atribute izriše še nabor grafov, ki še dodatno pripomorejo k lažji odločitvi, katere atribute je potrebno spremeniti, da dosežemo visoko kakovosten izdelek.

6 ZAKLJUČEK

Razvili in preizkusili smo nov pristop za klasifikacijo in vizualizacijo podatkov, dobljenih pri procesu. Pristop se odlikuje z visoko klasifikacijsko točnostjo kljub majhnemu številu podatkov.

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.