# Napovedovanje biorazgradljivosti z regresijskimi drevesi

Bernard Ženko
Fakulteta za elektrotehniko
Univerza v Ljubljani
Tržaška 25, 1001 Ljubljana, Slovenija
*bernardze@yahoo.com*

Sašo Džeroski
Oddelek za inteligentne sisteme
Institut Jožef Stefan
Jamova 39, 1001 Ljubljana, Slovenija
*saso.dzeroski@ijs.si*

**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 lahko zgradimo s klasično metodo linearne regresije ali z metodami strojnega učenja; običajno so to metode za gradnjo regresijskih dreves. Ta dva tipa modelov sta bila primerjana v tem prispevku. Za več različnih množic podatkov smo zgradili modele z orodjema za gradnjo regresijskih dreves Cubist in RETIS. Vsi zgrajeni modeli so bili prečno preverjeni; najboljše med njimi sta pregledala strokovnjaka s področja biorazgradljivosti. Za majhne množice strukturno sorodnih spojin so modeli zgrajeni z linearno regresijo običajno bolj točni kot modeli z regresijskimi drevesi, čeprav imajo slednji včasih primerljivo točnost in so lažje razumljivi. Za večje množice strukturno različnih spojin so modeli z regresijskimi drevesi bolj točni kot linearni regresijski modeli.*

## 1 Uvod

Biorazgradljivost spojine nam pove nam kako hitro se spojina v okolju razgradi na neškodljive snovi. Običajno je podana v obliki razpolovnega časa. Na hitrost razgradnje spojine gotovo vpliva njena struktura oz. z njo povezane lastnosti. V splošnem so spojine s strukturami, ki nastopajo tudi v naravnih spojinah, bolje razgradljive od spojin s strukturami, ki so v naravi neznane. S pomočjo QSAR modeliranja lahko raziščemo povezave med strukturo spojin ter 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 želimo, da bi bili povezani z mehanizmom proučevane aktivnosti.

Pri QSAR modeliranju ločimo dve stopnji. Prva je razvoj modela na osnovi 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) [4] s katero dobimo model v obliki ene linearne enačbe. V zadnjem času se uporabljajo tudi metode strojnega učenja [6], običajno 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 je zanimala točnost in uporabnost modelov z regresijskimi drevesi v primerjavi z modeli zgrajenimi s klasično PLS metodo.

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

## 2 Množice podatkov

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

1. Toksičnost ter biorazgradljivost anilinov in fenolov (15 spojin, 9 deskriptorjev).

2. Akutna toksičnost nasičenih in nenasičenih ali-

fatskih ogljikovodikov (19 spojin, 24 deskriptorjev).

3. Biorazgradljivost dioksinov in furanov (14 spojin, 55 deskriptorjev).

4. Biorazgradljivost haloalifatskih spojin (27 spojin, 9 deskriptorjev).

5. Biorazgradljivost mutantov haloalkanske dehalogenaze (16 spojin, 33 deskriptorjev aminokislin).

6. Aktivnost in stabilnost 4. skupin namensko spremenjenih proteinov (15, 19, 13 in 18 spojin, 9 deskriptorjev aminokislin).

7. Biorazgradljivost haloalkenov (13 spojin, 18 deskriptorjev)

8. Biorazgradljivost komercialnih kemičnih spojin opisanih z dvema različnima skupinama deskriptorjev (328 spojin, 31 deskriptorjev – P1 in 61 deskriptorjev – P2).

Prvih sedem množic podatkov (točke 1–7) je javno dostopnih na spletnem naslovu [5]. Vsebujejo majhno število bolj ali manj sorodnih spojin, katerih aktivnost (večinoma biorazgradljivost) nas zanima. Deskriptorji posameznih množic so različni; anilini in fenoli iz prve množice so npr. opisani z naslednjimi deskriptorji: *HOMO* – energija najvišje zasedene molekulske orbitale, *LUMO* – energija najnižje nezasedene molekulske orbitale, $r_w$ – Van der Waalsov radij, $V_w$ – Van der Waalsov volumen, *Dip* – dipolni moment, $M_w$ – molekulska teža, *sigma* – Hammettova sigma konstanta, $pK_a$ – ionizacijska konstanta, $\log K_{ow}$ – logaritem koeficienta 1-oktanol/vodne razdelitve. Poleg modelov na osnovi celotnih množic, so bili za te množice zgrajeni tudi nekateri modeli na osnovi njihovih izpeljank, ki vsebujejo zožen nabor deskriptorjev in/ali primerov in so bile dobljene na osnovi klasičnega QSAR modeliranja. Zadnji dve množici (točka 8, P1 in P2) so pripravili avtorji članka [6]. Vsebujeta bistveno več zelo raznovrstnih spojin (alkoholov, fenolov, pesticidov, kislin, ketonov, itd.). 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*, skupno torej 61 deskriptorjev.

## 3 Poskusi

Uporabili smo metodi gradnje regresijskih dreves, implementirani v sistemih Cubist in RETIS. Prvi je naslednik sistema M5 opisanega v [1] in nadgrajenega z izboljšavami opisanimi v [2]. 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 [3].

Za množice z malo učnimi primeri sta bila z orodjem Cubist zgrajena po dva modela. Eden s privzetimi ter 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 ter z dvema različnima vrednostima parametra $m$ za naknadno rezanje dreves (zaradi programskega hrošča v programu RETIS ni bilo mogoče zgraditi modela za množico P2 z vključeno regresijo v listih drevesa.). 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.

Poleg tega, kako točno model napoveduje vrednosti učnih primerov, nas pri oceni veljavnosti modelov zanima predvsem kako točno napoveduje vrednosti neznanih primerov. Točnost na učnih primerih nam podajata koeficienta $r$ in $R^2$. Prvi je korelacijski koeficient med dejanskimi (izmerjenimi) vrednostmi odvisne spremenljivke učnih primerov in vrednostmi, ki jih napove model. Vrednosti, ki jih lahko zavzame so med -1 in 1. Koeficient $R^2$ (Multiple Corelation

| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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** | | | | | | | | | | | | | | |
| 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 |

Tabela 1: Korelacijski koeficienti zgrajenih modelov.

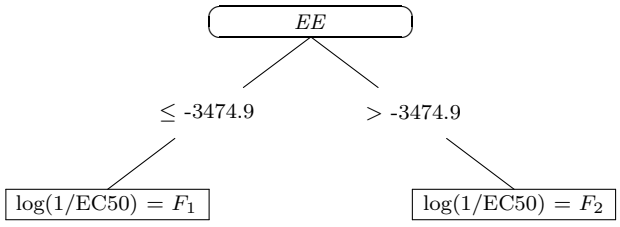
Coefficient, Explained Variance) je podan z enačbo:

$$R^2 = 1 - \frac{\sum (Y_d - Y_n)^2}{\sum (Y_d - \overline{Y}_d)^2}\,, \qquad (1)$$

kjer $Y_d$ pomeni dejansko vrednost ter $Y_n$ napovedano vrednost odvisne spremenljivke. Njegova zaloga vrednosti je navzgor omejena z 1, ki pomeni popolno ujemanje dejanskih in napovedanih vrednosti. Točnost napovedi modela na neznanih primerih (njegovo napovedno moč) smo ocenili s t.i. „izloči enega“ prečnim preverjanjem (Leave One Out Cross-Validation). Pri „izloči enega“ prečnem preverjanju sestavimo več spremenjenih učnih množic, tako da iz prvotne množice odstranimo en primer. Število teh spremenjenih učnih množic je enako številu učnih primerov (število primerov v teh množicah je za 1 manjše od števila primerov v prvotni učni množici). Za vsako tako dobljeno množico, zgradimo model in z njim napovemo vrednost odvisne spremenljivke za primer, ki v tej množici ne nastopa. 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). Koeficienta sta analogna koeficientoma $r$ in $R^2$ (za izračun uporabimo isti enačbi).

## 4 Rezultati

V tabeli 1 so zbrani korelacijski koeficienti PLS modelov ter koeficienti modelov, zgrajenih s sistemom Cubist (s privzetimi parametri) in sistemom RETIS (z in brez linearne regresije v listih 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). Iz-




$$F_1 = 4.66 - 0.0039\ EE - 0.72\ MR$$
$$F_2 = -2.72 + 0.000069\ EE + 0.1\ MR$$

Slika 1: Model za akutno toksičnost haloalkanov (samo deskriptorja *MR* in *EE*) zgrajen s sistemom RETIS, z vključeno linearno regresijo v listih, m=1.

brane modele malih množic z dovolj veliko napovedno močjo (poudarjene vrednosti v tabeli 1) je pregledal J. Damborsky in jih strokovno komentiral. Modele množic P1 in P2 je strokovno komentiral B. Kompare. Vse klasične PLS modele, ki so bili uporabljeni za primerjavo, je izdelal J. Damborsky.

Na tem mestu si oglejmo le dva zgrajena modela. Na sliki 1 vidimo model za akutno toksičnost haloalkanov zgrajen s sistemom RETIS. Na osnovi istih podatkov je bil s PLS metodo zgrajen model:

$$\log EC_{50}^{-1} = -0.0003\ EE + 0.0671\ MR - 2.6298. \quad (2)$$

Za biorazgradljivost dioksinov in furanov smo z sistemom Cubist dobili model:

$$\log k = 7.651 - 0.0424\ MV. \quad (3)$$

Kot vidimo vsebuje le en deskriptor, medtem ko analogni PLS model (enačba 4) vsebuje 9 deskriptorjev.

$$\begin{aligned}\log k \ = & -0.1581\ \log P - 0.0030\ MM - 0.0053\ SA - \\ & - 0.0063\ MV - 0.0011\ IM2s - 0.0011\ IM3s - \\ & - 0.0184\ MR + 0.0002\ te + 0.2638\ dip + 6.6951 \quad (4)\end{aligned}$$

Zaradi svoje enostavnosti je bil ta model, kljub manjši napovedni moči od PLS modela, ocenjen kot zelo dober.

## 5 Zaključki

Ugotovili smo, da je kvaliteta zgrajenega modela močno odvisna od števila učnih primerov, ki smo jih uporabili za njegovo gradnjo. Za velike učne množice dobimo z orodji za gradnjo 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 primerljive ali slabše natančnosti od linearne regresije. 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, zaradi česar 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 PLS metode 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.

Primerjava točnosti modelov, zgrajenih z obema orodjema pokaže, da je Cubist v rahli prednosti. Po drugi strani pa imamo pri sistemu RETIS večjo kontrolo nad gradnjo in predvsem nad naknadnim rezanjem dreves.

## 6 Zahvala

Delo opisano v tem prispevku je bilo opravljeno v okviru diplomske naloge Bernarda Ženka na Fakulteti za elektrotehniko Univerze v Ljubljani pod mentorstvom prof. dr. Nikole Pavešića in somentorstvom doc. dr. Saša Džeroskega. Zahvaljujem se dr. Jiřiju Damborskyju za izdelavo PLS modelov in komentar nekaterih zgrajenih modelov ter prof. dr. Borisu Komparetu za komentar modelov množic P1 in P2.

## Literatura

[1] J. R. Quinlan. Learning with continuous classes. V Adams, Sterling (ured.), *Proceedings AI'92*, strani 343–348. World Scientific, Singapore, 1992.

[2] 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.

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

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

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

[6] S. Džeroski, H. Blockeel, B. Kompare, S. Kramer, B. Pfahringer, W. Van Laer. Experiments in predicting biodegradability. V *Proceedings of the 9th International Workshop on Inductive Logic Programming*, strani 80–91. Springer, Berlin, 1999.