Sveučilište u Zagrebu
Fakultet Elektrotehnike i Računarstva

Ivan Sović, Matija Piškorec, Igor Čanadi

Nova robusna metoda za QSAR analizu temeljena na multivarijatnoj regresiji i normi L_1

3. svibnja 2010.

Ovaj rad izrađen je na Zavodu za elektroničke sustave i obradbu informacija Fakulteta elektrotehnike i računarstva pod vodstvom doc. dr. sc. Mile Šikića i dr. sc. Bone Lučića s "Instituta Ruđer Bošković" i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2009./2010.

Sadržaj

Popis slika

Popis tablica

Popis simbola

Slijedi popis korištenih kratica i simbola.

Skraćenice:

A-D	Anderson-Darling
K-S	Kolmogorov-Smirnov
LOO	Leave-One-Out
MLE	Maximum Likelihood Estimate
MLR	Multiple Linear Regression
PSA	Polar Surface Area
SAR	Structure-Activity Relationship
SLR	Simple Linear Regression
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship

D Statistika testa, stranica 26

d Gustoća spoja, stranica 21

E_i Pogreška, stranica 15

I Indikatorska varijabla, stranica 21

n Refraktivni indeks, stranica 21

$PRESS$ Prediktivna residualna suma kvadrata, stranica 25

q, q^2 Unakrsno validirani korelacijski koeficijent, stranica 24

r, r^2 Koeficijent korelacije, stranica 24

W_i Težinski koeficijent u težinskom medijanu, stranica 6

Poglavlje 1

Uvod

Utjecaj molekularne strukture na aktivnost pojedinih bioloških sustava danas još nije poznata, te stoga predstavlja područje velikog interesa za znanstvenike iz raznih područja: biologije, kemije, medicine, računarstva (bioinformatika, kemoinformatika)... Kada bi se otkrila točna povezanost između utjecaja spojeva i njihovih kemijskih struktura, imali bismo priliku stvoriti lijekove za velik broj bolesti te poboljšati kvalitetu življenja općenito. Formalni razvoj ovih veza predstavlja i osnovu razvoja prediktivnih modela (7). Ako uzmemo niz kemijskih spojeva i pokušamo im pronaći kvantitativnu vezu između strukture opisane strukturnim parametrima i bioloških efekata koje uzrokuju, u mogućnosti smo izgraditi model kvantitativne povezanosti strukture i aktivnosti, za što se rabi engleska kratica - *QSAR* (eng. *Quantitative Structure-Activity Relationship*).

Vrlo važan razlog za razvoj i primjenu ovakvih modela mogućnost je dobivanja saznanja o vrijednosti/iznosu biološki/farmakološki važnih svojstava molekule, a da fizički tu molekulu ne moramo imati (tj. da je ne moramo sintetizirati). Ova se istraživanja značajno financiraju iz državnih i europskih fondova (2), a njihovo nastavljanje i unapređenje podupiru i razni ekološki pokreti i pokreti za zaštitu životinja. Naime, povećanjem učinkovitosti metoda *in silico* u procjeni i predviđanju svojstava molekula, smanjuje se potreba za pokusima na životinjama u farmaceutskim istraživanjima, a mogu se dodatno predvidjeti neželjeni i toksični utjecaji molekula na okoliš. Iznesena je optimistična procjena vezana uz moguću korist od razvoja, unapređenja i primjene QSAR metoda - očekuje se kako bi značajnija upotreba tehnologija in silico mogla smanjiti troškove istraživanja lijekova čak do 50% (16). Metode za predviđanje toksičnosti i kancerogenosti primjenjuju se i u

analizi molekula u farmaceutskoj industriji, ali i u industriji herbicida i pesticida, kao i u međunarodnim regulatornim agencijama, agencijama vlada zemalja ili Europske unije. U posljednjih 20 godina redovito se izvode projekti u kojima se izrađuju pravila koja se moraju slijediti u razvoju i provjeri kvalitete modela QSAR/QSPR (eng. *Quantitative Structure-Property Relationship*), kako bi se standardizirala njihova uporaba i definirala pouzdanost za krajnjeg korisnika (1), (3).

Najveći broj QSAR istraživanja proveden je linearnim statističkim metodama (7), kao što je linearna regresija. Linearni su pristupi u ovom slučaju vrlo prikladni za upotrebu, jer je teorija izbora i dobivanja većine važnih atributa znatno razvijena za njihovu primjenu.

Najznačajniji utjecaj na kvalitetu rezultata QSAR modela temeljenog na regresiji imaju sami ulazni podatci. Zbog toga je potrebno konstruirati kvalitetan, robustan model koji će biti otporan na razne smetnje. Robusne metode su pogotovo bitne za QSAR analizu jer su manje osjetljive na povremene ekstremne devijacije (eng. *outliers*) u podatcima koje mogu nastati zbog loše kalibracije mjernih uređaja, same težine i nepouzdanosti mjerenja aktivnosti i farmakološki važnih svojstava molekula u (ili na) biološkim uzorcima kao i nepreciznih proračuna kojima se kvantificiraju i računaju strukturna svojstva (molekularni deskriptori) na temelju kemijske strukture molekula.

L_1 multivarijatna regresija je estimacija najveće vjerojatnosti (eng. *Maximum Likelihood Estimate*, *MLE*) pravca koji linearizira podatke s Laplaceovom distribucijom. U slučaju da distribucija podataka uistinu više naginje prema Laplaceovoj nego Gaussovoj (za koju je MLE L_2 linearna regresija, tj. minimizacija zbroja kvardratnih odstupanja) prikladnija je primjena L_1 regresije u odnosu na L_2 zbog manjeg utjecaja ekstremnih devijacija, što je jedno od svojstava Laplaceove razdiobe.

Dva ključna dijela rada upravo su:

1. Ispitivanju distribucije ulaznih podataka,
2. Usporedbi dobivenih rezultata s drugim metodama.

Pretražujući dostupne literature i znanstvene radove nismo naišli na postojeća rješenja QSAR analize primjenom L_1 regresije, te stoga ovim radom predlažemo novu robusnu metodu QSAR modeliranja.

U poglavlju L_1 *multivarijatna regresija* opisani su, poboljšani i prošireni algoritmi L_1 regresije te dana usporedba s L_2 regresijom. Poglavlje *QSAR modeliranje* daje pregled osnovnih pojmova i metoda, s naglaskom na robusne metode, dok se u poglavlju L_1 *multivarijatna regresija kao metoda za QSAR analizu* opisuje struktura ulaznih podataka te načini njihove analize. Zadnja cjelina rada, *Rezultati* sadrži analizu ulaznih podataka, provjeru hipoteze o

distribuciji ulaznih podataka, podatke o pogreškama i usporedbu s drugim metodama, te unakrsnu validaciju rezultata. Skup molekula na kojima je provedeno modeliranje sastoji se od 256 triazinskih molekula s molekularnim deskriptorima kojima je izmjeren inhibicijski učinak na stanice leukemijskih tumora. Taj je skup poslužio za uvođenje nelinearnih metoda (neuronskih mreža) u QSAR modeliranju, a skup je rabljen i znatno ranije od Corwina Hanscha, utemeljitelja modernih QSAR analiza (12).

Poglavlje 2

L_1 multivarijatna regresija

U ovom poglavlju dat ćemo kratki uvod u L_1 linearnu regresiji te povijesni pregled razvoja u kojem je presudnu ulogu igrao upravo hrvatski znanstvenik Ruđer Bošković.

Od mnogo algoritama koji rješavaju L_1 regresiju, izložit ćemo jedan od novijih, Li-Arce algoritam, koji rješava dvodimenzionalan slučaj. Algoritam ćemo ubrzati te prikazati usporedbu brzina s osnovnim algoritmom.

Proširit ćemo Li-Arce algoritam na višedimenzionalni slučaj. Taj algoritam nije optimalan, no predložit ćemo njegovo poboljšanje korištenjem *random restart* metode.

Bitno je odmah na početku napomenuti kako smo radili s linearnim ulaznim podatcima, te stoga u radu opisujemo metode *linearne* multivarijatne (ili višestruke) regresije. Naravno, prezentiramo li sustavu nelinearne kombinacije ulaza, poopćili smo metode te ne govorimo više samo o *linearnoj* regresiji. Iz tog razloga u radu se mogu naći izrazi kao što su "linearna regresija", "multivarijatna regresija" ili čak samo "regresija", a pri tome se svi odnose na upravo opisano.

2.1 Utjecaj Ruđera Boškovića

Prva definicija nekog L_1 kriterija potječe još iz 18. stoljeća, oko pola stoljeća prije nego su Gauss i Legendre razvili metodu za rješenje problema L_2 linearne regresije. Tada je Ruđer Bošković definirao kriterij za odabir linije koja najmanje odstupa od skupa točaka u ravnini. Kratki pregled razvoja L_1 regresije može se naći u (23).

Bošković definira L_1 kriterij na sljedeći način: Ako je $(\bar{x}, \bar{y})$ centroid navedenih točaka, potrebno je odabirati parametar c koji minimizira kriterijsku funkciju

$$\sum_{i=1}^{n} |y_i - \bar{y} - c(x_i - \bar{x})| \tag{2.1}$$

U tom slučaju parametar c definira liniju koja prolazi kroz centroid točaka i koja minimizira L_1 kriterij. Primjećujemo da je gornji kriterij jako sličan općenitom kriteriju za L_1 regresiju 2.3 definiranom u poglavlju 2.2.

Nekoliko godina nakon toga Bošković je predložio i jednostavan geometrijski algoritam za nalaženje parametra c. Taj algoritam bio je polazišna točka Laplaceu koji je 1789. razradio njegovu algebarsku varijantu. Sljedeći pomak stigao je s Edgeworthom koji je predložio općenitu numeričku metodu za rješavanje L_1 problema koja se temeljila na nalaženju *težinskog medijana* u svakoj iteraciji.

2.2 Ideja L_1 linearne regresije

Jednostavni L_1 linearni regresijski problem definiran je na sljedeći način. Promotrimo N parova (X_i, Y_i) koji su modelirani pomoću

$$Y_i = aX_i + b + U_i, \quad i = 1, 2, 3, ..., N \tag{2.2}$$

gdje su a i b nagib i pomak pravca, a U_i pogreška modelirana pomoću Laplaceove distribucije $f(U) = (1/2\lambda)e^{-|U|/\lambda}$ s varijancom $\sigma^2 = 2\lambda^2$. Problem L_1 linearne regresije je iz parova točaka (X_i, Y_i) pronaći odgovarajuće parametre a i b tako da greška,

$$F(a, b) = \sum_{i=1}^{N} |Y_i - aX_i - b|, \tag{2.3}$$

bude najmanja moguća. Ta je funkcija konveksna i kontinuirana (23).

Poopćavanjem sustava 2.2 (uz zanemarenje pogreške) na više dimenzija (multivarijatni slučaj), dolazimo do modela koji se sastoji od M ulaznih varijabli $X_{i,j}, j \in [1, M]$ i jedne izlazne varijable Y. Modeli linearne regresije također mogu imati i više od jednog izlaza, što je ovisno o konkretnoj primjeni modela. U QSAR analizi potreban nam je samo jedan izlaz kojeg procjenjujemo, te stoga Y nije poopćen, kao što je to bio slučaj s ulazima

X. Proširenu jednadžbu možemo zapisati u obliku:

$$Y_i = \sum_{j=1}^{M} k_j X_{i,j} + k_0, \quad i = 1, 2, 3, ..., N \tag{2.4}$$

gdje je koeficijent b zamijenjen s k_0, a koeficijent a proširen na dimenziju ulaza X i preimenovan u k_j.

Izraz 2.4 može se intuitivno proširiti u oblik s pogreškom U_i:

$$Y_i = \sum_{j=1}^{M} k_j X_{i,j} + k_0 + U_i, \quad i = 1, 2, 3, ..., N \tag{2.5}$$

2.2.1 Usporedba s L_2 linearnom regresijom

Druga, češće korištena regresija je L_2 linearna regresija, odnosno metoda najmanjih kvadrata (eng. *least squares, LS*). Razlika je u kriterijskoj funkciji, koja se definira na sljedeći način:

$$F(a, b) = \sum_{i=1}^{N} (Y_i - aX_i - b)^2, \tag{2.6}$$

Razlog češćeg korištenja L_2 linarne regresije je taj što za nju postoji jednostavno analitičko rješenje koje se može lagano implementirati. Ona je pogodnija za modele u kojima greške slijede normalnu (Gaussovu) razdiobu. Međutim, u praksi se često događa da distribucija greške ima deblji rep, tj. da više nalikuje na Laplaceovu razdiobu (vidi 5.2). Također, takozvani outlieri, točke s velikim odstupanjem od modela, znatno utječu na rješenje L_2 linearne regresije (vidi 3.3). Na skupu s takvim uzorcima pogodnija je L_1 linearna regresija (10).

2.3 Metode L_1 regresije

Postoji mnogo metoda računanja L_1 regresije. Budući da ne postoji analitičko rješenje problema, sve metode se zasnivaju na iterativnom pristupu. Neke najbitnije su:

- Metode bazirane na *Simplex* algoritmu (18)
- Wesolowsky-ev algoritam spuštanja (26)
- Li-Arce-ov algoritam baziran na *MLE*-u (eng. *Maximum Likelihood Estimation*) (4)

- Metoda središnje točke (29)

Trenutno najpopularnije metode su bazirane na Simplex algoritmu, budući da rade brže od ostalih, iako su implementacijski zahtjevnije.

2.4 Li-Arce algoritam

Jedan od novijih algoritama predložili su 2003. godine Yinbo Li i Gonzalo R. Arce (4). Oni su povezali problem L_1 regresije i problem procjene lokacije s maksimalnom vjerojatnošću. Opisat ćemo taj problem.

Neka je V slučajna varijabla definirana kao $V = U + \mu$ gdje je μ nepoznata konstantna lokacija, a U se ponaša po Laplaceovoj razdiobi. MLE lokacije na skupu $V_i|_{i=1}^N$ je:

$$\mu* = \arg\min_{\mu} \sum_{i=1}^{N} |V_i - \mu|. \tag{2.7}$$

Rješenje problema je zapravo medijan početnog skupa:

$$\mu* = MED\left(V_i|_{i=1}^N\right) \tag{2.8}$$

Rješenje MLE-a u kojemu se μ množi nekom konstantom rješava se analogno pomoću težinskog medijana, gdje je težina konstanta s kojom se množi μ, te su sve vrijednosti u skupu podijeljenje tom konstantom:

$$\mu* = \arg\min_{\mu} \sum_{i=1}^{N} |V_i - W_i \cdot \mu| \Rightarrow \mu* = MED\left(|W_i| \diamond \frac{V_i}{W_i}\Big|_{i=1}^{N}\right) \tag{2.9}$$

Operator $\diamond$ označava da je lijevi operand (u našem slučaju W_i) težina uzorka, odnosno desnog operanda. Detalji o računanju težinskog medijana mogu se pronaći u poglavlju 2.4.1.

Pomoću MLE problema, možemo optimizirati jedan od parametara a i b, tako da drugi parametar držimo konstantim. Intuitivno rješenje se nalaže samo od sebe: naizmjenice jedan parametar ugađamo, a drugi držimo konstantnim. Ako s a_k označimo vrijednost parametra a u k-tom koraku, te analogno za b, možemo pisati:

$$F(a_{k-1}, b_{k-1}) \geq F(a_{k-1}, b_k) \geq F(a_k, b_k) \tag{2.10}$$

Budući da u svakom koraku tražimo najbolju vrijednost jednog parametra, s obzirom na fiksiran drugi, sigurni smo da će funkcija cijene kroz korake biti nerastuća. To nas može dovesti do zaključka da će funkcija završiti u svom globalnom minimumu. Međutim, pokazalo se da funkcija može zaglaviti te algoritam neće pronaći minimalno rješenje.

Rješenje koje su predložili Li-Arce je da se parametri uvijek ugađaju duž nekog pravca. Naime, funkcija cijene je konveksna (23) te je svaki njen brid definiran parom točaka (X_i, Y_i) u parametarskom prostoru. Nadalje, vrhovi konveksnog lika definiranog funkcijom cijene su sjecišta tih pravaca koji definiraju bridove. Minimum se nalazi upravo u jednom od tih vrhova, odnosno, u prostoru uzoraka, rješenje L_1 regresije će prolaziti kroz dvije točke (20). Spuštanjem po bridovima izbjegli smo opasnost da se algoritam zaglavi, te će sigurno završiti u globalnom minimumu.

Spuštanje po bridovima implementiramo pomoću koordinatnih transformacija, zadržavajući funkciju cijene ekvivalentnom. Ideja koordinatnih transformacija je da u svakom koraku radimo ugađanje po horizontalnom pravcu, s nagibom 0.

Li-Arce-ov algoritam je (4):

1. $k = 0$. Inicijaliziraj b_0 na rješenje LS regresije.

   Izračunaj a_0 pomoću težinskog medijana

   $$a_0 = MED\left(|X_i| \diamond \frac{Y_i - b_0}{X_i} \Big|_{i=1}^{N} \right). \tag{2.11}$$

   Zapamti index j na kojem je težinski medijan. U parametarskom prostoru točka (a_0, b_0) leži na pravcu s nagibom $-X_j$ i odsječkom Y_j.

2. Postavi $k = k+1$. Transformiraj koordinate, tako da pravac po kojem ugađamo bude paralelan s apscisom. Transformacije su:

   $$X_i' = X_i - X_j, \quad Y_i' = Y_i. \tag{2.12}$$

   $$a_{k-1}' = a_k, \quad b_{k+1}' = b_k + a_{k-1} X_j \tag{2.13}$$

3. Ugodi a' pomoću težinskog medijana:

   $$a_k' = MED\left(|X_i'| \diamond \frac{Y_i' - b_k'}{X_i'} \Big|_{i=1}^{N} \right). \tag{2.14}$$

   Zapamti indeks t na kojem je pronađen težinski medijan.

4. Vrati u originalne koordinate:

$$a_k = a'_k, \quad b_k = b'_k - a'_k X_j. \tag{2.15}$$

5. Postavi $j = t$. Ako je a_k jednak a_{k-1} s određenom tolerancijom, prekini izvođenje programa. Inače, vrati se na korak 2.

Lako se provjeri da je funkcija cijene ostala ista:

$$\begin{aligned} F'(a') &= \sum_{i=1}^{N} |Y'_i - a'X'_i - b'_k| \\ &= \sum_{i=1}^{N} |Y'_i - a(X_i - X_j) - (aX_j + b_k)| \\ &= \sum_{i=1}^{N} |Y_i - aX_i - b_k| = F(a) \end{aligned} \tag{2.16}$$

2.4.1 Računanje težinskog medijana

Težinski medijan

$$Y = MED\left(W_i \diamond X_i|_{i=1}^{N}\right), \tag{2.17}$$

s pozitivnim težinama W_i računa se ovako:

1. Izračunaj prag $W_0 = (1/2)\sum_{i=1}^{N} W_i$.
2. Sortiraj vrijednosti $X_{(1)}, ..., X_{(n)}$ uzlazno zajedno s pripadajućim težinama $W_{(1)}, ..., W_{(n)}$.
3. Zbrajaj težine $W_{(i)}$ po redu, počevši od $W_{(1)}$.
4. Težinski medijan je onaj $X_{(j)}$ za kojeg će nejednakost $\sum_{i=1}^{j} W_{(i)} \geq W_0$ biti prva zadovoljena.

Računanje težinskog medijana je vremenski najzahtjevniji dio Li-Arce-ovog algoritma. Oni u svom članku (4) navode računanje medijana u složenosti $O(n \log n)$, što daje izravna implementacija gornje definicije. Međutim, razvijeno je mnogo metoda koje omogućavaju računanje težinskog medijana u očekivanoj složenosti $O(n)$ (11). Jednu od njih smo mi implementirali te smo brzinu usporedili s Li-Arcovim računanjem medijana u složenosti $O(n \log n)$.

Algoritam koji smo implementirali bazira se na parcijalnoj primjeni *quick sort* algoritma. Quick sort odabire pivota, presloži sve elemente tako da

oni manji po vrijednosti budu lijevo, a veći desno od njega. Nakon toga, postupak se rekurzivno ponovi za desni i lijevi dio. Jednostavno možemo provjeriti nalazi li se težinski medijan u lijevom ili desnom dijelu. Ukoliko se nalazi u lijevom dijelu, desni dio više ne diramo, a postupak nastavljamo samo za lijevi dio. Analogno vrijedi u slučaju da se medijan nalazi u desnom dijelu. Ako pretpostavimo da se u svakom koraku niz dijeli na dva jednaka dijela, složenost je:

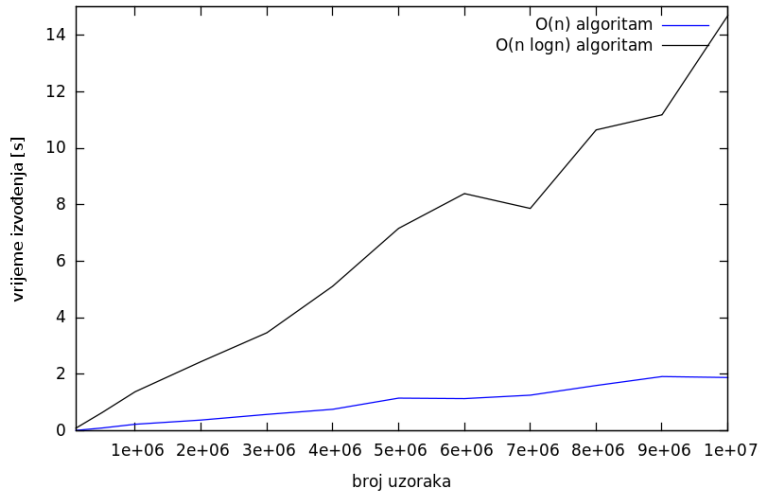
$$O\left(n + \frac{n}{2} + \frac{n}{2^2} + \frac{n}{2^3} + ...\right) = O\left(n \cdot \frac{1}{1 - \frac{1}{2}}\right) = O(n \cdot 2) = O(n) \qquad (2.18)$$

Najgori slučaj algoritma je $O(n^2)$, međutim u praksi gotovo uvijek završi u $O(n)$ vremenu, a pogotovo uz pametno odabiranje pivota - kao medijan prvog, srednjeg i zadnjeg elementa u nizu.

Pseudokod algoritma slijedi:

1. Izračunaj prag $W_0 = (1/2) \sum_{i=1}^{N} W_i$.
2. Postavi pivota kao medijana prvog, srednjeg i zadnjeg elementa u nizu.
3. Postavi l na početak niza, r na kraj niza.
4. Povećavaj l tako dugo dok je element na l-tom mjestu manji od pivota.
5. Smanjuj r tako dugo dok je element na r-tom mjestu veći od pivota.
6. Zamjeni elemente na l-tom i r-tom mjestu. Povećaj l za 1 i smanji r za 1.
7. Ponavljaj (4.-6.) tako dugo dok se l i r ne izjednače. Kad se to dogodi, onda su svi elementi lijevo od tog mjesta, sigurno manji ili jednaki pivotu, a desno veći ili jednaki pivotu.
8. Zbroji težine elemenata s lijeve strane $W_{lijevo} = (1/2) \sum_{i=1}^{l} W_i$.
9. Ako je W_{lijevo} veći ili jednak W_0, vrati se na 2. korak s lijevim dijelom niza. Inače, postavi $W_0 = W_0 - W_{lijevo}$ i vrati se na 2. korak s desnim dijelom niza.
10. Ako niz ima manje ili jednako 3 elemenata, pronađi težinski medijan jednostavnim postukom *selection sorta*.

Usporedba brzine izvođenja $O(n)$ i $O(n \log n)$ algoritma prikazana je na slici 2.1. Sva prikazana vremena nastala su kao aritmetička sredina tri pokretanja

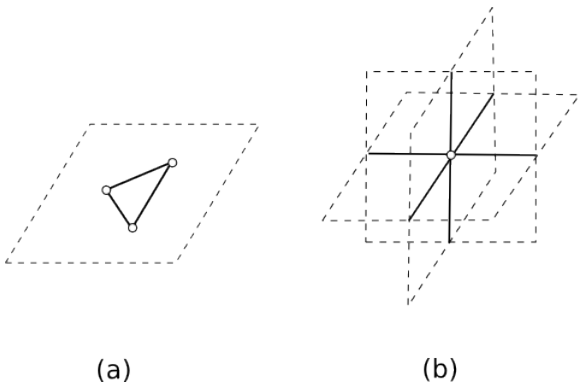


Slika 2.1: Usporedba brzine izvođenja $O(n)$ i $O(n \log n)$ algoritma.

algoritma na slučajno odabranim uzorcima na računalu s Intel Core2 Quad 2.40GHz procesorom te 4 GB radne memorije.

Postignuto je znatno ubrzanje koje nam je omogućilo efikasno proširivanje algoritma opisano u poglavlju 2.4.2.

2.4.2 Višedimenzionalni slučaj



Slika 2.2: Trodimenzionalni slučaj u L_1 regresiji. (a) prikazuje prostor uzoraka a (b) parametarski prostor. U dvodimenzionalnom slučaju u svakoj iteraciji je jednoznačno određeno po kojem pravcu će se optimizirati L_1 kriterij dok u trodimenzionalnom slučaju to više nije tako i treba se odlučiti za jednu od dvije ili više ravnina.

Li-Arce-ov algoritam je ograničen na dvije dimenzije. Mi smo razvili analogan algoritam na više dimenzija.

U višedimenzionalnom slučaju minimiziramo funkciju cijene:

$$F(\vec{a}, b) = \sum_{i=1}^{N} |Y_i - a_1 X_{i1} - a_2 X_{i2} ... - a_q X_{i2} - b| \tag{2.19}$$

Višedimenzionalan algoritam je:

1. Inicijaliziraj sve parametre na rješenje L_2 regresije.

   Izračunaj a_1 pomoću težinskog medijana

$$a_1 = MED \left(|X_{i1}| \diamond \frac{Y_i - a_2 X_{i2} - ... - a_q X_{iq} - b}{X_{i1}} \Bigg|_{i=1}^{N} \right). \tag{2.20}$$

   Zapamti index j na kojem je nađen težinski medijan. Postavi $k = 2$

2. Transformiraj koordinate

$$\vec{X_i}' = \vec{X_i} - \vec{X_j}, \quad \vec{Y_i}' = \vec{Y_i}. \tag{2.21}$$

$$b' = b + \vec{a}^T \vec{X_j} \tag{2.22}$$

3. Ugodi a'_k pomoću težinskog medijana:

$$a'_k = MED \left(|X'_{ik}| \diamond \frac{Y_i - \sum_{j=1, j\neq k}^{q} (a'_j X'_{ij}) - b'}{X'_{ik}} \Bigg|_{i=1}^{N} \right). \tag{2.23}$$

   Zapamti indeks t na kojem je pronađen težinski medijan.

4. Vrati u originalne koordinate:

$$b = b' - \vec{a}^T \vec{X_j}. \tag{2.24}$$

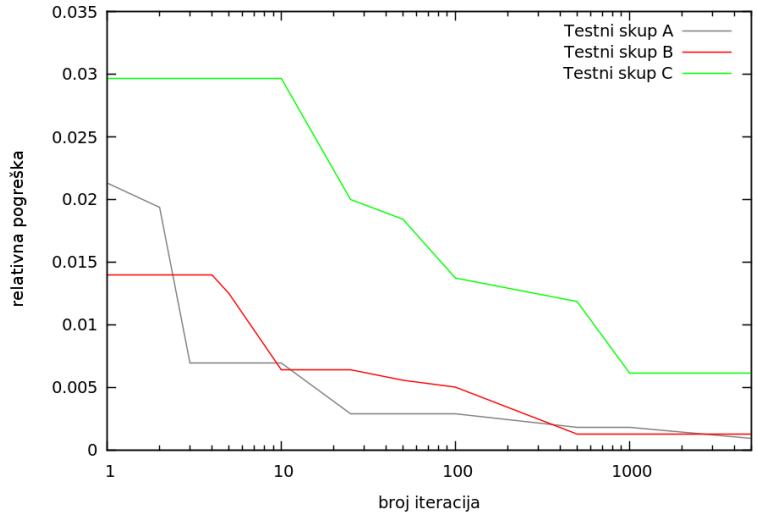
5. Postavi $j = t$. Ponavljaj korake $(2. - 5.)$ tako dugo dok se više ni jedan od parametara ne bude mogao ugoditi (nova i stara vrijednost budu jednake unutar određene tolerancije).

Umjesto ugađanja parametara po redu, mogu se ugađati i slučajno odabranim redosljedom. No, efikasnost u oba slučaja je približno ista.

Pokazalo se, očekivano, da Li-Arce algoritam završava u globalnom minimumu samo za dvodimenzionalni slučaj. Naš višedimenzionalni algoritam se često zaglavio u neminimalnim točkama. Prosječna relativna greška na tri trening skupa iz (13) je 0.0216756. Točnu minimalnu vrijednost dobili smo metodom *linearnog programiranja*. Informacije o ovoj metodi mogu se pronaći u (17).

Jedna od metoda kojom smo pokušali poboljšati algoritam je *random restart*. U svakoj iteraciji početna vrijednost parametara nam je rješenje L_2 regresije, s iznimkom da 3 slučajno odabrana parametra pomnožimo s uniformno odabranom slučajnom vrijednosti iz intervala $[-1.5, 1.5]$.

Ispitivanja su pokazala da se nakon određenog broja iteracija algoritam približava globalnom minimumu. Brzina konvergencije prikazana je na slici 2.3. Već nakon nekoliko iteracija postigli smo zanemarivo malu relativnu pogrešku, te na osnovu tih rezultata možemo zaključiti da je algoritam primjenjiv u praksi.



Slika 2.3: Relativna pogreška našeg algoritma nakon određenog broja iteracija. Algoritam je pokretan na trening skupovima iz (13).

Iako su postignuti vrlo dobri rezultati u razvoju algoritma za višestruku regresiju, u nastavku rada on ipak nije primjenjivan. Razlog tome je što su razvoj algoritma i ispitivanje razdioba distribucija rađeni istovremeno, te je algoritam prekasno dovršen kako bismo ga mogli efektivno iskoristiti za dobivanje rezultata izloženih u radu. Za računanje vrijednosti regresije

u nastavku korištena je implementacija temeljena na linearnom programiranju, a ispitivanje pogrešaka korištenjem opisanog algoritma predviđeno je u budućem radu na ovom projektu.

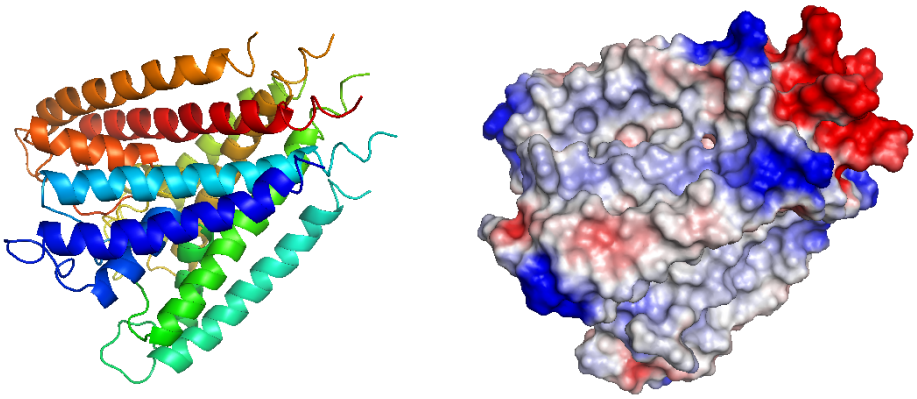
Poglavlje 3

QSAR modeliranje

QSAR danas je široko prihvaćen skup metoda koje polaze od osnovne ideje da postoji veza između kemijske strukture molekula i njihove biološke aktivnosti i da je tu vezu moguće kvantitativno opisati (5). Spada u skupinu takozvanih *in silico* metoda, a njenim razvojem i primjenom bave se različita pod-područja poput kemoinformatike, bioinformatike te teorijske i računalne biofizike, biokemije i biologije. C.D. Selassie daje dobar pregled povijesnog razvoja QSAR metoda u (21). Uvod u najnovije spoznaje i praktične primjene može se naći u (7).

Metode u QSAR-u mogu se dodatno podijeliti s obzirom na kompleksnost podataka na kojima se provodi izbor i optimizacija modela, a nakon toga i predviđanja (5):

- 2D-QSAR: Skup metoda koje su neosjetljive na konformacijske promjene atoma u prostoru.
- 3D-QSAR: Skup metoda koje za osnovnu informaciju uzimaju poziciju atoma u trodimenzionalnom prostoru.
- 4D-QSAR: Skup metoda gdje se za svaku molekulu nalazi skup prijanjajućih (eng. *docked*) orijentacija i konformacija.
- 5D-QSAR: Skup metoda koje iskorištavaju scenarije induciranog prijanjanja (eng. *induced fit*) na aktivno mjesto molekule.
- 6D-QSAR: Skup metoda koje u obzir uzimaju modele otapala u kojima se nalaze molekule.



Slika 3.1: Primjer različitih vrsta informacija pomoću kojih QSAR modeli donose svoje predikcije. Na slici lijevo vidi se sekundarna struktura 1BL8 kompleksa, a na slici desno elektrostatski potencijal na njegovoj površini. Slike su generirane uz pomoć programskog alata Pymol (8).

U tijeku su pojačani napori da QSAR postane relevantna metoda prilikom odlučivanja u legislativi, primjerice, prilikom donošenja odluka o toksičnosti pojedinih supstanci (14; 19; 27).

3.1 Metode u QSAR analizi

QSAR objedinjuje široki skup statističkih metoda za donošenje zaključaka o korelaciji strukture i aktivnosti molekula. Osnovni pregled može se naći u (21; 25; 28). Najvažnije su:

- Nadzirani prediktivni modeli
  - Višestruka linearna regresija (eng. *multiple linear regression (MLR)*)
  - Metoda parcijalnih najmanjih kvadrata (eng. *partial least squares regression (PLS)*)
  - Neuralne mreže (eng. *neural networks (NN)*)
  - Algoritam potpornih vektora (eng. *support vector machine (SVM)*)
  - Stabla odluke (eng. *decision trees*)
  - Genetski algoritmi (eng. *genetic algorithms*)
- Modeli raspoznavanja uzoraka
  - Analiza glavnih komponenti (eng. *principal component analysis (PCA)*)

- Nelinearno mapiranje (eng. *non-linear mapping (NLM)*)
- Analiza klastera (eng. *cluster analysis (CA)*)

U nastavku slijedi detaljniji opis višestruke linearne regresije.

Višestruka linearna regresija

Najčešće korištena matematička tehnika u QSAR analizi je višestruka linearna regresija (eng. *multiple linear regression, MLR*). Ona pretpostavlja da postoji linearna veza između skupa nezavisnih varijabli (eng. *independent variables*) koje opisuju strukturu molekule i zavisnih varijabli (eng. *dependent variables*) koje opisuju njihovu aktivnost.

U slučaju da imamo skup od N molekula koje opisujemo s M nezavisnih varijabli $X_{i,j}$ (struktura) i jednom zavisnom varijablom Y_i (aktivnost) linearni regresijski model možemo definirati kako je to prikazano izrazom 2.5.

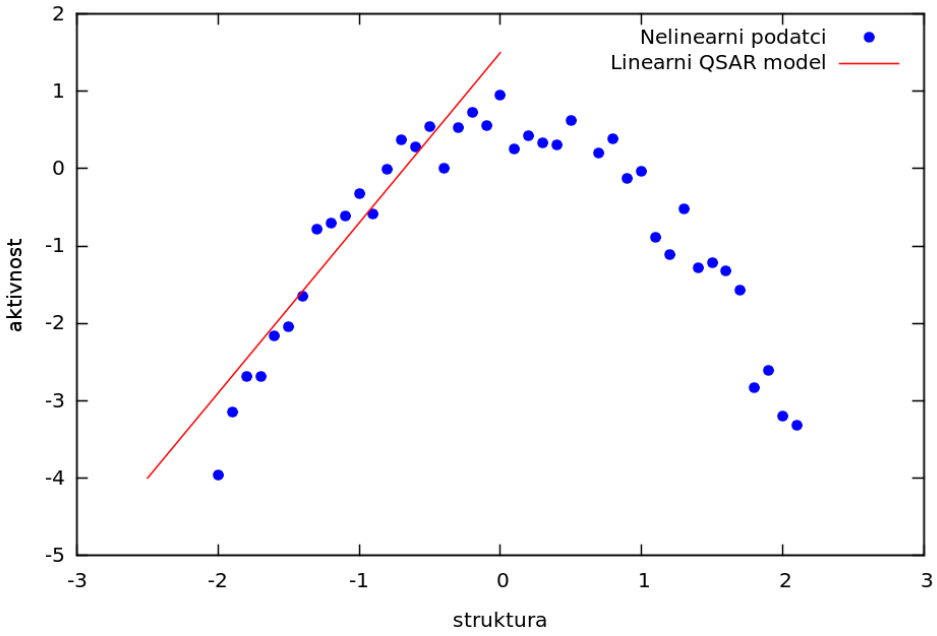
Kako bi ovakav model bio primjenjiv potrebno je usvojiti određene pretostavke (15):

1. Nezavisne varijable (fizikalno-kemijski parametri koji reprezentiraju strukturu molekule) su mjerene bez pogreške ili je barem ta pogreška zanemariva u odnosu na pogrešku zavisnih varijabli.
2. Za svaku danu vrijednost X vrijednost Y je nezavisna varijabla koja je normalno distribuirana. Pogreška U_i je također normalno distribuirana sa srednjom vrijednošću 0.
3. Očekivana srednja vrijednost varijable Y za sve vrijednosti leži na ravnoj liniji (hiperravnini). Ovo je takozvana *pretpostavka o linearnosti*.
4. Varijanca duž regresijske linije (hiperravnine) je konstantna.

U slučaju da neka od ovih pretpostavki nije zadovoljena potrebna su određena odstupanja od jednostavnog linearnog modela. Primjerice, *pretpostavka o linearnosti* se gotovo uvijek može uzeti za važeću ako se uzme dovoljno ograničen raspon parametara. Grafički prikaz se može vidjeti na slici 3.2. Također, kasnije ćemo vidjeti da pretpostavka o normalnoj distribuciji podataka nije zadovoljena na našem skupu podataka i da distribucija podataka naginje Laplaceovoj.

3.2 Molekularni deskriptori

Karakteristike molekula mogu se promatrati na različitim razinama pri čemu svaka nosi točno određenu vrstu informacija. Najjednostavnija reprezentacija



Slika 3.2: Shematski prikaz pretpostavke linearnosti. Nelinearnih podatci (plave točke) se unutar ograničenog raspona parametara opisuju linearnim modelom.

je kemijska formula molekule, no moguće su i kompleksnije reprezentacije - raspodjela elektrostatskog potencijala primjerice. Konačni rezultat logičkog i matematičkog postupka koji transformira kemijske informacije kodirane simboličkom reprezentacijom molekule u numeričku vrijednost naziva se kemijski deskriptor (6).

Za deskriptore su u početku bila korištena samo jednostavna globalna molekularna svojstva kao što su $\log P$, topljivost (eng. *solubility*) i ionizacijska konstanta (24). Naknadno su dodani i novi deskriptori zbog adresiranja specifičnog problema, ili kao bolji način enkripcije strukturalne informacije. Danas ih je poznato više od 3000. Mogu biti izračunati ili izmjereni eksperimentalno; dobiveni od 1D, 2D ili 3D reprezentacije molekule; ovisni o svojstvenim karakteristikama ili o vanjskom okruženju...

Glavne grupe molekularnih deskriptora su (25):

- Fizikalno-kemijski ($\log P$, $\log D$, molarna refraktivnost, pK_a, topljivost ...)
- Veličina/oblik (molekularne težine, momenti inercije ...)
- Topološki (indeksi povezanosti ...)

- Vodikove veze (broj vodikovih donora i akceptora, Abrahamovi α i β deskriptori ...)
- Elektrostatski (formalni naboj, parcijalni naboji atoma, dipolni momenti i vektori, HOMO i LUMO energija ...)
- Deskriptori temeljeni na prebrojavanju atoma određene vrste, funkcionalnih skupina ili strukturnih fragmenata

Svaki od ovih deskriptora obično je prikazan kao numerička vrijednost. Idealno, preferirani su oni deskriptori koje je lako fizikalno i strukturno interpretirati jer oni pružaju izravniju informaciju o mehanizmu aktivnosti.

Poseban je problem odlučiti koji skup deskriptora uključiti u QSAR analizu (25). Većina deskriptora je po prirodi korelirana što znači da skup sadrži redundantnu informaciju. Primjerice, molekularna težina, površina molekule i molarna refraktivnost uglavnom su jako korelirane pa je u tom slučaju opravdano u QSAR model uvrstiti samo jednu od njih. Korelacija, pa čak i ona slučajna, je to vjerojatnija što je veći omjer broja deskriptora naspram broja dostupnih uzoraka (5). Skup nekoreliranih svojstava je poželjan jer rezultira robusnim modelima koje je lakše interpretirati.

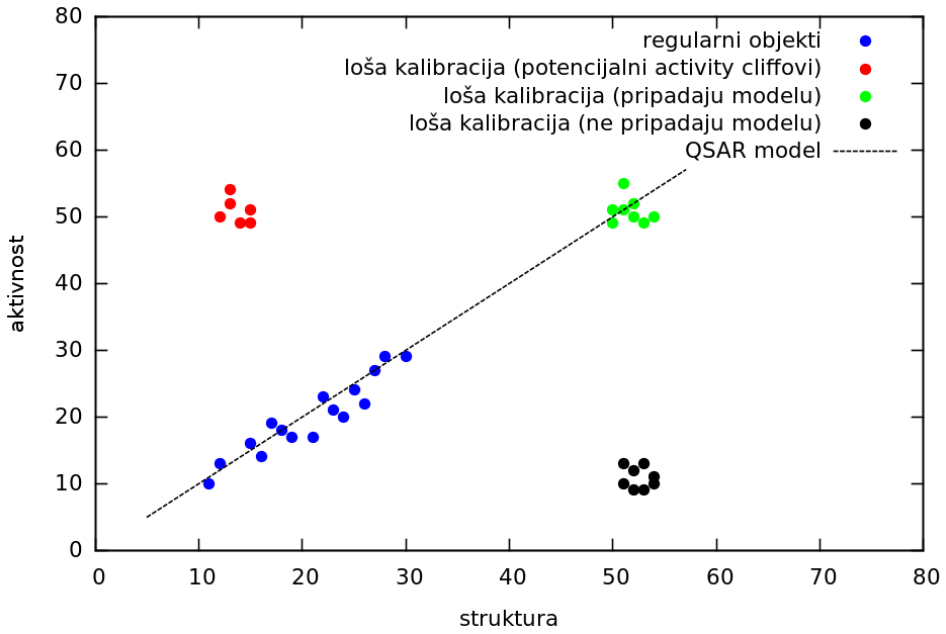
3.3 Robusne metode u QSAR modeliranju

U standardnoj QSAR terminologiji termin *robustan* se koristi za modele koji dobro predviđaju relativno širok spektar novih molekula. S druge strane, u terminologiji robusne statistike termin *robustan* je rezerviran za grupu modela koji pružaju dobra predviđanja za većinu podataka (22).

Najveći problem prilikom oblikovanja modela predstavljaju *outlieri* - opažanja koja svojim parametrima znatno odskaču od većine ostalih koje model pokušava opisati. S obzirom na nastanak, postoje dva osnovna tipa outliera (5): (i) *pravi outlieri* koji nastaju kao rezultat sustavne pogreške prilikom mjerenja ili proračuna nekog parametra i (ii) *prividni outlieri* koji nastaju zbog krivo modeliranih linearnih jednadžbi ili nelinearnosti u podatcima koja nije uzeta u obzir. Ovi potonji mogu pružiti korisnu informaciju o distribuciji podataka i ne smiju se zanemariti.

Dva su načina da se tretiraju outlieri (22). Prvi uključuje identifikaciju i sustavnu eliminaciju outliera iz podataka koji služe za modeliranje QSAR modela. Ovakav je pristup zavaravajući jer outlieri obično dosta snažno utječu na izgled samog modela korištenog u identifikaciji outliera[1]. Drugi

[1]Utjecaj pri tome imaju *masking* i *swamping* efekti. *Masking* efekt se događa kada u podatcima postoji relativno mnogo outliera koji si međusobno skrivaju utjecaj pri korištenje nerobustnih metoda. S druge strane, *swamping* efekt uzrokuje da se neki regularni uzorci identificiraju kao outlieri.



Slika 3.3: Razne vrste outliera koje se mogu naći u ulaznim podatcima. Nastaju kao rezultat loše kalibracije u mjerenju ili proračunu i dijele se na one izvan uobičajenog raspona parametara koji pripadaju (zeleni) ili ne pripadaju (crni) i na one unutar uobičajenog raspona parametara ali s izraženim visokim vrijednostima (crveni). Ovi potonji su značajni jer mogu ukazivati na skokove u aktivnosti (eng. *activity cliffs*) i mogu imati važnu ulogu u interpretaciji modela.

je pristup da se izgradi *robustan* model koji se dobro ponaša u prisutnosti outliera. U tom slučaju nije potrebno na umjetan način kompenzirati ili eliminirati utjecaj outliera.

Kako se L_1 linearna regresija dobro ponaša u prisutnosti outliera smatramo da čini dobar temelj za razvoj robusnog QSAR modela. U sljedećem poglavlju detaljnije razrađujemo moguću primjenu L_1 linearne regresije u QSAR analizi.

Poglavlje 4

L_1 multivarijatna regresija kao metoda za QSAR analizu

Svrha je ovog poglavlja pokazati metode kojima smo ispitali opravdanost upotrebe L_1 multivarijatne regresije u QSAR analizi. Pokrivena su područja strukture i analize ulaznih podataka, te metoda validacije rezultata. U sklopu potpoglavlja strukture ulaznih podataka prikazana je direktna veza ulaza i izlaza s matematičkim modelom. Osim toga, opisana je fizikalna reprezentacija ulaza i neki relevantniji deskriptori (koji se mogu naći u podatcima za provjeru). Prilikom analize ulaznih podataka, proveden je odabir relevantnih deskriptora, te su pomoću njih, kao ulaza u sustav regresije, izračunate pogreške. Opisan je postupak određivanja distribucije pogrešaka. Na kraju je dan pregled metoda validacije, kako bi se dobiveni rezultati mogli usporediti.

4.1 Struktura ulaznih podataka

Prije početka provođenja procesa računanja koeficijenata regresije po L_1 (pa tako i po bilo kojoj drugoj) normi, potrebno je razumjeti kako predočiti ulazne podatke. Njihovu strukturu najlakše je predočiti pogledamo li sam izraz za linearnu regresiju prikazan formulom 2.4. Možemo ga zapisati i u

matričnom obliku:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,M} \\ 1 & X_{2,1} & \cdots & X_{2,M} \\ \vdots & \ddots & \vdots & \\ 1 & X_{N,j} & \cdots & X_{N,M} \end{bmatrix} \begin{bmatrix} k_0 \\ k_1 \\ \vdots \\ k_M \end{bmatrix} \tag{4.1}$$

Može se vidjeti kako je izlaz Y dan u obliku vektora dimenzije N, čije vrijednosti predstavljaju izlaze sustava za svaku od obzervacija, dok je X matrica dimenzija NxM, i sadrži podatke o stanjima na svakom od M ulaza u svakoj od N obzervacija.

Bitno je napomenuti da smo za razliku od jednadžbe (2.5) u navedenim izrazima promatrali sustav bez pogreške. Utjecaj pogreške biti će detaljnije razmatran u narednim poglavljima.

Kako bismo mogli primjenjivati regresiju u QSAR analizi, ulazni podatci u sustav moraju sadržavati relevantne podatke. Ti podatci nazivaju se *deskriptori*, a predstavljaju realne brojeve koji opisuju kvantizirana svojstva molekula. Vektor ulaznih vrijednosti X svake obzervacije sadrži točno M ovakvih deskriptora, a vektor Y točnu (tj. uzorkovanu) ili predviđenu vrijednost modelirane veličine.

4.1.1 Korišteni QSAR deskriptori

Prilikom ispitivanja regresije po L_1 normi u QSAR analizi, češće smo se susretali sa skupinom deskriptora opisanih u nastavku.

Koeficijent razdjeljenja u sustavu oktanol-voda ($\log P$) Koeficijent razdjeljenja u sustavu oktanol-voda povezuje kemijsku strukturu s primjećenim ponašanjem kemikalije. $\log P$ je povezan s hidrofobnošću molekule (eng. *hydrophobic character*).

Molekularna masa (MW) Molekularna masa.

Molarna refraktivnost (MR) Molarna refraktivnost (MR) je mjera potpune polarizibilnosti jednog mola neke tvari, a definira se izrazom:

$$MR = \frac{n^2 - 1}{n^2 + 2} \frac{MW}{d} \tag{4.2}$$

gdje je n refraktivni indeks, MW molekularna težina, a d gustoća spoja.

Indikatorske varijable (I) Indikatorske varijable (I) su jedni od najjednostavnijih deskriptora, a koriste se kada se dani problem ne može

reprezentirati realnim numeričkim vrijednostima. I obično poprima pozitivnu, negativnu ili nultu cjelobrojnu vrijednost koja predstavlja stanje neke količine. (6)

Hidrofobnost (pi) Hidrofobnost (eng. *hydrophobic character*). Za određeni *substituent* X, *pi* označava razliku njegovog $\log P$ od vrijednosti $\log P$ za vodik.

Površina polarnog dijela molekule (PSA) Polarno područje površine (eng. *Polar Surface Area*) je definirano kao dio područja površine molekule koje međudjeluje s kisikom, dušikom, sumporom i vodikom vezanim za bilo koji od ovih atoma. Ovaj površinski deskriptor je povezan sa sposobostima spojeva da ostvaruju vodikove veze.

4.2 Analiza ulaznih podataka

Prije konstruiranja QSAR modela, potrebno je provesti analizu ulaznih podataka, s posebnim naglaskom na distribuciju pogreške. Prvo je potrebno odabrati relevantne deskriptore za provedbu postupka multivarijatne regresije, što se, na primjer, može izvesti jednostavnim izbacivanjem deskriptora čija vrijednost je konstantna u cijelom skupu ulaznih podataka, ili primjenom metoda kao što je unazadna eliminacija (eng. *backward elimination*). Nadalje, potrebno je odrediti nekoliko standardnih mjera za određivanje pogreške odnosno kvalitete rezultata regresije. Za ispitivanje distribucije pogrešaka moramo imati definirane statističke provjere, kao što su: *Kolmogorov-Smirnov*, *Anderson-Darling* i χ^2. Ispitivanje distribucije pogrešaka od ključne je važnosti za ovaj rad, jer pokazuje je li metoda L_1 multivarijatne regresije primjenjiva za rješavanje problema iz QSAR analize.

4.2.1 Odabir relevantnih deskriptora

Ako je prilikom modeliranja svojstva molekula broj deskriptora velik u odnosu na broj spojeva, od velike je važnosti napraviti odabir relevantnih deskriptora za razmatrani hipotetski mehanizam ili poznati farmakoforni model. (5) Na primjer, neki deskriptori su pogodniji za stanične ili *invivo* eksperimente zbog farmakokinetičkih procesa (difuzija, transport, $\log P$, pK_a, $\log D$, MW, V, PSA,...). Drugi su pak pogodniji u razmatranjima sličnosti lijekova-receptora (eng. *drug-receptor affinity studies*) (vodikove veze, parcijalni naboji, *pi* vrijednosti, reaktivne, prostorne i elektronske komponente,...).

Postoji nekoliko strategija za odabir odgovarajućeg podskupa deskriptora: unazadna eliminacija (eng. *backward elimination*), inverzna unaprijedna selekcija (eng. *inverse forward selection*), analiza glavnih komponenti, simulirano hlađenje, rangiranje po Bayes-u, evolucijski i genetički algoritmi, te

općenite tehnike raspoznavanja uzoraka. Zbog relativno malog broja deskriptora korištenih u ispitnim podatcima ovog rada, skup relevantnih deskriptora odabran je filtracijom onih koji zadovoljavaju barem jedan od uvjeta:

1. Vrijednosti deskriptora jednake su 0 za sve ulazne uzorke - u ovom slučaju jednadžbe regresije postaju linearno zavisne,
2. Vrijednosti deskriptora identične su za sve ulazne uzorke - u ovom slučaju deskriptor ne utječe na rješenje regresije, jer je jedan od parametara konstanta.

Bitno je napomenuti kako su indikatorske varijable vrlo značajne za točnost rezultata dobivenih primjenom regresije u QSAR analizi. Konkretnim ispitivanjem (rezultati nisu izneseni u radu) pokazalo se kako model koji ne koristi indikatorske varijable daje veću pogrešku u odnosu na model koji ih koristi. Do istog su zaključka došli i Hansch i Silipo u svom radu (12). Dodatno, zbog same definicije indikatorskih varijabli, pretpostavljamo kako one imaju velik utjecaj u tome da distribucija podataka više odgovara Laplaceovoj nego Gaussovoj, kao što će to biti pokazano u poglavlju 5.2.1.

4.2.2 Određivanje pogreške

Mogućnost primjene linearne regresije nad skupom ulaznih podataka povlači za sobom pretpostavku linearnosti u promjeni izlaznih podataka u odnosu na ulaze u sustav. Kako, osim u teoriji, promjena izlaza sustava nikad neće pratiti promjene ulaza savršeno po pravcu, definiramo mjeru pogreške. Pogreška predstavlja razliku točne [1] uzorkovane izlazne vrijednosti sustava od one dobivene uvrštavanjem ulaznih vrijednosti u izraz za linearnu regresiju. Pri tome su koeficijenti regresije prethodno određeni po jednoj od normi (npr. L_1, L_2,...), kao što je to opisano u poglavlju 2. Ilustracija pogreške prikazana je na slici 4.1.

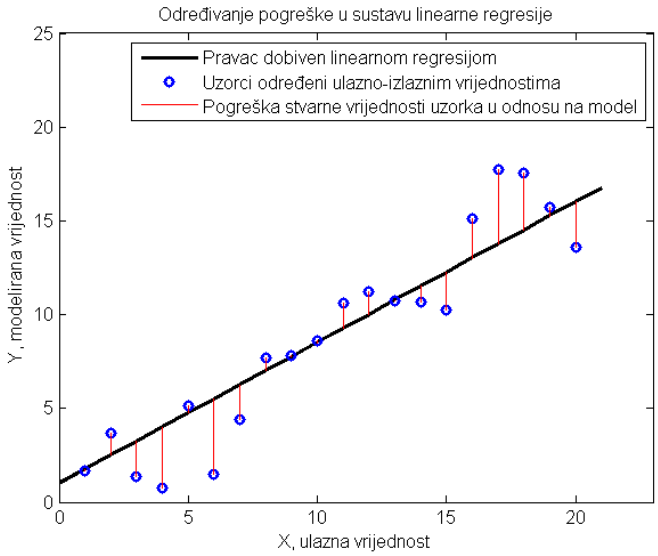
Dobivena pogreška je posljedica utjecaja vanjskog šuma prilikom mjerenja i uzorkovanja podataka, nesavršenosti mjernih instrumenata i samih svojstava opažanog procesa.

Ukupna pogreška

Pogreška i-tog uzorka definirana je izrazom:

$$E_i = Y_{uzork.,i} - Y_{predv.,i} = Y_{uzork.,i} - \vec{X}_i\vec{k}, \tag{4.3}$$

[1] Točnost je u ovom slučaju relativan pojam zbog utjecaja šuma prilikom mjerenja, preciznosti mjernih uređaja i aproksimacija prilikom obrade podataka.



Slika 4.1: Nesavršeno poklapanje uzoraka (točke) s pravcem dobivenim linearnom regresijom uzrokuje pogrešku kod regresije.

gdje je $Y_{uzork.,i}$ uzorkovana "točna" vrijednost modeliranog svojstva, $Y_{predv.,i}$ vrijednost svojstva predviđena preko modela dobivenog linearnom regresijom, a $\vec{X}_i$ vektor ulaznih vrijednosti i $\vec{k}$ vektor koeficijenata linearne regresije, kako je to opisano u poglavlju 4.1.

Ukupna pogreška jednaka je:

$$E = \sum_{i=1}^{N} E_i = \sum_{i=1}^{N} Y_{uzork.,i} - \vec{X}_i \vec{k}. \tag{4.4}$$

Srednja apsolutna pogreška

Srednja apsolutna pogreška (eng. *Mean Absolute Error*) jedna je od standardnih mjera pogreške, a definirana je izrazom:

$$MAE = \sum_{i=1}^{N} \frac{|Y_{uzork.,i} - Y_{predv.,i}|}{N} \tag{4.5}$$

Srednja kvadratna pogreška

Srednja kvadratna pogreška (eng. *Root Mean Square Error*) također je jedna od standardnih mjera pogreške, a računa se prema izrazu:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(Y_{uzork.,i} - Y_{predv.,i})^2}{N}} \tag{4.6}$$

Korelacijski koeficijent r^2

Pearsonov korelacijski koeficijent r statistička je mjera korelacije (linearne zavisnosti) između dvije varijable, a poprima vrijednost u granicama $[-1, +1]$. U procjeni kvalitete jednostavne i multivarijatne linearne regresije primjenjuje se kvadrat ove vrijednosti, a potpuni izraz za njegov proračun dan je sljedećom formulom (9):

$$r^2 = \frac{\sum_{i=1}^{N}(Y_{predv.,i} - \bar{Y}_{uzork.})^2}{\sum_{i=1}^{N}(Y_{uzork.,i} - \bar{Y}_{uzork.})^2}, \tag{4.7}$$

gdje je $\bar{Y}_{uzork.}$ srednja vrijednost izlaza dobivenih uzorkovanjem. Veća vrijednost r^2 odgovara većoj povezanosti varijacija podataka obzervacija s varijacijama nezavisnih varijabli. Statistika r^2 je vrlo korisna, ali ako se razmatra sama za sebe lako dovodi do pogrešnih zaključaka. Na primjer, za iste vrijednosti r^2 pravac dobiven linearnom regresijom može predstavljati najbolju linearnu reprezentaciju podataka, dok isto tako može predstavljati i pravac koji ne opisuje podatke dovoljno dobro, zbog utjecaja *outlier*-a ili drugih trendova unutar podataka.

Unakrsno validacijski korelacijski koeficijent q^2

Unakrsno validacijske metode omogućuju prevladavanje nekih problema svojstvenih pri korištenju samo r^2 mjere (9). Unakrsna validacija predstavlja proces izdvajanja nekih vrijednosti iz ukupnog skupa podataka, konstrukciju QSAR modela nad preostalim podatcima, te zatim korištenje tog modela za predviđanje izostavljenih podataka. Jedna od jednostavnijih metoda unakrsne validacije je $Leave-One-Out$ (LOO) gdje se se izdvaja samo jedan element iz skupa podataka. Ova metoda će biti detaljnije opisana u narednim poglavljima. Ponavljanjem ovog postupka za svaki podatak iz skupa, dolazimo do *unakrsno validirane* r^2 *mjere*, koja se uobičajeno označava s q^2. Vrijednost q^2 je obično manja od r^2. Konkretnije, unakrsno validacijski

korelacijski koeficijent definiran je izrazom:

$$q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{N}(Y_{uzork.,i} - \bar{Y}_{uzork.})^2}. \tag{4.8}$$

Pri tome $PRESS$ označava prediktivnu residualnu sumu kvadrata, danu izrazom:

$$PRESS = \sum_{i=1}^{N}(Y_{uzrok.,i} - Y_{predv.,i})^2 \tag{4.9}$$

4.2.3 Distribucija pogrešaka

Kako bismo opravdali našu pretpostavku o pogodnosti upotrebe L_1 norme za modeliranje regresije u QSAR analizi, potrebno je pokazati da podatci imaju zadovoljavajuću distribuciju pogreške. Način određivanja pogreške opisan je u poglavlju 4.2.2, a distribuciju određujemo na temelju izraza 4.3. Postoje brojni testovi za određivanje pripadnosti skupa podataka pojedinoj distribuciji (eng. *distribution fitting*), a u ovom radu korišteni su: *Kolmogorov-Smirnov*, *Anderson-Darling* i χ^2 testovi.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov (*K-S*) test primjenjuje se prilikom odlučivanja o tome dolazi li neki skup podataka iz određene distribucije. Temeljen je na empirijskoj distribucijskoj funkciji (eng. *empirical cumulative distribution function*, *ECDF*). Uz dani skup točaka $Y_1, Y_2, ..., Y_N$ iz neke distribucije s kumulativnom distribucijskom funkcijom (*CDF*) $F(Y_i)$, *ECDF* se definirana kao:

$$F_N(Y_i) = \frac{n(x)}{N}, \tag{4.10}$$

gdje je $n(x)$ broj točaka čija je vrijednost manja od Y_i, pri čemu su točke poredane po veličini od najmanje do najveće.

Dobro svojstvo *K-S* testa je to što sama distribucija njegove statistike ne ovisi o kumulativnoj distribucijskoj funkciji ispitivanih podataka. Druga prednost *K-S* testa je što je, za razliku od nekih drugih (npr. χ^2), to egzaktna provjera. *K-S* test također ima i neka važna ograničenja:

- Može se primjenjivati samo nad kontinuiranim distribucijama,
- Osjetljiviji je u centru distribucije nego na njezinim rubovima,

- Distribucija mora biti potpuno specificirana.

Kolmogorov-Smirnov test definiran je hipotezama:

- H_0 - nulta hipoteza, ako je prihvaćena podatci prate specificiranu distribuciju,
- H_A - alternativna hipoteza, ako je prihvaćena podatci ne prate specificiranu distribuciju,

te statistikom testa:

$$D = \max_{1 \leq i \leq N} \left(Y_i - \frac{i-1}{N}, \quad \frac{i}{N} - F(Y_i) \right). \tag{4.11}$$

Hipoteza vezana uz distribucijski oblik odbacuje se ako je statistika provjere D veća od kritične vrijednosti α očitane iz tablica.

Anderson-Darling test

Anderson-Darling test je modificirana verzija Kolmogorov-Smirnov testa. Za razliku od *K-S*, *A-D* test pridaje veću težinu rubovima distribucije. Osim toga, kod K-S testa, kritične vrijednosti ne ovise o konkretnoj distribuciji koja se ispituje, dok kod A-D ovise. To ima svoje prednosti i nedostatke. Nedostatak je da za svaku pojedinu distribuciju moramo posebno računati kritične vrijednosti, ali nam to omogućava osjetljiviji test.

Anderson-Darling test definiran je hipotezama:

- H_0 - nulta hipoteza, ako je prihvaćena podatci prate specificiranu distribuciju,
- H_A - alternativna hipoteza, ako je prihvaćena podatci ne prate specificiranu distribuciju,

te statistikom testa:

$$A^2 = -N - S, \tag{4.12}$$

gdje je

$$S^2 = \sum_{i=1}^{N} \frac{2i-1}{N} \left[\ln F(Y_i) + \ln \left(1 - F(Y_{N+1-i}) \right) \right], \tag{4.13}$$

a F je kulumativna distributivna funkcija specifične distribucije. Pri tome su podatci Y_i sortirani.

Ta statistika se uspoređuje s kritičnom vrijednošću α, koja ovisi o distribuciji koja se ispituje. Ukoliko je statistika veća od kritične vrijednosti, hipoteza se odbacuje.

χ^2 test

χ^2 test zahtijeva da se podatci prvo grupiraju. Broj podataka u određenoj grupi se tada uspoređuje s očekivanim brojem podataka u toj grupi, te je statistika definirana kao funkcija te razlike. Broj grupa i algoritam raspodjele podataka u grupe direktno utječu na svojstva testa. Iako ne postoji jasno definirana *najbolja metoda*, postoji nekoliko dobrih savjeta.

Jedan od savjeta je, na primjer, da se broj grupa postavi na $2n^{2/5}$. Također, u svakoj grupi ne bi trebalo biti manje od 5 podataka.

Za razliku od prije opisanih testova, χ^2 se može primijeniti i na diskretne distribucije.

χ^2 test definiran je hipotezama:

- H_0 - nulta hipoteza, ako je prihvaćena podatci prate specificiranu distribuciju,
- H_A - alternativna hipoteza, ako je prihvaćena podatci ne prate specificiranu distribuciju.

Nakon što su podatci grupirani u k grupa, statistika se računa kao:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{4.14}$$

gdje je O_i frekvencija podataka u grupi i, a E_i očekivana frekvencija podataka u grupi i. E_i možemo izračunati kao:

$$E_i = N(F(Y_g) - F(Y_d)), \tag{4.15}$$

a F je kulumativna distribucijska funkcija, Y_g je gornja granica grupe i, a Y_d donja granica. N je broj podataka.

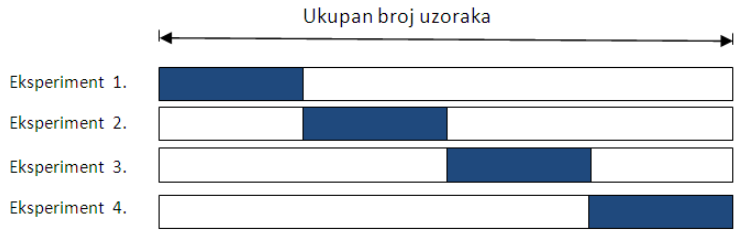
Kritična vrijednost s kojom se uspoređuje statistika, ovisi o $(k - c)$, gdje je k broj grupa, a c broj procijenjenih parametra za određenu distribuciju uvećan za jedan.

4.3 Metode validacije

Metode validacije rezultata primjenjivane u radu temelje se na unakrsnoj validaciji (eng. *cross-validation*). Korištene su k-struka i *Leave-One-Out* unakrsna validacija, a njihovi opisi mogu se pronaći u ovom poglavlju.

K-struka unakrsna validacija

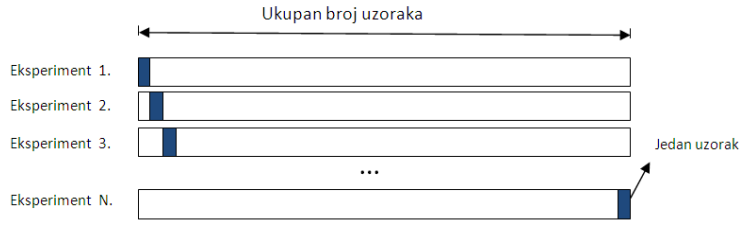
K-struka unakrsna validacija izvodi se tako da se početni skup podataka podijeli na k podskupova i onda k puta izvodi učenje modela i to tako da se jedan od skupova koristi za provjeru, a svi preostale za učenje. Shematski prikaz postupka može se vidjeti na slici 4.2.



Slika 4.2: Prikaz podjele skupa podataka za 4-struku unakrsnu validaciju.

LOO unakrsna validacija

Leave-One-Out unakrsna validacija krajnji je slučaj k-struke unakrsne validacije u kojoj je broj podskupova k jednak broju elemenata u skupu podataka. Time se model u svakoj od N iteracija uči na gotovo cijelom skupu podataka, a provjerava samo na jednom podatku. Shematski prikaz može se vidjeti na slici 4.3.



Slika 4.3: Prikaz podjele skupa podataka za LOO (Leave-One-Out) unakrsnu validaciju.

Poglavlje 5

Rezultati

Dobiveni rezultati primjenom metoda opisanih u poglavlju 4 prikazani su u nastavku. Prije svega definiramo ispitne skupove nad kojima vršimo analizu. Prema potpoglavljima postepeno se provode koraci analize i prikazuju rezultati. Numerički rezultati prikazani su u tablicama, a dio njih predočen je i grafički radi preglednosti. Uspoređujemo L_1 i L_2 regresiju prikazom rezultata distribucijskih provjera i statistike pogrešaka dobivenih analizom za svaki od skupova ulaznih podataka. Nakon toga provode se postupci validacije korištenjem metoda k-struke i LOO (eng. *Leave-One-Out*) unakrsne validacije, čiji rezultati su također dani u nastavku.

5.1 Ispitni skupovi

Podatci nad kojima smo vršili ispitivanje sustava podijeljeni su u dvije glavne cjeline: skup za učenje i skup za provjeru. Skup za učenje korišten je za određivanje koeficijenata regresije, dok je skup za provjeru korišten za ispitivanje prikladnosti tog regresivnog modela nad realnim podatcima, po principu proračuna pogreške. Za svaku od ovih cjelina na raspolaganju imamo niz deskriptora (pomoću kojih modeliramo naš sustav temeljen na multivarijatnoj regresiji) te točnu vrijednost modelirane veličine.

Konkretne vrijednosti nad kojima smo vršili ispitivanje preuzeti su iz (13). U trenutku pisanja ovog rada, izvor (13) je citiran čak 196 puta, te smatramo da predstavlja relevantan izvor podataka iz ovog područja.

Podatci su podijeljeni u tri skupa sastavljenih od podataka za učenje i prov-

jeru[1]:

- A skup - 132 parova ulaz-izlaz, 66 za učenje modela i 66 za provjeru
- B skup - 132 parova ulaz-izlaz istih kao i u A skupu, 100 za učenje modela i 32 za provjeru
- C skup - 113 parova ulaz-izlaz, 57 za učenje modela i 56 za provjeru

Ulazni vektor ima dimenziju $M = 13$, što znači da je svaka molekula opisana s 13 različitih deskriptora pomoću kojih modeliramo željeno svojstvo. Nadalje, izlaz predstavlja točnu vrijednost željenog svojstva za određenu kombinaciju ulaza.

5.2 Usporedba L_1 i L_2 multivarijatne regresije

Primjenom metoda opisanih u poglavlju 4.2.3 dobili smo rezultate distribucijskih provjera nad pogreškama iz opisanih skupova podataka. Pri tome su modeli L_1 i L_2 regresije izračunati pomoću podskupova za treniranje, dok su pogreške dobivene gledajući razliku poznate točne vrijednosti modelirane veličine i one predviđene izračunatim modelom. Rezultati su dani numerički u tablicama, a radi preglednosti i dobrog uvida, rezultati distribucijskih provjera prikazani su i grafički. Dalje, usporedba pogrešaka i mjera kvalitete pogodnosti između L_1 i L_2 regresije za svaki od skupova također je prikazana tablično.

5.2.1 Distribucijske provjere

Distribucija pogreške uvelike ovisi o položaju pravca u odnosu na koji je određena pogreška. Radi mogućnosti usporedbe utjecaja odabira pravca na pogrešku, provjere su provedene u dva slučaja, korištenjem pravca dobivenog L_1 i L_2 linearnom regresijom.

Slika 5.1 prikazuje histograme pogrešaka podataka s obzirom na pravce dobivene multivarijatnom regresijom prema L_1 odnosno L_2 normi. Na istoj slici prikazane su i distribucijske funkcije Gaussove i Laplaceove razdiobe radi mogućnosti usporedbe dobivenih rezultata. Tablice 5.1 - 5.3 prikazuju numeričke vrijednosti statistika svake od tri distribucijske provjere primijenjene nad podatcima pogrešaka dobivenih upotrebom regresije po L_1 odnosno L_2 normi. Vrijednosti označene plavom bojom predstavljaju razdiobu čija statistika bolje zadovoljava određenu distribuciju (generalno, odabrana je razdioba s manjom statistikom).

[1] Radi lakše usporedbe skupovi su imenovani isto kao i u (13).

Tablica 5.1: Rezultati ispitivanja distribucije pogrešaka skupa za provjeru A. KS - Kolmogorov-Smirnov, AD - Anderson-Darling.

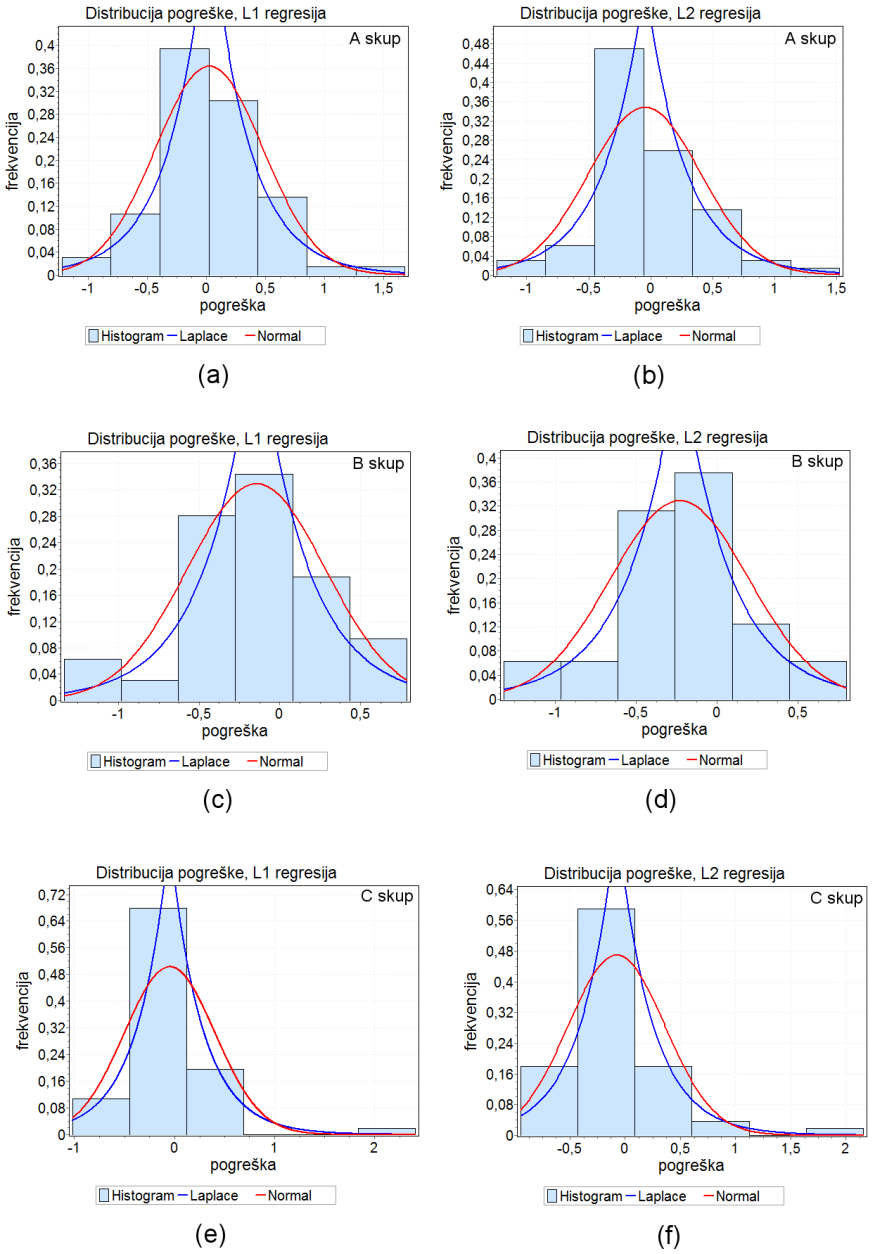
	L_1			L_2		
	KS	AD	χ^2	KS	AD	χ^2
Laplace	0.07	0.49	2.5	0.13	0.75	10.65
Gauss	0.12	0.74	9	0.11	1.07	10.54

Tablica 5.2: Rezultati ispitivanja distribucije pogrešaka skupa za provjeru B. KS - Kolmogorov-Smirnov, AD - Anderson-Darling.

	L_1			L_2		
	KS	AD	χ^2	KS	AD	χ^2
Laplace	0.08	0.18	1.36	0.09	0.18	0.21
Gauss	0.13	0.61	3.02	0.15	0.67	1.71

Tablica 5.3: Rezultati ispitivanja distribucije pogrešaka skupa za provjeru C. KS - Kolmogorov-Smirnov, AD - Anderson-Darling.

	L_1			L_2		
	KS	AD	χ^2	KS	AD	χ^2
Laplace	0.13	0.89	4.22	0.12	0.86	8.65
Gauss	0.19	2.3	15.56	0.18	1.81	12.47



Slika 5.1: Usporedba histograma pogreške s distribucijskim funkcijama Gaussove i Laplaceove razdiobe. Pogreške su određene iz podskupa podataka za provjeru u odnosu na pravac dobiven multivarijatnom regresijom po L_1 ili L_2 normi iz podskupa za učenje. Pripadnost skupu podataka i primijenjena norma regresije označeni su uz odgovarajući graf.

Iz rezultata možemo vidjeti kako je vjerojatnost Laplaceove raspodjele pogreške vrlo visoka u odnosu na Gaussovu. U slučaju računanja pogreške prema pravcu dobivenim L_1 regresijom, čak sva tri ispitna skupa zadovoljavaju kriterije provjera distribucija, i to uz veći stupanj sigurnosti nego za Gaussovu razdiobu. U slučaju L_2 regresije, Laplaceova razdioba je vjerojatnija od Gaussove u 2/3 skupa (B i C). Sveukupno možemo zaključiti kako pogreška realnih podataka zadovoljava Laplaceovu razdiobu, što potvrđuje našu pretpostavku o mogućnosti primjene L_1 regresije za robusnu QSAR analizu.

5.2.2 Usporedba pogrešaka

U tablici 5.4 prikazane su dvije mjere pogrešaka: *1-norma* (suma apsolutnih pogrešaka) i *RMSE* (srednja kvadratna pogreška), te dvije mjere kvalitete pogodnosti regresije: r^2 i q^2. Uz njih, prikazana je i standardna devijacija pogreške. Svrha ove provjere je usporedba pogrešaka dobivenih korištenjem modela L_1 i L_2 regresije, izračunatih pomoću skupa za učenje, pri redviđanju uzoraka iz skupa za provjeru. Plavom bojom označena je bolja između te dvije razmatrane regresije: u slučaju pogreške manja vrijednost prevladava, dok je kod mjera pogodnosti regresije obrnuto. Dodatno, radi usporedbe s radom (13) iz kojeg smo preuzeli podatke, rezultati iz članka su upisani u tablicu unutar odgovarajućeg skupa.

Tablica 5.4: Pogreška podataka skupova za provjeru u odnosu na pravce dobivene L_1 i L_2 multivarijatnom regresijom. *MLRI* (eng. *Multiple Linear Regression with Indicator variables*) i *NN* (eng. *Neural Networks*) su rezultati iz (13).

		1-norma	RMSE	r^2	q^2	σ
Skup A	L_1	21.674	0.452	0.757	0.604	0.455
	L_2	22.242	0.451	0.653	0.607	0.452
	$MLRI$	-	-	0.744	0.547	0.490
	NN	-	-	0.844	0.672	0.431
Skup B	L_1	10.412	0.447	0.728	0.707	0.430
	L_2	11.727	0.480	0.621	0.662	0.428
	$MLRI$	-	-	0.700	0.814	0.369
	NN	-	-	0.833	0.804	0.372
Skup C	L_1	15.574	0.451	0.720	0.146	0.453
	L_2	16.431	0.442	0.541	0.178	0.440
	$MLRI$	-	-	0.591	0.273	0.430
	NN	-	-	0.926	0.511	0.341

Tablica 5.4 pokazuje kako smo primjenom L_1 multivarijatne regresije dobili manju apsolutnu pogrešku u svakom od tri ispitna skupa, u odnosu na istu pogrešku po L_2 normi. *RMSE* je za dva od tri skupa bolji za L_2 regresiju. Ovakvi rezultati su očekivani, jer je apsolutna pogreška mjera koja se pobliže povezuje s L_1 normom, a *RMSE* s L_2. r^2 nam ovdje predstavlja kvalitetu, tj. pogodnost regresije za točke iz trening skupa, dok je q^2 mjera *pogodnosti predviđanja* (točnost regresije za podatke koji nisu bili u trening skupu). Mjera r^2 je ispala bolja u sva tri slučaja, dok je q^2 nešto lošiji za skupove A i C u odnosu na L_2 regresiju, ali zato daje bolji rezultat za skup B. Varijanca je ovdje dana radi usporedbe s rezultatima iz rada (13), a može se vidjeti kako se njezina vrijednost ne mijenja znatno između L_1 i L_2 regresije. Mjesta na kojima je napisan znak "-" označava podatak koji nije bio dostupan u razmatranom radu. Usporedba rezultata linearne regresije iz članka i rezultata dobivenih u ovom radu je samo simbolička, jer u članku nigdje nije određen točan tip linearne regresije (npr. norma, korišteni algoritam,...).

5.3 Rezultati validacije

Ovom provjerom razmatramo koja će nam regresija dati manju ukupnu pogrešku (*1-norma* i *RMSE*) u svrhu procjene kvalitete primjene tih regresija nad konkretnim podatcima.

Rezultati *k-struke* unakrsne validacije prikazani su u tablici 5.5. Ona prikazuje trening podskupove slučajnim odabirom razlomljene na 10 jednakih dijelova. Jedan redak tablice prikazuje ukupnu pogrešku po *1-normi* i *RMSE* u odnosu na pravac dobiven L_1 i L_2 multivarijatnom regresijom. Plavom bojom označene su regresije s manjom pogreškom.

Tablica 5.6 prikazuju dvije mjere pogrešaka (*1-norma* i *RMSE*), te dvije mjere kvalitete pogodnosti regresije (r^2 i q^2) kod *LOO* unakrsne validacije. Plavom bojom označena je bolja od dvije razmatrane norme: u slučaju pogreške manja vrijednost prevladava, dok je kod mjera pogodnosti regresije obrnuto.

Tablica 5.5: *10-struka* unakrsna validacija nad tri trening skupa. Plavom bojom su označene bolje vrijednosti u svakoj kategoriji.

	1-norma		RMSE	
	L_1	L_2	L_1	L_2
$\sum$ (Skup A)	32.77	34.76	6.80	6.95
$\sum$ (Skup B)	43.84	44.60	5.93	5.98
$\sum$ (Skup C)	16.96	20.21	3.95	4.60

Tablica 5.6: *LOO* unakrsna validacija nad trening skupovima.

		1-norma	RMSE	r^2	q^2
Skup A	L_1	51.662	0.562	0.710	0.608
	L_2	55.055	0.592	0.667	0.567
Skup B	L_1	51.662	0.562	0.710	0.608
	L_2	55.055	0.592	0.667	0.567
Skup C	L_1	32.998	0.451	0.630	0.262
	L_2	35.109	0.438	0.413	0.305

Promatrajući rezultate iz tablice 5.5 uočavamo kako su nam sume apsolutnih i *RMSE* pogreška manje nad svim ispitnim skupovima. Podatci su dobiveni podjelom cijelog ispitnog skupa (podskup za učenje + podskup za provjeru) na deset jednakih dijelova, pri čemu su skupovi sastavljeni od slučajno odabranih parova ulaz-izlaz. To znači da će se, ponovnim pokretanjem ove provjere, vrijednosti u tablici 5.5 promijeniti, iako ne znatno - k skupova je dobiveno slučajnim izborom parova ulaz-izlaz s jednolikom razdiobom, što znači da je podjednaka vjerojatnost da se svaki od parova nađe u bilo kojem od skupova. Iz toga proizlazi da bi ukupna izračunata pogreška trebala ostati približna vrijednostima iznesenim u tablici.

Tablica 5.6 pokazuje kako L_1 multivarijatna regresija daje bolje rezultate u većini slučaja i kod *LOO* unakrsne validacije. Bitno je napomenuti kako *skup A* i *skup B* sadrže iste rezultate u tablici, jer su ovi skupovi i temeljeni na istim podatcima koji su u drukčijem omjeru raspoređeni u podskupove za učenje i provjeru. Prilikom *LOO* validacije pojedini su skupovi za učenje i provjeru spojeni, što je uzrok identičnim rezultatima.

Poglavlje 6

Zaključak

U radu je prikazan poznati *Li-Arce* algoritam jednostavne L_1 linearne regresije. Postignuta su poboljšanje tog algoritma smanjivanjem prosječne složenosti jednog njegovog dijela s $O(n \log n)$ na $O(n)$. Dodatno, proširili smo algoritam kako bi mogao biti korišten i u višedimenzionalnom slučaju.

Prikazali smo novi pristup u QSAR analizi, temeljen na korištenju L_1 norme u razvoju multivarijatnih regresijskih modela. Potvrdili smo pretpostavke o distribucijama QSAR podataka iznesene u uvodnom poglavlju. Rezultati su prikazani u brojčanom obliku u tablicama 5.1, 5.2 i 5.3, a slikom 5.1 predočene su i distribucije podataka uspoređene s Gaussovom i Laplaceovom. Kako je pogreška stvarnih podataka u odnosu na predviđene, temeljem koje određujemo distribuciju tih podataka, ovisna o odabranom pravcu linearne regresije, dana je usporedba pogrešaka i njihovih distribucija u slučaju primjene i L_1 i L_2 norme. Određivanje pogrešaka napravljeno je na tri skupa za provjeru, a svaki od tih skupova rastavljen je na dio za učenje, s kojim se određuje regresivni model, i dio za provjeru izračunatog modela. Provjere pogodnosti distribucija provedene su postupcima *Kolmogorov-Smirnova*, *Anderson-Darlinga* i χ^2. Čak 5 od 6 slučaja (83.3%) rezultiralo je većom vjerojatnošću (u sve tri provjere) da je distribucija podataka Laplaceova nego Gaussova.

Na osnovu dobivenih rezultata provjera distribucija, ispitane su pogreške i mjere kvalitete modela temeljenog na L_1 multivarijatnoj regresiji. Dobivene vrijednosti prikazane su u tablici 5.4 i uspoređene s rezultatima primjene L_2 regresije. Uspoređivane pogreške su *1-norma* (suma apsolutnih pogrešaka) i *RMSE* (eng. *Root Mean Square Error*). Pri tome su rezultati 1-norme

bili u korist L_1 regresije (manja pogreška) za sva tri skupa podataka, dok je RMSE u dva od tri slučaja bila u korist L_2 regresije. Pri tome je najveća razlika RMSE koja je bila u korist regresije po L_2 normi bila manja od 2%.

Dodatna ispitivanja predložene metode provedena su primjenom LOO i k-struke unakrsne validacije, a rezultati su prikazani u tablicama 5.5 i 5.6. Pri tome su rezultati obje validacije prevladali u korist L_1 u odnosu na L_2 regresiju.

U ovom radu, modeli su temeljeni samo na linearnim ulaznim deskriptorima, no uz isti je formalizam moguće uključiti i nelinearne transformacije ulaznih deskriptora. Ispitivanje rezultata regresije za ovakav skup podataka predviđen je u nastavku istraživačkog rada. Na taj način omogućili bismo usporedbu metode temeljene na L_1 normi s nelinearnim modelima razvijenim na ovom skupu podataka (neuronske mreže).

Zaključujemo kako predloženi model daje bolje rezultate u usporedbi s postojećim i često primjenjivanim regresivnim modelima temeljenim na L_2 normi za iste skupove podataka. To omogućuje njegovu realnu primjenu u postupku istraživanja novih lijekova, određivanju toksičnosti spojeva, raznim drugim istraživanjima, kao i primjenu u industriji.

Bibliografija

[1] Report from the expert group on (q)sars on principles for the validation of (q)sars. http://appli1.oecd.org/olis/2004doc.nsf/linkto/env-jm-mono(2004)24, 2004.

[2] Regulation (ec) no 1907/2006 of the european parliament and of the council of 18 december 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (reach), establishing a european chemicals agency, amending directive 1999/45/ec and repealing council regulation (eec) no 793/93 and commission regulation (ec) no 1488/94 as well as council directive 76/769/eec and commission directives 91/155/eec, 93/67/eec, 93/105/ec and 2000/21/ec, prosinac 2006.

[3] Guidance document on the validation of (q)sar models. http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)2, 2007.

[4] Y. Li; G. R. Arce. A maximum likelihood approach to least absolute deviation regression. *Journal on Applied Signal Processing*, 2004.

[5] T. Scitor; J. L. Medina-Franco; Q.-T. Do; K. Martinez-Mayorga; J. A. Y. Rojas; P. Bernard. How to recognize and workaround pitfalls in qsar studies: A critical review. *Current Medicinal Chemistry*, 2009.

[6] R. Todeschini; V. Consonni. *Molecular Descriptors for Chemoinformatics*. WILEY-VCH, 2009.

[7] T. Puzyn; J. Lesczczynski; M. T. D. Cronin, editor. *Recent Advances in QSAR Studies - Methods and Applications*. Springer, 2010.

[8] W. L. DeLano. Pymol molecular viewer. http://www.pymol.org/, travanj 2010.

[9] A. R. Leach; V. J. Gillet. *An Introduction to Chemoinformatics*, chapter Chapter 4. Computational Models. Springer, 2007.

[10] J. G. Hunt; J. M. Dowling; F. R. Glahe. L1 estimation in small samples with laplace error distributions. *Decision Sciences*, 1974.

[11] C. Gurwitz. Weighted median algorithms for l1 approximation. *BIT Numerical Mathematics*, 1989.

[12] C. Silipo; C. Hansch. Correlation analysis. its application to the structure-activity relationship of triazines inhibiting dihydrofolate reductase. *Journal of the American Chemical Society*, 1975.

[13] T.A. Andrea; H. Kalayeh. Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *Journal of Medicinal Chemistry*, 1991.

[14] M. T. D. Cronin; D. J. Livingstone. *Predicting Chemical Toxicity and Fate*. CRC Press, 2004.

[15] Y. C. Martin. *Quantitative Drug Design: A Critical Introduction*. CRC Press, 2010.

[16] M. P. Matiheu. *Parexel's Pharmaceutical R&D Statistical Sourcebook 2002./2003.* Parexel International Corporation, 2002.

[17] J. E. Gentle; V. A. Sposito; S. C. Narula. Algorithms for unconstrained l1 simple linear regression. *Computational Statistics and Data Analysis*, 1988.

[18] I. Barrodale; F. D. K. Roberts. An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis*, 1973.

[19] M. Tichý; M. Rucki. Validation of qsar models for legislative purposes. *Interdisciplinary Toxicology*, travanj 2009.

[20] K. Sabo; R. Scitovski. The best least absolute deviations line-properties and two efficient methods for its derivation. *ANZIAM Journal*, 2008.

[21] C.D. Selassie. History of quantitative structure-activity relationships. In *Burger's Medicinal Chemistry and Drug Discovery*. John Wiley & Sons, Inc., 2003.

[22] B. Walczak; M. Daszykowski; I. Stanimirova. Robust methods in qsar. In *Recent Advances in QSAR Studies - Methods and Applications*. Springer, 2010.

[23] P. Bloomfield; W. L. Steiger. *Least Absolute Deviations - Theory, Applications and Algorithms*. Birkhäuser, 1983.

[24] K. Sumathy. Chemical descriptors from qsarworld - a strand life sciences web resource. http://www.qsarworld.com/insilico-chemistry-chemical-descriptors.php, travanj 2010.

[25] H. van de Waterbeemd; S. Rose. *The Practice of Medicinal Chemistry*, chapter Chapter 23. Quantitative Approaches to Structure–Activity Relationships. Academic Press, 2008.

[26] G. O. Wesolowsky. A new descent algorithm for the least absolute value regression problem. *Communications in Statistics - Simulation and Computation*, 1981.

[27] A. P. Worth. The role of qsar methodology in the regulatory assessment of chemicals. In *Recent Advances in QSAR Studies - Methods and Applications*. Springer, 2010.

[28] A. B. Richon; S. S. Young. An introduction to qsar methodology. http://www.netsci.org/Science/Compchem/feature19.html, travanj 2010.

[29] Y. Zhang. Primal-dual interior point approach for computing l_1-solutions, and l_∞-solutions of overdetermined linear systems. *Journal of Optimization Theory and Applications*, 1993.

Sažetak

Ivan Sović, Matija Piškorec, Igor Čanadi: Nova robusna metoda za QSAR analizu temeljena na multivarijatnoj regresiji i normi L_1

Ključne riječi: QSAR, L1, L2, linearna, multivarijatna, regresija, Laplace, Gauss

QSAR (eng. *Quantitative Structure-Activity Relationship*) danas je široko prihvaćen skup metoda koje polaze od osnovne ideje da postoji veza između kemijske strukture molekula i njihove biološke aktivnosti, te da je tu vezu moguće kvantitativno opisati. Vrlo popularna i razrađena metoda QSAR analize je multivarijatna linearna i nelinearna regresija po L_2 normi. Velik utjecaj na kvalitetu takvih modela čine *outlieri* - uzorci koji po vrijednostima parametara znatno odskaču od ostalih. Njihov se utjecaj može kompenzirati na dva načina: eksplicitnim izuzimanjem iz skupa podataka, ili konstruiranjem modela koji se dobro ponaša u njihovoj prisutnosti.

U radu opisujemo mogućnost primjene multivarijatne regresije temeljene na L_1 normi, koja minimizira zbroj apsolutnih pogrešaka umjesto zbroja kvadratnih pogrešaka, u svrhu kreiranja robusnog modela za QSAR. L_1 regresija predstavlja procjenu najveće vjerojatnosti za Laplaceovu distribuciju za koju su outlieri specifična pojava. Iz razloga što za nju ne postoji jednostavno analitičko rješenje već se rješenje mora tražiti računski zahtjevnim iterativnim metodama, povijesno je bila puno manje zastupljena nego L_2 regresija. Izložen je jedan od algoritama L_1 linearne regresije, modificiran tako da mu prosječna složenost bude $O(n)$ te proširen i poopćen za višedimenzionalan slučaj. Pokazujemo kako razmatrani QSAR podatci zaista pokazuju veliku sličnost Laplaceovoj distribuciji, što nam omogućuje primjenu L_1 norme u razvoju robusnog modela. Prikazani su rezultati multivarijatne regresije po normi L_1 u usporedbi s rezultatima temeljenim na normi L_2, iz čega se može vidjeti kako L_1 regresija daje manju pogrešku predviđanja. Dodatno je prikazana i usporedba s rezultatima rada iz kojeg je preuzet skup podataka korišten prilikom razvoja modela u ovom radu.

Summary

Ivan Sović, Matija Piškorec, Igor Čanadi: New robust method for QSAR analysis based on multivariate regression and L_1 norm

Keywords: QSAR, L1, L2, linear, multivariate, regression, Laplace, Gauss

QSAR (*Quantitative Structure-Activity Relationship*) is today a widely accepted set of methods. All of them originate from the same idea that there is a relationship between chemical structures of molecules and their biological activity, and that this relationship can be quantitatively described. Very popular and well developed and often used method for QSAR analysis is linear and non-linear multivariate regression based on the L_2 norm. Great effect on quality of such models is caused by outliers - samples that considerably diverge from others. Their influence can be compensated in two ways: by explicitly removing them from the data set, or by constructing a model that behaves well in their presence.

The possibility of using L_1 multivariate regression for creating robust QSAR models has been described in our work. L_1 regression represents the maximum likelihood estimate for the Laplace distribution. Appearance of outliers is specific for this type of distribution. Since there is no simple analytical solution to the L_1 regression, and the solution has to be reached using computationaly intensive iterative methods, this type of regression has been significantly less used in the past than the L_2 regression. One of the algorithms for linear regression based on L_1 norm has been described. We modified this algorithm in two ways: complexity of a part of the algorithm was reduced to $O(n)$, and it was expanded and generalized for the multivariate case. We show how the distribution of used QSAR data resemble the Laplace distribution, which gives us the possibility of using L_1 norm to construct a robust model. The results of using L_1 and L_2 norms in development of multivariate regression models were compared. The conclusion to the whole process is that the use of L_1 norm reduces the error of prediction. In addition, comparison to the results published in the article that was the source for data set used for modeling performed in this work was given.