Napovedovanje biorazgradljivosti z regresijskimi drevesi

Bernard Ženko, Sašo Džeroski
Bernard.Zenko@ijs.si, Saso.Dzeroski@ijs.si
Odsek za inteligentne sisteme, Institut Jožef Stefan
Jamova 39, SI-1000 Ljubljana, Slovenija

14. marec 2002

Povzetek

Biorazgradljivost spojine je ena pomembnejših lastnosti, ki jih moramo upoštevati pri ocenjevanju varnosti njene uporabe. Ker bi bilo eksperimentalno določanje biorazgradljivosti množice različnih kemikalij težko izvedljivo, se problema lotimo z modeliranjem količinskih odnosov med strukturo in določeno lastnostjo spojine (Quantitative Structure-Activity Relationships – QSAR). Za vzorčno množico spojin eksperimentalno določimo njihovo biorazgradljivost ter nato zgradimo model, ki zadovoljivo opisuje tako proučene kot neproučene spojine. Model ponavadi zgradimo s klasično metodo linearne regresije. V tem prispevku uporabimo za gradnjo tovrstnih modelov metode strojnega učenja, in sicer metode za gradnjo regresijskih dreves. Za več različnih množic podatkov smo zgradili modele z orodjema za gradnjo regresijskih dreves Cubist in RETIS. Točnost zgrajenih modelov je bila ocenjena s prečnim preverjanjem ter primerjana s točnostjo modelov, zgrajenih z metodo linearne regresije. Najboljše zgrajene modele sta pregledala strokovnjaka s področja biorazgradljivosti. Za majhne množice strukturno sorodnih spojin so modeli zgrajeni z metodo linearne regresije običajno bolj točni kot modeli z regresijskimi drevesi, čeprav imajo slednji včasih primerljivo točnost in so lahkó razumljivi. Za večje množice strukturno različnih spojin so modeli z regresijskimi drevesi bolj točni kot linearni regresijski modeli.

Ključne besede: strojno učenje, odločitveno drevo, regresijsko drevo, biorazgradljivost, modeliranje QSAR

Predicting biodegradability with regression trees

Abstract

The biodegradability of a chemical compound must be considered when estimating the safety of its use for the environment. It is usually given by its half-life and it certainly depends on the structure of a given chemical compound. Generally, compounds with structures similar to natural compounds are more biodegradable than compounds with structures not known in nature. Because of the huge number of various chemicals, it is practically impossible to experimentally determine biodegradability for all or at least for a significant number of them.

A possible solution to this problem is a quantitative structure-activity relationships (QSAR) analysis. Compounds are described in terms of various structural, physicochemical and quantum chemical descriptors which should affect its activity (biodegradability in our case). An example data set of anilines and phenols is given in Table 1. We experimentally test a representative group of chemicals for their activity and then build a model which satisfactorily describes activity of the tested as well as unknown chemicals. Usually the model is built with classical linear regression methods (e.g. Partial Least Squares – PLS method [9]) which gives us one linear equation as a model. A more recent approach to building QSAR models is the use of machine learning methods [8], typically regression tree building methods. A comparison between these two types of models in terms of accuracy and applicability is made in this paper.

For several data sets (described in Section 2), models with Cubist and RETIS regression tree building systems are built. The accuracy of all models is estimated by calculating their correlation coefficients (Equations 1 and 2) on the training data (r, R^2) and via leave-one-out cross validation (q, Q^2). Correlation coefficients are presented in Table 2.

The applicability and understandability of the best models (in terms of accuracy) is inspected by domain expert and are presented in Section 4 and Figures 1–3. For small sets of compounds with a similar structure, models built with linear regression are usually more accurate than models built with regression trees, although the latter sometimes have comparable accuracy and are easily understandable. For large sets of structurally diverse compounds, regression trees yield more accurate models than linear regression.

Key words: machine learning, decision tree, regression tree, biodegradability, QSAR modeling

1. Uvod

Biorazgradljivost spojine nam pove, kako hitro se spojina v okolju razgradi na neškodljive snovi. Običajno je podana v obliki razpolovnega časa. Na hitrost razgradnje spojine prav gotovo vpliva njena struktura oz. z njo povezane lastnosti. Na splošno so spojine s strukturami, ki nastopajo tudi v naravnih spojinah, bolje razgradljive kot spojine s strukturami, ki so v naravi neznane. S pomočjo modeliranja QSAR lahko raziščemo povezave med strukturo spojin in njihovo aktivnostjo. V našem primeru je proučevana aktivnost biorazgradljivost (Quantitative Structure-Biodegradability Relationships – QSBR), lahko pa je tudi kaj drugega: toksičnost (Quantitative Structure-Toxity Relationships – QSTR), stabilnost spojine (Quantitative Structure-Stability Relationships – QSSR) itn. Strukturo proučevanih spojin opišemo z množico strukturnih, fizikalno-kemijskih ali kvantno-kemijskih lastnosti, t.i. *molekulskih deskriptorjev*, za katere sumimo, da bi lahko bili povezani z mehanizmom proučevane aktivnosti.

Pri modeliranju QSAR ločimo dve stopnji. Prva je razvoj modela na podlagi spojin, katerih aktivnost je bila eksperimentalno določena. Druga stopnja je uporaba modela, dobljenega v prvi stopnji, za napovedovanje aktivnosti neznanih spojin. Neznane spojine so tiste, katerih aktivnost še ni bila eksperimentalno določena, znana pa je njihova struktura.

Običajno za gradnjo modela uporabimo metodo delnih najmanjših kvadratov (Partial Least Squares – PLS) [9], s katero dobimo model v obliki ene linearne enačbe. V zadnjem času se uporabljajo tudi metode strojnega učenja [8], po navadi metode za gradnjo regresijskih dreves. Regresijsko drevo je drevo, ki ima v vsakem notranjem vozlišču test, ki preverja vrednost nekega deskriptorja, v listih pa ima linearno enačbo, ki določa vrednost odvisne spremenljivke (npr. biorazgradljivosti). Regresijsko drevo lahko prepišemo v množico pravil, kjer vsakemu listu ustreza eno pravilo. Nas sta zanimali točnost in uporabnost modelov z regresijskimi drevesi v primerjavi z modeli, zgrajenimi s klasično metodo PLS.

V nadaljevanju najprej opišemo množice podatkov, poskuse gradnje regresijskih dreves iz teh množic ter rezultate le-teh, na koncu pa podamo sklepe, do katerih smo prišli z analizo rezultatov.

2. Množice podatkov

Za preskus uporabnosti regresijskih dreves pri napovedovanju biorazgradljivosti smo uporabili naslednje množice podatkov.

1. *Toksičnost in biorazgradljivost anilinov in fenolov* [4]. Toksičnost spojin je bila izmerjena s populacijsko rastjo mikroorganizma *Tetrahymena pyriformis*, soj GL-C, in je podana kot koncentracija, pri kateri je rast zavrta 50-odstotno (IGC_{50}). Hitrost oksidacije je podana s kinetično konstanto drugega reda (k_b). Množica je kot primer podana v tabeli 1; vsaka spojina je opisana z desetimi deskriptorji. Modeli so bili zgrajeni posebej za aniline (7 spojin), posebej za fenole (8 spojin) ter za oboje skupaj (15 spojin).

2. *Akutna toksičnost nasičenih in nenasičenih alifatskih ogljikovodikov* [1]. Ocenjena je bila z merjenjem zmanjšanja bioluminiscence morske bakterije *Photobacterium phosphoreum*, na podlagi katere je bila določena efektivna koncentracija ($EC50$) preverjane spojine. Množica vsebuje 19 spojin, opisanih s 24 deskriptorji. Modeli so bili zgrajeni posebej za haloalkane ter za haloalkane in haloalkene skupaj. Poleg osnovne množice z vsemi deskriptorji sta bili za gradnjo modelov uporabljeni še dve množici z manjšim številom deskriptorjev (2 oz. 5).

Tabela 1: Podatki o toksičnosti in biorazgradljivosti anilinov ter fenolov. Toxity and biodegradability of anilines and phenols data set.

Spojina	$\log K_{ow}$	HOMO $[ev]$	LUMO $[ev]$	r_w $[10nm]$	V_w $[\frac{cm^3}{Mol}]$	μ $[Debye]$	M_w	σ	pK_a	$\log IGC_{50}^{-1}$ $[mol.l^{-1}]$	$\log k_b$ $[L.org.^{-1}.h^{-1}]$
anilin	0.90	-8.61	0.42	0.12	56.38	1.5833	93.13	-0.16	4.58	0.109	-12.68
3-metilanilin	1.40	-8.57	0.42	0.17	70.05	1.4153	107.16	-0.23	4.75	-0.28	-13.77
3-metoksianilin	0.95	-8.69	0.26	0.26	73.75	2.3423	123.15	-0.04	4.26	0.085	-14.30
3-kloroanilin	1.88	-8.76	0.12	0.175	63.38	2.6026	127.57	0.21	3.51	0.092	-13.64
3-bromoanilin	2.10	-8.83	-0.01	0.185	71.50	2.7165	172.02	0.23	3.43	0.517	-13.43
3-cianoanilin	1.05	-9.03	-0.51	0.32	61.08	4.4301	118.14	0.40	2.79	-0.465	-15.67
3-nitroanilin	1.37	-9.28	-1.06	0.259	73.18	6.3025	138.13	0.55	2.45	0.026	-14.92
fenol	1.48	-9.17	0.29	0.12	53.88	1.2338	89.07	-0.16	9.92	-0.241	-11.16
4-metilfenol	2.12	-8.95	0.33	0.17	67.55	1.3602	108.14	-0.31	10.1	-0.162	-11.33
4-metoksifenol	1.57	-9.11	0.17	0.26	71.25	2.4059	124.14	-0.32	10.2	-0.143	-12.70
4-klorofenol	2.48	-9.01	0.05	0.175	65.88	1.4767	128.56	0.11	9.38	0.402	-11.77
4-bromofenol	2.63	-9.31	-0.03	0.185	69.00	1.5930	173.01	0.12	9.45	0.500	-11.80
4-cianofenol	1.60	-9.56	-0.54	0.32	69.58	3.3106	119.12	0.84	7.96	0.516	-13.82
4-nitrofenol	1.85	-10.71	-1.08	0.259	70.68	5.2640	139.11	1.08	7.15	1.420	-13.00
4-acetilfenol	1.45	-9.45	-0.4	0.25	78.25	3.8245	136.15	0.34	8.05	-0.093	-12.51

3. *Biorazgradljivost dioksinov in furanov* [7]. Aerobična degradacija proučevanih spojin je bila izvedena z bakterijo *Sphingomonas sp.*, soj RW1. Množica vsebuje 14 spojin s 50 deskriptorji [7]. Poleg te množice sta bili uporabljeni še dve z zmanjšanim številom deskriptorjev (15 oz. 9).

4. *Biorazgradljivost haloalifatskih spojin* [2]. Ocena dehalogenizacije je bila izvedena z nepoškodovanimi celicami bakterije *Rhodococcus erythropolis*, soj Y2. Ciljna vrednost je naklon premice v grafu časovnega poteka koncentracije klorovih ionov. Množica vsebuje 27 spojin z devetimi deskriptorji. Poleg celotne množice je bila za gradnjo modelov uporabljena tudi množica brez treh izstopajočih spojin.

5. *Biorazgradljivost mutantov haloalkanske dehalogenaze* [3]. Zanima nas dehalogenizacija nespremenjene haloalkanske dehalogenaze (*Xanthobacter autotrophicus GJ10*) in njenih 15 enotočkovnih (pozicija 172) mutantov (16 spojin). Izmerjena je bila spektrofotometrično z detekcijo sproščenih halidnih ionov in je izražena kot konstanta prvega reda (k). V nasprotju s prejšnjimi množicami, kjer smo spojino opisovali z njenimi deskriptorji,

tu spojino opišemo z deskriptorji njenega sestavnega dela (aminokisline). To različico modeliranja QSAR imenujemo QSFR (Quantitative Structure-Function Relationships). Aminokisline so opisane s 33 deskriptorji. Poleg te množice sta bili uporabljeni še dve z zmanjšanim številom deskriptorjev (14 oz. 4).

6. *Aktivnost in stabilnost namensko spremenjenih proteinov* [5]. Obravnavane so štiri skupine proteinov, dobljenih s pozicijsko usmerjenimi mutagenimi poskusi. Prva skupina proteinov je ista kot v prejšnjji množici (mutanti haloalkanske dehalogenaze na poziciji 172 – Dhla-Phe172, 15 spojin), vendar so uporabljeni drugi deskriptorji za opis aminokislin. V drugi skupini so mutanti subtilizina na poziciji 222 (Subt-Met222, 19 spojin), tretjo skupino sestavljajo mutanti lizozima faga T4 na poziciji 157 (Lyso-Thr175, 13 spojin). V zadnji skupini so mutanti α-podenot triptofan sintaze, ki je proizvod bakterije *Escherichia coli,* na poziciji 49 (Synth-Glu49, 18 spojin). Aminokisline so opisane z 9 deskriptorji.

7. *Biorazgradljivost haloalkenov* [6]. Množica vsebuje 13 spojin. Merjenje biorazgradljivosti je bilo izvedeno s celicami bakterije *Rhodococcus erythropolis*, soj Y2, oz. haloalkansko dehalogenazo, ki so jo le te proizvedle. Spojine so opisane z 18 deskriptorji.

8. *Biorazgradljivost komercialnih kemičnih spojin* [8]. Degradacijske stopnje spojin so bile zbrane iz literature. Glavni vir je bil *Handbook of Environmental Degradation Rates* (Howard et al. 1991). Ta za primere, kjer ni izmerjenih podatkov, upošteva strokovne ocene, ki so lahko pristranske. Uporabljene so bile biodegradacijske stopnje spojin v površinskih vodah, kjer živi organizmi niso prilagojeni onesnaženju s proučevano spojino. Biorazgradljivost je bila izračunana kot naravni logaritem povprečja spodnjih in zgornjih ocen razpolovnega časa v urah (*HLT*). Množica vsebuje dve različni predstavitvi (P1 in P2) za 328 spojin. Razlikujeta se po deskriptorjih, s katerimi so opisane spojine. Prva (P1) je opisana z 31 deskriptorji: poleg molekulske teže (*mweight*) in hidrofobičnosti (*logP*) še prisotnost oz. število 29 podstruktur – funkcijskih skupin, ki so bile določene na podlagi predznanja o obravnavanem problemu. Deskriptorji druge množice (P2) so bili dobljeni s štetjem vseh podstruktur z dvema ali tremi atomi ter podstruktur s štirimi atomi zvezdaste topologije (brez verig). Upoštevane so bile vse podstrukture, ki so bile prisotne v vsaj treh spojinah, ne glede na njihov pomen. Deskriptorji množice P2 so število vsake od podstruktur ter *logP* in *mweight*, skupaj torej 61 deskriptorjev.

Vse množice lahko v grobem razdelimo na dve skupini. Prvih sedem množic podatkov vsebuje majhno število bolj ali manj sorodnih spojin, katerih aktivnost (večinoma biorazgradljivost) nas zanima. Zadnji dve množici (točka 8, P1 in P2) vsebujeta bistveno več zelo raznovrstnih spojin (alkoholov, fenolov, pesticidov, kislin, ketonov itd.). Vrednosti deskriptorjev (razen *logP* in *mweight*) za spojine teh množic so bile dobljene avtomatsko iz strukturnih opisov spojin in ne z merjenjem, kot za spojine množic iz prve skupine. Množice iz prve skupine so javno dostopne na spletu [14].

3. Poskusi

3.1 Metoda delnih najmanjših kvadratov

Največkrat pri modeliranju QSAR predpostavimo linearno odvisnost med deskriptorji (neodvisnimi spremenljivkami) in ciljno vrednostjo (odvisno spremenljivko). V tem primeru dobimo

sistem linearnih enačb: vsakemu učnemu primeru ustreza ena enačba, število členov v enačbah je za ena večje od števila deskriptorjev. Neznanke so parametri modela, ki ga iščemo. Če je število primerov (enačb) večje od števila deskriptorjev (kar skoraj vedno velja), lahko parametre modela izračunamo z metodo najmanjših kvadratov. Problem nastane, če so posamezni deskriptorji med seboj močno korelirani in postane tak izračun numerično neugoden. To je žal pri modeliranju QSAR pogosto. Problemu se lahko izognemo z metodo delnih najmanjših kvadratov (Partial Least Squares Method – PLS). Metoda temelji na analizi glavnih komponent (Principal Component Analysis), katere bistvo je, da prvotne deskriptorje nadomestimo s t.i. *glavnimi komponentami.* Glavne komponente so linearne kombinacije prvotnih deskriptorjev in so med seboj ortogonalne in nekorelirane. Metoda računanja glavnih komponent, uporabljena v postopku PLS, nam vrne glavne komponente, urejene po "pomembnosti". To pomeni, da vsebuje prva glavna komponenta največji del informacije, vsebovane v vseh deskriptorjih, zadnja pa najmanj. Večinoma dobimo boljše rezultate (modele), če ne upoštevamo vseh glavnih komponent, ampak le nekaj najpomembnejših. Optimalno število upoštevanih glavnih komponent določimo za vsak primer posebej s poskušanjem in z morebitnim dodatnim znanjem o problemu (npr. kateri deskriptorji bolj vplivajo na ciljno vrednost ipd.). To nam onemogoča avtomatično grajenje (dobrih) modelov. Opis celotnega postopka delnih najmanjših kvadratov lahko najdemo v [9].

3.2 Sistema Cubist in RETIS ter njune nastavitve

Za gradnjo regresijskih dreves smo uporabili metodi, implementirani v sistemih Cubist in RETIS. Prvi je naslednik sistema M5, opisanega v [11] in nadgrajenega z izboljšavami, opisanimi v [12]; zgrajeno regresijsko drevo nam vrne prepisano v pravila. Demonstracijska različica je na voljo na spletni strani podjetja *RuleQuest* (`www.rulequest.com`). Sistem RETIS je bil razvit na Institutu Jožef Stefan v Ljubljani in je opisan v [10].

Za množice z malo učnimi primeri sta bila z orodjem Cubist zgrajena po dva modela. Eden s privzetimi in eden z optimiziranimi parametri. Optimizirani parametri so tisti, pri katerih je imel z njimi dobljen model največji korelacijski koeficient prečnega preverjanja q; za vsako množico podatkov so različni. Za množici z večjim številom primerov sta bila zgrajena le modela s privzetimi parametri.

S sistemom RETIS je bilo za množice z malo primeri zgrajenih po šest modelov: z vključeno in izključeno linearno regresijo v listih dreves ter s tremi različnimi vrednostmi parametra m za naknadno rezanje dreves (0, 0.5 in 1). RETIS lahko upošteva največ 30 deskriptorjev, zato modeli za množice z večjim številom deskriptorjev niso bili zgrajeni. Ta omejitev je onemogočila modeliranje celotnih množic P1 in P2. Modeli so bili zato zgrajeni z deskriptorji, izbranimi na naslednji način. Iz vsake množice je bilo 10-krat naključno izbranih po 197 primerov (spojin). Za vseh 10 (pod)množic so bili zgrajeni modeli s sistemom Cubist. Vsi deskriptorji, ki so se vsaj enkrat pojavili v teh modelih, so bili nato uporabljeni za gradnjo regresijskih dreves s sistemom RETIS. Za tako dobljeni množici so bili zgrajeni po štirje modeli: z vključeno in izključeno linearno regresijo v listih dreves in z dvema različnima vrednostima parametra m za naknadno rezanje dreves. Prva vrednost parametra m je bila vedno 1, druga pa je bila interaktivno določena tako, da je imelo porezano drevo največ osem listov. Tako veliko drevo je namreč še mogoče strokovno interpretirati. V vseh primerih je bila vrednost parametra m za učenje enaka 0, najmanjše dovoljeno število primerov v listih drevesa pa 1.

3.3 Ocenjevanje veljavnosti modelov

Prvi pogoj za veljavnost modela je, da zadovoljivo opisuje primere iz učne množice. To pomeni, da se dejanske vrednosti odvisnih spremenljivk ne razlikujejo "preveč" od vrednosti, ki jih napove model. Mera za linearno odvisnost dveh nizov števil je korelacijski koeficient

$$r = \frac{\overline{(Y_d - \overline{Y_d})(Y_n - \overline{Y_n})}}{\sqrt{\overline{(Y_d - \overline{Y_d})^2}\,\overline{(Y_n - \overline{Y_n})^2}}}\,, \tag{1}$$

kjer Y_d pomeni dejansko vrednost ter Y_n napovedano vrednost odvisne spremenljivke, črta nad izrazom pomeni njegovo povprečno vrednost. Korelacijski koeficient lahko zavzame vrednosti med -1 in 1. Za dober model si želimo vrednost r čim bliže 1. V modeliranju QSAR je razširjena še ena mera za "podobnost" dejanskih in napovedanih vrednosti. To je koeficient R^2 (multiple corelation coefficient, explained variance), podan z enačbo:

$$R^2 = 1 - \frac{\sum(Y_d - Y_n)^2}{\sum(Y_d - \overline{Y_d})^2}\,. \tag{2}$$

Zaloga vrednosti koeficienta R^2 je navzgor omejena z 1, kar pomeni popolno ujemanje dejanskih in napovedanih vrednosti.

Veliki vrednosti koeficientov r in R^2 še ne zagotavljata veljavnosti modela, saj upošteva-ta le učne primere. Navadno si niti ne želimo prevelikega ujemanja modela z učnimi primeri zaradi možnosti pretiranega prilagajanja (overfitting). Vrednosti spremenljivk učnih primerov so izmerjene in zato vsebujejo šum. Če bi z modelom dosegli popolno ujemanje, bi to pomenilo, da model opisuje tudi šum, česar pa ne želimo.

Drugi pogoj za veljavnost modela je, da nam da zadovoljive napovedi za neznane primere (eden od ciljev QSAR modeliranja) oz. da ima zadovoljivo napovedno moč. To smo ocenili s postopkom *prečnega preverjanja*, kjer sestavimo več spremenjenih učnih množic, tako da iz prvotne množice odstranimo manjšo skupino primerov. Vsak primer moramo izvzeti natanko enkrat. Uporabili smo t.i. prečno preverjanje "izloči enega" (leave one out): v vsaki novi množici manjka natanko en primer iz celotne učne množice. Za vsako tako dobljeno množico zgradimo model in z njim napovemo vrednost odvisne spremenljivke za primer(e), ki v tej množici ne nastopa(jo). Tako dobljene napovedane vrednosti primerjamo z dejanskimi in njihovo ujemanje ovrednotimo s korelacijskim koeficientom prečnega preverjanja q in koeficientom Q^2 (cross-validated multiple corelation coefficient, predicted variance), ki ju izračunamo z enačbama 1 in 2. Za veljavni model še velja, da so vrednosti r in q (ali R^2 in Q^2) približno enake, kar pomeni, da model ni preveč prilagojen učnim primerom. Več o preverjanju veljavnosti QSAR modelov najdemo v [13].

4. Rezultati

V tabeli 2 so zbrani korelacijski koeficienti PLS modelov in koeficienti modelov, zgrajenih s sistemom Cubist (s privzetimi parametri) in sistemom RETIS (z linearno regresijo v listih in brez nje ter z vrednostjo parametra za naknadno rezanje dreves m=1 oz. za množico P1: m=5 z regresijo in m=8 brez regresije ter za množico P2: m=11 brez regresije). Izbrane modele malih množic z dovolj veliko napovedno močjo (poudarjene vrednosti v tabeli 2) je strokovno ocenil J. Damborský. Izdelal je tudi vse klasične PLS modele, ki so bili uporabljeni za primerjavo. Modele množic P1 in P2 je strokovno ocenil B. Kompare.

Tabela 2: Korelacijski koeficienti zgrajenih modelov. Correlation coefficients of models built in this paper.

Model	PLS		Cubist				RETIS							
							Z regresijo				Brez regresije			
	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2	r	q	R^2	Q^2
1. Toksičnost ter biorazgradljivost anilinov in fenolov (7+8 spojin, 10 desk.)														
Toksičnost anilinov	–	–	0.00	-0.39	0.00	-0.96	1.00	0.24	1.00	-9.02	0.99	-0.11	0.75	-0.47
Toksičnost fenolov	0.96	0.83	0.83	-0.15	0.69	-0.44	1.00	0.96	1.00	0.18	0.98	-0.08	0.81	-0.25
Toksičnost anilinov in fenolov	–	–	0.51	0.05	0.26	-0.24	0.97	-0.39	0.94	-4.05	0.94	0.18	0.79	-0.10
Biorazgradljivost anilinov	0.95	0.89	0.97	-0.40	0.93	-0.77	1.00	-0.59	1.00	-135	0.85	0.55	0.68	0.28
Biorazgradljivost fenolov	0.99	0.93	0.98	0.48	0.96	0.16	1.00	-0.43	1.00	-670	0.95	-0.24	0.80	-0.91
Biorazgradljivost anilinov in fenolov	0.96	0.95	**0.98**	**0.91**	**0.96**	**0.82**	**1.00**	**0.92**	**1.00**	**0.84**	0.94	0.72	0.84	0.49
2. Akutna toksičnost nasičenih in nenasičenih alifatskih ogljikovodikov (19 spojin, 24 desk.)														
Haloakani (vsi desk.)	0.90	0.77	**0.92**	**0.79**	**0.84**	**0.61**	0.97	0.36	0.83	0.06	0.97	0.36	0.83	0.06
Haloakani (desk. MR in EE)	0.90	0.88	**0.93**	**0.83**	**0.86**	**0.65**	**0.98**	**0.93**	**0.95**	**0.85**	0.95	0.59	0.78	0.31
Haloakani in haloalkeni (desk. MR in EE)	0.42	0.30	0.71	0.62	0.51	0.38	0.87	0.35	0.75	-0.23	0.91	0.22	0.73	-0.10
Haloakani in haloalkeni brez dveh spojin (desk. MR in EE)	0.89	0.88	**0.94**	**0.92**	**0.88**	**0.84**	**0.98**	**0.94**	**0.97**	**0.88**	**0.96**	**0.81**	**0.88**	**0.65**
Haloakani in haloalkeni (desk. MR, EE, BO, Hf in CR)	0.85	0.68	0.71	0.62	0.51	0.38	**0.99**	**0.78**	**0.97**	**0.59**	0.95	0.36	0.79	0.07
3. Biorazgradljivost dioksinov in furanov (14 spojin, 50 desk.)														
Model z vsemi deskriptorji	0.94	0.78	**0.98**	**0.78**	**0.97**	**0.60**	–	–	–	–	–	–	–	–
Model s 15 deskriptorji	0.95	0.88	**0.93**	**0.82**	**0.85**	**0.64**	**1.00**	**0.89**	**1.00**	**0.71**	**0.96**	**0.75**	**0.88**	**0.55**
Model z 9 deskriptorji	0.94	0.92	**0.89**	**0.87**	**0.79**	**0.76**	1.00	0.55	1.00	-0.01	**0.96**	**0.75**	**0.88**	**0.55**
4. Biorazgradljivost haloalifatskih spojin (27 spojin, 9 desk.)														
Model z vsemi spojinami	0.34	0.20	0.55	0.12	0.30	-0.38	0.94	0.01	0.88	-2.60	0.97	0.19	0.89	-0.13
Model brez dveh spojin	0.92	0.87	**0.89**	**0.80**	**0.80**	**0.63**	**0.99**	**0.88**	**0.98**	**0.74**	0.96	0.74	0.88	0.54
5. Biorazgradljivost mutantov haloalkanske dehalogenaze (16 spojin, 33 desk.)														
Model z vsemi deskriptorji	0.50	0.35	0.84	0.35	0.71	-0.08	–	–	–	–	–	–	–	–
Model s 14 deskriptorji	0.86	0.60	–	–	–	–	1.00	0.14	1.00	-48.0	0.99	0.16	0.89	-0.30
Model s 4 deskriptorji	0.84	0.75	0.84	0.46	0.71	0.07	1.00	0.74	0.99	0.49	0.99	0.32	0.89	0.01
6. Aktivnost in stabilnost namensko spremenjenih proteinov (15, 19, 13 in 18 spojin, 9 desk.)														
Dhla-Phe172	0.83	0.77	0.95	0.60	0.91	0.28	0.99	0.76	0.98	0.19	0.99	0.34	0.89	0.08
Subt-Met222	0.86	0.81	0.70	0.25	0.46	0.01	0.99	0.45	0.98	-0.13	0.96	0.14	0.76	-0.28
Lyso-Thr175	0.87	0.85	0.93	0.70	0.87	0.47	0.99	0.58	0.98	-0.21	0.94	0.38	0.81	0.05
Synth-Glu49	0.76	0.71	**0.87**	**0.81**	**0.76**	**0.65**	0.99	0.67	0.97	0.33	**0.97**	**0.80**	**0.87**	**0.63**
7. Biorazgradljivost haloalkenov (13 spojin, 18 desk.)														
Edini model	0.92	0.81	0.88	-0.54	0.78	-1.37	1.00	0.10	1.00	-177	0.93	-0.55	0.80	-1.39
8. Biorazgradljivost komercialnih spojin (328 spojin, 31 in 61 desk.)														
P1	0.27	0.26	0.76	0.67	0.57	0.44	0.78	0.58	0.60	0.30	0.65	0.58	0.41	0.33
P2	0.36	0.35	0.77	0.63	0.59	0.38	–	–	–	–	0.69	0.61	0.46	0.36

Oglejmo si nekaj boljših zgrajenih modelov. Klasični model, dobljen z metodo PLS za biorazgradljivost anilinov in fenolov (množica iz točke 1)

$$\log k_b = -11.233\ r_w + 0.315\ pK_a - 12.738, \tag{3}$$

in model, zgrajen s sistemom Cubist

$$\log k_b = -10.5\ r_w + 0.328\ pK_a - 12.983, \tag{4}$$

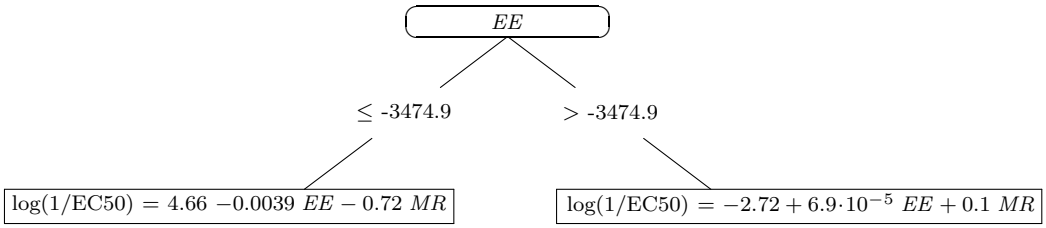
sta si zelo podobna. Oba vsebujeta iste deskriptorje, razlika je le v utežeh. Vrednost koeficienta R^2=0.98 (glej tabelo 2) sicer kaže na boljše ujemanje učnih primerov kot pri modelu PLS

(R^2=0.955), vendar smemo zaradi nižje vrednosti Q^2 (Cubist: 0.82; PLS: 0.949) pričakovati, da se bo model pri ocenjevanju novih primerov slabše obnesel kot model PLS.

Model, dobljen z metodo PLS za akutno toksičnost haloalkanov (množica iz točke 2, 2 deskriptorja), podaja naslednja enačba:

$$\log EC_{50}^{-1} = -0.0003\ EE + 0.0671\ MR - 2.6298, \tag{5}$$

z orodjem RETIS pa je bil dobljen model na sliki 1. Molekule razdeli glede na velikost (deskriptor *EE*). Napovedna moč obeh modelov je primerljiva (glej tabelo 2).



Slika 1: Model za akutno toksičnost haloalkanov (samo deskriptorja *MR* in *EE*), zgrajen s sistemom RETIS (vključena linearna regresija v listih, m=1). Model for acute toxity of haloalkanes (descriptors *MR* and *EE* only) built with RETIS system (linear regression in leaves turned on, m=1).

Referenčni model PLS za biorazgradljivost dioksinov in furanov (množica iz točke 3, 9 deskriptorjev) vsebuje vseh 9 deskriptorjev (enačba 6),

$$\begin{aligned}\log k = &- 0.1581\ \log P - 0.0030\ MM - 0.0053\ SA - \\ &- 0.0063\ MV - 0.0011\ IM2s - 0.0011\ IM3s - \\ &- 0.0184\ MR + 0.0002\ te + 0.2638\ dip + 6.6951,\end{aligned} \tag{6}$$

medtem ko model, zgrajen z orodjem Cubist (enačba 7), vsebuje le deskriptor *MV* (molekulski volumen). Zaradi svoje preprostosti je bil ta model kljub manjši napovedni moči ocenjen kot zelo dober; poudaril je deskriptor, o katerem je znano, da zelo vpliva na biotransformacijo dioksinov in furanov z bakterijo *Sphingomonas sp.*, soj RW1.

$$\log k = 7.651 - 0.0424\ MV \tag{7}$$

Za biorazgradljivost haloalifatskih spojin (množica iz točke 4, brez treh spojin) je referenčni model PLS podan z enačbo

$$\begin{aligned}Y2 = &- 0.1692\ Mw - 1.0745\ IX - 3.8575\ \log P + 0.0708\ Hf + \\ &+ 0.0059\ TE + 0.001\ EE - 37.817\ HOMO - 15.072\ Dip - \\ &- 673.68\ BCLU - 1581.5.\end{aligned} \tag{8}$$

Z orodjem Cubist smo tudi v tem primeru dobili preprostejši model z le dvema deskriptorjema:

$$Y2 = -679.3 + 0.0562\ TE - 76.9\ HOMO. \tag{9}$$

Celotna energija (*TE*) je verjetno povezana z velikostjo molekule (pomembno za vezavo substrata na aktivno mesto encima), medtem ko energija najvišje zasedene molekulske orbitale

(*HOMO*) opisuje elektronske lastnosti, pomembne za reakcijo dehalogenizacije. Model je preprost, kar je dobro za interpretacijo, vendar so njegove napovedi manj natančne kot napovedi kompleksnejšega modela PLS (enačba 8). Zaradi preprostosti je strokovnjakova ocena tega modela zelo dobra.

Za biorazgradljivost mutantov haloalkanske dehalogenaze (množica iz točke 5) z orodji za gradnjo regresijskih dreves nismo dobili uporabnih modelov. Vrednosti koeficientov R^2 (glej tabelo 2) so sicer visoke, vendar bistveno nižje vrednosti koeficientov Q^2 kažejo, da so modeli preveč prilagojeni učnim primerom.

Model, dobljen z metodo PLS za stabilnost mutantov α-podenot triptofan sintaze (množica iz točke 6, Synth-Glu49) podaja enačba

$$\Delta_d G \;=\; 32.263 - 24.036\ H311. \tag{10}$$

Ustrezni model, zgrajen z orodjem Cubist (enačba 11), je tako rekoč; nekoliko različni uteži sta vzrok za majhno razliko vrednosti koeficientov Q^2.

$$\Delta_d G \;=\; 32.29 - 24.1\ H311. \tag{11}$$

Orodja za gradnjo regresijskih dreves nam za biorazgradljivost haloalkenov (množica iz točke 7) niso dala uporabnega modela; vsi modeli so bili preveč prilagojeni učnim primerom.

S klasičnim postopkom PLS dobimo za biorazgradljivost komercialnih spojin (množici iz točke 8) modela, podana z enačbo 12 za P1 oz. 13 za P2.

$$\begin{aligned}\log HLT = \; & 5.651 + 0.101\ alkyl\_halide + 0.1286\ ar\_halide + \\ & + 0.1530\ benzene + 0.1615\ six\_ring + \\ & + 0.8462\ carbon\_5\_ar\_ring + 0.1679\ five\_ring + \\ & + 0.1107\ \log P + 0.001746\ mweight\end{aligned} \tag{12}$$

$$\begin{aligned}\log HLT = \; & 6.040 + 0.1542\ \log P + 0.002432\ mweight + \\ & + 0.1797\ c1cl - 0.1108\ c1o - 0.3223\ c2o - \\ & + 0.3059\ h1o + 0.09968\ c1n2o - 0.2960\ c1o1h\end{aligned} \tag{13}$$

Velja omeniti, da pri gradnji teh dveh modelov ni bilo upoštevano dodatno strokovno znanje o problemu, tako da bi bilo modele verjetno mogoče še izboljšati. S sistemom Cubist je bil za množico P1 zgrajen model na sliki 2. Uporabljeni so bili privzeti parametri. Zgrajeni model vsebuje relativno malo pravil (6) za tako raznoliko množico spojin, zato je bil ocenjen kot zelo dober. Med vsemi zgrajenimi modeli ima največjo napovedno moč (glej tabelo 2). Za množico P2 je bil s sistemom RETIS zgrajen model na sliki 3, ki se ujema s strokovnim znanjem o problemu. V primerjavi z modeli za druge množice podatkov imajo vsi modeli za množici P1 in P2 bistveno slabše vrednosti koeficientov R^2 in Q^2; eden od vzrokov za to je prav gotovo velika raznolikost modeliranih spojin.

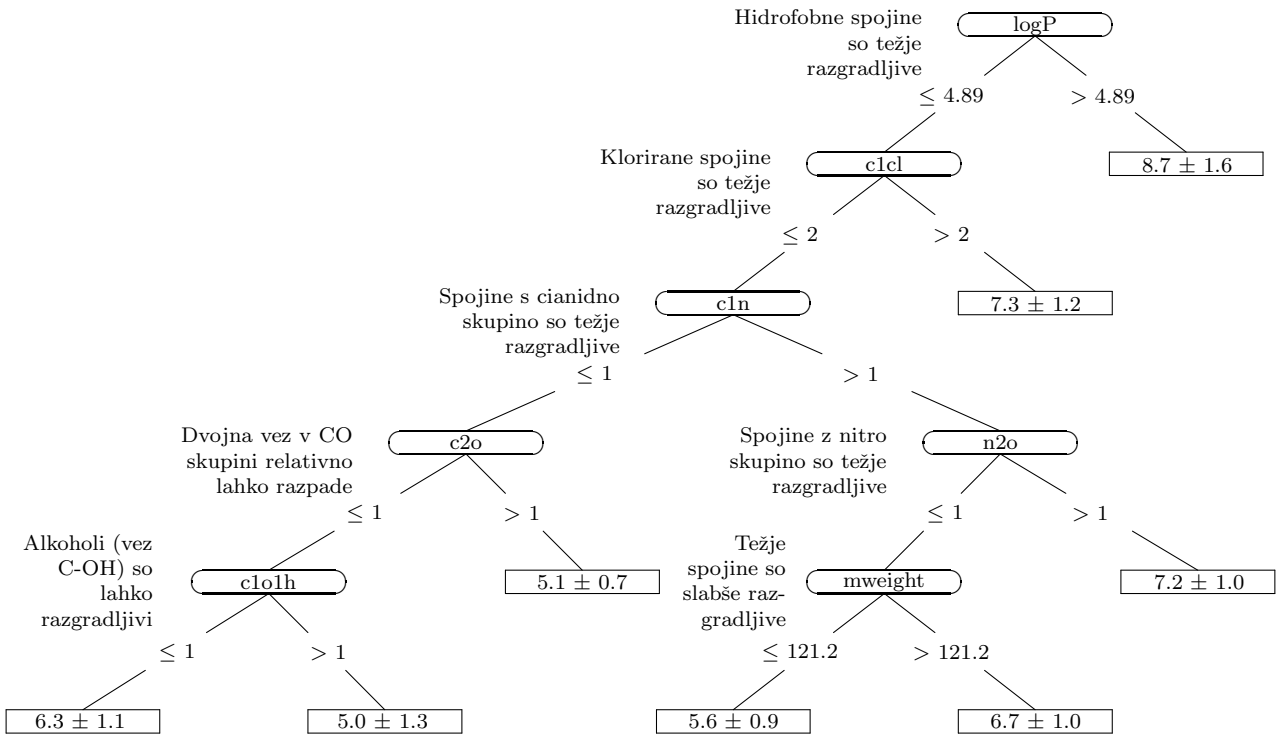
5. Sklep

Ugotovili smo, da je kakovost zgrajenega modela močno odvisna od števila učnih primerov, ki smo jih uporabili za njegovo gradnjo. Za velike učne množice dobimo z orodji za gradnjo

```
Rule 1: [95 cases, mean 5.630266, range 2.140066 to 7.822445, est err 1.210264]
  if  alkyl_halide <= 1        % Spojine brez halogenih elementov in z
      mweight <= 110.971       % manjso tezo so lazje razgradljive
  then  logHLT = 5.29 - 0.947 benzene + 0.00915 mweight - 1.22 ester + 0.15 logP
               - 0.76 phenol - 0.64 alcohol - 1.28 ketone - 0.75 aldehyde
               + 0.108 ar_halide + 0.052 alkyl_halide + 0.14 nitro
               + 0.046 six_ring + 0.07 amine - 0.09 non_ar_6c_ring
               - 0.09 carboxylic_acid + 0.05 methoxy + 0.08 imine
               + 0.013 methyl + 0.01 five_ring

Rule 2: [7 cases, mean 6.079297, range 6.040255 to 6.313548, est err 0.000741]
  if  alkyl_halide > 1         % Manj hidrofobne halogenirane spojine so
      logP <= 1.43             % lahko razgradljive
  then  logHLT = 5.494 + 0.273 alkyl_halide

Rule 3: [139 cases, mean 6.484813, range 4.533674 to 9.054388, est err 1.078450]
  if  alkyl_halide <= 1
      logP <= 4.84
      mweight > 117.107        % Spojine z vecjo tezo so nekoliko tezje razgradljive
  then  logHLT = 6.248 + 0.43 ar_halide + 0.42 amine + 0.54 nitro
               - 0.00195 mweight + 0.052 alkyl_halide + 0.046 six_ring
               + 0.021 logP - 0.11 phenol - 0.08 ester - 0.09 non_ar_6c_ring
               + 0.05 methoxy - 0.018 benzene + 0.08 imine
               - 0.06 carboxylic_acid - 0.08 aldehyde + 0.013 methyl
               + 0.01 five_ring

Rule 4: [41 cases, mean 7.551500, range 4.564348 to 9.768354, est err 1.767965]
  if  alkyl_halide > 1         % Skupina halogeniranih spojin s hidrofobnostjo
      logP > 1.43              % znotraj dolocenega intervala
      logP <= 4.84
  then  logHLT = 7.515 - 0.00647 mweight + 0.34 alkyl_halide + 0.142 ar_halide
               + 0.129 six_ring + 0.059 logP - 0.098 benzene + 0.2 nitro
               - 0.26 non_ar_6c_ring + 0.12 amine - 0.15 phenol + 0.12 methoxy
               + 0.21 imine - 0.1 ester - 0.21 aldehyde + 0.032 methyl
               + 0.01 five_ring

Rule 5: [5 cases, mean 7.831836, range 7.822445 to 7.869402, est err 0.035290]
  if  alkyl_halide <= 1        % Zelo ozko definirana skupina sorodnih
      mweight > 110.971        % biorazgradljivih spojin (verjetno gre za izomere)
      mweight <= 117.107
  then  logHLT = 7.832

Rule 6: [41 cases, mean 8.624575, range 5.81413 to 11.27416, est err 2.209445]
  if  logP > 4.84              % Hidrofobne spojine so tezje razgradljive
  then  logHLT = 8.385 - 0.549 methyl + 0.145 alkyl_halide + 0.144 ar_halide
               - 0.0014 mweight + 0.101 six_ring + 0.049 logP + 0.17 nitro
               - 0.2 non_ar_6c_ring + 0.08 amine + 0.046 benzene - 0.15 phenol
               - 0.1 ester + 0.11 methoxy + 0.19 imine + 0.07 five_ring
               - 0.19 aldehyde
```

Slika 2: Cubistov model za množico P1. Model for P1 dataset built with Cubist.

regresijskih dreves natančnejše modele kot s klasično linearno regresijo. Njihova prednost pride še posebej do izraza pri modeliranju množice različnih vrst spojin, ker nam drevo omogoča za vsako vrsto spojin svoj model. Pri modeliranju manjših množic podatkov so regresijska drevesa primerljivo ali manj natančna kot lineara regresija. Kljub temu regresijska drevesa, zgrajena iz majhnih učnih množic, niso neuporabna. Dobili smo namreč nekaj modelov, ki so se ob slabši natančnosti odlikovali po svoji preprostosti in razumljivosti, zato so bili zelo dobro ocenjeni. Pri tem ne gre pozabiti, da je s stališča uporabnika gradnja regresijskih dreves z omenjenimi orodji, bistveno hitrejša in preprostejša kot uporaba metode PLS linearne regresije. Pri slednji gre za interaktiven postopek modeliranja, ki zahteva veliko strokovnega znanja, medtem ko strojno učenje poteka avtomatično, ko imamo že pripravljene podatke.



Slika 3: Regresijsko drevo, zgrajeno s sistemom RETIS (brez linearne regresije v listih) za množico P2. Regression tree built with the RETIS system (no linear regression in leaves) for dataset P2.

6. Zahvala

Zahvaljujeva se dr. Jiřiju Damborskýju za izdelavo modelov PLS in komentar nekaterih zgrajenih modelov ter prof. dr. Borisu Komparetu za komentar modelov množic P1 in P2.

Literatura

[1] L. Blaha, J. Damborský in M. Nemec. QSAR for acute toxicity of saturated and unsaturated halogenated aliphatic compounds. V *Chemosphere*, št. 36, strani 1345–1365, 1998.

[2] J. Damborský, K. Manova in M. Kuty. A mechanistic approach to deriving QSBR - A case study: dehalogenation of haloaliphatic compounds. V W.J.G.M. Peijnenburg in J. Damborský, *Biodegradability Prediction*. Kluwer Academic Publishers, Dordrecht, 1996.

[3] J. Damborský. Quantitative structure-function relationships of the single-point mutants of haloalkane dehalogenase: A multivariate approach. V *Quantitative Structure-Activity Relationships*, št. 16, strani 126–135, 1997.

[4] J. Damborský in T. W. Schultz. Comparison of the QSAR models for toxicity and biodegradability of anilines and phenols. V *Chemosphere*, št. 34, strani 429–446, 1997.

[5] J. Damborský. Quantitative structure-function and structure-stability relationships of purposely modified proteins. V *Protein Engineering*, št. 11, strani 21–30, 1998.

[6] J. Damborský, A. Berglund, M. Kuty, A. Ansorgova, Y. Nagata, in M. Sjostrom. Mechanism-based Quantitative Structure-Biodegradability Relationships for hydrolytic dehalogenation of chloro- and bromo-alkenes. V *Quantitative Structure-Activity Relationships*, št. 17, strani 450–458, 1998.

[7] J. Damborský, M. Lynam in M. Kuty. Structure-biodegradability relationships for chlorinated dibenzo-p-dioxins and dibenzofurans. V R.-M. Wittich, *Biodegradation of dioxins and furans*. R.G. Landes Company, Austin, 1998.

[8] S. Džeroski, H. Blockeel, B. Kompare, S. Kramer, B. Pfahringer in W. Van Laer. Experiments in predicting biodegradability. V *Proceedings of the Nineth International Workshop on Inductive Logic Programming*, strani 80–91. Springer, Berlin, 1999.

[9] P. Geladi in B. R. Kowalski. Partial Least-Squares Regression: A Tutorial. V *Analytica Chimica Acta*, št. 185, strani 1–17, 1986.

[10] A. Karalič. *Avtomatsko učenje regresijskih dreves iz nepopolnih podatkov*. Magistrsko delo, Fakulteta za elektrotehniko in računalništvo, Ljubljana, 1991.

[11] J. R. Quinlan. Learning with continuous classes. V *Proceedings AI'92*, strani 343–348. World Scientific, Singapore, 1992.

[12] J. .R. Quinlan. Combining instance-based and model-based learning. V *Proceedings of the Tenth International Conference on Machine Learning*, strani 236–243. Morgan Kaufmann, San Fransisco, 1993.

[13] S. Wold. Validation of QSAR's. V *Quantitative Structure-Activity Relationships*, št. 10, strani 191–193, 1991.

[14] `www.chemi.muni.cz/~jiri/`. Spletna stran z množicami podatkov in njihovimi viri.

Bernard Ženko je diplomiral leta 2000 na Fakulteti za elektrotehniko Univerze v Ljubljani, smer Avtomatika. Trenutno je podiplomski študent na Fakulteti za računalništvo Univerze v Ljubljani. Njegovi raziskave potekajo na področju kombiniranja različnih metod za strojno učenje in uporabe strojnega učenja za analizo podatkov o okolju.

Doc. dr. Sašo Džeroski je višji znanstveni sodelavec Odseka za inteligentne sisteme Instituta Jožef Stefan ter pridruženi profesor Šole za znanosti o okolju Politehnike Nova Gorica. Bil je gostujoči raziskovalec na Turingovem inštitutu v Glasgowu (Velika Britanija), na Katoliški univerzi v Leuvenu (Belgija) ter na GMD (Nemški nacionalni raziskovalni center za informacijske tehnologije), Sankt Augustin. Njegov poglavitni raziskovalni interes je na področju strojnega učenja in njegove uporabe na praktičnih problemih analize podatkov, še posebej podatkov o okolju. Je znanstveni koordinator evropske mreže odličnosti na področju induktivnega logičnega programiranja (ILPnet2). Je soavtor oz. sourednik šestih knjig in zbornikov na področju strojnega učenja, izdanih pri uglednih tujih založbah.