UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Alen Jakovac

Napovedovanje lastnosti papirja iz spektrometričnih podatkov s strojnim učenjem

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: prof. dr. Igor Kononenko

Ljubljana, 2011

Univerza
v Ljubljani

Fakulteta za računalništvo
in informatiko

Tržaška 25
1000 Ljubljana, Slovenija
telefon: 01 476 84 11
faks: 01 426 46 47
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01754/2011
Datum: 01.04.2011

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **ALEN JAKOVAC**

Naslov: **NAPOVEDOVANJE LASTNOSTI PAPIRJA IZ SPEKTROMETRIČNIH PODATKOV S STROJNIM UČENJEM**

THE PREDICTION OF PAPER PROPERTIES FROM SPECTROMETRIC DATA WITH MACHINE LEARNING

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

Za približno 1000 vzorcev papirja je bilo izmerjenih 15 kemijsko fizikalnih lastnosti papirja. Poleg kemijsko fizikalnih lastnosti je bil s pomočjo spektrometra za vsak vzorec izmerjen tudi spekter v območju NIR (Near Infra Red). Absorbcija svetlobe pri neki valovni dolžini je odvisna sestave papirja. Spekter je bil izmerjen pri 256 približno enako oddaljenih valovnih dolžinah v območju od 1450 nm do 1950 nm. Ker imajo kemijsko fizikalne lastnosti numerične vrednosti, gre za 15 regresijskih problemov. Spektri so predstavljeni z 256 zveznimi atributi, vsaka kemijsko fizikalna lastnost pa predstavlja eno zvezno odvisno spremenljivko. Cilj diplomske naloge je poiskati čim boljše modele strojnega učenja za napovedovanje 15 kemijsko fizikalnih lastnosti papirja iz vrednosti izmerjenih spektrov. Kandidat naj preveri, kako različne metode predobdelave podatkov spremenijo točnost napovedi modelov.

Mentor:

prof. dr. Igor Kononenko

Dekan:

prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani **Alen Jakovac**,

z vpisno številko **63020060**,

sem avtor diplomskega dela z naslovom:

Napovedovanje lastnosti papirja iz spektrometričnih podatkov s strojnim učenjem

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom
  prof. dr. Igorja Kononenka
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 14. 10. 2011 Podpis avtorja:

Zahvala

Zahvaljujem se mentorju prof. dr. Igorju Kononenku za strokovno pomoč in usmerjanje pri izdelavi diplomskega dela.

Iskrena hvala tudi sodelavcem za podporo, razumevanje in potrpežljivost pri pisanju diplomskega dela.

Posebna zahvala gre moji družini, saj so me ves čas študija spodbujali in mi stali ob strani.

Kazalo

Povzetek

V diplomskem delu je predstavljeno reševanje problema napovedovanja kemijskih in fizikalnih lastnosti papirja iz spektrometričnih podatkov. Za napovedovanje lastnosti papirja je bila uporabljena podatkovna baza, ki vsebuje podatke o več kot 1000 vzorcih papirja. Za vsak vzorec papirja je bilo v laboratoriju izmerjenih 15 kemijskih in fizikalnih lastnosti ter spekter v bližnjem infrardečem območju.

Za napovedovanje lastnosti papirja so bile uporabljene različne metode strojnega učenja. Preizkušene so bile linearna regresija, regresija pace, metoda najbližjih sosedov, regresijska drevesa, metoda podpornih vektorjev, regresija glavnih komponent, regresija delnih najmanjših kvadratov, večnivojski perceptron in mreža radialnih baznih funkcij. Izkaže se, da je problem linearen. Najboljše rezultate zato dajejo linearna regresija, regresija glavnih komponent in regresija delnih najmanjših kvadratov.

Na spekter vpliva veliko zunanjih dejavnikov, ki povzročajo različne motnje v spektru. Z metodami predprocesiranja spektrov smo poskusili odstraniti motnje in tako izboljšati napovedi lastnosti papirja. Preizkušene so bile absorpcijska transformacija, Kubelka-Munkova transformacija, multiplikativna korekcija razpršenosti, metoda Standard Normal Variate, odvajanje spektrov in ortogonalna korekcija signala. Pogledali smo tudi, kako posamezna metoda predprocesiranja spektrov vpliva na posamezne metode strojnega učenja. Ugotovljeno je bilo, da večina metod izboljša napovedi modelov, najbolje pa se obneseta metoda Standard Normal Variate in multiplikativna korekcija razpršenosti.

Napovedi regresijskih modelov smo poskusili izboljšati še s kalibracijo regresijskih modelov, vendar se izkaže, da kalibracija ne izboljša napovedi modelov.

Ključne besede:

spektroskopija NIR, kemijske in fizikalne lastnosti papirja, predprocesiranje spektrov NIR, strojno učenje, kalibracija

Abstract

In this thesis we present a solution for the problem of predicting the chemical and physical properties of paper from spectrometric data. We used a data set that consists of over 1000 samples of paper. For each sample 15 chemical and physical properties and its near-infrared spectra were measured.

We used the following machine learning methods to predict the properties of paper: linear regression, pace regression, a nearest neighbor-based model, regression trees, a support vector machine, principal component regression, partial least squares regression, a multi-layer perceptron, and a radial basis function network. The prediction task turned out to be linear. Therefore, linear regression, principal component regression, and partial least squares regression gave the best results.

Many outside factors affect the spectra and cause different types of interference. We used the following spectra preprocessing methods to remove the interference and improve the predictions: absorbance transformation, Kubelka-Munk transformation, multiplicative scatter correction, standard normal variate transformation, spectra derivation and orthogonal signal correction. We also investigated how preprocessing affects the machine learning methods. The results show that most preprocessing methods improve the models' predictions. The standard normal variate transformation and multiplicative scatter correction gave the best results.

We tried to further improve the predictions with calibration. However, calibration did not improve the predictions.

Key words:

NIR spectroscopy, chemical and physical properties of paper, preprocessing NIR spectra, machine learning, calibration

Poglavje 1

Uvod

Papir ima že stoletja zelo pomembno vlogo pri shranjevanju podatkov in ohranjanju našega znanja za prihodnje generacije. Čeprav ga v zadnjem času vse bolj nadomešča informacijska tehnologija, se še vedno večina podatkov in znanja shranjuje v pisni obliki. V knjižnicah in arhivih je na milijone različnih knjig, časopisov, znanstvenih člankov, posterjev, fotografij in drugih oblik papirja ter se dnevno povečuje. Če želimo ohraniti svoje znanje in svojo kulturno dediščino, moramo papir ustrezno shranjevati. Knjižnice in arhivi zato shranjujejo knjige in druge vrste papirja pri točno določenih klimatskih pogojih. Prostori, v katerih se shranjujejo, morajo imeti konstantno temperaturo in vlago. Najstarejše knjige in zapisi, ki so še posebej zgodovinsko dragocene, pa imajo za shranjevanje določene še posebno stroge pogoje.

Čeprav so ustrezni pogoji za shranjevanje zelo pomembni pri ohranjanju papirja, pa na staranje in posledično razkroj papirja vpliva predvsem kemijska sestava papirja. Na staranje so bolj občutljive zlasti starejše knjige, pri katerih so se pri izdelavi papirja uporabljale drugačne sestavine in drugačni postopki kot danes. Za starejše knjige je namreč značilno, da vsebujejo večjo količino kislin, ki povzročajo razkrajanje papirja.

S kemijsko in fizikalno analizo različnih vzorcev papirja želimo pridobiti čim več podatkov o sestavi papirja. Na podlagi teh podatkov bi potem lahko z različnimi postopki povečali odpornost knjig na razkrajanje. S postopkom razkislinjenja papirja lahko na primer nevtraliziramo vrednost pH papirja, če pred tem s kemijskim postopkom izmerimo, kolikšno količino kisline vsebuje vzorec papirja.

Ker je kemijska analiza posamezne knjige zelo časovno potratna in ker lahko kemijsko analizo naredimo le na vzorcu papirja, ki ga vzamemo iz te knjige, se za določitev kemijskih in fizikalnih lastnosti papirja namesto kemijske analize uporablja spektroskopija NIR [2, 12]. Prednost spektroskopije NIR je v tem, da pri merjenju spektrov ne poškoduje knjige, saj se le za kratek čas osvetli določen del papirja. Pri kemijski analizi pa je treba vzeti neki delček papirja iz knjige, kar pa je pri dragocenih knjigah nesprejemljivo.

V kemijskih in fizikalnih laboratorijih je bilo izmerjenih 15 kemijskofizikalnih lastnosti za več kot 1000 vzorcev papirja. Kemijske lastnosti se nanašajo na sestavo papirja, fizikalne pa na mehanske lastnosti papirja. Poleg tega se je s spektrometrom NIR za vsak vzorec izmeril tudi spekter. Podatki so bili vneseni v podatkovno bazo [18].

Cilj diplomskega dela je ugotoviti, kako dobro se iz spektrov vzorcev papirja dajo

napovedati njegove kemijske in fizikalne lastnosti. Preveriti smo, katere metode strojnega učenja [7, 15] so najbolj primerne za uporabo na spektroskopskih podatkih. Ker pa na meritve spektrov vpliva tudi veliko zunanjih dejavnikov, smo pogledali še, s katerimi metodami predprocesiranja spektrov [2, 10] je možno najbolj ublažiti njihove posledice in s tem izboljšati napovedi kemijskofizikalnih lastnosti papirja.

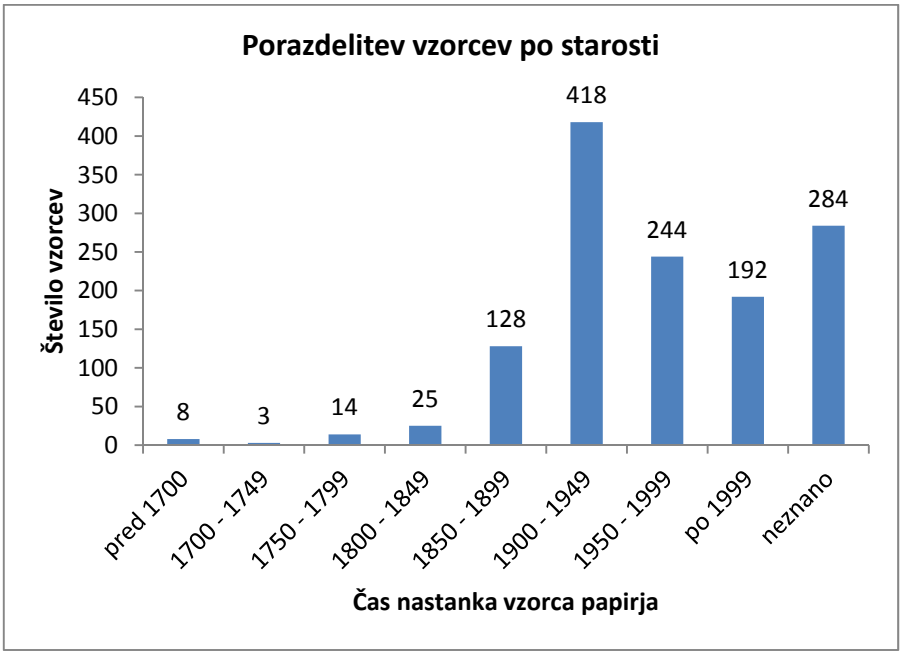
V nadaljevanju je v poglavju 2 najprej predstavljena podatkovna baza. V poglavju 3 je opisano, za kakšno vrsto problema gre. Sledita mu poglavje 4, v katerem so opisane uporabljene metode predprocesiranja spektrov, in poglavje 5, v katerem so zbrani opisi vseh preizkušenih metod strojnega učenja. V poglavju 6 so navedeni vsi eksperimenti, ki so bili narejeni, opisano pa je tudi, kako je potekalo ocenjevanje uspešnosti modelov. V poglavju 7 so predstavljeni rezultati napovedi posameznih metod strojnega učenja in ugotovitve, kako posamezne metode predprocesiranja spektrov vplivajo na različne metode strojnega učenja. Na koncu pa so v poglavju 8 navedene še sklepne ugotovitve.

Poglavje 2

Opis podatkovne baze

2.1 Vzorci papirja

Podatkovna baza [18] vsebuje podatke o vzorcih papirja, ki so bili izbrani tako, da predstavljajo večino vrst papirja, ki jih najdemo v evropskih knjižnicah, arhivih, muzejih in galerijah. Zbranih je bilo 1316 vzorcev, katerih starost se razteza vse do 18. stoletja, nekaj vzorcev pa je tudi starejših. Porazdelitev vzorcev papirja po starosti je prikazana na sliki 2.1.



Slika 2.1: Porazdelitev vzorcev papirja po starosti.

Kot lahko vidimo, je največ vzorcev iz dvajsetega stoletja, zlasti iz prve polovice stoletja. Veliko vzorcev pa je tudi iz sedanjosti. Ker so knjige iz obdobja pred 19. stoletjem bolj redke in zelo dragocene, je bilo vzorce iz tega obdobja težje dobiti. Za velik delež približno petino vzorcev pa ni znan čas nastanka.

Vzorci so bili večinoma vzeti iz knjig, nekaj pa je bilo vzetih tudi iz arhivskih map, posterjev, risb in drugih oblik papirja. Vzorci so bili skrbno izbrani, tako da upoštevajo čim večjo raznolikost v sestavi papirja. Zato je bila določena tudi tako velika podatkovna baza. Namen podatkovne baze je bil pridobiti podatke o takih vzorcih papirja, ki čim bolj odsevajo tipično zbirko knjig v neki knjižnici ali arhivu. V podatkovni bazi niso prisotni vzorci prosojnega ali plastificiranega papirja ter vzorci papirja, ki imajo določene poškodbe zaradi ognja, vode ali katerega drugega dejavnika. Podatkovna baza prav tako ne vsebuje vzorcev izven Evrope.

2.2 Kemijskofizikalne lastnosti

Podatkovna baza vsebuje za vsak vzorec podatke o 15 kemijskofizikalnih lastnostih papirja. Kemijskofizikalne lastnosti papirja so bile izmerjene v laboratoriju s klasičnimi kemijskimi in fizikalnimi postopki oziroma metodami. Meritve so bile za vsako kemijsko ali fizikalno lastnost narejene večkrat. V podatkovni bazi sta za vsak vzorec papirja zbrani povprečna vrednost in standardna deviacija večkratnih meritev posamezne kemijske oz. fizikalne lastnosti.

Izmerjene so bile naslednje kemijskofizikalne lastnosti:

1. vsebnost lignina,
2. vsebnost proteinov,
3. vsebnost aluminija,
4. vsebnost pepela,
5. vsebnost smole,
6. vsebnost karbonilne skupine,
7. natezna trdnost,
8. natezna trdnost po prepogibanju,
9. stopnja polimerizacije celuloze,
10. molekulska masa,
11. vrednost pH,
12. sestava vlaken – celuloza,
13. sestava vlaken – bombaž,
14. sestava vlaken – lesovina,
15. prisotnost optičnih belilnih sredstev.

Prisotnost optičnih belilnih sredstev je kemijskofizikalna lastnost, ki ima lahko le dve vrednosti, in sicer vrednost ”da” in vrednost ”ne”. Ker ta kemijskofizikalna lastnost predstavlja diskretno spremenljivko, gre pri napovedi te spremenljivke za klasifikacijski problem, s katerim pa se tukaj ne ukvarjam.

V tabeli 2.1 so zbrani osnovni podatki o posamezni kemijskofizikalni lastnosti (KFL) papirja.

KFL	Enota	MIN	MAX	STD	RSTD
Vsebnost lignina	%	0,00	35,44	0,84	2,36
Vsebnost proteinov	%	0,00	9,22	0,04	0,43
Vsebnost aluminija	mg/g	0,00	32,42	0,08	0,24
Vsebnost pepela	%	0,00	39,11	0,65	1,66
Vsebnost smole	mg/g	0,00	12,28	0,12	1,01
Vsebnost karbonilne skupine	mol/g	0,003	0,206	0,003	1,49
Natezna trdnost	N	11,15	106,53	1,76	1,85
Natezna trdnost po prepogibanju	N	0,00	136,75	2,11	1,54
Stopnja polimerizacije celuloze		258,6	4223,4	24,6	0,62
Molekulska masa		13967	1967350	9248	0,47
Vrednost pH		3,72	9,20	0,06	1,09
Sestava vlaken - celuloza		0	1		
Sestava vlaken - bombaž		0	1		
Sestava vlaken - lesovina		0	1		

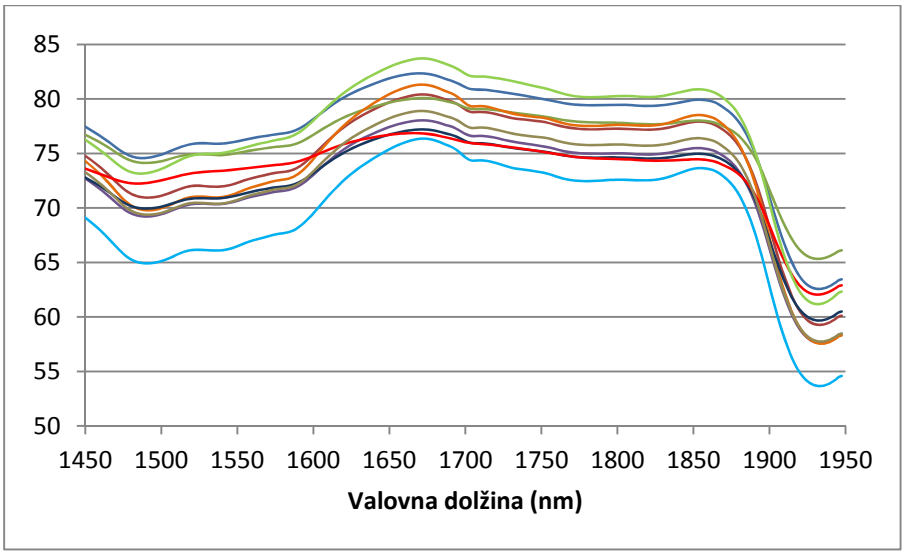
Tabela 2.1: Osnovi podatki o kemijskofizikalnih lastnostih papirja.

Obrazložitev oznak stolpcev tabele:

- **Enota** = merska enota KFL
- **MIN** = minimalna vrednost KFL vzorcev v podatkovni bazi
- **MAX** = maksimalna vrednost KFL vzorcev v podatkovni bazi
- **STD** = povprečna standardna deviacija večkratnih meritev KFL vzorca preko vseh vzorcev v podatkovni bazi
- **RSTD** = (STD * 100 %) / (MAX - MIN) = relativna povprečna standardna deviacija večkratnih meritev glede na razpon vrednosti KFL vzorcev (v odstotkih)

2.3 Spektri

Poleg kemijskofizikalnih lastnosti je bil za vsak vzorec s spektrometrom izmerjen tudi povprečni spekter v bližnjem infrardečem (angl. Near Infrared - NIR) območju [12]. Pri vsakem vzorcu je bil spekter izmerjen večkrat, in sicer na različnih mestih na površini papirja. Iz teh spektrov se je nato izračunal povprečni spekter in se shranil v podatkovno bazo. Spekter je bil izmerjen pri 256 približno enako oddaljenih valovnih dolžinah v območju od 1440 nm do 1950 nm. Kako izgledajo spektri, si lahko ogledamo na sliki 2.2, ki prikazuje spektre 10 vzorcev papirja.



Slika 2.2: Prikaz 10 naključno izbranih spektrov vzorcev papirja.

2.4 Predstavitev podatkovne baze

2.4.1 Atributna predstavitev

Na področju strojnega učenja se za predstavitev podatkovne baze največkrat uporablja atributna predstavitev [7] učnih primerov. Atribut je spremenljivka, ki ima določeno množico možnih vrednosti. V našem primeru so vsi atributi zvezni in njihovo množico možnih vrednosti lahko opišemo z intervalom $[Min, Max]$, kjer je Min minimalna možna vrednost in Max maksimalna možna vrednost, ki jo ima lahko atribut. Atributi so uporabljeni tako za predstavitev spektrov kot za predstavitev kemijskofizikalnih lastnosti. Vsak učni primer (vzorec papirja) je torej opisan z vektorjem vrednosti atributov.

Spekter lahko opišemo z 256 zveznimi atributi, pri čemer vsak atribut ustreza eni valovni dolžini. Atribut, ki predstavlja vrednost spektra pri k-ti valovni dolžini, je označen z oznako x_k. Njegov interval možnih vrednosti je interval $[0, 100]$.

Spremenljivkam, katerih vrednosti napovedujemo, pravimo odvisne spremenljivke. V tem primeru so odvisne spremenljivke posamezne kemijskofizikalne lastnosti. Vsako kemijskofizikalno lastnost lahko zato predstavimo z enim zveznim atributom, katerega interval možnih vrednosti je za vsako spremenljivko drugačen. Atribut, ki predstavlja vrednost j-te kemijskofizikalne lastnosti, označimo z oznako y_j.

2.4.2 Matrična predstavitev

Druga predstavitev, ki je značilna za predstavitev podatkov na področju spektroskopije in kemometrije, pa je matrična predstavitev [10]. Pri tej predstavitvi so podatki predstavljeni z matrikami. Ta predstavitev je, kot je opisano v nadaljevanju, predvsem uporabljena pri opisu metod predprocesiranja spektrov in nekaterih metod strojnega učenja.

Podatke v tej podatkovni bazi lahko predstavimo z dvema različnima matrikama, in

sicer spektralno matriko in matriko oz. vektorjem kemijskofizikalne lastnosti.

Spektralna matrika X je matrika, ki vsebuje podatke o spektrih ter ima N vrstic in K stolpcev. N je število učnih primerov (vzorcev papirja), K pa je število valovnih dolžin v spektru. Grafično jo lahko predstavimo kot:

$$X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,K} \end{bmatrix}$$

kjer je $x_{i,k}$ vrednost i-tega spektra pri k-ti valovni dolžini.

Vrednosti j-te kemijskofizikalne lastnosti vseh vzorcev pa lahko zapišemo z matriko velikosti $N \times 1$ oziroma s stolpčnim vektorjem y_j. Grafično lahko vektor predstavimo kot:

$$y_j = \begin{bmatrix} y_{1,j} \\ \vdots \\ y_{N,j} \end{bmatrix}$$

kjer je $y_{i,j}$ vrednost j-te kemijskofizikalne lastnosti i-tega vzorca.

2.5 Podatkovne množice

Ker podatki o nekaterih kemijskofizikalnih lastnostih za nekatere vzorce papirja v podatkovni bazi manjkajo in ker spektri nekaterih vzorcev niso bili izmerjeni, je bilo narejenih 14 ločenih podatkovnih množic, in sicer za vsako kemijskofizikalno lastnost ena.

Podatkovna množica za j-to kemijskofizikalno lastnost je podana kot množica vektorjev vrednosti atributov $u_i = \langle x_{i,1}, x_{i,2}, \ldots, x_{i,256}, y_{i,j} \rangle$, pri čemer vektor u_i predstavlja i-ti vzorec papirja.

KFL	Število vzorcev	Povprečna vrednost
KFL 1	903	0,055
KFL 2	637	0,279
KFL 3	947	0,474
KFL 4	671	0,253
KFL 5	951	0,265
KFL 6	533	0,382
KFL 7	806	0,246
KFL 8	920	0,436
KFL 9	951	0,520
KFL 10	896	0,037
KFL 11	864	0,114
KFL 12	944	0,199
KFL 13	950	0,214
KFL 14	626	0,254

Tabela 2.2: Število vzorcev in povprečna vrednost kemijskofizikalne lastnosti v posamezni podatkovni množici.

Podatki o številu učnih primerov (vzorcev papirja) in njihovi povprečni vrednosti kemijskofizikalne lastnosti so za posamezno podatkovno množico zbrani v tabeli 2.2.

Ker je podatkovna baza poslovna skrivnost, so bili podatki o kemijskofizikalnih lastnostih nekoliko zakriti. Vrstni red kemijskofizikalnih lastnosti je bil premešan tako, da podatkovne množice, ki so navedene v tabeli 2.2, ne ustrezajo istoležnim kemijskofizikalnim lastnostim, ki so naštete v začetnem seznamu. Poleg tega so bile vrednosti posameznih podatkovnih množic normalizirane, tako da imajo vrednosti na intervalu med 0 in 1. Povprečna vrednost v tabeli 2.2 je izračunana na normaliziranih podatkih. Pri računanju modelov pa so bili seveda uporabljeni originalni nenormalizirani podatki.

Porazdelitve vrednosti posameznih kemijskofizikalnih lastnosti si lahko pogledate v dodatku A, kjer so zbrane grafične predstavitve porazdelitev za vse podatkovne množice.

Poglavje 3

Opis problema

Poskusimo najprej ugotoviti, za kakšno naravo problema gre. Zanima nas, kakšna je povezava med spektralnimi vrednostmi in vrednostmi kemijskih oz. fizikalnih lastnosti. Če želimo to ugotoviti, si moramo najprej pogledati, kaj je spektroskopija NIR in na kakšen način deluje.

3.1 Osnove spektroskopije NIR

Spektroskopija NIR je hitra, preprosta, cenovno ugodna in zanesljiva metoda za ugotavljanje sestave neke snovi. Spektroskopija NIR se danes uporablja na mnogih področjih za določanja kvalitativnih in kvantitativnih lastnosti snovi, še posebej pa bioloških snovi, kot so žita, meso, zelenjava in sadje [2, 12].

Spektroskopija NIR meri spekter snovi v bližnjem infrardečem območju (angl. Near Infrared - NIR), ki predstavlja le majhen delček celotnega spektra elektromagnetnega valovanja. Bližnje infrardeče območje se razteza od 800 do 2500 nm valovne dolžine.

Spektrometer NIR deluje tako, da neko snov osvetli s svetlobo v bližnjem infrardečem območju in izmeri, koliko svetlobe se od vzorca odbije nazaj v spektrometer. Za svetlobo NIR je značilno, da prodre globoko v vzorec, kar je zelo pomembno, saj pri tem v snovi doseže več molekul. Nekaj svetlobne energije se v snovi absorbira in povzroči povečanje energije nihanja molekulskih vezi molekul v snovi. Pri katerih valovnih dolžinah svetlobe se bo energija absorbirala, je odvisno od vrste molekule in vrste molekulskih vezi.

Obstaja več tipov nihanja atomov oz. molekulskih vezi. Večinoma se razlikujejo po tem, ali se vezi krčijo in raztezajo ali pa se upogibajo. Poleg različnega gibanja atomov pa na tip nihanja (vibracije) molekulskih vezi vpliva tudi simetrija molekule. Za molekulo z N atomi velja, da ima 3N - 6 tipov nihanj, ki jim pravimo osnovna nihanja. Za vsako osnovno nihanje velja, da niha z določeno frekvenco, ki ji pravimo osnovna frekvenca. Osnovna frekvenca nihanja molekulske vezi določa frekvenco elektromagnetnega valovanja, ki ga bo molekula absorbirala [17].

Absorpcija svetlobne energije povzroči povečanje energije nihanja v molekuli. Osnovno nihanje molekulske vezi se lahko poveča na večkratnike osnovnega nihanja. Molekula tako pride na višje energijske nivoje. Večkratnikom osnovnega nihanja pravimo

nadtoni (angl. overton). Nadtonom ustrezajo frekvence, ki so večkratniki osnovne frekvence. Intenziteta absorpcije vsakega višjega nadtona je od 10 do 100-krat manjša od intenzitete nižjega nadtona.

Za večino molekul velja, da se njihove frekvence osnovnih nihanj nahajajo v srednjem infrardečem območju. V bližnjem infrardečem območju pa se pojavijo nadtoni osnovnih nihanj in kombinacije osnovnih nihanj. Pri kombinacijah osnovnih nihanj se frekvence dveh ali več osnovnih nihanj seštevajo. To pomeni, da spekter v bližnjem infrardečem območju sestavljajo t. i. absorpcijski pasovi, katerih frekvence oz. valovne dolžine ustrezajo nadtonom in kombinacijam osnovnih nihanj.

Ker se svetloba pri osvetlitvi vzorca največkrat difuzno razprši, pravimo spektru tudi difuzno refleksijski spekter, spektrometriji pa difuzno refleksijska spektrometrija.

Kako je absorpcija svetlobe, ki jo posredno izmeri difuzni spektrometer NIR, povezana s sestavo snovi, v našem primeru papirja, ponazarja Beerov zakon. Pred tem pa lahko predpostavimo, da so kemijske lastnosti papirja linearno odvisne od količin njegovih sestavin. Podobno verjetno velja tudi za fizikalne lastnosti, kot je natezna trdnost, saj so fizikalne značilnosti papirja pogosto odraz njegove kemijske sestave.

3.2 Beerov zakon

Tip nihanja molekul določa frekvenco, pri kateri molekule absorbirajo energijo svetlobe NIR. Amplituda absorpcije molekul pri neki valovni dolžini je določena z njeno absorptivnostjo in številom molekul, ki so osvetljene s svetlobnim snopom, ki ga oddaja spektrometer. Predpostavimo lahko, da je spektralni odziv (vrednosti spektra) povezan s koncentracijami sestavin vzorca. Bouguerov, Lambertov in Beerov zakon (na kratko mu pravimo kar Beerov zakon) [17] namreč pravi, da je absorpcija sestavine vzorca enaka produktu absorptivnosti določenega tipa molekulskega nihanja, koncentracije molekul, ki so obsevane s svetlobnim snopom, in dolžine poti prodora svetlobnega snopa v vzorec. Relacijo med izmerjenim spektralnim odzivom in koncentracijo molekul v vzorcu podaja enačba:

$$A = \varepsilon c l \tag{3.1}$$

kjer je ε molekulska absorptivnost ($liter \cdot mol^{-1} \cdot cm^{-1}$), c koncentracija molekul v vzorcu v področju znotraj svetlobnega snopa ($mol \cdot liter^{-1}$) in l dolžina poti prodora svetlobe v vzorec (cm^{-1}). Za dolžino poti lahko vzamemo tudi debelino vzorca, kadar svetloba prodre skozi celoten vzorec. Količina A označuje absorpcijo molekul neke sestavine snovi in nima enot. Absorptivnost določenega tipa molekule je izračunana z natančnimi meritvami absorpcije sestavine in z uporabo enačbe:

$$\varepsilon = \frac{A}{cl}$$

Pri refleksijski spektrometriji [2] se spekter izmeri s pomočjo odboja svetlobe, s katero osvetljujemo neki vzorec. Svetloba se bodisi difuzno odbije od površine vzorca bodisi pa prodre v vzorec in se odbije šele v notranjosti vzorca. Relacijo med absorpcijo in refleksijskim spektrom opisuje enačba:

$$R = \frac{I}{I_0} = 10^{-\varepsilon cl} \Rightarrow A = -log_{10}\left(\frac{I}{I_0}\right) = -log_{10}R = \varepsilon cl \quad (3.2)$$

kjer je I_0 začetna svetlobna energija, s katero osvetljujemo vzorec in I svetlobna energija, ki jo detektor v spektrometru zazna po interakciji z vzorcem. Pri interakciji z vzorcem se nekaj svetlobe absorbira, nekaj pa odbije nazaj. Refleksijski spekter, ki ga izmeri spektrometer, predstavlja količina R, ki nam pove, kolikšen del svetlobe se je v vzorcu absorbiral. Količina odbite svetlobe R je torej definirana kot razmerje med količino odbite svetlobne energije I in količino svetlobne energije I_0, ki pade na vzorec.

Spektrometri NIR ponavadi merijo relativni odboj [12] namesto absolutnega, ki se uporablja v enačbi (3.2). Relativni odboj merijo glede na neko referenco. Za referenco se največkrat uporabi keramična ploščica, za katero je značilno, da ne absorbira nobene svetlobe. Pri merjenju nekega vzorca moramo najprej s spektrometrom izmeriti refleksijski spekter referenčne ploščice, v tem primeru je to keramična ploščica. Potem pa naredimo enako meritev še na vzorcu, ki ga damo na mesto keramične ploščice. Relativni odboj svetlobe $R(\lambda)$ pri valovni dolžini λ izračunamo z enačbo:

$$R(\lambda) = \frac{I_S(\lambda)}{I_R(\lambda)}$$

kjer je $I_S(\lambda)$ intenziteta odbite svetlobe od vzorca pri valovni dolžini λ in $I_R(\lambda)$ intenziteta odbite svetlobe od referenčne ploščice. Spekter referenčne ploščice se izmeri samo enkrat in ga ni treba opraviti pri meritvi vsakega vzorca. Če pa želimo, da bodo meritve spektrov vzorcev med seboj konsistentne, moramo pri merjenju vedno uporabljati isto referenco.

Predpostavka, da velja Beerov zakon, pa ni vedno pravilna. Glavna težava je, da molekule pogosto med seboj interaktirajo. Pri tem pride do nastajanja novih molekul, katerih absorptivnost pa je drugačna. Spremembe pa lahko povzročajo tudi temperatura, pritisk in drugi dejavniki. Izkušnje kažejo, da relacija, ki jo opisuje Beerov zakon, v večini primerov vseeno drži precej dobro.

Iz Beerovega zakona lahko sklepamo, da ima napovedovanje kemijskofizikalnih lastnosti linearno naravo problema. Kemijske lastnosti papirja so torej linearno povezane spektralnimi vrednostmi, pri čemer moramo refleksijski spekter najprej z neko transformacijo pretvoriti v absorpcijskega. To lahko naredimo bodisi z enačbo (3.2), ki je navedena zgoraj, bodisi s Kubelka-Munkovo transformacijo. Ker se fizikalne lastnosti ne razlikuje veliko od kemijskih, lahko sklepamo, da so tudi fizikalne lastnosti linearno odvisne od absorpcijskega spektra.

3.3 Pregled obstoječih rešitev

Poglejmo si, katere rešitve problema napovedovanja kemijskofizikalnih lastnosti papirja in drugih snovi že obstajajo ter katere metode se pri tem uporabljajo.

Uporaba spektroskopije NIR za analizo kemijskih in drugih lastnosti neke snovi se je v zadnjem času zelo razširila. Največ se uporablja v prehrambni industriji, v farmacevtski industriji, na področju biomedicine in še na mnogih drugih področjih. Uporablja se predvsem za merjenje vsebnosti neke sestavine v snovi. Narejene so bile številne raziskave in poskusi napovedovanja kemijskih lastnosti snovi iz spektrov NIR. Izkaže se, da je uspešnost uporabe

spektroskopije NIR pri napovedovanju kemijskih lastnosti snovi zelo različna. V nekaterih primerih daje zelo dobre rezultate, v nekaterih pa bolj slabe. Največkrat daje dobre rezultate pri napovedovanju vsebnosti vlage, maščob in proteinov v žitu, mesu in drugih prehrambnih izdelkih. Vrednost korelacijskega koeficienta teh napovedi je večja od 0,9. V nekaterih primerih pa je celo večja 0,98, kar je zelo dober rezultat [2].

Pri pregledu obstoječih raziskav zasledimo, da se pri večini poskusov ugotavljanja kemijskofizikalnih lastnosti uporabljata le 2 metodi strojnega učenja. To sta regresija glavnih komponent in regresija delnih najmanjših kvadratov. Predvsem slednja se uporablja kot standardna metoda za napovedovanje kemijskih lastnosti snovi, saj največkrat daje najboljše rezultate.

Tudi na področju predprocesiranja spektrov so se nekatere metode predprocesiranja bolj uveljavile kot druge. To velja predvsem za absorpcijsko transformacijo, ki je postala skoraj obvezna metoda, ki se uporablja za linearizacijo refleksijskega spektra. Poleg te metode pa se pogosto uporabljajo še multiplikativna korekcija razpršenosti, metoda SNV (angl. Standard Normal Variate) in odvajanje spektrov [2, 12].

Na področju analize papirja je bilo najdenih le nekaj raziskav, ki sta jih naredili dve raziskovalni skupini. Večino raziskav je naredila prva skupina [9, 13, 14]. V teh raziskavah je bila za napovedovanje lastnosti papirja uporabljena le regresija delnih najmanjših kvadratov. Ali so se pri tem uporabljale tudi metode predprocesiranja, pa ni navedeno. Spektri so bili izmerjeni v območju od 1540 do 20000 nm. Korelacijski koeficienti (R) napovedi posameznih kemijskofizikalnih lastnosti so zbrani v tabeli 3.1.

KFL	**R**
Vsebnost lignina	0,992
Vsebnost aluminija	0,875
Vsebnost pepela	0,995
Vsebnost smole	0,930
Vrednost pH	0,966
Stopnja polimerizacije celuloze	0,990
Natezna trdnost	0,761
Natezna trdnost po prepogibanju	0,885

Tabela 3.1: Korelacijski koeficienti napovedi kemijskofizikalnih lastnosti papirja pri prvi raziskovalni skupini.

Druga skupina ljudi [6] pa je preučevala molekulsko maso in vsebnost karbonilne skupine v papirju. Tudi tukaj je bila za napovedovanje obeh lastnosti papirja uporabljena regresija delnih najmanjših kvadratov. Preizkušene pa so bile tudi različne metode predprocesiranja spektrov, in sicer multiplikativna korekcija razpršenosti, odvajanje spektrov in normalizacija vektorjev. Korelacijski koeficient za molekulsko maso je bil 0,9, za vsebnost karbonilne skupine pa 0,86.

Poglavje 4

Metode predprocesiranja spektrov

4.1 Splošno o predprocesiranju spektrov

Metode predprocesiranja spektrov v splošnem delimo v dve skupini. V prvo skupino spadajo metode za linearizacijo spektrov. V drugo skupino pa spadajo metode, ki poskušajo odstraniti vplive različnih dejavnikov, ki povzročajo napake v spektrih [2, 10, 12].

Za difuzni refleksijski spekter je značilno, da spektralne vrednosti niso linearno odvisne od vsebnosti (koncentracije) neke sestavine v snovi, zato je priporočljivo, da se pred uporabo metod strojnega učenja na spektrih izvede neka transformacija, ki bi pripomogla k boljši linearni odvisnosti spektralnih vrednosti od vrednosti kemijskofizikalnih lastnosti. Obstajata dve transformaciji, ki se uporabljata v ta namen. To sta Kubelka-Munkova transformacija in absorpcijska transformacija. Obe sta opisani v nadaljevanju.

Poleg uporabe metod za linearizacijo spektrov pa je še bolj pomembna uporaba metod za odstranjevanje neželenih zunanjih vplivov, ki v spektru povzročajo različne motnje. Te motnje se kažejo kot premiki spektrov v navpični ali vodoravni smeri, kot spremembe naklona spektra, lahko pa tudi kot spremembe oblike spektra. Vzroki oz. dejavniki [2, 10], ki povzročajo motnje, so:

- medsebojna interakcija različnih sestavin vzorca,
- razpršenost svetlobe pri merjenju vzorcev, ki imajo difuzno površino,
- slaba ponovljivost meritev, npr. zaradi različne oddaljenosti vzorca od spektrometra,
- strojna oprema spektrometra, npr. šum detektorja,
- klimatski pogoji, npr. temperatura, vlaga.

Glavni problem pri difuznem refleksijskem spektru povzroča vpliv razpršenosti svetlobe zaradi delcev v snovi. Zaradi različne porazdelitve delcev v snovi pride tudi merjenju istega vzorca do različnih spektrov. Prihaja predvsem do zamaknjenosti spektrov v navpični smeri ali pa do spremembe naklona spektra. Stopnja razpršenosti svetlobe pa je odvisna tudi od valovne dolžine svetlobe. Pri daljših valovnih dolžinah je stopnja razpršenosti večja. Zato velja, da vpliv razpršenosti ni enakomerno prisoten po celotnem spektru.

V nadaljevanju so opisane naslednje metode za zmanjševanje oz. odstranjevanje neželenih motenj v spektru:

- multiplikativna korekcija razpršenosti,
- metoda Standard Normal Variate,
- odvajanje spektrov,
- ortogonalna korekcija signala.

V razdelku 4.8 na koncu poglavja so zbrani grafi transformiranih spektrov, ki ponazarjajo, kako posamezne metode predprocesiranja spektrov vplivajo na originalne spektre. Najprej je prikazan graf 10 spektrov pred uporabo predprocesiranja, sledijo pa mu grafi spektrov po uporabi ene od metod ali ene od kombinacij metod predprocesiranja. Prikazane so vse tiste metode in kombinacije metod, ki so bile preizkušene pri napovedovanju kemijskofizikalnih lastnosti papirja.

4.2 Absorpcijska transformacija (A)

Najpogosteje uporabljena metoda za linearizacijo spektrov v refleksijski spektroskopiji je absorpcijska transformacija [2, 12]. Absorpcijska transformacija se na refleksijskem spektru izvede s pomočjo enačbe:

$$A = log_{10}(1/R)$$

Osnova za to transformacijo je izpeljana iz Beerovega zakona, ki pravi, da so koncentracije neke sestavine vzorca, ki absorbira svetlobo, povezane z logaritmom relativne intenzitete monokromatskega sevanja, ki se odbije od vzorca, glede na intenziteto vpadnega sevanja. Količina odbite svetlobe R je definirana kot razmerje med količino svetlobne energije, ki se difuzno odbije od vzorca, in količino svetlobne energije, ki pade na vzorec. Čeprav je v mnogih primerih težko ali celo nemogoče doseči zahteve Beerovega zakona, se ta transformacija še vedno uporablja bolj ali manj kot standardna metoda za linearizacijo spektrov in kot kaže, v večini primerov dobro deluje.

4.3 Kubelka-Munkova transformacija (KM)

Pri osvetljevanju neke difuzne trdne površine snovi pride do difuznega in zrcalnega odboja svetlobe. Intenziteta difuzne svetlobe je odvisna od kota vpadne svetlobe, gostote in porazdelitve velikosti delcev v snovi, kristalne strukture, lomnega količnika in absorpcije svetlobe. Idealno difuzno snov v praksi redkokdaj srečamo, zato skoraj vedno obstaja tudi zrcalni odboj svetlobe. Za vpadno svetlobo, ki prodre v vzorec in se potem iz notranjosti vrača nazaj proti površini vzorca, lahko predpostavimo, da je izotropna (enakomerno se širi v vse smeri). Del svetlobne energije, ki prodre v vzorec, se absorbira oziroma spremeni v nihajno energijo atomov. To pomeni, da je zmanjšanje difuzno odbite svetlobe odvisno od absorpcijskega koeficienta materiala v vzorcu. Absorpcijski koeficient K pa je proporcionalen s količino materiala v snovi, ki je absorbiral svetlobo.

Kubelka in Munk [2] sta razvila teorijo, ki opisuje difuzen odboj svetlobe pri prašnih

vzorcih in trdnih vzorcih z grobo (neravno) površino. Teorija pravi, da se svetloba, s katero osvetljujemo neki homogeni vzorec bodisi absorbira bodisi razprši v okolico. Z analizo svetlobnega toka v smeri prodora svetlobe v vzorec in svetlobnega toka v nasprotni smeri (v smeri odbite svetlobe) sta Kubelka in Munk prišla do enačbe (4.1), ki se zdaj uporablja za transformacijo refleksijskih spektrov.

$$\frac{K}{S} = \frac{(1 - R^2)}{2R} \tag{4.1}$$

K je absorpcijski koeficient, S je koeficient razpršenosti in R je difuzno odbita svetloba. $R = I/I_0$ je definiran kot količnik med intenziteto difuzno odbite svetlobe I in intenziteto vpadne svetlobe I_0. Koeficient razpršenosti je neodvisen od absorpcijskega koeficienta.

S pomočjo enačbe (4.1) lahko predpostavimo, da je difuzno odbita svetloba R proporcionalna z absorpcijskim koeficientom K in posledično s količino materiala, ki je absorbiral svetlobo. Difuzno odbita svetloba je torej odvisna od vsebnosti (koncentracije) nekega materiala v vzorcu, pri čemer pa ta odvisnost ni linearna.

Glede na to teorijo lahko sklepamo, da ima difuzno odbita svetloba multiplikativen vpliv na spekter. To pomeni, da lahko v primeru, ko je spekter pravilno transformiran v Kubelka-Munkove enote, razliko med dvema spektroma istega vzorca kompenziramo tako, da vrednost spektra pri vsaki valovni dolžini pomnožimo z isto konstanto. Podoben multiplikativen vpliv lahko zasledimo tudi pri absorpcijski transformaciji.

Kubelka in Munk sta pri izpeljavi poenostavljene enačbe (4.1) naredila kar nekaj predpostavk:

- Svetlobna toka potujeta v nasprotnih smereh.
- Vzorec je osvetljen z monokromatskim sevanjem.
- Porazdelitev razpršenega sevanja je izotropna (to pomeni, da zrcalni odboj ni upoštevan).
- Delci v vzorcu so naključno porazdeljeni.
- Delci so veliko manjši od debeline vzorca.
- Delci, ki razpršijo svetlobo, so porazdeljeni homogeno preko celotnega vzorca.

Zaradi omenjenih predpostavk velja, da ima Kubelka-Munkova enačba omejeno uporabo. Uporabljena naj bi bila samo pri šibkih absorpcijskih pasovih. Ker se v spektru NIR pojavljajo pasovi nadtonov nihanj in kombinacij osnovnih nihanj, za katere je značilno, da so veliko šibkejši od pasov osnovnih nihanj, lahko enačbo (4.1) v večini primerov uporabljamo brez skrbi.

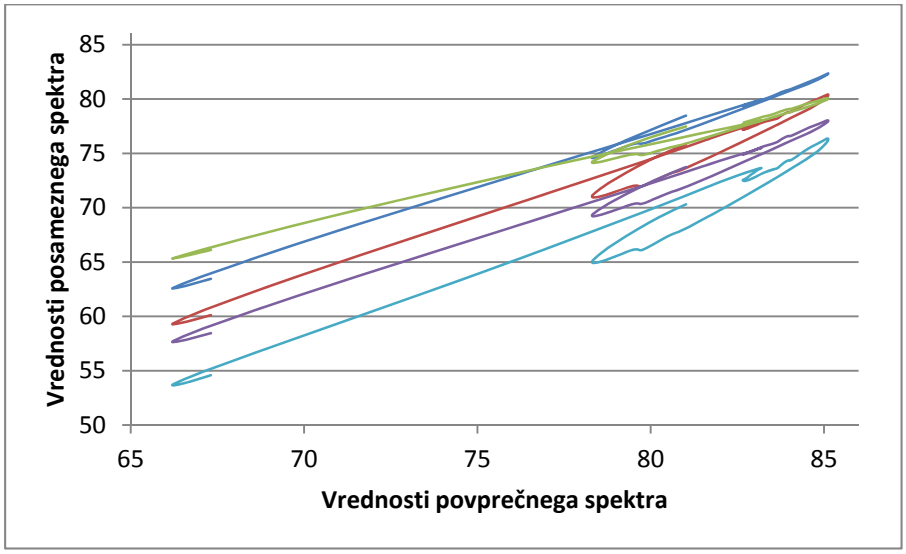
Poudariti pa je potrebno tudi, da spektrometri običajno merijo relativni odboj namesto absolutnega. Relativni odboj je enak količniku I_S/I_R, kjer je I_S intenziteta odbite svetlobe od vzorca in I_R inteziteta odbite svetlobe od referenčne ploščice (največkrat se za referenčno ploščico uporablja keramična ploščica), za katero je značilno, da ne absorbira nobene svetlobe. Strogo povedano Kubelka-Munkova enačba zahteva absolutni odboj, ki je definiran kot količnik I_S/I_0, kjer je I_0 intenziteta svetlobe, ki pade na vzorec. Meritve absolutnih odbojev so možne samo z uporabo integracijskih krogel, ki pa se v praksi zelo redko uporabljajo.

4.4 Multiplikativna korekcija razpršenosti (MSC)

Multiplikativna korekcija razpršenosti (angl. Multiplicative Scatter Correction - MSC) [2, 10, 12] je metoda, ki je bila razvita v osemdesetih letih prejšnjega stoletja. Prvi jo je predstavil Martens leta 1983, pozneje pa sta se z njo ukvarjala še Geladi in Næs. Originalno je bila razvita za zmanjšanje vpliva razpršenosti na difuzni refleksijski spekter, kasneje pa je prišla tudi v bolj splošno uporabo in se zato včasih imenuje tudi multiplikativna korekcija signala (angl. Multiplicative Signal Correction). Tukaj je predstavljena osnovna verzija metode MSC, obstajajo pa tudi bolj izpopolnjene verzije, ki predpostavljajo drugačen model razpršenosti za različne regije v spektru.

Princip metode MSC temelji na dejstvu, da ima razpršenost svetlobe drugačno odvisnost v valovnih dolžinah kot absorpcija svetlobe, ki jo povzroča nihanje molekul. Z uporabo podatkov pri več valovnih dolžinah je teoretično možno razlikovati med razpršenostjo in absorpcijo svetlobe.

Namig za to metodo je Martens našel s primerjavo vrednosti posameznih spektrov z vrednostmi nekega idealnega spektra. Idealni spekter je tisti spekter, ki vsebuje samo kemijsko informacijo v spektru. Očitno je, da perfektnega spektra, ki bi predstavljal vse možne vzorce, ni mogoče dobiti in ga je zato treba na neki način oceniti. Za idealni spekter se zato vzame kar povprečni spekter vseh spektrov v učni množici. Na sliki 4.1 je prikazan graf vrednosti posameznih spektrov proti vrednostim povprečnega spektra. Opazimo lahko, da vrednosti posameznega spektra ležijo zelo blizu neke regresijske premice. Takšne različne regresijske premice je Martens interpretiral kot razlike zaradi vpliva razpršenosti svetlobe, medtem ko je deviacije vrednosti spektra od regresijske premice interpretiral kot kemijsko informacijo v spektru.



Slika 4.1: Graf vrednosti spektrov petih vzorcev proti vrednostim povprečnega spektra vseh vzorcev.

Vidimo lahko, da imajo regresijske premice, ki gredo skozi posamezne spektre, različne odmike in naklone od povprečnega spektra. Ta empirična spoznanja nakazujejo, da ima vpliv razpršenosti svetlobe tako aditivno kot multiplikativno komponento.

Linearen regresijski model za vsak posamezen spekter lahko zapišemo z enačbo:

$$x_{i,k} = a_i + b_i \bar{x}_k + e_{i,k} \qquad (i = 1, \ldots, N;\ k = 1, \ldots, K) \tag{4.2}$$

kjer je i številka vzorca (učnega primera) in k indeks v zaporedju valovnih dolžin. Koeficient a_i predstavlja aditiven vpliv, koeficient b_i pa multiplikativen vpliv na spekter i-tega vzorca. Povprečje $\bar{x}_k$ izračunamo kot povprečna vrednost k-te spektralne spremenljivke vseh vzorcev v učni množici.

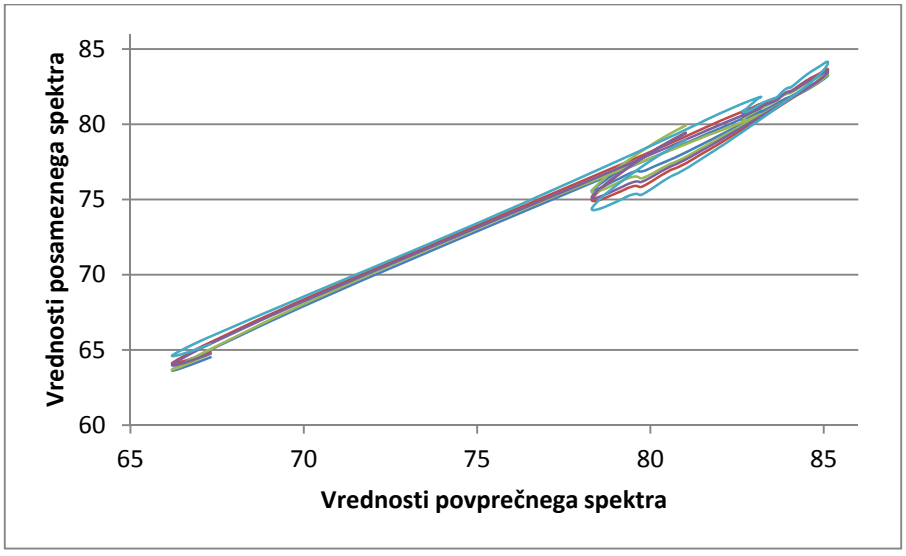
$$\bar{x}_k = \frac{1}{N} \sum_{i=1}^{N} x_{i,k}$$

Napaka $e_{i,k}$ v modelu ustreza vsem tistim vplivom v spektru, ki jih ne moremo modelirati le z aditivno in multiplikativno konstanto. Koeficienta a_i in b_i morata biti izračunana za vsak spekter posebej z uporabo vseh ali le podmnožice spektralnih meritev (valovnih dolžin). Koeficienta a_i in b_i izračunamo z linearno regresijo po principu najmanjših kvadratov.

Transformiran spekter potem dobimo tako, da od vrednosti originalnega spektra pri vsaki valovni dolžini odštejemo koeficient a_i in nato razliko delimo s koeficientom b_i.

$$x^*_{i,k} = \frac{x_{i,k} - a_i}{b_i}$$

Ključno pri tej metodi je, da morajo biti vsi spektri v učni množici in tudi novi spektri transformirani z uporabo istega idealnega spektra. Učinek te transformacije je prikazan na sliki 4.2. Vidimo, da je bilo večino variacije med spektri odstranjene.



Slika 4.2: Graf vrednosti spektrov petih vzorcev proti vrednostim povprečnega spektra vseh vzorcev po uporabi metode MSC.

Metoda MSC večinoma deluje dobro v primerih, ko je vpliv razpršenosti svetlobe prevladujoči izvor spremenljivosti v spektrih. To pa je značilno ravno za difuzni refleksijski spekter v spektroskopiji NIR. V kolikor pa to ne velja, potem ima kemijska informacija v

spektru, ki ji ustreza napaka v enačbi (4.2), velik vpliv na naklon in odmik regresijske premice. Transformacija bo v tem primeru pogostokrat preveč odvisna od kemijske informacije. To pa pomeni, da lahko metoda MSC iz spektrov odstrani tudi nekaj koristne informacije, kar pa lahko poslabša napovedi kemijskofizikalnih lastnosti. Za varno uporabo metode MSC bi morala biti vsota sestavin vzorca, ki absorbirajo svetlobo, enaka konstanti (npr. 100 %). Če to ne drži, potem je težko ali celo nemogoče razlikovati med kemijsko informacijo in informacijo, ki jo povzroča razpršenost svetlobe, saj imata oba pojava multiplikativen vpliv na spekter.

Za izračun koeficientov a_i in b_i lahko uporabimo vrednosti pri vseh valovnih dolžinah ali pa uporabimo samo podmožico. Če obstaja regija v spektru, ki je manj odvisna od kemijske informacije kot druge regije, potem lahko koeficienta a_i in b_i izračunamo le na tej regiji. Na ta način postaneta koeficienta bolj neodvisna od kemijske informacije. Transformacija MSC pa je seveda izvedena pri vseh valovnih dolžinah.

S spremembo naklona in odmika spektra vzorca se želimo čim bolj približali idealnemu povprečnemu spektru, vendar to še ne pomeni, da transformirani spekter predstavlja ”resnični” spekter vzorca, čeprav je glavni izvor naključne variacije med spektri odstranjen.

Metoda MSC se običajno uporablja le na spektrih, katerih vrednosti so približno linearno povezane s kemijskofizikalnimi lastnosti vzorca. Vsak refleksijski spekter mora biti zato pred uporabe metode MSC pretvorjen v absorpcijskega (A = log(1/R)) ali pa v spekter, ki uporablja Kubelka-Munkove enote. Poleg tega je metodo MSC smiselno uporabljati le na spektrih vzorcev, ki so kemijsko podobni. Če pa so vzorci po sestavi med seboj precej različni, potem ta metoda, ki teži k približevanju idealnemu spektru, ne bo dala želenih rezultatov.

Metoda MSC je bila uporabljena že pri mnogih aplikacijah v spektroskopiji NIR. V večini primerov je izboljšala rezultate napovedi modela. Poleg tega pa je tudi poenostavila model (npr. zmanjšala je število potrebnih komponent pri metodah PCR in PLSR) in izboljšala linearnost modela.

Primerjavo med metodo MSC in drugimi podobnimi metodami za predprocesiranje spektrov je naredil Helland. Ena od ugotovitev primerjave pravi, da je MSC konceptualno in empirično močno povezana z metodo SNV.

Prednost metode MSC pred metodami odvajanja spektra je v tem, da transformiran spekter ohrani obliko originalnega spektra.

4.5 Metoda Standard Normal Variate (SNV)

Metodo Standard Normal Variate (SNV) [10, 12] je prvič predstavil Barnes in se podobno kot metoda MSC uporablja za odstranjevanje oz. korekcijo vpliva razpršenosti svetlobe in vpliva velikosti delcev v snovi. Metoda SNV najprej centrira spekter okoli vrednosti nič, potem pa vsako vrednost spektra še normalizira s standardno deviacijo celotnega spektra. Centriranje spektra se izvede z odštevanjem povprečne vrednosti spektra preko vseh valovnih dolžin. Povprečna vrednost in standardna deviacija se izračunata na spektru, ki ga želimo transformirati in zato ne potrebujemo nobenega povprečnega spektra, izračunanega na učni množici, kot je to značilno za metodo MSC.

Naj bo $x_{i,k}$ vrednost i-tega spektra pri k-ti valovni dolžini. Metoda SNV izračuna transformiran spekter $x_{i,k}^*$ z naslednjo enačbo:

$$x_{i,k}^* = \frac{x_{i,k} - \bar{x}_i}{s_i}$$

$$\bar{x}_i = \frac{1}{K}\sum_{k=1}^{K} x_{i,k}$$

$$s_i = \sqrt{\frac{\sum_{k=1}^{K}(x_{i,k} - \bar{x}_i)^2}{K-1}}$$

kjer je $\bar{x}_i$ povprečje in s_i standardna deviacija vseh spektralnih vrednosti i-tega vzorca. Za transformiran spekter je značilno, da ima povprečno vrednost vedno enako 0 in varianco vedno enako 1, kar pomeni, da je neodvisen od originalnih vrednosti spektra.

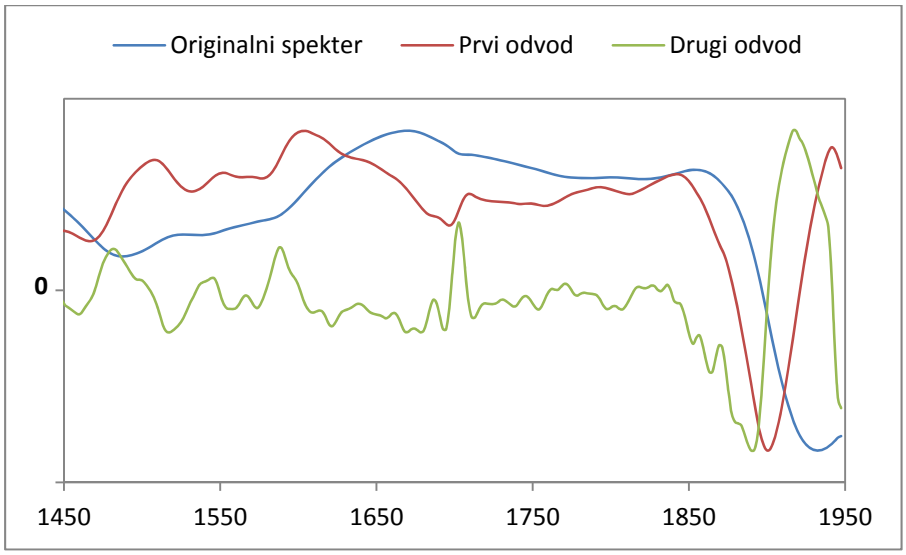
Rezultat metode SNV je v večini primerov zelo podoben rezultatu metode MSC, kar nam včasih povzroča preglavice pri izbiri med obema metodama. V praksi je zato najbolje preizkusiti obe metodi in pri končnem modelu izbrati tisto, ki vrača boljše rezultate. Prikaz spektrov, transformiranih z metodo SNV, je prikazan na sliki 4.9 na koncu poglavja. Opazimo lahko, da se transformiran spekter nahaja okrog abscisne osi, kar je posledica odštevanja povprečne vrednosti spektra. Zaradi normalizacije spektra s standardno deviacijo pa je spremenjen tudi razpon spektralnih vrednosti. Čeprav sta rezultata obeh metod podobna, vidimo, da se metodi razlikujeta v razponu vrednosti transformiranega spektra. Poleg tega pa je še ena razlika ali prednost metode SNV. Metodi SNV namreč za transformacijo spektra ni treba poznati vseh spektrov v učni množici, ampak podatke, ki so potrebni za transformacijo, izlušči iz samega spektra, ki ga želi transformirati. Metoda torej ne potrebuje povprečnega spektra, kot to zahteva metoda MSC. Ker je vsak spekter korigiran neodvisno od drugih, metoda SNV dovoljuje večje variabilnosti spektrov v učni množici kot metoda MSC.

Podobno kot pri metodi MSC, se tudi ta metoda običajno uporablja samo na spektrih, katerih vrednosti so linearno odvisne od vsebnosti sestavin v snovi. Vsak refleksijski spekter moramo zato prej linearizirati.

4.6 Odvajanje spektrov (SGD)

Ena od metod, ki se pogosto uporablja za odstranjevanje vpliva razpršenosti svetlobe oz. odstranjevanje premika osnovne črte (angl. baseline), je tudi odvajanje spektrov [2, 10, 12]. Ta metoda je ena od najzgodnejših, ki so bile uporabljene za korekcijo osnovne črte.

Na sliki 4.3 je prikazan originalni spekter ter prvi in drugi odvod istega spektra. Krivulje so bile povečane oz. zmanjšane in premaknjene tako, da so vse vidne na istem grafu. Prvi odvod spektra v neki točki predstavlja naklon krivulje originalnega spektra v tej isti točki. Vrhove ima tam, kjer ima originalni spekter največji naklon, ničle pa ima pri vrhovih originalnega spektra. Drugi odvod spektra pa predstavlja naklon prvega odvoda in meri ukrivljenost originalnega spektra. Drugi odvod je na neki način bolj podoben originalnemu spektru, saj ima vrhove na približno enakih mestih, čeprav imajo nasprotno smer.



Slika 4.3: Prikaz originalnega spektra ter prvega in drugega odvoda originalnega spektra.

Prvi odvod odstrani aditivno osnovno črto. Spekter, ki je vzporeden s spektrom na sliki 4.3 in premaknjen navzgor ali navzdol, ima enak prvi odvod, saj je naklon pri obeh enak. Drugi odvod pa odstrani linearno osnovno črto. Linearni premik osnovne črte ni enak multiplikativnemu učinku, vendar je v praksi videti zelo podobno.

Do zdaj smo se izogibali dejstvu, da izmerjeni spekter ni zvezna matematična krivulja, ampak zaporedje meritev na enako medsebojno oddaljenih diskretnih točkah. Obstaja več načinov za izračun odvodov. Najbolj enostaven način za izračun odvoda na takih podatkih je uporaba metode enostavnih razlik. Pri tej metodi se odvod v neki točki izračuna z razliko vrednosti dveh sosednjih točk. Prvi odvod $x'_{i,k}$ i-tega spektra pri k-ti valovni dolžini lahko tako izračunamo z enačbo:

$$x'_{i,k} = x_{i,k} - x_{i,k-1}$$

kjer je $x_{i,k}$ vrednost i-tega originalnega spektra pri k-ti valovni dolžini. Da bi dobili pravi naklon, bi praviloma morali deliti še z razliko med valovnima dolžinama dveh sosednjih točk. Ker pa to ne vpliva na napoved odvisne spremenljivke, se tega ponavadi ne naredi. Drugi odvod $x''_{i,k}$ i-tega spektra pri k-ti valovni dolžini pa je razlika med sosednjima točkama prvega odvoda:

$$x''_{i,k} = x_{i,k-1} - 2x_{i,k} + x_{i,k+1}$$

Problem metode enostavnih razlik je v tem, da računanje razlik med sosednjima točkama zmanjšuje signal in hkrati povečuje šum v spektru. Pred računanjem odvodov je zato v praksi potrebno uvesti še neko glajenje spektra.

Drugi bolj eleganten način pa sta predstavila Savitzky in Golay, s katerim se v večini primerov izognemo problemu povečevanja šuma in se poleg tega naredi še nekaj glajenja na podatkih. Vzamemo ozko okno okoli valovne dolžine, pri kateri želimo izračunati odvod. Na točke znotraj okna aproksimiramo polinom nizke stopnje po metodi najmanjših kvadratov. Tukaj smo uporabili okno širine 7, se pravi po tri točke na levi in desni strani, in aproksimirali kvadratno krivuljo. Kot vidimo na sliki 4.4, krivulja ne gre skozi točke, ampak je zelo blizu

vsem, saj je spekter na tako majhnem razponu valovnih dolžin skoraj kvadraten. S tem smo dosegli glajenje spektra okrog točke, ki nas zanima.

Kvadratna funkcija je zvezna krivulja, ki jo zapišemo z enačbo:

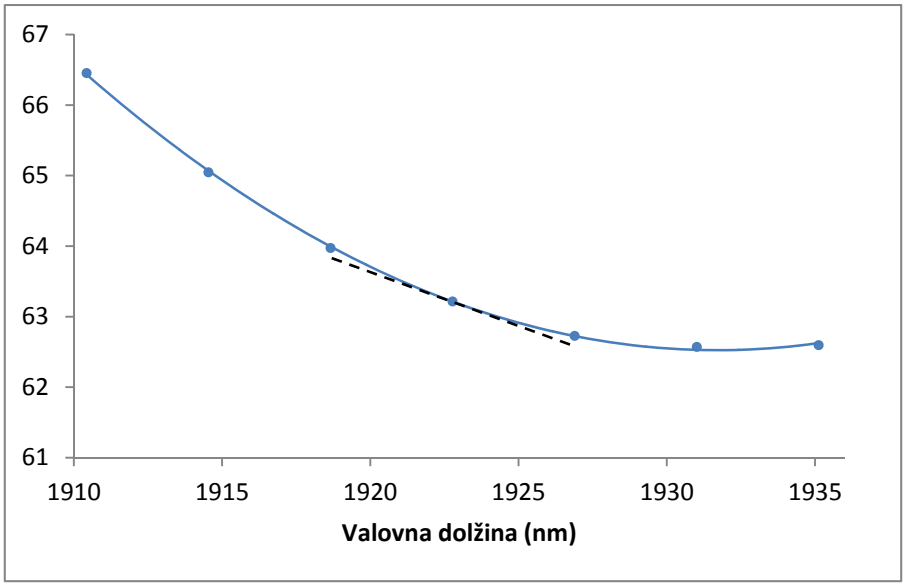
$$x = a + bw + cw^2$$

kjer je w valovna dolžina in x vrednost spektra pri valovni dolžini w. Koeficiente a, b in c dobimo z aproksimacijo kvadratne krivulje po metodi najmanjših kvadratov na točkah znotraj okna. Naklon oz. prvi odvod je enak:

$$\frac{dx}{dw} = b + 2cw$$

Drugi odvod pa je enak:

$$\frac{d^2x}{dw^2} = 2c$$



Slika 4.4: Primer aproksimacije kvadratne funkcije na točkah spektra znotraj okna širine 7.

Osnovna ideja, ki sta jo uporabila Savitzky in Golay za izračun odvodov, je ta, da namesto razlik sosednjih točk za odvod spektra vzamemo kar odvod aproksimirane krivulje v točki, ki predstavlja sredino okna.

Izračun odvoda je zelo preprost. Ko aproksimiramo polinomsko (npr. kvadratno) krivuljo na točkah znotraj okna, dobimo koeficiente, ki so linearne kombinacije spremenljivk x z utežmi, ki so funkcije spremenljivk w. Ker pa so valovne dolžine med seboj enako oddaljene, lahko odvod enostavno izračunamo z uteženo vsoto točk znotraj okna. Če uporabljamo kvadratno funkcijo in okno širine 7, so uteži za izračun prvega odvoda enake (-3, -2, -1, 0, 1, 2, 3)/28, za izračun drugega odvoda pa (5, 0, -3, -4, -3, 0, 5)/42. Odvajanje celotnega spektra poteka tako, da najprej izračunamo uteži, potem pa okno premikamo vzdolž spektra in odvode izračunavamo z uteženimi vsotami vrednosti spektra znotraj okna. Ta način računanja je znan tudi pod imenom linearno filtriranje.

Podobno kot pri metodi preprostih razlik, tudi tukaj ne dobimo pravega odvoda. Izračuni namreč temeljijo na predpostavki, da se sosednji točki razlikujeta za 1 enoto valovne dolžine. Pravi odvod bi sicer lahko izračunali z upoštevanjem dejanske razlike med dvema sosednjima točkama, vendar pa s tem ne bi nič pridobili, saj to ne vpliva na napovedovanje kemijskofizikalnih lastnosti.

Pri izbiri velikosti okna (ali dolžine filtra) tehtamo med zmanjšanjem šuma ali glajenjem in popačenjem krivulje, kadar je okno preširoko. Učinek preširokega okna je preveliko zaokroževanje vrhov, s čimer lahko izgubimo nekaj pomembne informacije. Problem ozkega okna pa je v tem, da se šum v originalnem spektru z odvajanjem le še poveča. Eden od načinov za iskanje primernega okna je, da začnemo s tremi točkami in okno povečujemo, dokler ne najdemo ustreznega odvoda, ki na pogled ne vsebuje preveč šuma.

Poleg izbire ustrezne velikosti okna je pomembna tudi izbira stopnje polinoma, ki ga aproksimiramo znotraj okna. Polinomi višjih stopenj so sicer bolj prilagodljivi, vendar pa za dosego enakega zmanjšanja šuma potrebujemo širše okno.

Savitzky-Golayeva metoda (angl. Savitzky-Golay Derivate - SGD) je sicer bolj elegantna kot metoda enostavnih razlik, vendar ali je sploh kaj boljša? Če primerjamo metodi na umetnih primerih, potem je jasno, da se Savitzky-Golayeva metoda bolj približa pravemu odvodu. Če pa razmišljamo o odvodih kot metodi predprocesiranja spektrov, še ne pomeni, da bo bolj natančen odvod pripomogel k boljšim napovedim. Za napovedovanje kemijskofizikalnih lastnosti papirja je bila vseeno uporabljena Savitzky-Golayeva metoda.

Slabost metode odvajanja spektrov je v tem, da spremeni obliko originalnih spektrov. Drugi odvod ima sicer prednost pred prvim, saj se vrhovi v drugem odvodu pojavijo na podobnih mestih kot vrhovi v originalnem spektru, le smer imajo nasprotno. Za drugi odvod je značilno tudi, da ima več značilnosti (vrhov). Kadar originalni spekter ne vsebuje nekih značilnosti, je to lahko prednost, saj nam omogoča identifikacijo šibkih vrhov, ki v originalnem spektru niso vidni. Odvod namreč vrhove v originalnem spektru zoži in bolj naostri. Kadar pa je originalni spekter že tako ali tako kompleksen, pa vsako povečevanje kompleksnosti ni dobrodošlo, saj otežuje interpretacijo modelov.

4.7 Ortogonalna korekcija signala (OSC)

V zadnjem času je bila kot alternativa metodama MSC in SNV razvita metoda, ki se imenuje ortogonalna korekcija signala (angl. Orthogonal Signal Correction - OSC) [2, 10, 19]. Uspešno je bila uporabljena že pri kar nekaj analizah spektrov NIR.

Osnovna ideja metode OSC je zelo preprosta in lahko razumljiva. V spektru želimo obdržati samo tisto informacijo, ki prispeva pri napovedovanju kemijskofizikalne lastnosti snovi, preostalo informacijo pa iz spektra odstraniti. Metoda OSC iz spektralne matrike X odstrani močno strukturirano varianco, ki ni korelirana s spremenljivko y, oziroma varianco, ki je ortogonalna (pravokotna) na spremenljivko y. S filtriranjem ortogonalne variance želimo odstraniti šum, vpliv razpršenosti svetlobe, premike osnovne črte in druge pojave, ki nič ne prispevajo pri napovedovanju kemijskofizikalnih lastnosti, in tako izluščiti samo tisto varianco, ki prispeva pri napovedi.

Cilj metode OSC [5] je popolnoma odstraniti tisto variacijo v spektralni matriki X, ki je ortogonalna na neko spremenljivko y. Poglejmo si naslednjo enačbo.

$$X = R + Z + E$$

Matrika R predstavlja želeni signal, Z označuje matriko, ki vsebuje neželeno variacijo in E označuje matriko belega šuma. Ker je matrika Z ločena od matrike R, je možno najti tako projekcijo v podprostor prostora X, ki bo eliminirala Z. Obstajajo različni algoritmi za iskanje tega podprostora. Če matrika Z vsebuje spektralno variacijo, ki je ortogonalna na y, potem lahko ortogonalno komponento dobimo z enačbo:

$$Z = TP^T$$

Matrika P predstavlja množico baznih vektorjev, matrika T pa uteži v podprostoru, ki ga razteza množica baznih vektorjev P. Množica baznih vektorjev mora biti ortogonalna na y. Ker pa v originalnem prostoru največkrat najdemo le eno dve stabilni ortogonalni komponenti, ostale komponente pa nimajo stabilnih smeri, lahko opuščanje omejitve stroge ortogonalnosti privede do boljših rezultatov napovedi.

Osnovni model OSC za eno komponento lahko zapišemo z enačbama:

$$X = tp^T + E$$

$$t = Xw$$

$$\|w\| = 1$$

Vektor w je vektor uteži, matrika E pa predstavlja preostanek informacije, ki je ni mogoče modelirati z ortogonalno komponento. Dodatno komponento lahko odstranimo tako, da namesto matrike X uporabimo matriko E.

Model OSC mora izpolnjevati tri zahteve:

- komponenta mora vsebovati sistematično (močno strukturirano) variacijo v spektralni matriki X,
- komponento je mogoče izračunati samo iz spektralne matrike X (zato, da jo lahko uporabimo na novih spektrih),
- komponenta mora biti ortogonalna na spremenljivko y.

Za izračun ortogonalnih komponent (matrike T) se uporabljata dva različna načina, in sicer indirekten (originalen) in direkten način. Delimo ju glede na to, na kakšen način ocenijo oz. izračunajo ortogonalne komponente. Indirekten način ima kar nekaj problemov. Ker je indirekten algoritem iterativen in ker za izračun komponent uporablja regresijski model, ima težave s hitrostjo, z ortogonalnostjo komponent in s prevelikim prileganjem učnim podatkom. Direkten način pa ne uporablja iterativnega algoritma ter notranjega regresijskega modela in ima zato manj težav. Originalni (indirekten) algoritem je razvil Wold, kasneje pa je bilo objavljenih še veliko alternativnih algoritmov (Sjöblom, Andersson, Fearn, Westerhuis, Trygg in Wold) [19]. Dosedanje raziskave so pokazale, da so rezultati različnih algoritmov podobni.

V dosedanjih raziskavah je bilo pokazano tudi, da predprocesiranje z metodo OSC ne prinaša pomembnih izboljšav pri napovedih z regresijo delnih najmanjših kvadratov (PLSR). Metoda OSC v večini primerov samo zmanjša število komponent v modelu PLSR, ne izboljša pa same napovedi. Število komponent v modelu PLSR se zmanjša za toliko, kolikor je bilo odstranjenih ortogonalnih komponent. Pomaga pa nam pri razumevanju modelov, saj je modele z manjšim številom komponent lažje interpretirati.

Razlog za ta pojav je verjetno v tem, da podobno kot metoda PLSR tudi metoda OSC modelira globalno varianco v podatkih. Postopek, po katerem dobimo globalne komponente z metodo OSC, je analogen postopku za izračun globalnih komponent pri metodi PLSR. Če vektorje v matriki X, ki so ortogonalni na y, odstranimo z metodo OSC, potem zna te vektorje upoštevati tudi metoda PLSR.

Drugi razlog, da filtriranje z metodo OSC ne izboljša napovedi pa je v tem, da večina metod OSC za odstranjevanje variance zahteva strogo ortogonalnost komponent. Westerhuis je pokazal, da stroga ortogonalnost privede do prevelikega prileganja komponent k učnim podatkom. To se zgodi zato, ker metode ne upoštevajo možnosti, da lahko pri merjenju spremenljivke y prihaja tudi do napak.

Feudale, Tan in Brown [4] so predlagali novi algoritem, ki deluje po korakih. To pomeni, da se pri vsaki točki oz. valovni dolžini spektra vzame neko okno, ki vključuje nekaj sosednjih točk. Ortogonalna korekcija signala se potem za vsako valovno dolžino posebej izvede samo na točkah znotraj okna. S tem namesto globalne variance odstranimo lokalno varianco. Pokazano je, da predprocesiranje s to metodo izboljša rezultate modela PLSR. Pri metodi OSC, ki jo izvajamo po korakih, je bil uporabljen novejši algoritem, za katerega velja, da ni iterativen, ne vsebuje notranjega regresijskega modela in nima omejitve stroge ortogonalnosti variance.

Algoritem za ortogonalno korekcijo signala po korakih je malenkost spremenjena verzija Fearnovega algoritma:

1. Izračunamo večine variacije v X, ki je povezana z Y (tukaj je spremenljivka y predstavljena z matriko Y).
   $$M = 1 - X^T Y (Y^T X X^T Y)^{-1} Y^T X$$
2. Izračunamo matriko Z, tako da je produkt ZZ^T simetrična matrika.
   $$Z = XM$$
3. Izračunamo prvo glavno komponento (angl. first principal komponent) produkta ZZ^T. S pomočjo lastnega vektorja p in lastne vrednost y produkta ZZ^T izračunamo vektor uteži w.
   $$w = (\lambda)^{-1/2} M X^T p$$
4. Z množenjem matrike X in vektorja uteži w izračunamo še drugi vektor uteži t.
   $$t = Xw$$
5. Ortogonaliziramo t na Y in izračunamo vektor p^*.
   $$t^* = t - Y(Y^T Y)^{-1} Y^T t$$
   $$p^* = X^T t^* / (t^{*T} t^*)$$
6. Z odštevanjem ortogonalne komponente od matrike X dobimo matriko E.
   $$E = X - t^* p^{*T}$$
7. Če želimo odstraniti še dodatno komponento, potem namesto matrike X uporabimo matriko E in ponovimo korake od 1 od 6.
8. Na novih spektrih lahko naredimo ortogonalno korekcijo s pomočjo naslednje enačbe, pri čemer matriki W in P dobimo v prejšnjih korakih.
   $$X^*_{test} = X_{test} - X_{test} W (P^T W)^{-1} P^T$$

Dobra lastnost zgornjega algoritma je, da odstranjena varianca ne rabi biti strogo ortogonalna na Y, saj algoritem ne uporablja iterativnega postopka, ki bi konvergiral k končni rešitvi. Druga prednost pa je, da med ortogonalizacijo ne potrebuje notranjega regresijskega modela PLSR in zato ni treba optimizirati števila komponent v notranjem modelu PLSR.

V zgornjem algoritmu pa nastopi problem, kadar v spektru odstranimo več kot eno komponento. Po filtriranju ene komponente se za izračun naslednjih komponent namesto matrike X uporablja matrika E. Ker pa pride do zmanjšanja ranga matrike, so lastne vrednosti, ki jih izračunamo v tretjem koraku, blizu 0 in deljenje s temi vrednostmi povzroči neželene motnje v filtriranem spektru. Da bi se izognili temu problemu, so algoritem Feudale, Tan in Brown [5] še izboljšali.

Izboljšan algoritem za ortogonalno korekcijo signala po korakih:

1. Izračunamo prvo glavno komponento matrike X in s pomočjo lastnega vektorja p izračunamo vektor uteži w.
   $$w = p/(p^T p)$$
2. Ortogonaliziramo vektor t na Y in izračunamo nov vektor p^*.
   $$t^* = t - Y(Y^T Y)^{-1} Y^T t$$
   $$p^* = X^T t^*/(t^{*T} t^*)$$
3. Z odštevanjem ortogonalne komponente od matrike X dobimo matriko E.
   $$E = X - t^* p^{*T}$$
4. Če želimo odstraniti še dodatno komponento, potem namesto matrike X uporabimo matriko E in ponovimo korake od 1 od 3.
5. Na novih spektrih lahko naredimo ortogonalno korekcijo s pomočjo naslednje enačbe, pri čemer matriki W in P dobimo prejšnjih korakih.
   $$X^*_{test} = X_{test} - X_{test} W (P^T W)^{-1} P^T$$

Glavna razlika med algoritmoma je v korakih od 1 do 4 originalnega algoritma in koraku 1 novega algoritma. Podroben pregled razkrije, da sta matematično podobna. V originalnem algoritmu je v korakih od 1 do 4 najprej izračunan vektor največje kovariance (prva komponenta modela PLSR), v koraku 5 pa je potem odstranjen del lastnega vektorja, ki je koreliran z Y. V novem algoritmu pa je v koraku 1 izračunan vektor maksimalne variance (prva glavna komponenta matrike X) in potem v koraku 2 ortogonaliziran na Y. Ker je glavni cilj metode OSC izračun vektorjev, ki minimizirajo kovarianco med X in Y, je drugi algoritem bolj ustrezen.

Ortogonalna korekcija signala po korakih se izračuna le za eno valovno dolžino naenkrat. Pri vsakem koraku premaknemo okno za eno mesto naprej. Ta postopek potem ponavljamo, dokler ne filtriramo celotnega spektra.

Naj bo $O(X,Y)$ ortogonalno korigiran signal matrike X z uporabe matrike Y. Pri ortogonalni korekciji signala dobimo matriki W in P, ki ju uporabimo za korekcijo novih spektrov.

$$[W, P] = O(X, Y)$$

Naj bo $2r + 1$ velikost spektralnega območja (velikost okna), ki ga bomo ortogonalizirali. Če želimo narediti korekcijo na celotnem spektru (vseh K valovnih dolžinah), moramo matriki W in P izračunati za vsako valovno dolžino posebej.

$$[W_j, P_j] = O(X_{j-r:j+r}, Y) \qquad (j = r+1, r+2, \ldots, K-r)$$

Črka j označuje indeks valovne dolžine.

Korekcijo na novem spektru izvedemo z enačbo:

$$X^*_{test,j-r:j+r} = X_{test,j-r:j+r} - X_{test,j-r:j+r} W_j (P_j^T W_j)^{-1} P_j^T$$

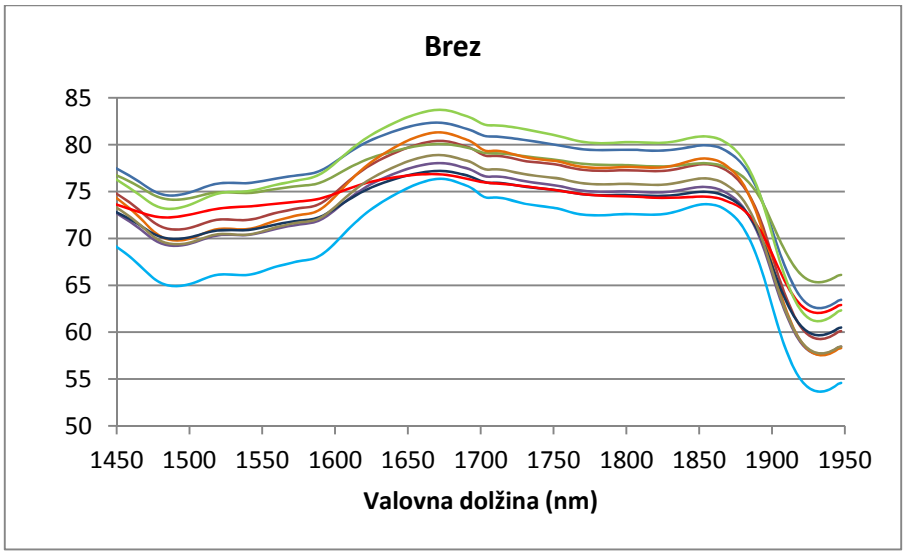
Za korekcijo signala pri j-ti valovni dolžini vzamemo sredinsko točko okna $(j-r:j+r)$. Če želimo narediti korekcijo signala tudi na začetnih $(1, 2, \ldots, r)$ in končnih $(K-r, K-r+1, \ldots, K)$ točkah spektra, vzamemo za točke, ki so pred začetkom ali po koncu spektra, kar zrcalno sliko spektra glede na njegov začetek oziroma konec.

Najboljša lastnost metode OSC glede na ostale metode predprocesiranja spektrov je, da varianco oz. informacijo, ki vpliva na napoved, pusti popolnoma nedotaknjeno. Metoda OSC v nasprotju z ostalimi metodami predprocesiranja zahteva optimizacijo števila komponent za vsako kemijskofizikalno lastnost posebej. Določiti je potrebno optimalno število komponent, ki jih bomo odstranili. Več komponent kot jih odstranimo, bolj zmanjšamo ortogonalno varianco. Uspeh metode OSC je delno odvisen tudi od natančnosti meritev spremenljivke, ki jo napovedujemo. Druge metode predprocesiranja spektrov pa niso odvisne od vrednosti kemijskofizikalne lastnosti in lahko spektre obdelamo tudi, če vrednosti ne poznamo. Poleg tega se predprocesiranje spektra pri drugih metodah naredi samo enkrat in ne za vsako kemijskofizikalno lastnost posebej. Ker metoda OSC optimizira predprocesiranje spektrov za vsako spremenljivko posebej, lahko vodi do boljših rezultatov napovedi posameznih kemijskofizikalnih lastnosti.

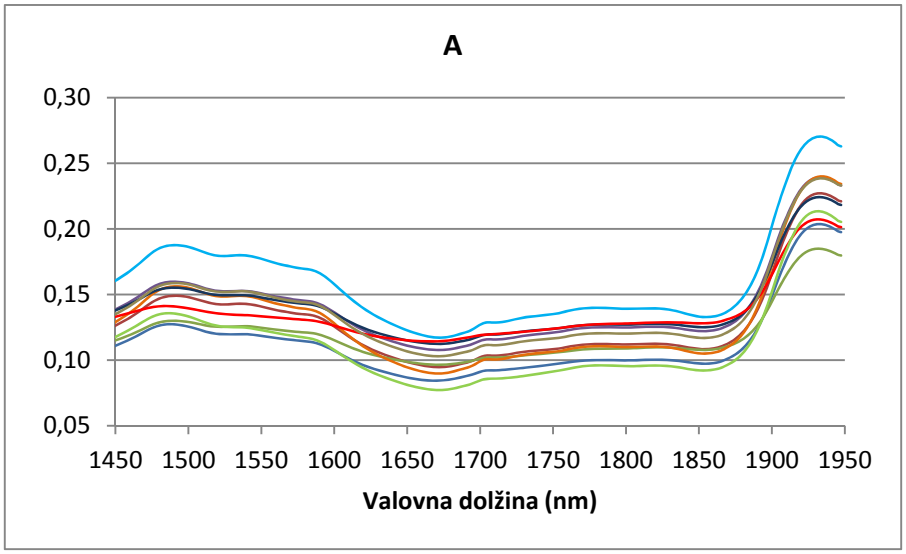
Ortogonalna korekcija signala pa ima tudi slabosti. Prva slabost metode OSC v primerjavi z metodo MSC je, da spektri, obdelani z metodo OSC, ne izgledajo več kot originalni spektri. To je posledica tega, da je bil del resničnega signala spektra odstranjen skupaj z delom, ki ga predstavljajo vpliv razpršenosti svetlobe in vplivi drugih pojavov. Druga pomanjkljivost pa je, da se intenziteta preostalega signala zmanjšuje s številom odstranjenih ortogonalnih komponent.

4.8 Grafični prikaz metod predprocesiranja spektrov

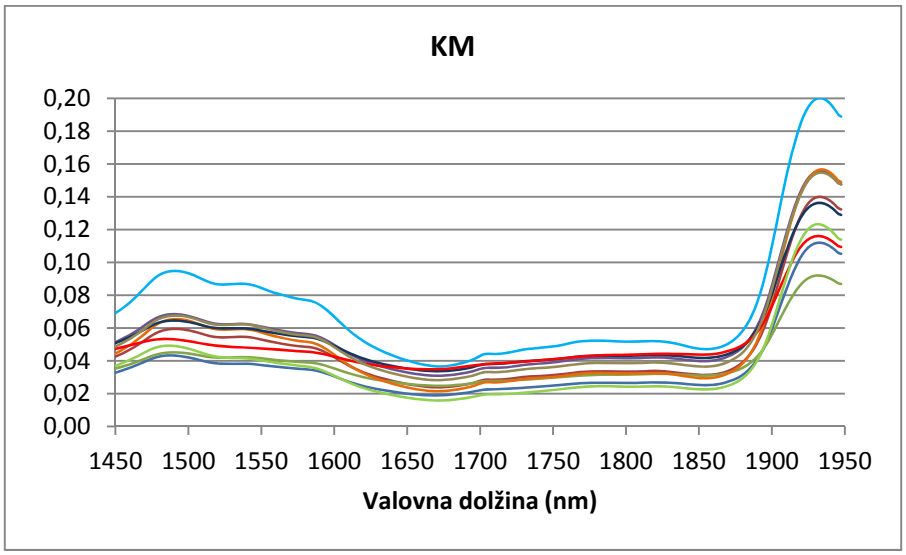
V nadaljevanju so prikazani grafi transformiranih spektrov, ki ponazarjajo, kako posamezne metode predprocesiranja spektrov vplivajo na originalne spektre. Prvi graf prikazuje originalne spektre 10 vzorcev papirja, v nadaljevanju pa so prikazani še grafi transformiranih spektrov za vse metode oz. kombinacije metod predprocesiranja spektrov, ki so bile preizkušene pri napovedovanju kemijskofizikalnih lastnosti papirja.



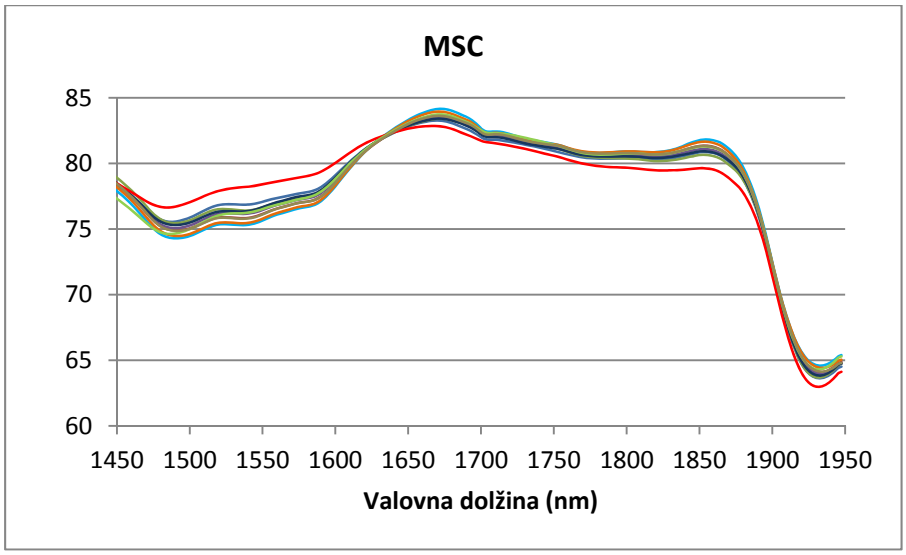
Slika 4.5: Originalni spektri 10 vzorcev papirja.



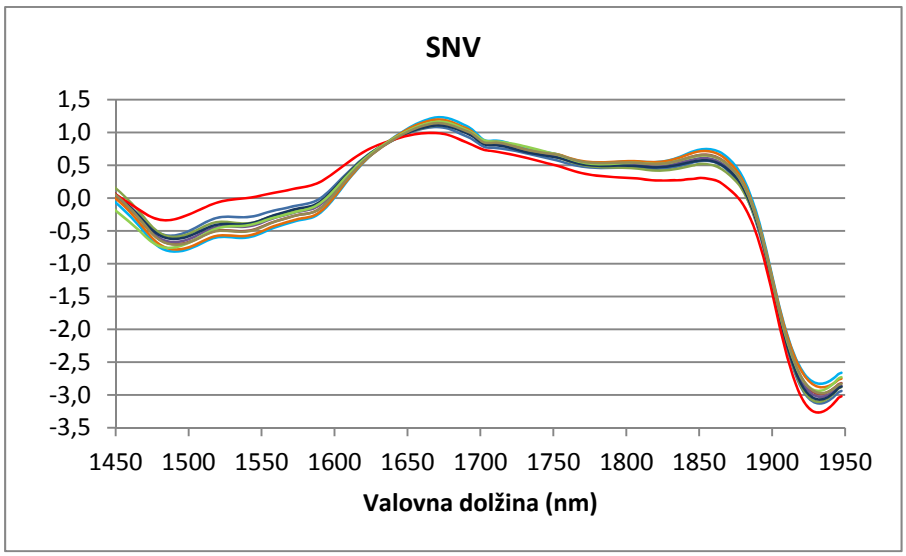
Slika 4.6: Spektri 10 vzorcev papirja po uporabi absorpcijske transformacije.



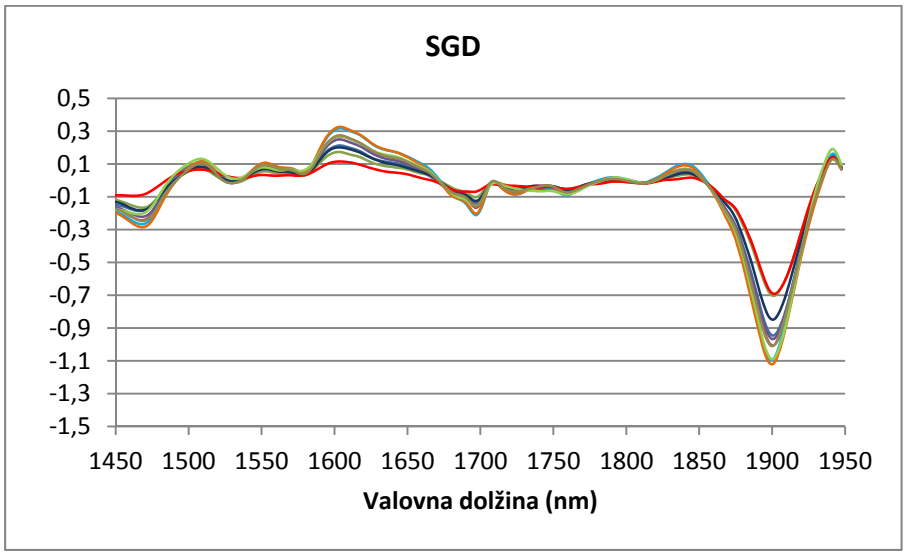
Slika 4.7: Spektri 10 vzorcev papirja po uporabi Kubelka-Munkove transformacije.



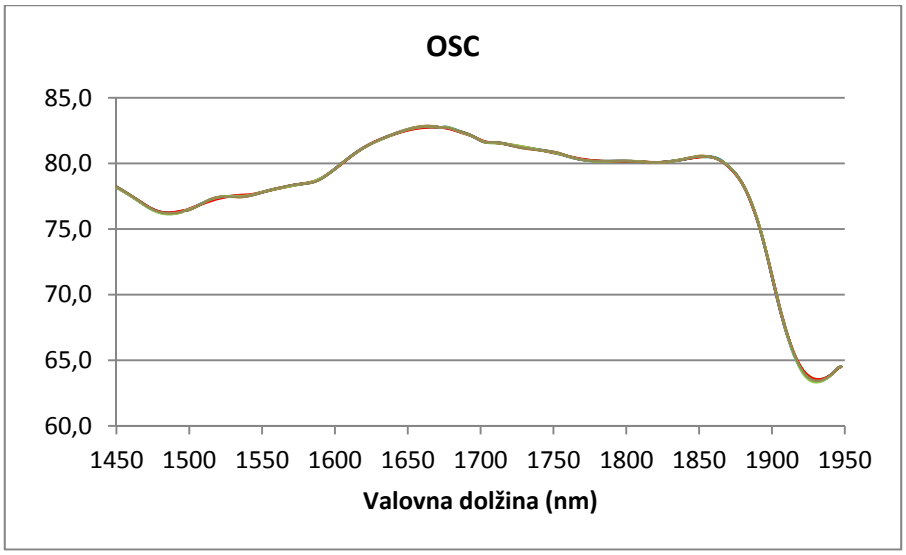
Slika 4.8: Spektri 10 vzorcev papirja po uporabi multiplikativne korekcije razpršenosti.



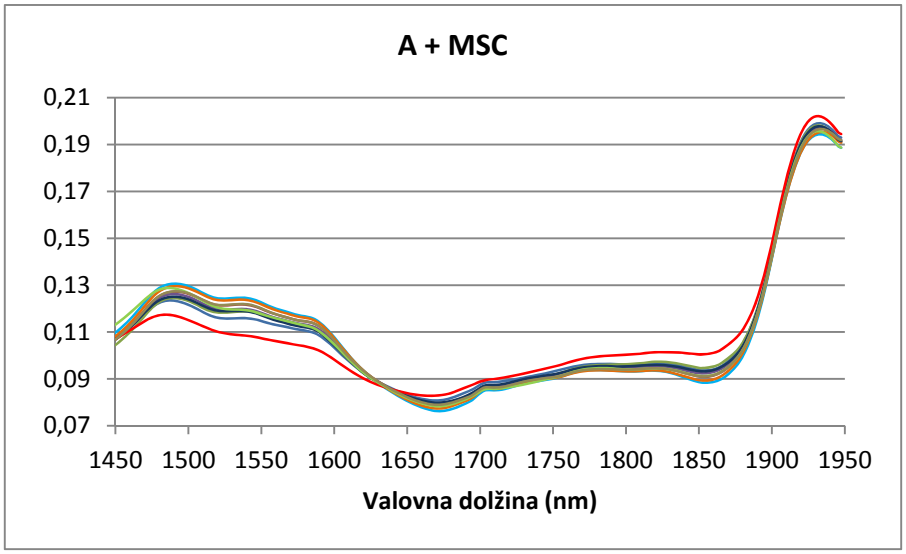
Slika 4.9: Spektri 10 vzorcev papirja po uporabi metode Standard Normal Variate.



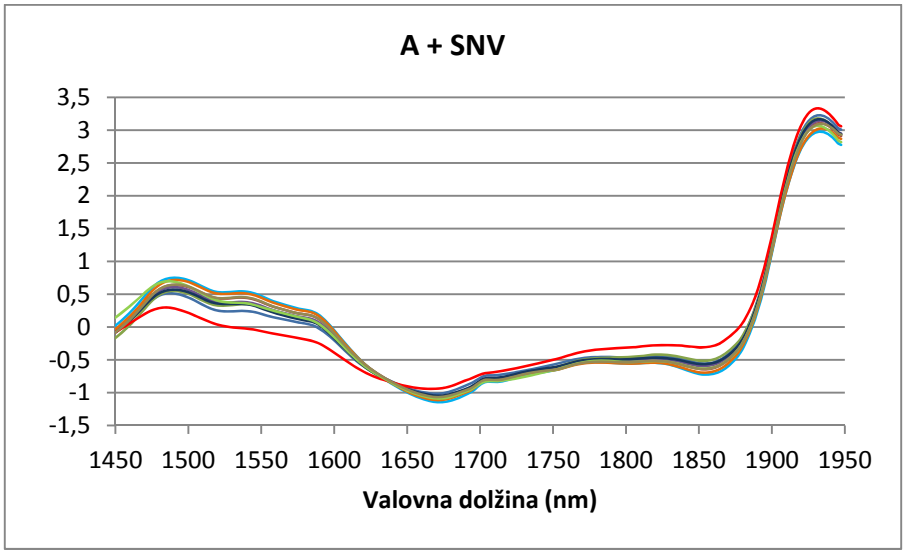
Slika 4.10: Spektri 10 vzorcev papirja po odvajanju s Savitzky-Golayevo metodo.



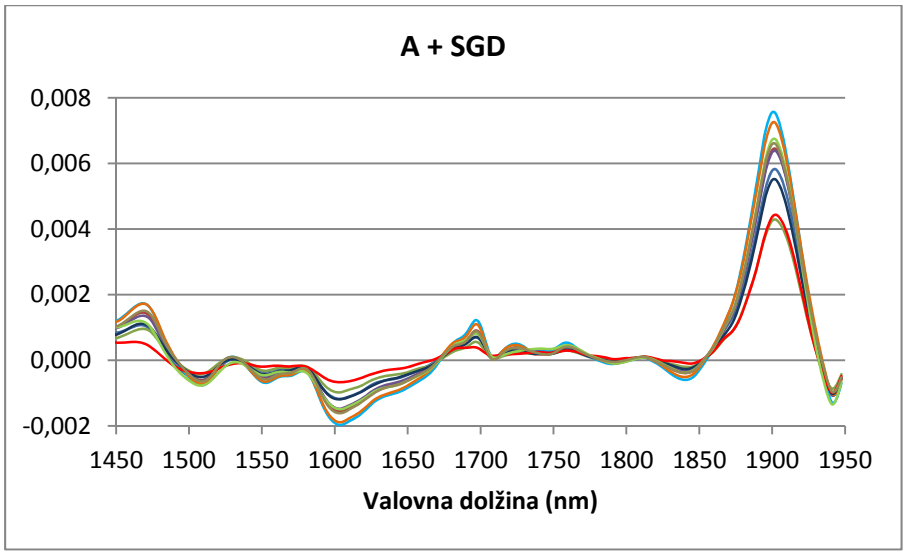
Slika 4.11: Spektri 10 vzorcev papirja po uporabi ortogonalne korekcije signala po korakih.



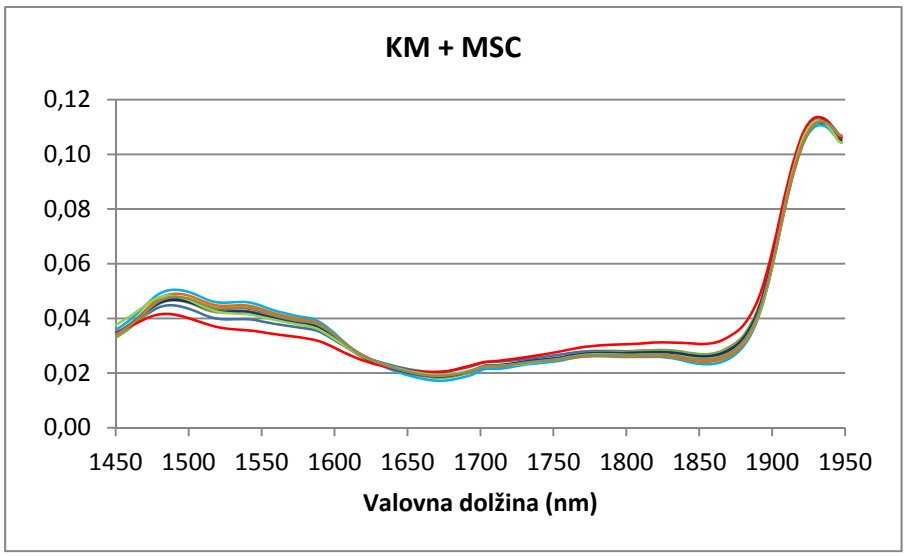
Slika 4.12: Spektri 10 vzorcev papirja po uporabi metode A in metode MSC.



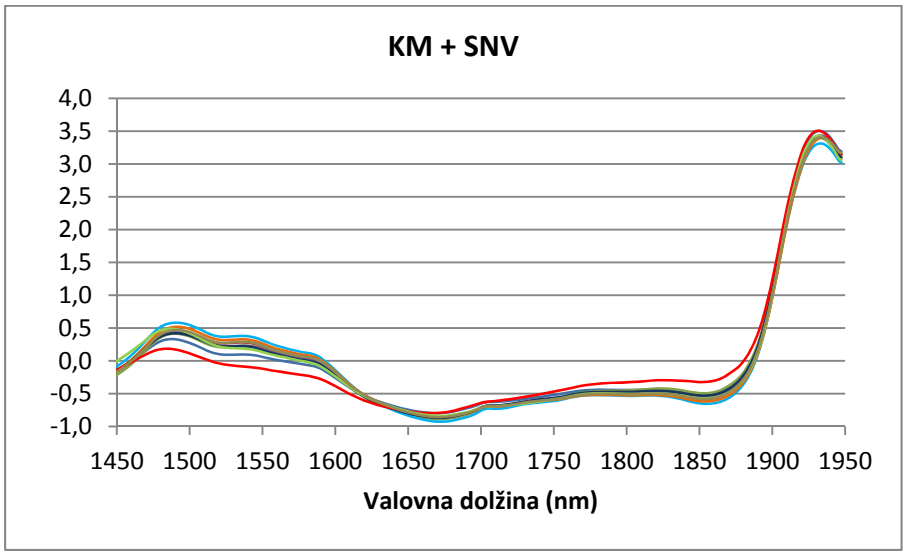
Slika 4.13: Spektri 10 vzorcev papirja po uporabi metode A in metode SNV.



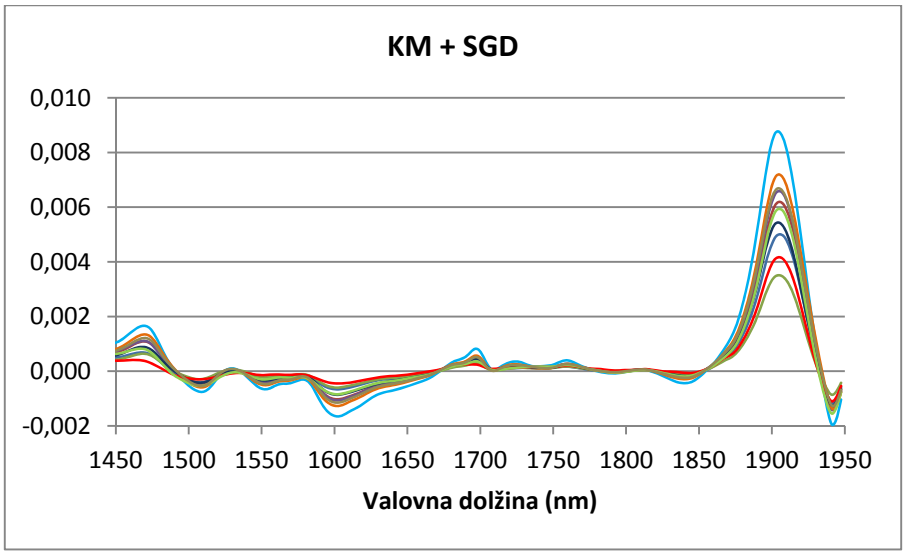
Slika 4.14: Spektri 10 vzorcev papirja po uporabi metode A in metode SGD.



Slika 4.15: Spektri 10 vzorcev papirja po uporabi metode KM in metode MSC.



Slika 4.16: Spektri 10 vzorcev papirja po uporabi metode KM in metode SNV.



Slika 4.17: Spektri 10 vzorcev papirja po uporabi metode KM in metode SGD.

Poglavje 5

Metode strojnega učenja

5.1 Enostavna linearna regresija

Enostavna linearna regresija [7, 15] pri napovedovanju kemijskofizikalne lastnosti uporablja le en atribut (spektralno spremenljivko). Vrednost kemijskofizikalne lastnosti $\hat{y}$ napove tako, da aproksimira linearno funkcijo skozi vse primere (vzorce) v učni množici. Linearna funkcija je podana z enačbo:

$$\hat{y} = ax_k + b$$

kjer je x_k k-ta spektralna spremenljivka, a in b pa sta koeficienta, ki ju moramo oceniti. Koeficienta a in b določimo tako, da minimiziramo vsoto kvadratov napake (SSE) napovedi kemijskofizikalne lastnosti vseh učnih primerov.

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Enostavna linearna regresija preizkusi vse atribute in izbere tistega, pri katerem je vsota kvadratov napake napovedi najmanjša.

5.2 Linearna regresija

Linearna regresija [7, 15] je posplošitev enostavne linearne regresije in pri napovedovanju kemijskofizikalne lastnosti uporablja enega ali več atributov. Linearna regresija aproksimira funkcijo, ki je linearna kombinacija vseh ali izbrane podmnožice atributov.

Naj bo $v^T = \langle 1, x_1, \ldots, x_K \rangle$ vektor vrednosti vseh spektralnih spremenljivk razširjen z elementom 1. Aproksimirano funkcijo podaja enačba:

$$\hat{y} = w_0 + \sum_{k=1}^{K} w_k x_k = w^T v$$

Naloga je določiti vektor w parametrov w_k, $k = 0, \ldots, K$, tako, da minimiziramo vsoto kvadratov napake (SSE) napovedi kemijskofizikalne lastnosti preko vseh učnih primerov.

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{k=1}^{K} w_k x_{i,k} \right)^2$$

Z V označimo razširjeno spektralno matriko:

$$V = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,K} \\ 1 & x_{2,1} & \cdots & x_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,K} \end{bmatrix}$$

in z $\boldsymbol{y}^T = \langle v_1, \ldots, v_N \rangle$ vektor vrednosti kemijskofizikalne lastnosti vseh učnih primerov. Minimum vsote kvadratov napake dobimo takrat, ko velja naslednja enakost:

$$w = (V^T V)^{-1} V^T y$$

Pri tem moramo paziti, da je število učnih primerov vedno večje od števila spektralnih spremenljivk, ki jih uporabimo pri aproksimiranju linearne funkcije, ker drugače ne bo možno izračunati inverza produkta $V^T V$.

5.3 Regresija pace

Regresija pace je metoda za aproksimacijo linearnih modelov. Razvil jo je Wang in odpravlja mnoge pomanjkljivosti navadne linearne regresije, predvsem problem izbire podmnožice atributov, ki se uporabljajo v regresijski funkciji. Regresija pace izboljša navadno linearno regresijo tako, da izračuna vpliv vsakega atributa na linearen model. Najprej se naredi ortogonalna dekompozicija linearnega modela, nato pa se izračunajo pričakovani prispevki posameznih atributov v tej ortogonalni dekompoziciji.

Razvitih je bilo 6 različnih metod, imenovanih $PACE_1$ do $PACE_6$. Dosedanji eksperimenti so pokazali, da regresija pace bolje napoveduje od navadne linearne regresije. Algoritmi metod in druge podrobnosti o regresiji pace so opisani v članku [16].

5.4 Metode najbližjih sosedov

Metode najbližjih sosedov [7, 15] so metode, pri katerih ni skoraj nobenega učenja, saj se večina izračunov naredi šele pri napovedi. Za te metode je značilno, da kot model shranijo samo učno množico. Na napoved novega primera vpliva le nekaj najbližjih sosedov. V učni množici se torej poišče nekaj učnih primerov, ki so najbolj podobni novemu primeru. Te učne primere se potem uporabi za napoved novega primera. Ker se večina računanja opravi šele pri napovedovanju novih primerov, so te metode bolj časovno potratne kot druge metode strojnega učenja.

Tukaj sta predstavljeni dve glavni metodi najbližjih sosedov, in sicer metoda k-najbližjih sosedov in linearna lokalno utežena regresija.

5.4.1. Metoda k-najbližjih sosedov

Metoda k-najbližjih sosedov [7, 15] deluje tako, da za vsak primer (vzorec papirja), katerega vrednost kemijskofizikalne lastnosti želimo napovedati, poišče k najbližjih primerov iz učne množice. Teh k najbližjih sosedov se potem uporabi pri napovedi vrednosti kemijskofizikalne lastnosti novega primera.

Najenostavnejša varianta te metode je, da za vrednost kemijskofizikalne lastnosti l-tega vzorca papirja napovemo povprečno vrednost te kemijskofizikalne lastnosti vseh k najbližjih sosedov. To pomeni, da poiščemo k najbolj spektralno podobnih vzorcev in izračunamo povprečno vrednost KFL teh vzorcev.

$$\hat{y}_l = \frac{1}{k}\sum_{i=1}^{k} y_i$$

Pri enostavni varianti ne upoštevamo razdalje med primerom, katerega vrednost napovedujemo, in drugimi učnimi primeri. Druga varianta te metode pa upošteva razdaljo. Razdalja se namreč uporabi za uteževanje učnih primerov. S tem želimo doseči, da bližnji primeri bolj vplivajo na napoved kot bolj oddaljeni primeri. Bližnji primeri zato dobijo večjo utež kot tisti, ki so bolj stran.

Razdalja med dvema primeroma se lahko izračuna na veliko načinov. Največkrat se za izračun razdalje uporabi evklidska razdalja. Evklidska razdalja je običajna razdalja med dvema točkama v večdimenzionalnem koordinatnem sistemu. Obstajajo pa tudi druge vrste razdalj, kot so Čebiševa, Manhattanska in Minovskijeva razdalja. Evklidska razdalja med dvema vzorcema papirja je podana z:

$$d(x_j, x_l) = \sqrt{\sum_{i=1}^{K} (x_{i,j} - x_{i,l})^2}$$

Razdalja lahko različno vpliva na uteževanje učnih primerov. Funkciji vpliva učnega primera glede na razdaljo pravimo tudi jedrna funkcija $K(d)$. Tukaj je prikazanih nekaj jedrnih funkcij [1], ki jih lahko uporabimo za uteževanje učnih primerov.

Inverzna jedrna funkcija:

$$K(d) = \frac{1}{1-d}$$

Gaussova jedrna funkcija:

$$K(d) = e^{-d^2}$$

Kvadratna jedrna funkcija:

$$K(d) = \begin{cases} (1-d^2), & |d| < 1 \\ 0 \quad , & |d| \geq 1 \end{cases}$$

Če pri napovedovanju uporabljamo uteževanju učnih primerov, potem lahko napoved $\hat{y}_l$ izračunamo kot uteženo povprečje vrednosti kemijskofizikalne lastnosti najbližjih sosedov.

$$\hat{y}_l = \frac{\sum_{i=1}^{k}[yK(d(x_i, x_l))]}{\sum_{i=1}^{k} K(d(x_i, x_l))}$$

Za iskanje najbližjih sosedov lahko uporabimo več različnih algoritmov, ki se razlikujejo predvsem po hitrosti iskanja. Največkrat se uporablja linearno iskanje. Kadar pa je linearno iskanje prepočasno, pa lahko uporabimo tudi kD-drevesa ali prekrivna drevesa (angl. cover tree) [15]. Če uporabljamo uteževanje učnih primerov, potem najbližjih sosedov ponavadi sploh ni treba iskati, saj bodo oddaljeni primeri imeli zaradi majhnih uteži zanemarljiv vpliv na napoved.

Zelo pomembno je, da pravilno izberemo parameter k, ki določa število najbližjih sosedov. Parameter k je odvisen predvsem od tega, koliko napak vsebujejo podatki. Kadar je napak malo, bo zadostovalo že majhno število sosedov. Kadar pa je napak veliko, bo potrebno vzeti več sosedov, saj s povprečenjem zmanjšamo napake. Pri tem pa moramo paziti, ker lahko s povečevanjem parametra k zajamemo tudi primere, ki se precej razlikujejo od bližnjih, kar pa potem slabo vpliva na napoved. Parameter k moramo določiti eksperimentalno.

5.4.2. Linearna lokalno utežena regresija

Linearna lokalno utežena regresija [7, 15] je metoda, ki je zelo podobna metodi k-najbližjih sosedov. Tudi ta metoda v učni množici poišče k najbližjih učnih primerov, ki so najbolj podobni novemu primeru. Podobno kot pri drugi različici metode k-najbližjih sosedov, se lahko učne primere uteži glede na razdaljo do novega primera. Za napoved vrednosti novega primera se namesto povprečenja uporabi linearno funkcijo skozi k najbližjih sosedov. Linearna lokalno utežena regresija torej aproksimira linearno funkcijo v okolici novega primera. Namesto linearne funkcije se lahko uporabi tudi poljubna druga funkcija. Ker pa je v okolici običajno malo učnih primerov, bolj kompleksnih funkcij ni priporočljivo uporabljati, ker bi se preveč prilegale učnim podatkom. Časovna zahtevnost te metode je v primerjavi z metodo k-najbližjih sosedov še nekoliko večja, saj mora za vsak novi primer aproksimirati

drugačno linearno funkcijo na k najbližjih sosedih.

5.5 Regresijska drevesa

Regresijsko drevo je sestavljeno iz notranjih vozlišč, vej in listov. Notranje vozlišče predstavlja en atribut, v tem primeru je to ena spektralna spremenljivka. Vejam, ki izhajajo iz notranjega vozlišča, ustrezajo podmnožice vrednosti atributa v notranjem vozlišču. Listi pa predstavljajo funkcije, ki preslikajo vrednosti atributov v vrednost kemijskofizikalne lastnosti.

Drevo gradimo tako, da izberemo tisti atribut, s katerim lahko množico učnih primerov najbolje razdelimo na več podmnožic. Potem naredimo enako še za vsako podmnožico. Postopek ponavljamo, dokler ni podmnožica dovolj majhna oz. dokler funkcija v listu drevesa dovolj dobro ne modelira primerov, ki so še ostali.

Za izbiro najboljšega atributa se uporablja razlika variance:

$$ds^2(T) = s^2(T) - \sum_i \frac{n_i}{n} s^2(T_i)$$

kjer je T množica primerov v vozlišču, T_i množica primerov v i-tem podrevesu, n število primerov v množici T in n_i število primerov v množici T_i. Varianca je definirana kot povprečen kvadrat napake:

$$s^2(T) = \frac{1}{T}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

kjer je $\bar{y}$ povprečna vrednost kemijskofizikalne lastnosti primerov v množici T.

Funkcija, ki modelira podatke v listu, je lahko poljubna. Največkrat pa se uporablja konstanta ali linearna funkcija. Konstanta se izračuna kot povprečna vrednost kemijskofizikalne lastnosti vseh primerov (vzorcev) v listu.

Zgrajeno drevo uporabimo za napovedovanje kemijskofizikalne lastnosti tako, da od korena drevesa potujemo po ustreznih vejah do lista in za napoved uporabimo funkcijo, ki se nahaja v tem listu.

Zaradi majhnega števila primerov v listih postanejo napovedi nezanesljive. Da bi nezanesljivost zmanjšali, se drevo naknadno še poreže. Drevo se poreže tako, da se za vsako notranje vozlišče od spodaj navzgor oceni pričakovano napako vozlišča in povprečno pričakovano napako poddreves. Če je povprečna pričakovana napaka poddreves večja od pričakovane napake vozlišča, potem poddrevesa porežemo in vozlišče spremenimo v list. Postopek ponavljamo, dokler obstajajo vozlišča, ki se jih da porezati. Za ocenjevanje pričakovane napake e v vozlišču lahko uporabimo prirejeno m-oceno:

$$e = \frac{n}{N+M} e_v + \frac{m}{N+m} e_k$$

kjer je N število primerov v vozlišču, e_v povprečna napaka na teh primerih modela, ki bi ga dobili, če bi to vozlišče bilo list, in e_k povprečna napaka tega istega modela na vseh učnih primerih.

Drevesom, ki v listih uporabljajo linearno ali katero drugo funkcijo, pravimo tudi drevesa modelov. Drevesa modelov imajo ponavadi v vsakem listu drugačno linearno funkcijo. Posledica tega je, da pride na mejah med dvema sosednjima linearnima funkcijama do nezveznosti. Nezveznost lahko zmanjšamo tako, da tudi v vsakem notranjem vozlišču modeliramo primere z linearno funkcijo. Vrednost napovedi potem izračunamo tako, da od lista potujemo nazaj proti korenu drevesa. Pri vsakem notranjem vozlišču pa vrednost nekoliko popravimo. Vrednost v notranjem vozlišču izračunamo z:

$$p' = \frac{np + kq}{n + k}$$

kjer je p' vrednost napovedi, ki jo pošljemu zgornjemu vozlišču, p je napoved, ki jo smo jo dobili od spodnjega vozlišča, in q je vrednost, ki jo napove to vozlišče. Konstanta k je parameter, s katerih lahko uravnamo povprečenje, n pa je število primerov v spodnjem vozlišču. Izkušnje kažejo, da tako drevo poveča točnost napovedi. Podrobnosti o gradnji in rezanju drevesa modelov si lahko pogledate v knjigi [7] ali knjigi [15].

5.6 Metoda podpornih vektorjev

Osnovna ideja metode podpornih vektorjev je, podobno kot pri linearni regresiji, z minimizacijo napake napovedi poiskati funkcijo, ki najbolje aproksimira učne primere. Glavna razlika med obema metodama je, da metoda podpornih vektorjev pri računanju napake napovedi ne upošteva odstopanja primerov od regresijske funkcije, ki so manjša od neke meje, ki je določena s parametrom ε. Druga razlika pa je, da se namesto kvadratne napake računa absolutna napaka napovedi.

Parameter ε določa širino okolice okrog regresijske funkcije, v kateri se napake napovedi primerov ne upoštevajo. Če je aproksimirana funkcija linearna, si lahko okolico predstavljamo kot valj okoli linearne premice, ki gre skozi središče tega valja. Širina valja je enaka 2ε. Če so vsi učni primeri znotraj valja, potem metoda podpornih vektorjev poišče tisto funkcijo, pri kateri je valj širine 2ε najbolj položen. V tem primeru je napaka enaka 0.

S parametrom ε lahko nadziramo, kako dobro se bo funkcija prilegala učnim podatkom. Če je vrednost parametra prevelika, bo regresijska premica vodoravna, kar ustreza ravno napovedi povprečne vrednosti kemijskofizikalne lastnosti vseh učnih primerov. Taka napoved pa ni najbolj uporabna. Po drugi strani pa majhna vrednost parametra ne bi zajela vseh učnih primerov. Ker bodo v tem primeru učni primeri, ki so ostali zunaj, imeli neničelno napako, je treba poiskati kompromis med napako napovedi in naklonom funkcije.

Linearno funkcijo lahko aproksimiramo z naslednjo enačbo:

$$y = b + \sum_{i \text{ je podporni vektor}} a_i x_i \cdot x$$

kjer je x_i podporni vektor, x vektor testnega primera, a_i in b pa so koeficienti, ki jih moramo določiti. $x_i \cdot x$ predstavlja skalarni produkt dveh vektorjev. Podporni vektorji so tisti učni primeri, ki ležijo zunaj valja.

V učnem procesu sodelujejo le tisti primeri, ki so zunaj omejenega območja okrog funkcije. Kot je bilo že omenjeno, metode podpornih vektorjev poskušajo poleg minimizacije

napake minimizirati tudi naklon regresijske funkcije. Kompromis med napako in naklonom uravnamo s parametrom kompleksnosti C. Ta omejuje absolutno maksimalno vrednost koeficientov a_i. S tem omejimo vpliv podpornih vektorjev na obliko regresijske funkcije. Večja kot je vrednost parametra kompleksnosti, bolj se funkcija prilega učnim primerom.

Algoritem in ostale podrobnosti o metodi podpornih vektorjev lahko najdete v knjigi [15].

5.7 Metode, ki temeljijo na podatkovni kompresiji

Osnovna ideja metod, ki temeljijo na podatkovni kompresiji [10, 11], je, da informacijo, ki je shranjena v spektralnih spremenljivkah $x_1, x_2, \ldots, x_K$, stisnemo oziroma opišemo z novimi spremenljivkami, katerih število je manjše od števila originalnih spremenljivk. Na ta način večino nepomembne in nestabilne informacije odstranimo in pri regresiji upoštevamo le tisti del informacije, ki najbolj prispeva pri napovedi kemijskofizikalnih lastnosti.

Novim spremenljivkam pravimo komponente ali faktorji in jih označimo s $t_1, t_2, \ldots, t_A$. Komponente so linearne kombinacije spektralnih spremenljivk in predstavljajo sistematično variacijo v spektralni matriki X, ki prispeva pri napovedovanju spremenljivk $y = \langle y_1, \ldots, y_J \rangle$.

$$\langle t_1, \ldots, t_A \rangle^T = h_1[\langle x_1, \ldots, x_K \rangle^T]$$

Pri napovedi kemijskofizikalnih lastnosti se namesto spektralnih spremenljivk uporabijo te komponente.

$$\langle y_1, \ldots, y_J \rangle^T = h_2[(t_1, \ldots, t_A)^T] + f^T$$

Tukaj f predstavlja tisto informacijo v y, ki je ne moremo razložiti s faktorji $t = \langle t_1, \ldots, t_A \rangle$. Funkciji h_1 in h_2 skupaj tvorita želeni model $\hat{y} = h_2(h_1(x))$.

V praksi se za aproksimacijo relacij med podatki zelo pogosto uporablja linearno modeliranje. Naj X in Y predstavljata centrirane vhodne podatke, pri čemer matrika Y predstavlja kemijskofizikalne lastnosti.

$$X = X_{orig} - 1\bar{x}^T$$
$$Y = Y_{orig} - 1\bar{y}^T$$

Linearni model podatkovne kompresije lahko potem zapišemo z dvema enačbama:

$$T = XV$$
$$Y = TQ^T + F$$

Po določitvi matrike V in ocenitvi matrike Q lahko zapišemo končni model z enačbo $\hat{Y} = XV\hat{Q}$, pri čemer sta X in Y centrirani. Za necentrirani X in Y pa model podajajo naslednje enačbe:

$$\hat{Y} = 1\hat{b}_0^T + X\hat{B}$$

$$\hat{B} = V\hat{Q}^T$$
$$\hat{b}_0^T = \bar{y}^T - \bar{x}^T\hat{B}$$

S podatkovno kompresijo velikega števila spektralnih spremenljivk v le nekaj komponent t poenostavimo model, saj moramo oceniti manjše število parametrov. Enostavnejši modeli pa pripomorejo k lažji interpretaciji napovedi kemijskofizikalnih lastnosti. Poleg tega pa rešimo tudi problem kolinearnosti originalnih spremenljivk in tako dobimo bolj stabilne regresijske napovedi. Dosedanje raziskave so pokazale, da takšni načini pogosto dajejo boljše rezultate.

Povedati je treba, da linearne metode, ki temeljijo na podatkovni kompresiji in jih opisujemo tukaj, konvergirajo k navadni linearni regresiji, ko se število faktorjev (A) približuje številu spektralnih spremenljivk (K). S tem tudi izgubijo zmožnost reševanja problema kolinearnosti in izogibanja prevelikemu prileganju se učnim podatkom. Optimalno število komponent je v praksi skoraj vedno manjše od števila originalnih spremenljivk. Kako določiti optimalno število komponent, je eden od najpomembnejših faktorjev za napovedovanje novih primerov. Največkrat se število komponent določi s preverjanjem modela na testni množici. Pomembna pa je tudi grafična interpretacija, saj lahko z grafično predstavitvijo komponent in uteži opazimo, kdaj faktorji vsebujejo šum oziroma kdaj ne vsebujejo nobene koristne informacije.

Obstaja veliko različnih metod, ki uporabljajo ta način kompresije podatkov. Metode se razlikujejo v definiciji matrike V. Učinkovitost posameznih metod je lahko zelo različna in je odvisna od različnih dejavnikov. Tukaj dajemo poudarek predvsem na bilinearne metode, ki zahtevajo zelo malo predznanja o relacijah med podatki. Te metode ocenijo matriko V kar iz učnih podatkov, tako da so podatki sami odgovorni za izločanje relevantnih informacij. Dve glavni metodi, ki spadata v ta razred, sta regresija glavnih komponent in regresija delnih najmanjših kvadratov.

5.7.1. Bilinearno modeliranje

Pred uporabo bilinearnega modela je običajno potrebno narediti predprocesiranje podatkov. Bilinearno modeliranje je namreč smiselno uporabiti le takrat, ko so relacije med spremenljivkami vsaj približno linearne ali pa ko je populacija primerov dovolj ozka. Zato se spektralne spremenljivke in spremenljivke y centrira in po potrebi normalizira, kadar nimajo približno enakega razpona vrednosti.

Za centrirani matriki X in Y lahko zapišemo bilinearni model z enačbama:

$$X = TP^T + E$$
$$Y = TQ^T + F$$

kjer je $T = XV$. Matriki E in F predstavljeta variacijo v X in Y, ki je ne moremo razložiti z bilinearnim modelom.

Ime bilinearno kodiranje izhaja iz načina aproksimacije spektralne matrike X, saj jo aproksimiramo z modelom $X = TP^T$. Oceniti je torej treba dve množic linearnih parametrov, matriko komponent T in matriko uteži P.

Ocenjevanje parametrov bilinearnega modela

Poglejmo si, kako pri bilinearnem modelu ocenimo parametre V, T, P in Q. Najprej z optimizacijo nekega kriterija ocenimo matriko $\hat{V}$. Glede na ta kriterij ostajajo več različnih metod bilinearnega modeliranja. Komponente $\hat{T}$ potem izračunamo z enačbo $\hat{T} = X\hat{V}$. S pomočjo linearne regresije spremenljivk x_k in y na dobljenih faktorjih $\hat{T} = \{\hat{t}_a, a = 1, 2, \ldots, A\}$ ocenimo še matriki P in Q. V matrični obliki lahko to zapišemo z dvema enačbama:

$$\hat{P}^T = (\hat{T}^T\hat{T})^{-1}\hat{T}^T X$$
$$\hat{Q}^T = (\hat{T}^T\hat{T})^{-1}\hat{T}^T Y$$

Napaki $\hat{E}$ in $\hat{F}$ izračunamo z naslednjima enačbama:

$$\hat{E} = X - \hat{T}\hat{P}^T$$
$$\hat{F} = Y - \hat{T}\hat{Q}^T$$

Predikcija pri bilinearnem modelu

Predikcija kemijskofizikalnih lastnosti iz spektralnih spremenljivk novega primera je lahko narejena na dva načina. Imenujemo ju polna in kratka predikcija. Oba načina vračata identične napovedi, vendar z različno količino dodatne informacije.

Pri obeh načinih je treba najprej transformirati spremenljivke x_k in y. Spremenljivke centriramo in normaliziramo na enak način kot učne podatke.

Polna predikcija

Ta tehnika nam pri napovedovanju nudi največ dodatne informacije, ki jo lahko uporabimo za interpretacijo napovedi. Napoved spremenljivke $\hat{y}$ se izračuna s pomočjo komponent $\hat{t}^T = \langle \hat{t}_1, \ldots, \hat{t}_A \rangle$. Pri vsakem novem primeru najprej od vektorja spektralnih spremenljivk x odštejemo povprečno vrednost $\bar{x}$, ki smo jo izračunali na učnih primerih. Centriran vektor nato pomnožimo z matriko V in dobimo komponente $\hat{t}^T = \langle \hat{t}_1, \ldots, \hat{t}_A \rangle$.

$$\hat{t}^T = (x^T - \bar{x}^T)\hat{V}$$

Napoved $\hat{y}$ pa potem dobimo tako, da povprečni vrednosti $\bar{y}$, izračunani na učnih podatkih, prištejemo produkt $\hat{t}^T\hat{Q}^T$.

$$\hat{y}^T = \bar{y}^T + \hat{t}^T\hat{Q}^T$$

Preostanek spektra e po odštetju komponent lahko izračunamo z enačbo:

$$\hat{e}^T = x^T - \bar{x}^T - \hat{t}^T\hat{P}^T$$

Kratka predikcija

Druga možnost pa je, da napoved spremenljivke $\hat{y}$ zapišemo direktno kar z linearno funkcijo spektralnih spremenljivk. Tukaj x in y nista centrirana.

$$\hat{y}^T = \hat{b}_0^T + x^T \hat{B}^T$$
$$\hat{B} = \hat{V}\hat{Q}'$$
$$\hat{b}_0^T = \bar{y}^T - \bar{x}^T \hat{B}$$

V nasprotju s polno predikcijo nam kratka predikcija ne omogoča primerjave komponent novega primera s komponentami učni primerov.

Metode, ki uporabljajo bilinearno modeliranje

Tukaj si bomo pogledali dve glavni metodi, ki za napoved uporabljata bilinearni model. Prva je regresija glavnih komponent, druga pa je regresija delnih najmanjših kvadratov.

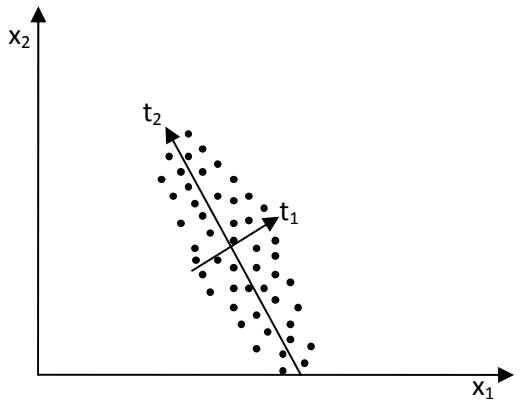
5.7.2. Regresija glavnih komponent

Regresija glavnih komponent (angl. Principal Component Regression - PCR) [10, 11] temelji na analizi glavnih komponent. Zato si najprej poglejmo, kaj je analiza glavnih komponent. Potem pa opišemo še uporabo glavnih komponent, ki jih dobimo pri analizi, za napovedovanje kemijskofizikalnih lastnosti.

Analiza glavnih komponent

Analiza glavnih komponent (angl. Principal Component Analysis - PCA) se največkrat uporablja v primerih, ko so spremenljivke x_k med seboj kolinearne. Namen analize glavnih komponent je izraziti večino informacije, ki je vsebovana v spremenljivkah x_k, z manjšim število novih spremenljivk $T = \{t_1, \ldots, t_A\}$, ki jim pravimo glavne komponente matrike X. Pred uporabe analize glavnih komponent, je treba spremenljivke x_k centrirati in normalizirati, tako da imajo približno enake razpone vrednosti.

Princip analize glavnih komponent je prikazan na sliki 5.1. Slika prikazuje oblak točk v koordinatnem sistemu, ki je določen z dvema spektralnima spremenljivkama x_1 in x_2 (pri K različnih spremenljivkah x_k bi graf vseboval K koordinatnih osi). Vsaka točka predstavlja en učni primer. Točke lahko enako opišemo v prostoru, ki je določen z dvema spektralnima spremenljivkama x_1 in x_2, ali pa v kateremkoli drugem prostoru, ki ga določa par ortogonalnih osi. Eden takih prostorov je prostor, ki ga določata dve glavni komponenti. Prva ustreza najdaljši možni osi, ki gre skozi oblak točk v obliki elipsoida. Druga pa ustreza osi, ki je pravokotna na smer prve komponente in gre skozi središče oblaka točk. Prva komponenta je linearna kombinacija spremenljivk x_k z največjo možno varianco. Druga komponenta je prav tako linearna kombinacija spremenljivk x_k z največjo možno varianco, vendar pod pogojem, da je pravokotna (ortogonalna) na prvo komponento.



Slika 5.1: Prikaz dveh glavnih komponent (t_1 in t_2) v dvodimenzionalnem prostoru.

Matematično je vektor uteži $\hat{p}_1$ določen kot normaliziran vektor, ki maksimizira varianco vektorja $\hat{p}_1^T x$, kar z drugimi besedami pomeni, da maksimizira $\hat{p}_1^T X^T X \hat{p}_1 = \hat{t}_1^T \hat{t}_1$. Naslednji vektor $\hat{p}_2$ maksimizira $\hat{p}_2^T X^T X \hat{p}_2 = \hat{t}_2^T \hat{t}_2$ z dodatnim pogojem, da mora biti $\hat{t}_2$ pravokoten na $\hat{t}_1$ ($\hat{t}_1^T \hat{t}_2 = 0$). Postopek na enak način nadaljujemo, pri čemer mora biti vsaka naslednja komponenta pravokotna na vse prejšnje.

Lastne vrednosti, ki nam povedo, koliko variabilnosti komponente odstranijo iz matrike X, označimo z $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_K$. Izračunamo jih z enačbo $\hat{\tau}_a = \hat{t}_a^T \hat{t}_a$, $a = 1, 2, \ldots, K$.

Dokazano je, da so poleg komponent tudi vektorji uteži $\hat{p}_1, \ldots, \hat{p}_K$ med seboj ortogonalni. V matrični obliki lahko obe ortogonalnosti zapišemo kot:

$$\hat{P}^T \hat{P} = I$$
$$\hat{T}^T \hat{T} = diag(\hat{\tau}_a)$$

Poleg tega pa velja tudi naslednja trditev. Če so spremenljivke x_k centrirane, potem so vektorji $\hat{p}_1, \ldots, \hat{p}_K$ lastni vektorji produkta $X^T X$ z lastnimi vrednostimi $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_K$. To pomeni, da vsi vektorji $\hat{p}_a$ zadoščajo enačbi:

$$X^T X \hat{p}_a = \hat{p}_a \hat{\tau}_a$$

Pokazati pa je mogoče tudi, da vektorji $\hat{t}_a$, $a = 1, .., A$, ustrezajo lastnim vrednostim produkta $X^T X$, katerih dolžina je enaka $\sqrt{\hat{\tau}_a}$.

Če izračunamo vse lastne vektorje ($A = K$), potem lahko X izrazimo z enačbo $X = \hat{T}\hat{P}^T$. Kadar pa vzamemo samo prvih nekaj glavnih komponent ($A < K$), pa lahko matriko X aproksimiramo z enačbo $X = \hat{T}\hat{P}^T + E$, kjer E predstavlja preostalo informacijo.

Algoritmi za izračun dekompozicije matrike na lastne vrednosti

Glavne komponente je mogoče enostavno izračunati s pomočjo dekompozicije matrike $X^T X$ na lastne vektorje. Pri tem se vsi lastni vektorji izračunajo naenkrat. Ker pa pogostokrat potrebujemo le prvih nekaj glavnih komponent, ki imajo največje lastne vrednosti, je bolje, da izračunamo le eno komponento naenkrat. Obstaja namreč preprost algoritem za izračun komponente, ki ji ustreza največja lastna vrednost. Imenuje se NIPALS algoritem. Več o tem

algoritmu si lahko prebere v knjigi [11].

Regresija glavnih komponent

Napovedovanje vrednosti y za nove primere je možno narediti na dva načina. Ena možnost je izračun vektorja t za vsak novi primer z uporabo formule $\hat{t}^T = x^T\hat{P}$ (vektor x mora biti centriran) in uporaba tega vektorja v enačbi za napoved $\hat{y} = \bar{y} + \hat{t}^T\hat{q}$. Drugi način pa je direktna uporaba enačbe $\hat{y} = \bar{y} + x^T\hat{b}$, kjer je vektor regresijskih koeficientov $\hat{b}$ izračunan kot produkt matrike $\hat{P}$ in vektorja $\hat{q}$.

$$\hat{b} = \hat{P}\hat{q}$$

Matriko $\hat{P} = \{P_{ka};\ k = 1, 2, \ldots, K\ in\ a = 1, 2, \ldots, A\}$ sestavljajo uteži, ki smo jih dobili pri analizi glavnih komponent, vektor q pa izračunamo z linearno regresijo spremenljivke y na glavnih komponentah t.

$$y = Tq + f$$

Ker so glavne komponente v $\hat{T}$ med seboj ortogonalne, je rešitev zgornje enačbe enaka:

$$\hat{q} = (diag(1/\hat{\tau}_a))\hat{T}^T y$$

Z vstavitvijo tega v prejšnjo enačbo in zamenjavo T z XP dobimo končno enačbo za izračun koeficientov b:

$$\hat{b} = \hat{P}(diag(1/\hat{\tau}_a))\hat{P}^T x^T y$$

Kadar je število faktorjev A enako K, potem regresija glavnih komponent vrača enake koeficiente $\hat{b}$ kot navadna linearna regresija. Ker pa so spektralne spremenljivke pogostokrat medsebojno korelirane, je optimalno število manjše od K. V takih primerih linearna regresija uporablja deljenje z lastnimi vrednostmi, ki so zelo blizu 0, kar pa povzroča, da so rešitve $\hat{b}$ zelo nestabilne. V nasprotju z linearno regresijo pa so koeficienti pri regresiji glavnih komponent ocenjeni zelo stabilno, saj nezanesljivih lastnih vrednosti pri izračunu ne upoštevamo.

5.7.3. Regresija delnih najmanjših kvadratov

Regresija glavnih komponent uporablja za izbiro komponent samo informacijo o varianci v matriki X, ne upošteva pa informacije, ki jo vsebuje spremenljivka y. Posledica tega je, da imajo glavne komponente, ki jih na ta način dobimo, lahko zelo majhno povezavo z napovedovanjem spremenljivke y. Namesto uporabe glavnih komponent v matriki T metoda delnih najmanjših kvadratov (angl. Partial Least Squares Regression - PLSR) [10, 11] uporablja komponente, ki jih določi z uporabo informacije, ki jo vsebujeta tako matrika X kot vektor y. Regresija delnih najmanjših kvadratov izračuna komponente tako, da maksimiziramo kovarianco med y in vsemi linearnimi kombinacijami spremenljivk x_k. S tem pridemo do komponent, ki so bolj direktno povezane s spremenljivko y, kot pa to velja za

osnovne komponente.

Algoritem

Za vsako komponento $a = 1, 2, \ldots, A$ ponovi korake od 1 do 5:

1. Z maksimiziranjem kovariance med linearno kombinacijo $X_{a-1}\widehat{w}_a$ in y ter upoštevanjem pogoja $\widehat{w}_a^T\widehat{w}_a = 1$ poišči vektor uteži $\widehat{w}_a$. To ustreza iskanju enotnega vektorja $\widehat{w}_a$, ki maksimizira $\widehat{w}_a^T X_{a-1}^T y_{a-1}$.

   $$X_{a-1} = y_{a-1}w_a^T + E$$
   $$\widehat{w}_a = cX_{a-1}^T y_{a-1}$$
   $$c = (y_{a-1}^T X_{a-1} X_{a-1}^T y_{a-1})^{-1/2}$$

2. S projekcijo X_{a-1} na $\widehat{w}_a$ poišči komponento $\hat{t}_a$.

   $$\hat{t}_a = X_{a-1}\widehat{w}_a$$

3. Z linearno regresijo izračunaj vektor uteži $\hat{p}_a$.

   $$\hat{p}_a = X_{a-1}^T\hat{t}_a / \hat{t}_a^T\hat{t}_a$$

4. Z linearno regresijo izračunaj utež $\hat{q}_a$.

   $$\hat{q}_a = y_{a-1}^T\hat{t}_a / \hat{t}_a^T\hat{t}_a$$

5. Vpliv komponente t_a odštej od X_{a-1} in y_{a-1}.

   $$X_a = X_{a-1} - \hat{t}_a\hat{p}_a^T$$
   $$y_a = y_{a-1} - \hat{t}_a\hat{q}_a$$

Napoved $\hat{y}$ novega primera poteka s pomočjo naslednjih enačb:

$$\hat{y} = \hat{b}_0 + x\hat{b}$$
$$\hat{b} = \widehat{W}(\hat{P}^T\widehat{W})^{-1}\hat{q}$$
$$\hat{b}_0 = \bar{y} - \bar{x}^T\hat{b}$$

Pri regresiji delnih najmanjših kvadratov velja, da so ortogonalne tako uteži $\widehat{w}_a$ kot komponente $\hat{t}_a$.

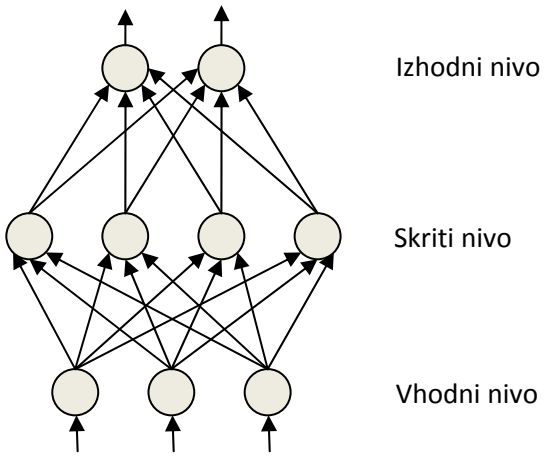
5.7.4. Primerjava metod

Izkušnje na spektroskopskih podatkih kažejo, da metoda PLSR velikokrat daje dobre napovedi z manj komponentami kot metoda PCR. Ker je število komponent manjše, je interpretacija modela bolj enostavna in pregledna. Pri uporabi optimalnega števila komponent dajeta obe metodi primerljive rezultate. V nekaterih primerih je boljša metoda PLSR, v nekaterih drugih primerih pa metoda PCR. Metoda PLSR je običajno računsko hitrejša od

metode PCR. Če pa pogledamo iz teoretičnega vidika, pa je metoda PCR bolj razumljiva od metode PLSR

5.8 Večnivojski perceptron

Večnivojski perceptron (angl. Multilayer Perceptron) je večnivojska usmerjena umetna nevronska mreža. Mrežo sestavljajo nevroni, ki so postavljeni v več nivojev. Večnivojski perceptron je sestavljen iz vhodnega nivoja, izhodnega nivoja in poljubnega števila skritih nivojev, ki se nahajajo med vhodnim in izhodnim nivojem. Zaporedna nivoja sta med seboj povezana tako, da je vsak nevron prvega nivoja z enosmerno vezjo povezan z vsakim nivojem drugega nivoja. Vezem med nevroni pravimo tudi sinapse. Vsaka vez ima pridruženo realno vrednost, ki ji pravimo utež. Nevron je procesni element, ki izračuna uteženo vsoto vhodov nevrona in dobljeno vsoto preslika v izhodno vrednost. Uteženi vsoti pravimo funkcija aktivacije, preslikavi aktivacije v izhod pa izhodna funkcija. Večnivojski perceptron ima toliko vhodnih nevronov, kolikor je atributov (spektralnih spremenljivk). Primer tronivojskega perceptrona je predstavljen na sliki 5.2.



Slika 5.2: Primer tronivojskega perceptrona.

Izračun nevronske mreže poteka tako, da se vhodnim nevronom vsilijo vrednosti atributov učnega primera (vzorca papirja). Vsi nevroni na naslednjem nivoju potem v enem koraku vzporedno in neodvisno drug od drugega izračunajo svoje izhodne vrednosti. Postopek se potem ponovi za vsak naslednji nivo vse do izhodnega nivoja, kjer izhodni nevroni vračajo končne vrednosti. Končna vrednost je v tem primeru vrednost kemijskofizikalne lastnosti papirja.

Izhod $H_{k,i}$ i-tega nevrona na k-tem nivoju se izračuna z enačbo:

$$H_{k,i} = f\left(\sum_{j=1}^{n_k} W_{j,i}^{(k)} H_{k-1,j}\right)$$

kjer je $W_{j,i}^{(k)}$ utež na j-ti povezavi med nevroni na prejšnjem nivoju in i-tim nevronom na k-tem nivoju, n_k število vhodnih povezav i-tega nevrona in f izhodna funkcija.

Poglejmo si, kako poteka učenje mreže. Na začetku so vse uteži naključne. Za učni primer, ki ga dobi mreža na vhodu, se najprej izračunajo izhodne vrednosti mreže. Potem se pri vsakem izhodnem nevronu izračuna razlika med dejanskim in želenim izhodom. Glede na to razliko se spremenijo uteži na povezavah med zadnjim in predzadnjim nivojem. Nato se izračunajo želene vrednosti nevronov na predzadnjem nivoju. Na enak način se izračunajo razlike med dejanskimi in želenimi vrednostmi na predzadnjem nivoju in potem ustrezno spremenijo uteži. Postopek ponavljamo vse do vhodnega nivoja.

Za spreminjanje uteži na vezeh se uporablja posplošeno pravilo delta, ki mu pravimo tudi pravilo vzvratnega razširjanja napake. Zanj je značilno, da uteži spremeni v smeri manjše napake (v smeri negativnega odvoda kvadratne razlike med dejanskimi in želenimi izhodnimi vrednostmi nevronov). Ker posplošeno pravilo delta pri izračunih uporablja odvode, mora biti izhodna funkcija zvezno odvedljiva. V večnivojskem perceptronu se zato za izhodno funkcijo uporablja sigmoidna funkcija:

$$f(X) = \frac{1}{1 + e^{-X}}$$

Učenje mreže zahteva veliko število (nekaj tisoč) prehodov preko učnih primerov. Uspešnost napovedovanja mreže je zelo odvisna od števila prehodov preko učnih primerov in topologije mreže (števila skritih nivojev, števila nevronov na posameznem skritem nivoju). Te parametre je treba eksperimentalno določiti s preverjanjem na testni množici.

Več informacij o učenju večnivojskega preceptrona in drugih podrobnostih lahko najdete v knjigi [7] ali knjigi [15].

5.9 Mreža radialnih baznih funkcij

Mreža radialnih baznih funkcij (angl. Radial Basis Function Network) [20] je ena od vrst večnivojskih usmerjenih umetnih nevronskih mrež. Tipično je sestavljena iz treh nivojev nevronov, in sicer iz vhodnega, skritega in izhodnega nivoja. Aktivacijske funkcije nevronov v skritem nivoju so izbrane iz razreda funkcij, ki jim pravimo radialne bazne funkcije. Čeprav je podobna umetni nevronski mreži, ki uporablja pravilo vzvratnega razširjanja napake, ima kar nekaj prednosti. Uči se veliko hitreje kot umetna nevronska mreža z vzvratnim razširjanjem napake in je zaradi obnašanja radialnih baznih funkcij manj podvržena problemom z nestacionarnimi vhodi.

Glavna razlika med mrežami radialnih baznih funkcij in večnivojskimi usmerjenimi nevronskimi mrežami, kot je večnivojski perceptron, je v obnašanju enega samega skritega nivoja nevronov. Namesto uporabe sigmoidnih aktivacijskih funkcij, nevroni v skritem nivoju uporabljajo Gaussove ali katere druge bazne jedrne funkcije. Vsak nevron v skritem nivoju se obnaša kot lokalno uglašen procesor, ki računa ujemanje vhodnega vektorja z utežmi na povezavah med vhodnim in skritim nivojem. Utežem na povezavah med nevroni vhodnega in skritega nivoja pravimo tudi centri ali središča. Uteži na povezavah med nevroni skritega nivoja in izhodnimi nevroni pa se uporabljajo pri računanju izhodov mreže. Vrednost izhodnega nevrona se namreč izračuna kot linearna kombinacija izhodov nevronov skritega nivoja, pri čemer se za koeficiente linearne kombinacije vzamejo uteži povezav med nevroni skritega nivoja in izhodnim nevronom.

Radialne bazne funkcije so se najprej uporabljale pri reševanju problemov

multivariantne interpolacije. Šele konec osemdesetih let prejšnjega stoletja so jih Broomhead in Lowe ter Moody in Darken uporabili pri načrtovanju nevronskih mrež.
Za mrežo radialnih baznih funkcij velja, da zna s poljubno natančnostjo aproksimirati katerokoli zvezno funkcijo.

V nadaljevanju je naprej predstavljena definicija radialnih baznih funkcij in struktura mreže radialnih baznih funkcij. Nato pa je še razloženo, kako lahko mrežo uporabimo za aproksimacijo ali interpolacijo funkcij in kako se mreža uči.

5.9.1. Radialna bazna funkcija

Radialna bazna funkcija [3] je definirana kot funkcija razdalje:

$$G(r) = G(\|x - c\|); x \epsilon R^n; r \geq 0$$

pri čemer je $G(r)$ zvezna funkcija na intervalu $(0, \infty)$, ki monotono pada (ali narašča) z naraščajočo razdaljo od centra. $\|\cdot\|$ je razdalja (največkrat evklidska ali Mahalanobisova) in c je center ali povprečje radialne bazne funkcije.

Gaussova radialna bazna funkcija

Pri umetnih nevronskih mrežah se največkrat uporablja Gaussova radialna bazna funkcija. Splošna oblika Gaussove radialne bazne funkcije je:

$$G(r) = e^{-r^2}$$

Gaussova radialna bazna funkcija je lokalna, saj velja $\lim_{\|x\| \to \infty} G(\|x - c\|) = 0$.
To pomeni, da ima največji vpliv v centru, z oddaljevanjem od centra pa njen vpliv pada.

Normalizirana radialna bazna funkcija

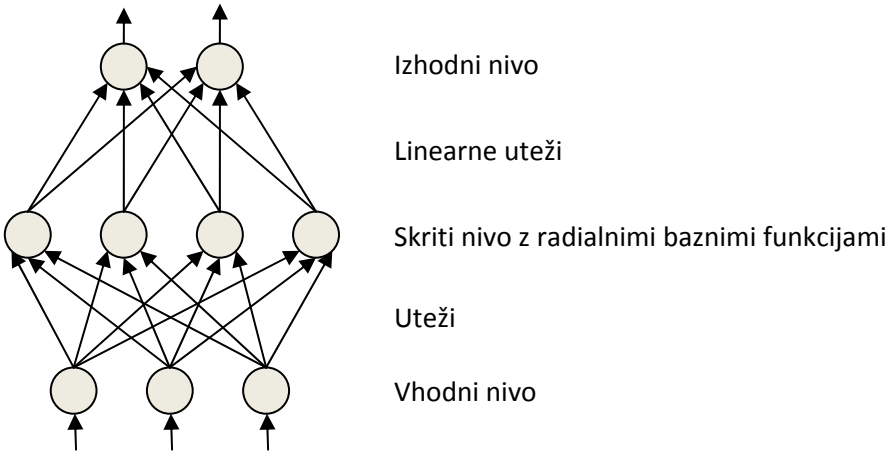
Radialna bazna funkcija pa je lahko tudi normalizirana z vsoto vseh baznih funkcij, ki nastopajo v skritem nivoju mreže radialnih baznih funkcij.

$$H(\|x - c_i\|) \stackrel{\text{def}}{=} \frac{G(\|x - c_i\|)}{\sum_{j=1}^{L} G(\|x - c_j\|)}$$

Center c_i označuje center bazne funkcije pri i-tem nevronu v skritem nivoju, L pa je število skritih nevronov.

5.9.2. Struktura mreže radialnih baznih funkcij

Struktura mreže radialnih baznih funkcij v njeni najbolj enostavni obliki sestavljena iz treh popolnoma različnih nivojev.



Slika 5.3: Primer mreže radialnih baznih funkcij.

Vhodni nivo

Vhodni nivo sestavljajo nevroni, ki sprejemajo vrednosti vhodnega vektorja. Število vhodnih nevronov je enako velikosti K vhodnega vektorja x.

Skriti nivo

Drugi nivo nevronov, ki ga imenujemo tudi skriti nivo, pa sestavljajo nelinearni nevroni. Vsak nevron skritega nivoja je povezan z vsakim nevronom vhodnega nivoja. Skriti nevroni v mreži radialnih baznih funkcij služijo popolnoma drugačnemu namenu kot skriti nevroni v večnivojskem perceptronu.

Vsak nevron v skritem nivoju dobi na vhodu vrednosti vhodnih nevronov. Kot je bilo že prej omenjeno, skriti nevroni za aktivacijsko funkcijo uporabljajo t. i. radialno bazno funkcijo, ki ima dva parametra, in sicer center (središče) in širino. Center bazne funkcije za i-ti nevron v skritem nivoju je vektor c_i, katerega velikost je enaka velikosti vhodnega vektorja x. Vsak skriti nevron ima ponavadi različen center.

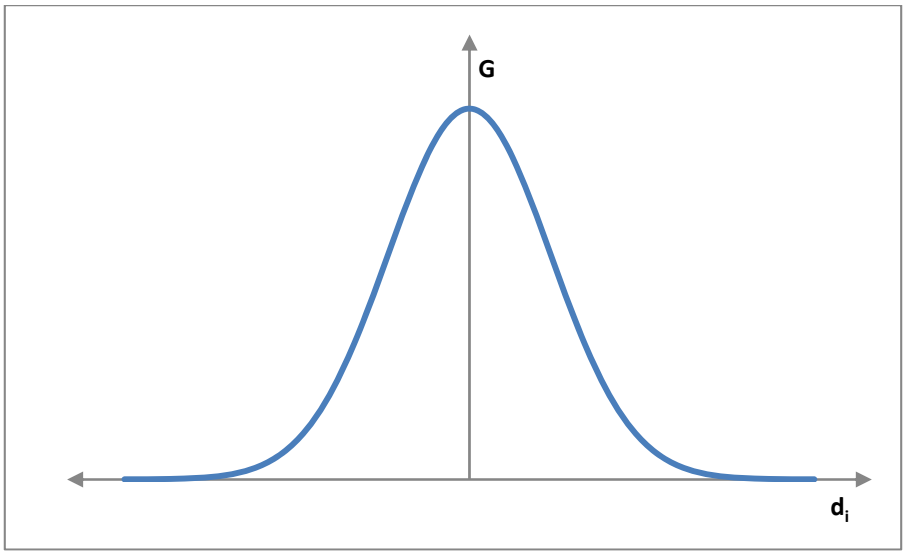
Izračun izhodov skritih nevronov poteka tako, da se najprej pri vsakem skritem nevronu izračuna t. i. radialna razdalja d_i med vhodnim vektorjem x in centrom bazne funkcije c_i. Za izračun radialne razdalje se uporabi kar evklidska razdalja.

$$d_i = \|x - c_i\|$$

Izhod h_i skritega nevrona pa se potem izračuna z uporabo bazne funkcije G na razdalji d_i.

$$h_i = G(d_i)$$

Na sliki 5.4 lahko vidimo, da je bazna funkcija zvezna krivulja, ki ima vrh pri ničelni razdalji in monotono pada s povečevanjem razdalje od centra.



Slika 5.4: Prikaz Gaussove radialne bazne funkcije.

Izhodni nivo

Transformacija prostora vhodnega nivoja v prostor skritega nivoja je nelinearna, medtem ko je transformacija iz prostora skritega nivoja v prostor izhodnega nivoja linearna.

Vrednost oziroma izhod j-tega nevrona v izhodnem nivoju izračunamo kot uteženo vsoto izhodov nevronov v skritem nivoju.

$$y_j = f_j(x) = w_{0,j} + \sum_{i=1}^{L} w_{i,j} h_i \qquad (j = 1, 2, \dots, M)$$

5.9.3. Aproksimacija funkcije

Matematični model

Matematični model mreže radialnih baznih funkcij lahko napišemo z naslednjo enačbo:

$$y = f(x), \quad f: R^N \rightarrow R^M$$

$$y_j = f_j(x) = w_{0,j} + \sum_{i=1}^{L} w_{i,j} G(\|x - c_i\|) \qquad (j = 1, 2, \dots, M)$$

kjer je $\|x - c_i\|$ evklidska razdalja med x in c_i.

Aproksimacija funkcije

Naj bo dana funkcija $z = g(x)$, $x \epsilon R$, $y \epsilon R$, $g: R \rightarrow R$, in naj bo G_i, $i = 1, 2, \dots, L$, končna množica radialnih baznih funkcij. Funkcijo g lahko zapišemo z uporabo baznih funkcij na naslednji način:

$$z = g(x) = \sum_{i=1}^{L} w_i G_i(x) + r(x)$$

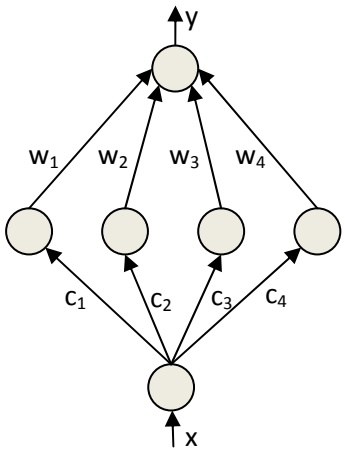
kjer je $r(x)$ napaka oz. preostala informacija, ki je ne moremo modelirati z baznimi funkcijami.

Funkcijo z lahko potem aproksimiramo z enačbo:

$$z = g(x) \cong \sum_{i=1}^{L} w_i G_i(x)$$

Z izbiro ustreznih parametrov baznih funkcij G_i želimo minimizirati napako $r(x)$.

Aproksimacija funkcije z mrežo radialnih baznih funkcij



Slika 5.5: Primer mreže radialnih baznih funkcij z enim vhodom in enim izhodom.

Naj ima mreža en vhodni in en izhodni nevron, kot je prikazano na sliki 5.5. Izhod y izračunamo z enačbo:

$$y = f(x) = \sum_{i=1}^{L} w_i G_i(\|x - c_i\|)$$

Z uporabe te mreže radialnih baznih funkcij lahko funkcijo z aproksimiramo z naslednjo enačbo:

$$z = \sum_{i=1}^{L} w_i G_i(\|x - c_i\|) + r(x) = f(x) + r(x)$$

kjer je $f(x)$ izhod mreže, $r(x)$ pa je napaka. Z ustrezno izbiro centrov c_i in uteži w_i lahko napako minimiziramo.

Funkcijo $g: R \to R$ lahko posplošimo na funkcijo $g: R^N \to R^M$ z uporabo mreže, ki

ima N vhodnih in M izhodnih nevronov.

5.9.4. Učenje mreže radialnih baznih funkcij

Učenje mreže radialnih baznih funkcij lahko formuliramo z naslednjim nelinearnim optimizacijskim problemom.

Danih je N učnih primerov (x_k, y_k), $k = 1, 2, \dots, N$. Izberi $w_{i,j}$ in c_i, $i = 1, 2, \dots, L$, $j = 1, 2, \dots, M$, tako, da bo spodnji izraz minimalen.

$$J(w, c) = \sum_{k=1}^{N} \|y_k - f(x_k)\|^2$$

Pri učenju mreže moramo določiti naslednje parametre:

1. število nevronov v skritem nivoju,
2. centre radialnih baznih funkcij skritih nevronov,
3. širino baznih funkcij (če nimajo fiksne širine),
4. uteži na povezavah med skritim in izhodnim nivojem.

Določanje centrov

Spomnimo se, da so pri umetnih nevronskih mrežah, ki uporabljajo pravilo vzvratnega razširjanja napake, vse uteži na vseh nivojih spremenjene istočasno. Pri mrežah radialnih baznih funkcij pa so uteži povezav med vhodnim in skritim nivojem ponavadi določene, preden so spremenjene uteži povezav med skritim in izhodnim nivojem. Ko se vhodni vektor premika stran od uteži povezave na prvem nivoju, se vrednost aktivacijske funkcije zmanjšuje. Zaradi takšnega vedenja pravimo utežem povezav med vhodnim in skritim nivojem tudi centri (ta izraz je bil uporabljen do sedaj). Uteži oz. centre določimo z uporabo metode K-Means Clustering ali kakšne druge metode. Ko jih enkrat določimo, se potem ne spreminjajo več.

Določanje uteži

Ko so uteži (centri) na povezavah med nevroni vhodnega in nevroni skritega nivoja določene, je v treba drugi fazi učenja določiti še uteži na povezavah med nevroni skritega in nevroni izhodnega nivoja. Za določanje teh uteži se uporablja isto pravilo kot pri večnivojskem perceptronu.

5.10 Kalibracija regresijskih modelov

Kalibracija regresijskih modelov je posebna metoda strojnega učenja, ki se uporablja za izboljšanje napovedi že obstoječih regresijskih modelov. Metoda vzame neki regresijski model, ki je bil zgrajen z eno od metod strojnega učenja, in poskuša izboljšati napovedi modela s primerjavo napovedanih vrednosti modela in dejanskih vrednosti. S podatki o

dejanskih in napovedanih vrednostih regresijske spremenljivke učnih primerov zgradi novi model, ki napovedane vrednosti popravi tako, da se bolj približajo dejanskim vrednostim. Za kalibracijo obstoječega modela se uporablja izotonična regresija. Cilj izotonične regresije je poiskati monotono naraščajočo funkcijo, ki minimizira kvadratno napako med napovedano in dejansko vrednostjo.

Učenje poteka tako, da učne primere najprej uredimo v naraščajočem vrstnem redu glede na napovedano vrednost. Potem pa gremo skozi urejen seznam primerov od najmanjše proti največji napovedani vrednosti. Pri vsakem primeru preverimo, ali je njegova dejanska vrednost večja od dejanske vrednosti prejšnjega primera. Če to drži, potem gremo na naslednji primer. V kolikor pa ta trditev ne velja, zamenjamo dejanski vrednosti obeh primerov z njuno povprečno vrednostjo. Na enak način potem preverimo, ali je povprečna vrednost večja od dejanske vrednosti predpreteklega primera. Če to ne drži, potem zopet zamenjamo vrednosti s povprečno vrednostjo vseh treh. Postopek ponavljamo, dokler ni povprečna vrednost primera večja od dejanske vrednosti prejšnjega primera. Tako dobimo izotonično funkcijo, ki jo potem uporabimo za napovedovanje novih primerov.

Dosedanje raziskave so pokazale, da kalibracija izboljša napovedi slabih modelov, pri dobrih modelih pa včasih daje tudi slabše rezultate. Več o metodi si lahko preberete v članku [8].

Poglavje 6

Eksperimenti

6.1 Uporabljena programska orodja

Za preizkušanje različnih metod predprocesiranje spektrov in metod strojnega učenja je bilo uporabljeno odprtokodno programsko okolje Eclipse [21], ki se uporablja za razvijanje programov v programskem jeziku Java, v povezavi s programskim paketom za strojno učenje WEKA. Odprtokodno programski paket WEKA [22] je razvila Univerza Waikato iz Nove Zelandije. V WEKI je implementirana večina algoritmov oziroma metod, ki se uporabljajo za reševanje klasifikacijskih in regresijskih problemov. Metode predprocesiranja spektrov so bile razvite v obliki filtrov, ki jih paket WEKA uporablja za filtriranje podatkov.

Pred preizkušanjem je bilo treba podatkovno bazo pretvoriti v primerne datoteke, ki jih je mogoče naložiti v program WEKA. Za vsako kemijskofizikalno lastnost papirja je bila narejena ena podatkovna množica, ki je bila predstavljena z datoteko v formatu ARFF, kot to zahteva program WEKA. Datoteke ARFF zahtevajo atributno predstavitev učnih primerov (vzorcev papirja).

6.2 Uporabljene metode predprocesiranja podatkov

Pred uporabo metod strojnega učenja so bile uporabljene naslednje metode in kombinacije metod predprocesiranje spektrov:

- brez predprocesiranja spektrov (Brez),
- absorpcijska transformacija (A),
- Kubelka-Munkova transformacija (KM),
- multiplikativna korekcija razpršenosti (MSC),
- metoda Standard Normal Variate (SNV),
- odvajanje spektrov s Savitzky-Golayevo metodo (SGD),
- ortogonalna korekcija signala (OSC),
- metoda A in metoda MSC (A + MSC),

- metoda A in metoda SNV (A + SNV),
- metoda A in metoda SGD (A + SGD),
- metoda KM in metoda MSC (KM + MSC),
- metoda KM in metoda SNV (KM + SNV),
- metoda KM in metoda SGD (KM + SGD).

Kratice imen, ki uporabljajo v nadaljevanju, so navedene v oklepajih za imenom metode.

6.3 Uporabljene metode strojnega učenja

Preizkušenih je bilo večino regresorjev, ki so implementirani v programskem paketu WEKA. V tabeli so predstavljeni vsi preizkušeni regresorji. Navedena so imena regresorjev, ki jih uporablja program WEKA, metoda strojnega učenja, ki jo regresor predstavlja, in kratice imen, ki se uporabljajo v nadaljevanju.

Ime regresorja	Metoda strojnega učenja	Kratica
SimpleLinearRegression	Enostavna linearna regresija	SLR
LinearRegression	Linearna regresija	LR
PaceRegression	Regresija pace	PR
IBk	Metoda k-najbližjih sosedov	IBK
LWL	Linearna lokalno utežena regresija	LWL
SMOreg	Metoda podpornih vektorjev	SMO
M5P	Drevo modelov	M5P
REPtree	Regresijsko drevo	REP
PCR	Regresija glavnih komponent	PCR
PLSClassifier	Regresija delnih najmanjših kvadratov	PLSR
MultilayerPerceptron	Večnivojski perceptron	MP
RBFNetwork	Mreža radialnih baznih funkcij	RBFN

Tabela 6.1: Seznam preizkušenih metod strojnega učenja.

Preizkušene so bile vse možne kombinacije metod predprocesiranja spektrov in metod strojnega učenja. Pri nekaterih metodah strojnega učenja je bilo narejenih več modelov, vsakokrat z drugačnimi parametri.

6.4 Preverjanje uspešnosti modelov

Za učenje modelov so bili uporabljeni vsi atributi (spektralne spremenljivke). Za ocenjevanje uspešnosti modelov je bilo uporabljeno 10-kratno prečno preverjanje. Podatkovne množice so bile razdeljene na 10 približno enako močnih podmnožic. Preverjanje je bilo izvedeno tako, da se je ena podmnožica vzela za testno množico, vse ostale skupaj pa so sestavljale učno množico. Model je bil naučen s pomočjo učne množice, uspešnost modela pa se je potem preverila na testni množici. Ta postopek je bil ponovljen za vsako izmed 10 podmnožic. Končna ocena uspešnosti je bila izračunana kot povprečje uspešnosti vseh 10 modelov.

Pri kalibraciji regresijskih modelov je preverjanje uspešnosti potekalo malenkost drugače. Celotna podatkovna množica je bila razdeljena na tri dele v razmerju 2:1:1. Prva polovica podatkovne množice se je uporabila za gradnjo regresijskega modela. Naslednja četrtina se je uporabila za kalibracijo tega modela, zadnja četrtina po za preverjanje uspešnosti modela, ki smo ga dobili s kalibracijo. Ta postopek je bil 10-krat ponovljen, vsakokrat na drugače naključno premešani podatkovni množici. Končna ocena uspešnosti je bila izračunana kot povprečje uspešnosti vseh 10 kalibracij.

Uspešnost napovedi modela je bila ocenjena s korenom srednje kvadratne napake (RMSE) in relativnim korenom srednje kvadratne napake (RRSE) napovedi testnih primerov. Poleg teh dveh pa je bil za ocenjevanje uspešnosti uporabljen tudi korelacijski koeficient (R).

Koren srednje kvadratne napake je definiran kot koren povprečne razlike med napovedano vrednostjo $\hat{y}$ in dejansko vrednostjo y.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2}$$

Relativni koren srednje kvadratne napake podaja spodnja enačba, pri čemer je $\bar{y}$ povprečna dejanska vrednost (y_U) učnih primerov.

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(y - \hat{y})^2}{\sum_{i=1}^{N}(y - \bar{y})^2}} \cdot 100\ \%$$

$$\bar{y} = \frac{1}{N_U}\sum_{i=1}^{N_U} y_U$$

Korelacijski koeficient meri statistično korelacijo med dejanskimi vrednostmi y in napovedanimi vrednostmi $\hat{y}$.

$$R = \frac{S_{y\hat{y}}}{\sqrt{S_y S_{\hat{y}}}}$$

$$S_{y\hat{y}} = \frac{\sum_{i=1}^{N}[(y - \bar{y})(\hat{y} - \bar{\hat{y}})]}{N - 1}$$

$$S_y = \frac{\sum_{i=1}^{N}(y - \bar{y})^2}{N - 1} \qquad S_{\hat{y}} = \frac{\sum_{i=1}^{N}(\hat{y} - \bar{\hat{y}})^2}{N - 1}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y \qquad \bar{\hat{y}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}$$

Poglavje 7

Rezultati

7.1 Rezultati najboljših modelov

Tabela 7.1 prikazuje rezultate najboljšega modela za posamezno kemijskofizikalno lastnost, pri čemer so pri izbiri najboljšega modela upoštevani najboljši modeli vseh možnih kombinacij, ki so bile preizkušene. V tabeli so prikazani koren srednje kvadratne napake (RMSE), relativni koren srednje kvadratne napake (RRSE) in korelacijski koeficient (R). Rezultati kemijskofizikalnih lastnosti so urejeni glede na RRSE, ker ta ocena najbolje prikazuje uspešnost modela. Lahko bi jih uredili tudi glede na RMSE, vendar to ni najbolj primerno, ker je RMSE odvisna od porazdelitve in razpona vrednosti posamezne kemijskofizikalne lastnosti. Primer, ki to potrjuje, je napoved KFL 10, ki ima sicer najmanjši koren srednje kvadratne napake, vendar, če si pogledamo graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti (glej sliko 7.10), vidimo, da je model praktično neuporaben. To dokazuje tudi relativni koren srednje kvadratne napake, katerega vrednost za KFL 10 je največja v primerjavi z drugimi kemijskofizikalnimi lastnostmi. Težavo namreč povzroča en vzorec papirja, ki zelo odstopa od ostale populacije in je zato razpon vrednosti te kemijskofizikalne lastnosti večji od tistega, pri katerem bi upoštevali samo populacijo brez tega primera.

Pri vseh kemijskofizikalnih lastnostih je relativni koren srednje kvadratne napake manjši od 100 odstotkov. To pomeni, da spekter vsebuje koristno informacijo o prav vseh kemijskofizikalnih lastnostih, ki so bile izmerjene. Iz tabele 7.1 je razvidno, da se rezultati posameznih kemijskofizikalnih lastnosti precej razlikujejo. Pri prvih treh kemijskofizikalnih lastnostih so rezultati dobri, pri čemer smo za oceno dobro določili, da mora biti korelacijski koeficient večji od 0,9. Napovedi zadnjih štirih kemijskofizikalnih lastnosti so slabe, saj je njihov korelacijski koeficient manjši od 0,8. Ostale kemijskofizikalne lastnosti pa imajo povprečne rezultate napovedi.

Rezultati napovedi kemijskofizikalnih lastnosti papirja so v primerjavi z rezultati obeh raziskovalnih skupin (glej razdelek 3.3) malenkost slabši, vendar je treba poudariti, da so bili spektri pri raziskovalnih skupinah izmerjeni na veliko večjem razponu valovnih dolžin, zato rezultati med seboj niso najbolj primerljivi. Glede na to, da sta raziskovalni skupini za napovedovanje lastnosti uporabljali metodo PLSR, ki je bila preizkušeni tudi tukaj in je ena

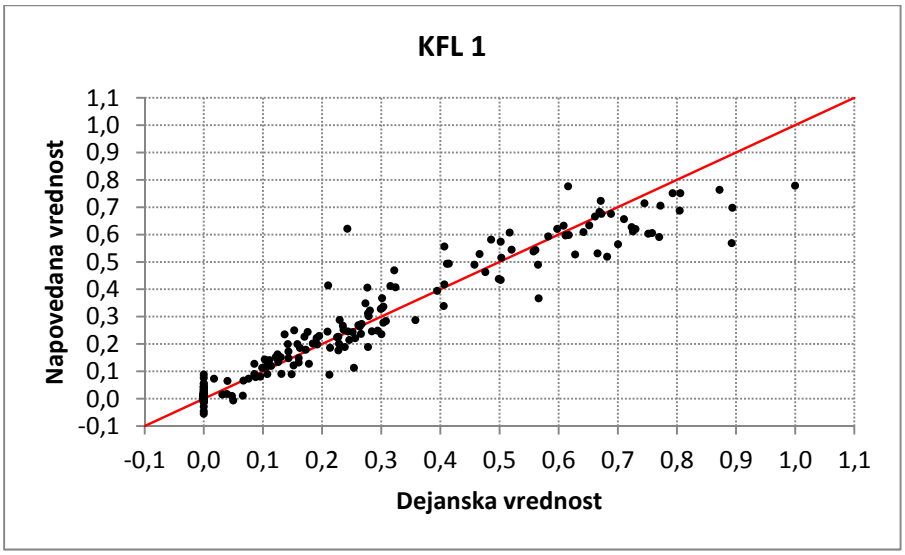
izmed najboljših metod, lahko sklepamo, da pri napovedovanju kemijskofizikalnih lastnosti prispevajo tudi spektralne spremenljivke pri večjih valovnih dolžinah (1950 – 20000 nm).

	RMSE	RRSE (%)	R
KFL 4	0,068	29,73	0,952
KFL 1	0,047	29,81	0,957
KFL 13	0,104	33,38	0,941
KFL 5	0,174	43,98	0,896
KFL 8	0,096	46,19	0,883
KFL 9	0,186	46,91	0,882
KFL 6	0,092	48,66	0,875
KFL 2	0,073	50,59	0,868
KFL 11	0,077	54,55	0,841
KFL 14	0,103	59,54	0,847
KFL 3	0,162	65,03	0,760
KFL 12	0,099	65,34	0,758
KFL 7	0,101	75,28	0,664
KFL 10	0,038	85,20	0,533

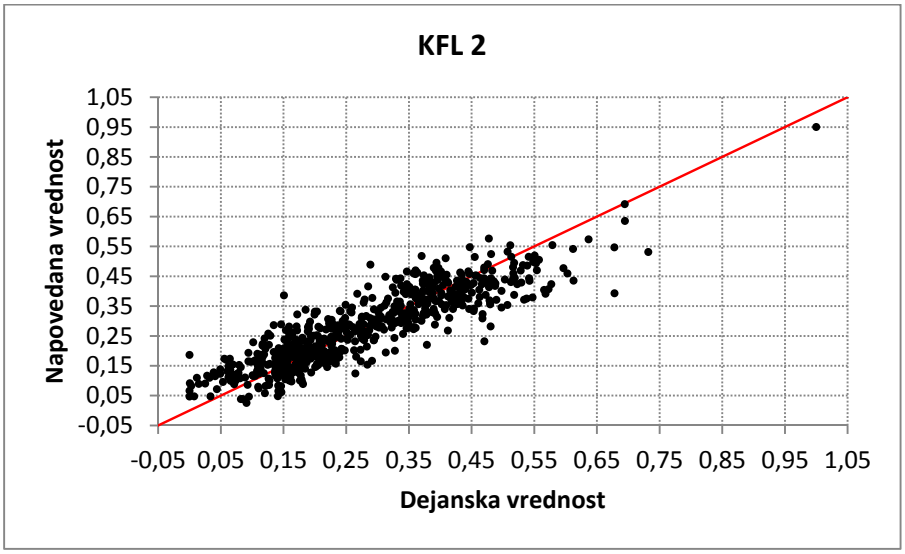
Tabela 7.1: Rezultati napovedi najboljših modelov.

Spodaj je prikazan še grafični prikaz napovedi za vsako kemijskofizikalno lastnost posebej. Graf prikazuje vrednosti napovedi v odvisnosti od dejanskih vrednosti kemijskofizikalne lastnosti. Prikazuje torej, kako se napovedane vrednosti razlikujejo od laboratorijsko izmerjenih vrednosti. Boljša napoved je tista, ki se čim bolj se približa poševni rdeči črti, ki označuje optimalno napoved.

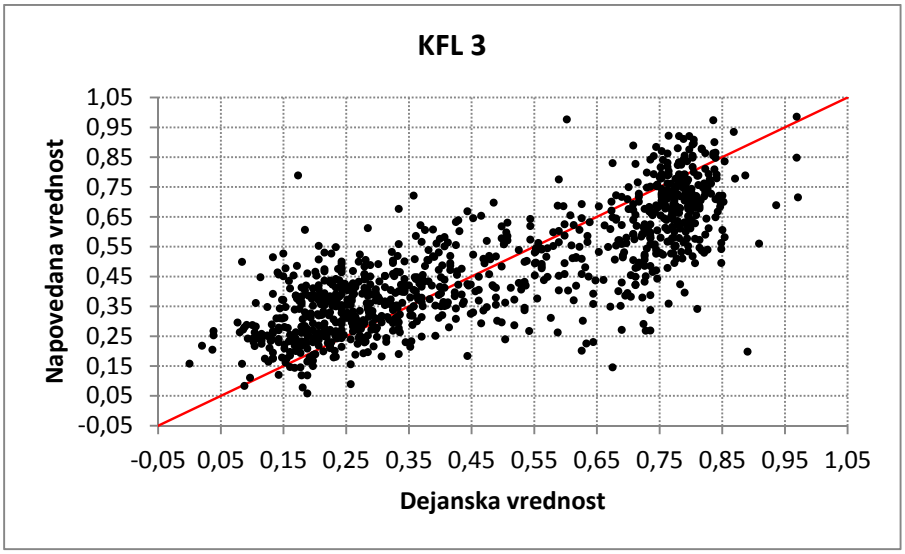
Če pogledamo grafe, vidimo, da imata najboljše napovedi KFL 1 in KFL 4, ker se točke najbolj približajo rdeči poševni črti. Zanimivo je, da grafični prikaz za KFL 13 ne daje občutka dobre napovedi, čeprav je glede na RRSE napoved KFL 13 tretja najboljša od vseh. Vidimo lahko tudi, da so napovedi KFL 3, KFL 7, KFL 10 in KFL 12 praktično neuporabne. Točke pri KFL 3 so dokaj enakomerno oddaljene od idealne rdeče črte, medtem ko se točke pri KFL 7, KFL 10 in KFL 12 od idealne črte bolj oddaljujejo proti koncu intervala dejanskih vrednosti. To še posebej velja za KFL 10, kar potrjuje tudi vrednost relativnega korena srednje kvadratne napake, ki je pri tej kemijskofizikalni lastnosti največja. Podobno kot za KFL 13 tudi za KFL 5 in KFL 9 velja, da sta napovedi na videz slabši od napovedi kemijskofizikalnih lastnosti, ki imajo slabši RRSE (npr. KFL 7, KFL 14 in KFL 11). Razlog je verjetno v tem, da KFL 5 in KFL 9 dejanskih vrednosti nimata razporejenih zvezno preko celotnega razpona vrednosti, ampak so dejanske vrednosti diskretno porazdeljene na določene intervale.



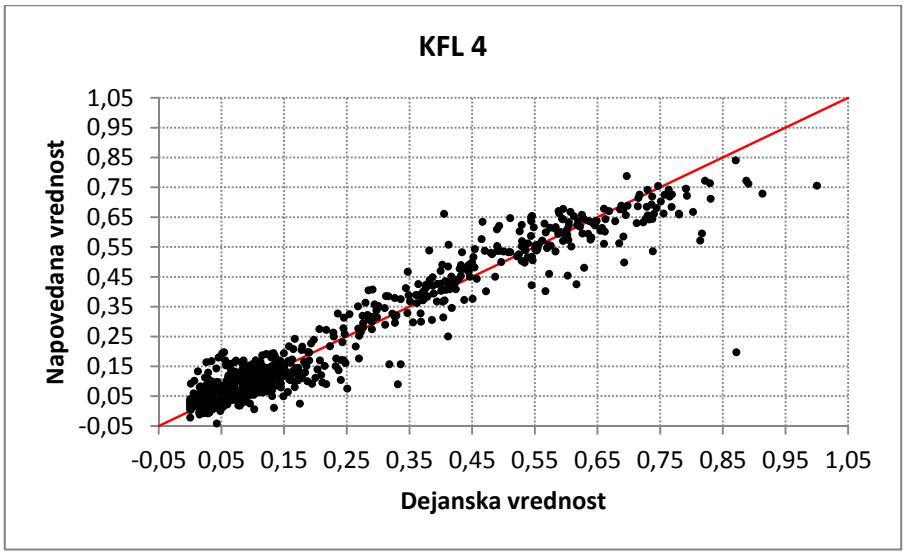
Slika 7.1: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 1.



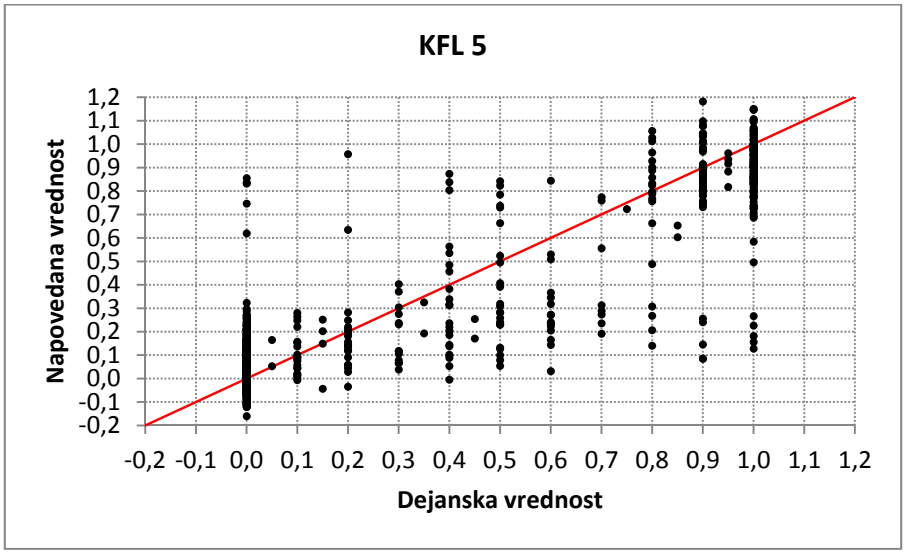
Slika 7.2: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 2.



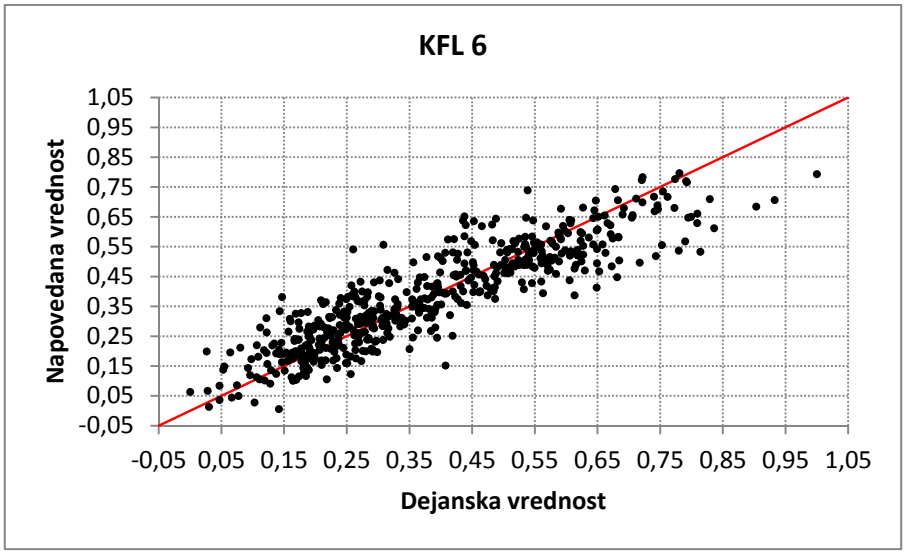
Slika 7.3: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 3.



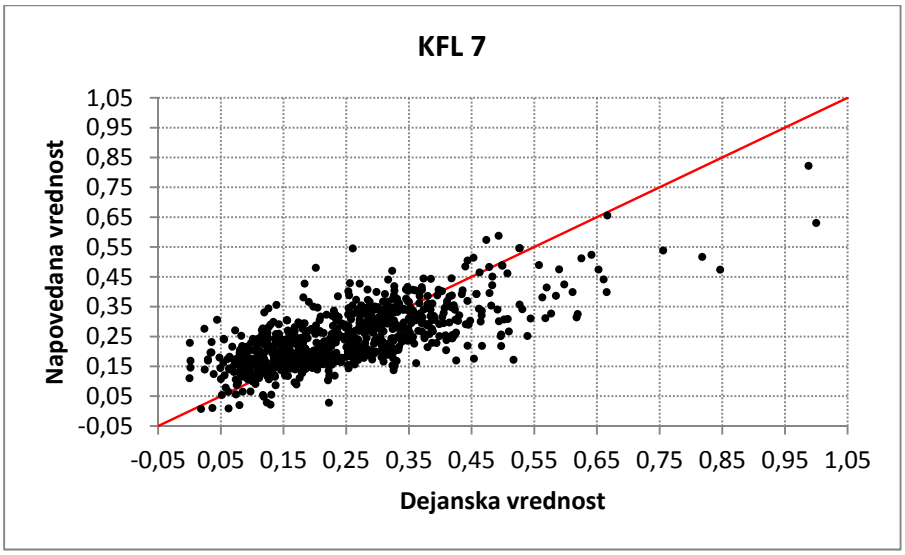
Slika 7.4: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 4.



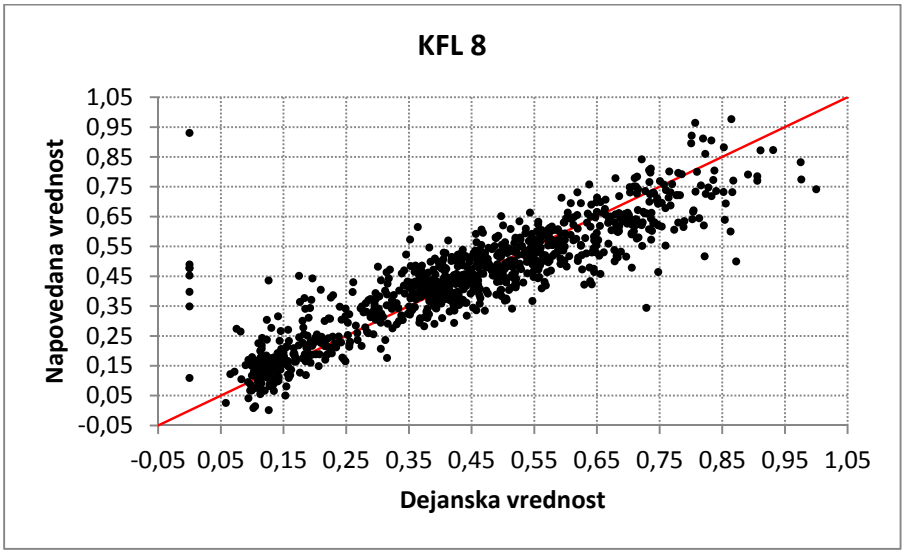
Slika 7.5: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 5.



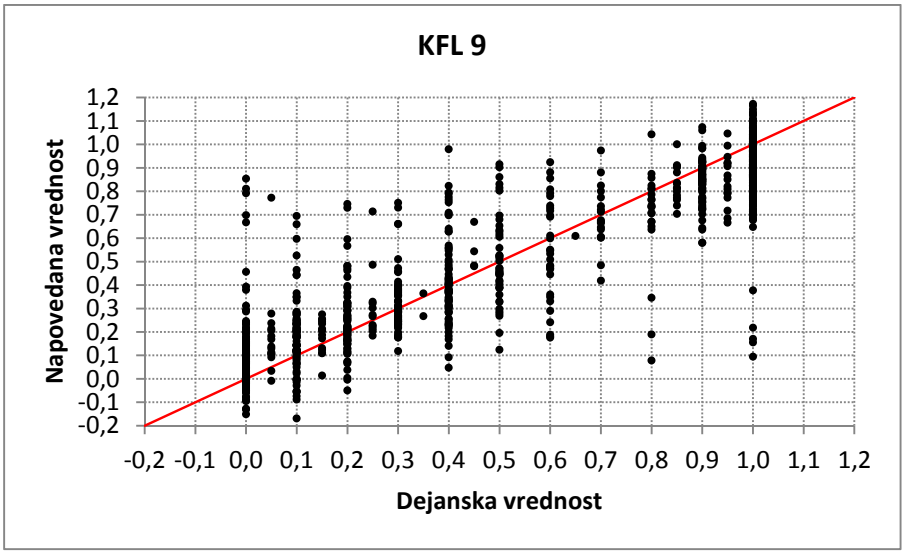
Slika 7.6: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 6.



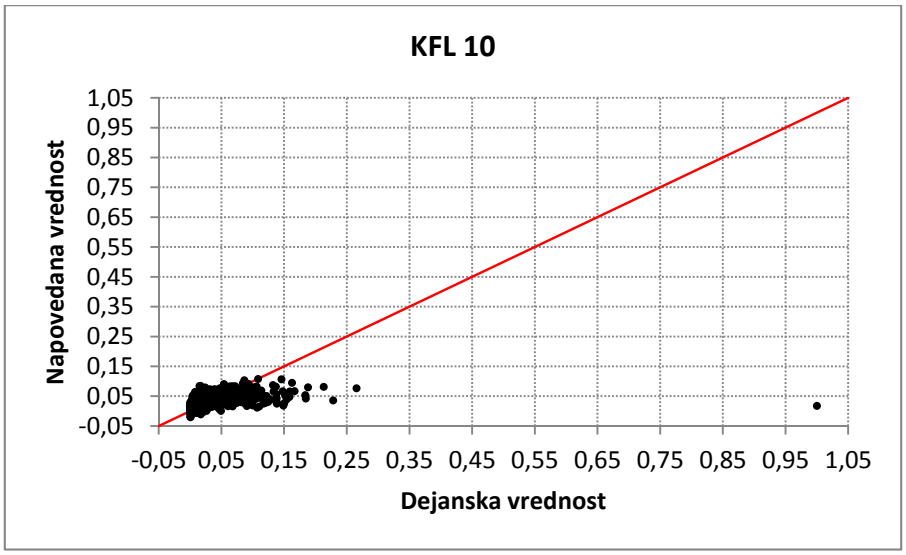
Slika 7.7: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 7.



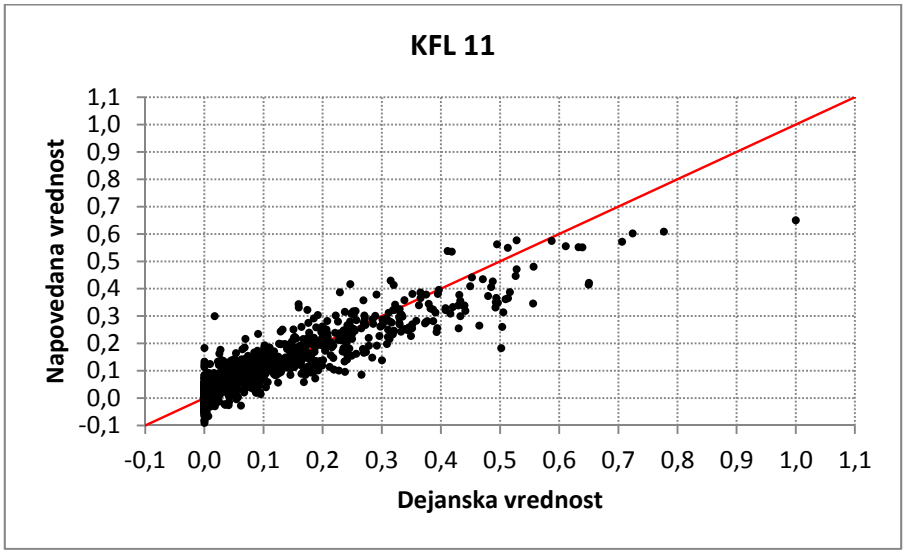
Slika 7.8: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 8.



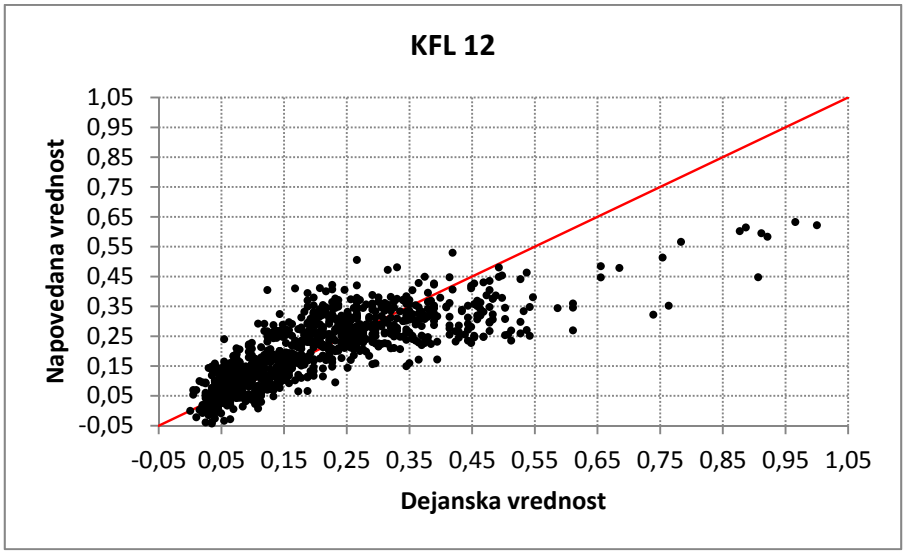
Slika 7.9: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 9.



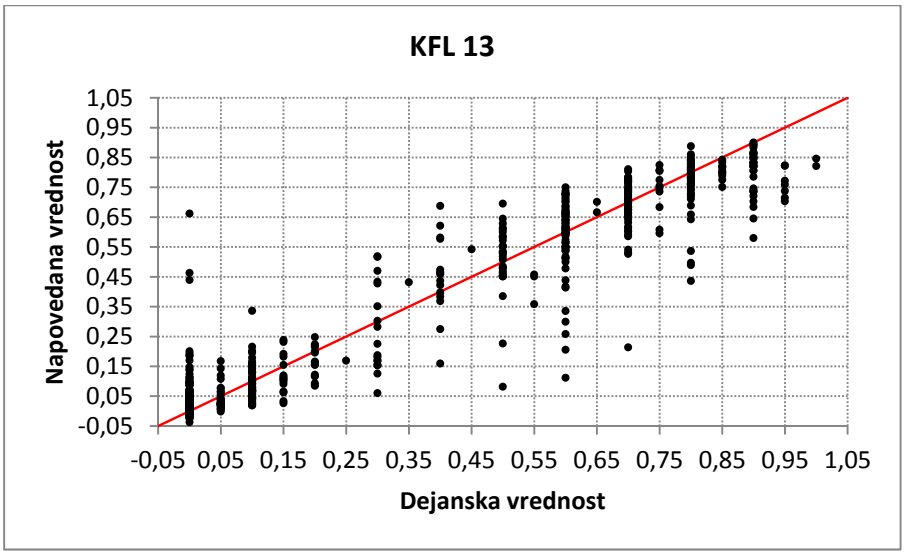
Slika 7.10: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 10.



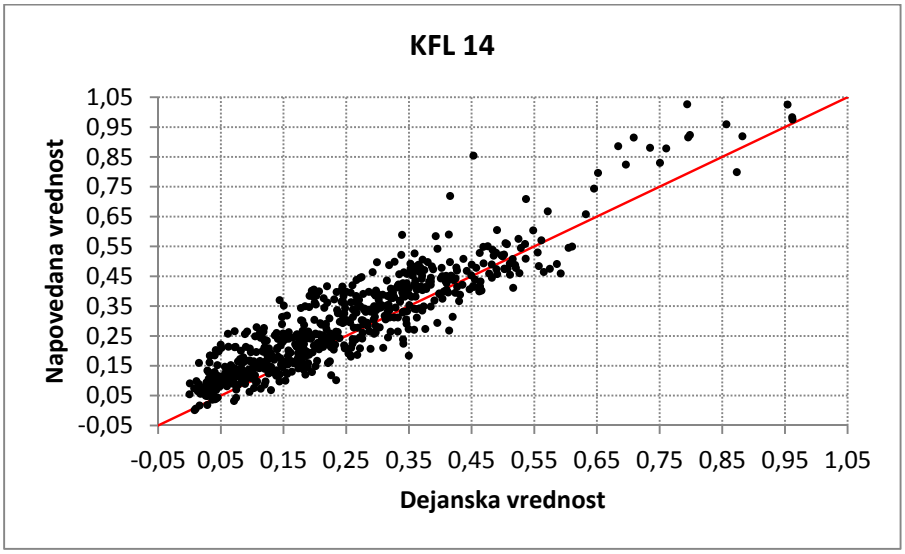
Slika 7.11: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 11.



Slika 7.12: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 12.



Slika 7.13: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 13.

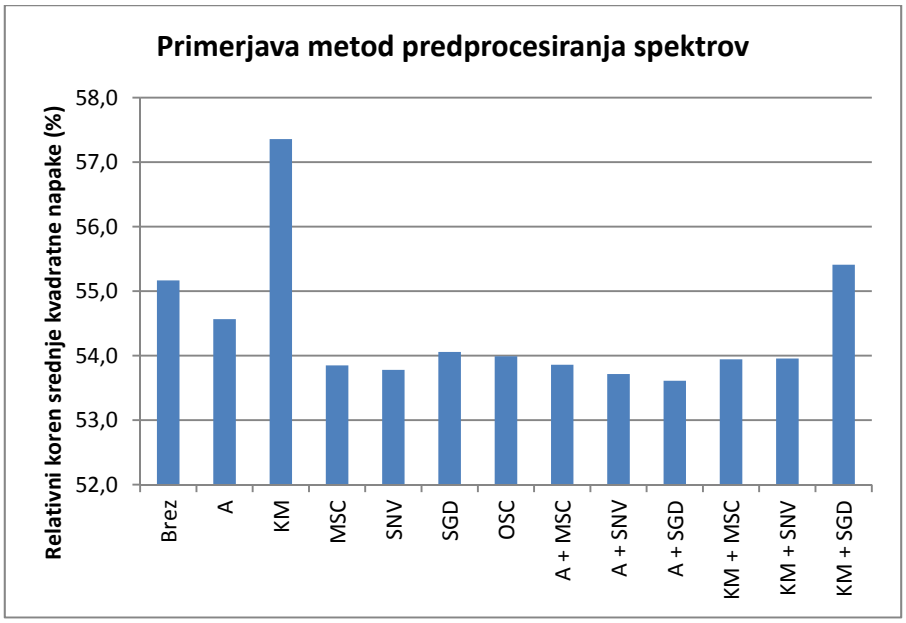


Slika 7.14: Graf napovedanih vrednosti v odvisnosti od dejanskih vrednosti KFL 14.

7.2 Primerjava metod predprocesiranja spektrov

Poglejmo si najprej, kako posamezne metode predprocesiranja spektrov vplivajo na rezultate modelov. Naslednji graf (glej sliko 7.15) prikazuje povprečne rezultate modelov za različne metode predprocesiranja spektrov, pri čemer je bil povprečni rezultat, ki ga prikazuje posamezen stolpec, izračunan tako, da se je naprej poiskal najboljši model za posamezno kemijskofizikalno lastnost in nato izračunalo povprečje relativnih korenov srednje kvadratne napake teh najboljših modelov.

Na grafu lahko vidimo, da večina metod predprocesiranja spektrov v povprečju izboljša napovedi kemijskofizikalnih lastnosti. Izboljšanje napovedi sicer ni veliko, saj večina metod zmanjša relativni koren srednje kvadratne napake le od 1 do 1,5 odstotka, vendar je vsekakor boljše uporabiti kakšno metodo predprocesiranja, kot pa nobene.



Slika 7.15: Primerjava metod predprocesiranja spektrov.

Metoda, ki se najslabše obnese, je Kubelka-Munkova transformacija. Napoved v povprečju poslabša za okoli 2 odstotka. Glede na to, da Kubelka-Munkova transformacija bolj upošteva fizikalne zakonitosti razpršenosti svetlobe kot absorpcijska transformacija, je presenetljivo, da se slabše obnese v primerjavi z absorpcijsko transformacijo. Še bolj nenavadno pa je, da ne izboljša napovedi v primerjavi z uporabo originalnih spektrov. Kubelka-Munkova transformacija naj bi namreč pomagala linearizirati vrednosti spektrov glede na vrednosti kemijskofizikalnih lastnosti. Očitno so originalni spektri že dovolj linearno odvisni od vsebnosti sestavin papirja, čeprav transformacija absorpcije, ki se tudi uporablja za linearizacijo spektrov, malenkost izboljša napovedi modelov.

Opazimo lahko, da imata obe transformaciji za linearizacijo spektrov (A in KM) manjši vpliv na izboljšanje napovedi kot ostale metode predprocesiranja, ki se uporabljajo predvsem za odstranjevanje vpliva razpršenosti svetlobe in vpliva premika osnovne črte. Vse ostale 4 metode predprocesiranja (MSC, SNV, SGD in OSC) v povprečju povečajo uspešnost napovedi modelov v primerjavi z uporabo originalnih spektrov. Prav tako se poveča povprečna uspešnost napovedi modelov glede na transformiran spekter, na katerem je bila predhodno izvedena bodisi absorpcijska transformacija bodisi Kubelka-Munkova transformacija. To pomeni, da vse 4 metode predprocesiranja v povprečju izboljšajo napovedi ne glede na to, katera transformacija je bila pred tem uporabljena. Razlog je v tem, da metode za linearizacijo spektrov spremenijo samo obliko spektrov, ne odstranijo pa navpične zamaknjenosti spektrov. Za slednje se namreč uporabljajo ostale metode predprocesiranja.

Kot smo lahko videli na grafih (glej sliko 4.8 in sliko 4.9), ki prikazujeta transformirane spektre, imata metodi MSC in SNV na pogled zelo podoben učinek. Razlikujeta se le v različnem razponu spektralnih vrednosti. Tudi ta graf prikazuje, da imata obe metodi praktično enak vpliv na napoved modelov. Metoda SNV je v vseh treh kombinacijah le malenkost boljša od metode MSC.

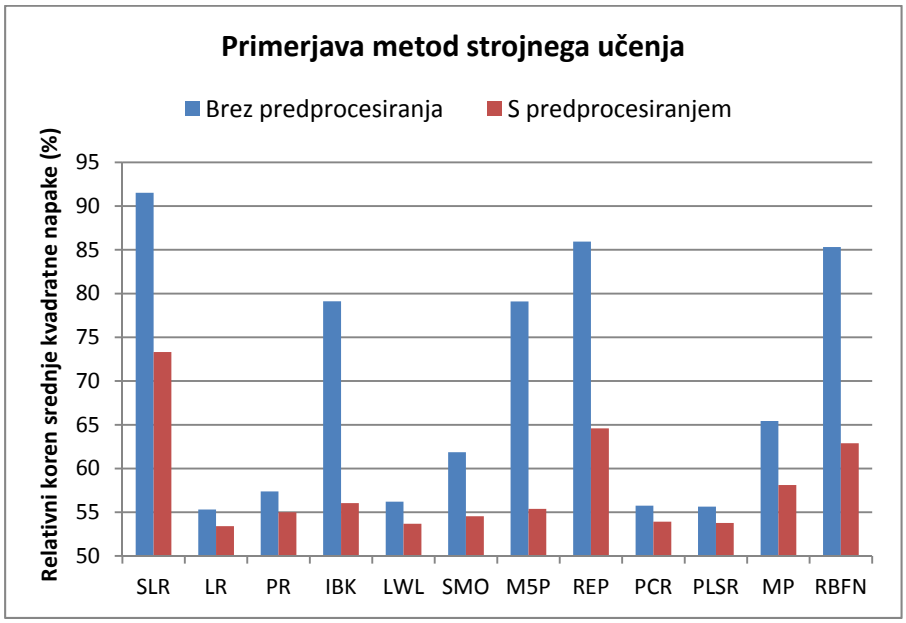
Odvajanje spektrov (SGD) pa ima različen vpliv na uspešnost napovedi glede na predhodno uporabo različnih metod za linearizacijo spektrov. Pri odvajanju originalnih spektrov je napoved v povprečju slabša od napovedi modelov na spektrih, na katerih je bila uporabljena metoda MSC ali SNV. Podobno velja za spektre, ki so bili predhodno

transformirani s Kubelka-Munkovo transformacijo. Razlika pa se pokaže pri spektrih, ki so bili pred tem transformirani z absorpcijsko transformacijo, kjer odvajanje daje boljše rezultate kot metodi MSC in SNV. Izkaže se, da odvajanje spektrov, transformiranih z absorpcijsko transformacijo, daje glede na vse preverjene kombinacije celo najboljše rezultate.

Ortogonalna korekcija signala (OSC) ima približno enak relativni koren srednje kvadratne napake kot metoda SGD. Čeprav metoda OSC odstrani le tisto informacijo, ki ni povezana z napovedovanjem kemijskofizikalne lastnosti, izgleda, da ni nič boljša od drugih metod predprocesiranja.

7.3 Primerjava metod strojnega učenja

Poglejmo si še, katere metode strojnega učenja se najbolje obnesejo pri napovedovanju kemijskofizikalnih lastnosti papirja. Naslednji graf (glej sliko 7.16) prikazuje povprečne rezultate modelov za različne metode strojnega učenja, pri čemer je bil povprečni rezultat, ki ga prikazuje posamezen stolpec, izračunan tako, da se je naprej poiskal najboljši model posamezne metode strojnega učenja za posamezno kemijskofizikalno lastnost in nato izračunalo povprečje relativnih korenov srednje kvadratne napake teh najboljših modelov. Za vsako metodo strojnega učenja sta prikazana 2 stolpca. Prvi stolpec prikazuje povprečne rezultate metod strojnega učenja na originalnih spektrih, drugi stolpec pa prikazuje povprečne rezultate metod strojnega učenja na spektrih, na katerih je bila izvedena ena od metod predprocesiranja. Pri računanju povprečja se vedno vzame najboljši model vseh metod predprocesiranja.



Slika 7.16: Primerjava metod strojnega učenja.

Prva zakonitost, ki nam pade v oči, je, da se rezultati modelov pri prav vseh metodah strojnega učenja izboljšajo pri predhodni uporabi vsaj ene izmed metod predprocesiranja spektrov. Vidimo lahko, da predprocesiranje spektrov na nekatere metode strojnega učenja bolj vpliva, na nekatere druge pa manj.

Predprocesiranje spektrov ima največji vpliv na tiste metode strojnega učenja, ki pri

gradnji modela uporabljajo le manjše število atributov, oziroma na tiste metode, pri katerih je napoved novega primera odvisna le od manjšega števila podobnih učnih primerov. V prvo skupino spada enostavna linearna regresija (SLR). V drugo skupino pa spadajo metode k-najbližjih sosedov, regresijska drevesa (REP), drevesa modelov (M5P) in mreže radialnih baznih funkcij (RBFN).

Izkaže se, da enostavna linearna regresija (SLR) daje najslabše rezultate tako brez uporabe kot z uporabo predprocesiranja spektrov. To je bilo tudi za pričakovati, saj je v spektru težko izluščiti le eno valovno dolžino (atribut), ki bi ustrezala absorpciji sestavine (kemijski lastnosti) papirja. Za spekter NIR je namreč značilno, da so absorpcijski pasovi posameznih sestavin široki (raztezajo se preko več valovnih dolžin) in se pogosto med seboj prekrivajo oz. seštevajo.

Predprocesiranje spektrov močno vpliva tudi na metodo k-najbližjih sosedov (IBK). Relativni koren srednje kvadratne napake se izboljša za več kot 20 odstotkov. Pri metodi k-najbližjih sosedov se napoved izračuna kot uteženo povprečje vrednosti kemijskofizikalne lastnosti najbolj podobnih vzorcev papirja. Ker pa so spektri zaradi vpliva razpršenosti svetlobe in vpliva drugih dejavnikov med seboj precej navpično zamaknjeni, se pri izbiri najboljših sosedov izbere tiste vzorce, katerih spektri so na približno isti višini kot spekter vzorca, katerega vrednost KFL napovedujemo. Vzorci s spektri na enakih višinah pa običajno po sestavi papirja niso podobni vzorcu, katerega vrednost KFL napovedujemo. Ko z metodami predprocesiranja zamaknjenost spektrov vsaj približno odpravimo, se pri iskanju podobnih vzorcev bolj upošteva njihova kemijska sestava. Napovedi metode k-najbližjih vzorcev so zato veliko bolj učinkovite na spektrih, na katerih je bilo predhodno izvedeno predprocesiranje.

Zaradi istega razloga kot pri metodi k-najbližjih sosedov, se tudi pri regresijskih drevesih (REP) in drevesih modelov (M5P) pomembno izboljšajo napovedi modelov, če pred tem spektre predprocesiramo. Tudi tukaj se povprečna napoved izboljša za več kot 20 odstotkov. Drevo razdeli vzorce papirja na več podmnožic. Ker je vpliv zamaknjenosti vzorcev večji od same kemijske sestave vzorcev, drevo razdeli vzorce na podmnožice vzorcev s podobnimi spektri, kar pa ne ustreza vedno podobni sestavi vzorcev. V listih drevesa ponavadi ostane majhno število vzorcev in lokalne napovedi na manjšem številu vzorcev, katerih spektri so zamaknjeni, niso najbolj zanesljive. S predprocesiranjem se pri gradnji drevesa bolj upošteva kemijska informacija v spektrih in v listih drevesa tako dobimo podmnožice spektrov, ki so po sestavi med seboj bolj podobne, kar pa vpliva tudi na boljšo napoved kemijskofizikalnih lastnosti.

Predprocesiranje ima precej dober učinek tudi pri mreži radialnih baznih funkcij. Podobno kot metoda k-najbližjih sosedov, tudi mreža radialnih baznih funkcij (RBFN) pri napovedovanju kemijskofizikalne lastnosti uporablja razdaljo, ki temelji na spektrih. Ker se spektri podobnih vzorcev zaradi zamaknjenosti in drugih vplivov med seboj precej razlikujejo, se tudi izračunane razdalje teh spektrov do nekega centra radialne bazne funkcije precej razlikujejo. Posledica tega je, da mreža pri učenju na originalnih spektrih bolj upošteva vpliv razpršenosti svetlobe in vplive drugih pojavov kot pa kemijsko informacijo v spektrih. S predprocesiranjem se ta vpliv zmanjša in mreža se zna bolje naučiti napovedovati kemijskofizikalne lastnosti papirja.

Predprocesiranje ima zelo majhen vpliv na linearno regresijo (LR), regresijo pace (PR), linearno lokalno uteženo regresijo (LWL), regresijo glavnih komponent (PCR) in regresijo delnih najmanjših kvadratov (PLSR). Pri teh metodah se napoved izboljša le za nekaj odstotkov.

Vidimo lahko, da se na originalnih spektrih najbolje obnese linearna regresija (LR). Z

odstotek slabšo povprečno napovedjo ji sledita metodi (PCR in PLSR), ki temeljita na podatkovni kompresiji, za odstotek slabše pa še linearna lokalno utežena regresija (LWL) in regresija pace (PR). Metoda podpornih vektorjev (SMO) daje približno 5 odstotkov in večnivojski perceptron (MP) približno 10 odstotkov slabšo napoved. Sledijo ji metode, ki imajo več kot 20 odstotkov slabšo napoved. To so metoda k-najbližjih sosedov (IBK), drevesa modelov (M5P), mreža radialnih baznih funkcij (RBFN), regresijska drevesa (REP) in enostavna linearna regresija (SLR).

Pri predhodni uporabi predprocesiranja spektrov pa se najbolje obnesejo linearna regresija (LR), linearno lokalno utežena regresija (LWL) in obe metodi (PCR in PLSR), ki temeljita na podatkovni kompresiji. Rezultati teh so približno enaki. Za kakšen odstotek slabše so metoda podpornih vektorjev (SMO), regresija pace (PR), drevesa modelov (M5P) in metoda k-najbližjih sosedov (IBK). Ostale metode (MP, RBFN, REP, SLR) pa so v povprečju precej slabše.

V opisu problema je bilo ugotovljeno, da je napovedovanje kemijskofizikalnih lastnosti linearen problem. To namreč pravi Beerov zakon. Rezultati, ki smo jih dobili, to tudi potrjujejo. Najbolje se namreč obnesejo tiste metode strojnega, ki temeljijo na linearni regresiji. To so predvsem linearna regresija (LR), linearna lokalno utežena regresija (LWL), regresija glavnih komponent (PCR) in regresija delnih najmanjših kvadratov (PLSR).

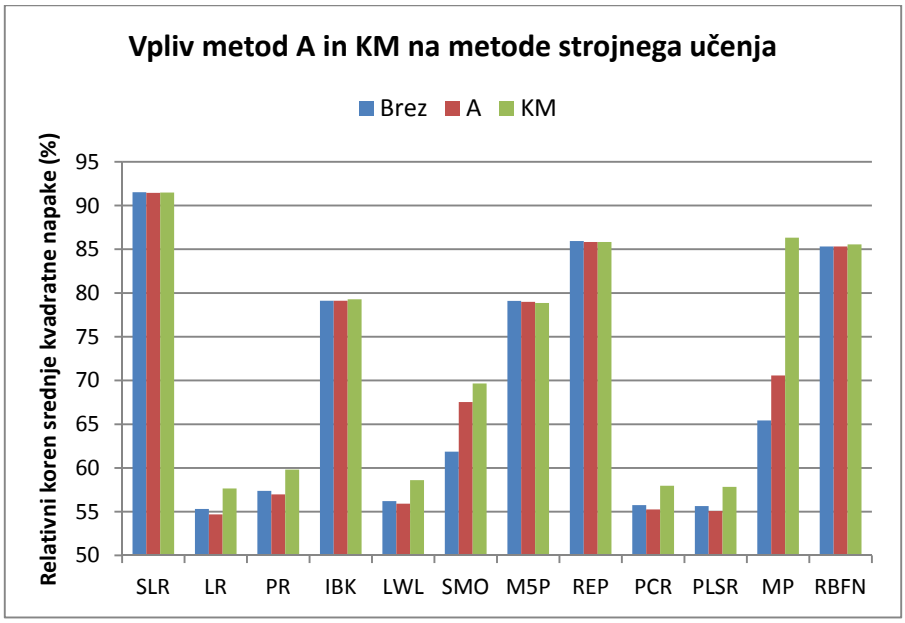
7.4 Vpliv posameznih metod predprocesiranja na metode strojnega učenja

Vpliv absorpcijske (A) in Kubelka-Munkove (KM) transformacije na metode strojnega učenja

Najprej si poglejmo, kako na uspešnost metod strojnega učenja vplivata obe transformaciji za linearizacijo spektrov. Graf (glej sliko 7.17) prikazuje za vsako metodo strojnega učenja povprečje najboljših napovedi vseh kemijskofizikalnih lastnosti pri uporabi posamezne metode predprocesiranja spektrov.

Izkaže se, da obe transformaciji nimata nobenega vpliva na metode strojnega učenja, ki za napoved bodisi uporabljajo le en atribut bodisi na napoved vpliva le nekaj podobnih vzorcev. Te metode so enostavna linearna regresija (SLR), metoda k-najbližjih sosedov (IBK), drevesa modelov (M5P), regresijska drevesa (REP) in mreža radialnih baznih funkcij (RBFN).

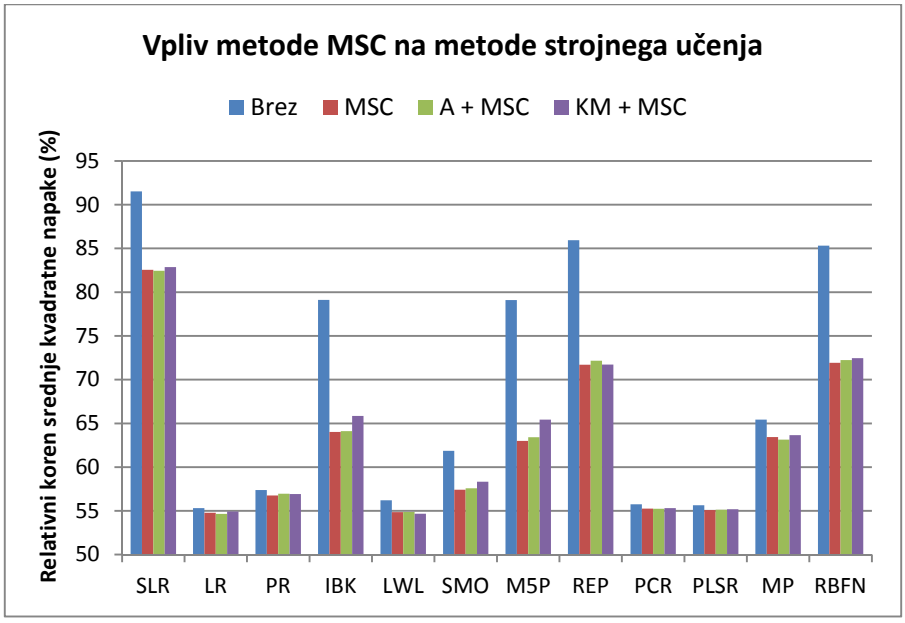
Na vse druge metode strojnega učenja je vpliv Kubelka-Munkove transformacije (KM) negativen, saj večino napovedi poslabša. Za absorpcijsko transformacijo pa je značilno, da malenkost izboljša napovedi najboljših metod strojnega učenja. Pri metodah strojnega učenja (SMO in MP), ki vrednosti kemijskofizikalnih lastnosti napovedujeta slabše, pa napovedi še poslabša.



Slika 7.17: Vpliv absorpcijske (A) in Kubelka-Munkove (KM) transformacije na posamezne metode strojnega učenja.

Vpliv multiplikativne korekcije razpršenosti (MSC) na metode strojnega učenja

Poglejmo si, kako metoda MSC vpliva na metode strojnega učenja. Vrednosti na naslednjem grafu (glej sliko 7.18) so izračunane na enak način kot pri prejšnjem grafu.



Slika 7.18: Vpliv multiplikativne korekcije razpršenosti (MSC) na posamezne metode strojnega učenja.

Opazimo lahko, da se metoda MSC v kombinaciji z absorpcijsko transformacijo obnese približno enako dobro, medtem ko se v kombinaciji s Kubelka-Munkovo transformacijo obnese malce slabše. Drugače pa velja, da v vsakem primeru izboljša napovedi

v primerjavi s tem, če metode MSC ne uporabimo.

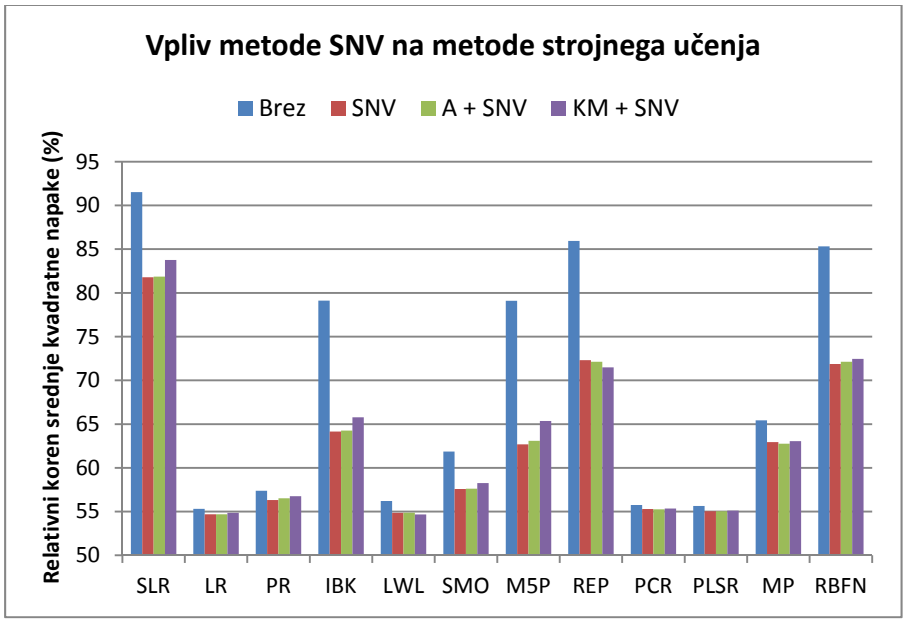
Kot je bilo ugotovljeno že zgoraj, metoda MSC, ki odpravlja zamaknjenost spektrov, vpliva predvsem na tiste metode strojnega učenja, ki so na to bolj občutljive. To pa so enostavna linearna regresija (SLR), metoda k-najbližjih sosedov (IBK), drevesa modelov (M5P), regresijska drevesa (REP) in mreža radialnih baznih funkcij (RBFN).

Metoda MSC ima zelo majhen vpliv na metode strojnega učenja, kot so linearna regresija (LR), linearna lokalno utežena regresija (LWL), regresija glavnih komponent (PCR) in regresija delnih najmanjših kvadratov (PLSR). Te metode znajo same odstraniti vpliv zamaknjenosti spektrov oz. vpliv premika osnovne črte.

Linearna regresija (LR) lahko vpliv zamaknjenosti spektrov kompenzira tako, da poišče tisto valovno dolžino v spektru, ki vsebuje zelo malo kemijske informacije. Pri tej valovni dolžini je potem vsebovana samo informacija o razpršenosti svetlobe in drugih pojavih, ki vplivajo na premik spektra v navpični smeri. Vpliv zamaknjenosti lahko potem odstrani z odštevanjem spektralnih vrednosti pri eni ali več takih valovnih dolžinah. Na enak način zna tudi linearna lokalno utežena regresija (LWL) odstraniti vpliv zamaknjenosti spektrov.

Tudi metodi (PCR in PLSR), ki temeljita na podatkovni kompresiji, sta sposobni upoštevati zamaknjenost spektrov. Izkaže se, da ima zamaknjenost spektrov največji prispevek pri varianci spektrov. Ker pa regresija glavnih komponent (PCR) išče ortogonalne komponente, ki imajo največjo varianco, potem s prvo glavno komponento poišče ravno tisto variacijo v spektrih, ki je posledica zamaknjenosti spektrov. Prva glavna komponenta bo tako po obliki enaka povprečnemu spektru. Ker pa regresija glavnih komponent napove vrednost kemijskofizikalne lastnosti z uporabo linearne regresije na glavnih komponentah, lahko z regresijskim koeficientom prve komponente vpliv zamaknjenosti spektrov odstranimo. Na enak način deluje tudi regresija delnih najmanjših kvadratov (PLSR).

Vpliv metode Standard Normal Variate (SNV) na metode strojnega učenja



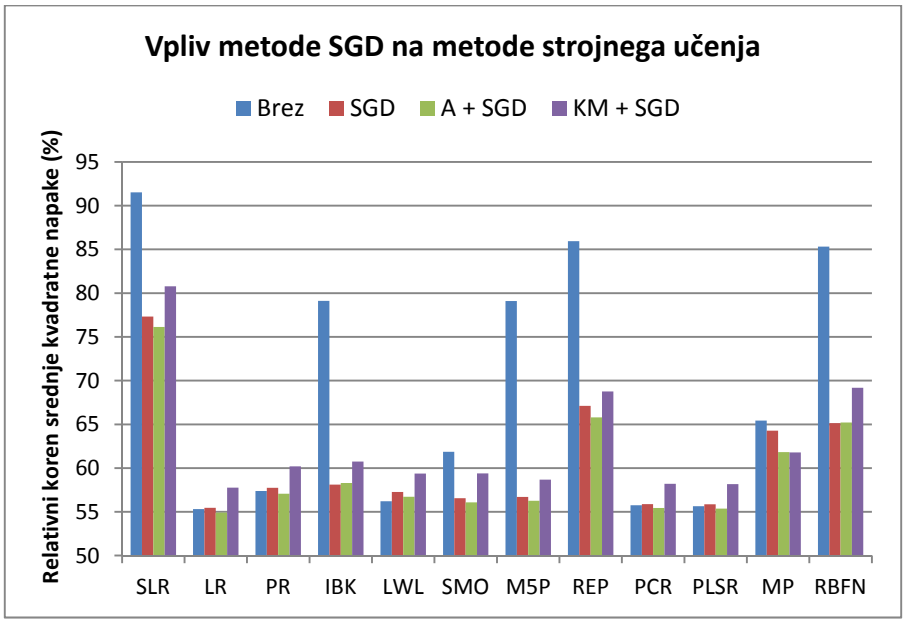
Slika 7.19: Vpliv metode Standard Normal Variate (SNV) na posamezne metode strojnega učenja.

Kot lahko vidimo iz grafa (glej sliko 7.19), ima metoda SNV praktično identičen učinek na uspešnost napovedi posameznih metod strojnega učenja kot metoda MSC. Če primerjamo graf spektrov (glej sliko 4.8), na katerih je bila uporabljena metoda MSC, in graf spektrov (glej sliko 4.9), na katerih je bila uporabljena metoda SNV, vidimo, da sta po obliki skoraj enaka, razlikujeta se le v različnem razponu vrednosti spektralnih spremenljivk. To pa je tudi vzrok za zelo podobne rezultate, saj različen razpon vrednosti spektralnih spremenljivk na napoved ne vpliva. Vpliv razpršenosti svetlobe in vplive drugih pojavov, ki povzročajo zamaknjenost spektrov, posamezne metode strojnega učenja odstranjujejo na enak način, kot je opisano pri metodi MSC.

Vpliv odvajanja spektrov (SGD) na metode strojnega učenja

Poglejmo si še, kakšen je učinek odvajanja spektrov (SGD) na uspešnost napovedi določenih metod strojnega učenja. Na naslednjem grafu (glej sliko 7.20) lahko opazimo, da ima odvajanje spektrov v primerjavi z metodama MSC in SNV precej večji vpliv na nekatere metode strojnega učenja. To predvsem velja za enostavno linearno regresijo (SLR), metodo k-najbližjih sosedov (IBK), drevesa modelov (M5P), regresijska drevesa (REP) in mrežo radialnih baznih funkcij (RBFN). Povprečni relativni koren srednje kvadratne napake se pri teh metodah izboljša za več kot 5 odstotkov v primerjavi z metodama MSC in SNV. Pri ostalih metodah strojnega učenja pa ima enak ali pa celo nekoliko negativen vpliv na uspešnost napovedi.

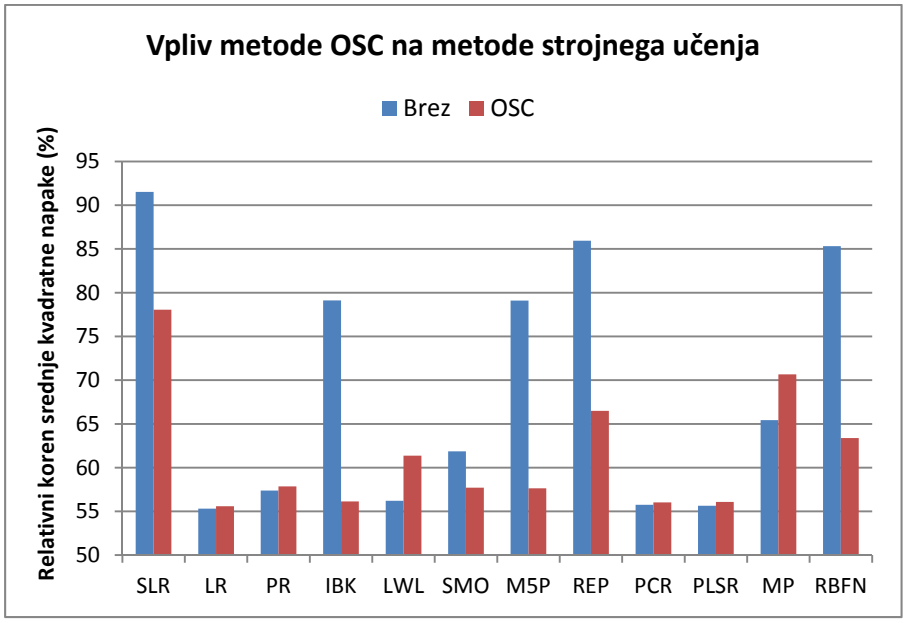
Opazimo pa lahko še nekaj. Odvajanje spektrov, na katerih je bila predhodno izvedena Kubelka-Munkova transformacija, prav pri vseh metodah strojnega učenja poslabša povprečno napoved za vsaj 3 odstotke v primerjavi z odvajanjem originalnih spektrov. Za predhodno uporabljeno absorpcijsko transformacijo pa velja, da napovedi malenkost izboljša.



Slika 7.20: Vpliv odvajanja spektrov s Savitzky-Golayevo metodo na metode strojnega učenja.

Vpliv ortogonalne korekcije signala (OSC) na metode strojnega učenja

Kakšen je vpliv ortogonalne korekcije (OSC) signala, pa prikazuje naslednji graf (glej sliko 7.21). Vidimo lahko, da metoda OSC daje zelo podobne rezultate kot odvajanje spektrov. Največja razlika je pri linearni lokalno uteženi regresiji (LWL) in večnivojskem perceptronu (MP), kjer se napoved poslabša za okrog 5 odstotkov. Poleg tega pa se izkaže, da pri metodah strojnega učenja, ki dajejo najboljše rezultate, nekoliko poveča relativni koren srednje kvadratne napake. To pomeni, da metode OSC ni priporočljivo uporabljati pri linearni regresiji (LR), regresiji pace (PR), regresiji glavnih komponent (PCR) in regresiji delnih najmanjših kvadratov (PLSR).



Slika 7.21: Vpliv ortogonalne korekcije signala (OSC) na metode strojnega učenja.

7.5 Najboljši pari metod predprocesiranja in metod strojnega učenja

Tabela 7.2 prikazuje povprečne rezultate vseh preizkušenih parov metod predprocesiranja in metod strojnega učenja. Povprečen rezultat za posamezen par je izračunan kot povprečje relativnega korena srednje kvadratne napake modelov vseh kemijskofizikalnih lastnosti. Polja v tabeli so obarvana tako, da zelena barva predstavlja najboljši par, rdeča barva pa najslabši par.

Iz tabele 7.2 je razvidno, da se najbolje obnese linearna regresija (LR) skupaj v paru bodisi z metodo MSC bodisi z metodo SNV, pri čemer je lahko pred tem izvedena tudi absorpcijska transformacija (A) ali Kubelka-Munkova transformacija (KM). Sledijo ji linearno lokalno utežena regresija (LWL) in metodi (PCR in PLSR), ki temeljita na podatkovni kompresiji, prav tako v paru bodisi z metodo MSC bodisi z metodo SNV. Naštete metode strojnega učenja delujejo malenkost slabše, ampak še vedno dobro tudi v kombinaciji z metodo SGD ali OSC ali pri neuporabi nobene metode predprocesiranja.

	SLR	LR	PR	IBK	LWL	SMO	M5P	REP	PCR	PLSR	MP	RBFN
Brez	91,52	55,31	57,38	79,12	56,20	61,87	79,10	85,94	55,74	55,64	65,42	85,33
A	91,45	54,69	56,98	79,12	55,92	67,53	78,99	85,82	55,27	55,08	70,57	85,32
KM	91,48	57,66	59,81	79,28	58,60	69,66	78,86	85,82	57,96	57,84	86,33	85,56
MSC	82,56	54,79	56,76	64,03	54,86	57,41	63,01	71,72	55,26	55,09	63,43	71,93
SNV	81,78	54,70	56,31	64,14	54,86	57,59	62,68	72,31	55,29	55,03	62,94	71,87
SGD	77,32	55,47	57,74	58,12	57,28	56,56	56,70	67,11	55,89	55,85	64,28	65,14
OSC	78,05	55,58	57,85	56,13	61,37	57,71	57,64	66,49	56,02	56,09	70,67	63,38
A + MSC	82,44	54,65	56,97	64,11	54,87	57,58	63,42	72,17	55,23	55,14	63,14	72,24
A + SNV	81,86	54,68	56,52	64,26	54,88	57,62	63,09	72,13	55,25	55,05	62,76	72,12
A + SGD	76,13	54,91	57,06	58,29	56,72	56,08	56,26	65,79	55,44	55,38	61,82	65,21
KM + MSC	82,86	54,92	56,92	65,85	54,67	58,33	65,43	71,72	55,31	55,19	63,65	72,46
KM + SNV	83,77	54,86	56,76	65,78	54,68	58,26	65,36	71,49	55,35	55,10	63,06	72,45
KM + SGD	80,78	57,77	60,20	60,75	59,38	59,39	58,69	68,77	58,20	58,17	61,79	69,17

Tabela 7.2: Povprečni rezultati najboljših modelov za vse preizkušene pare metod predprocesiranja in metod strojnega učenja.

7.6 Najboljše metode pri posamezni kemijskofizikalni lastnosti

Poglejmo še, katere metode predprocesiranja so najboljše pri posamezni kemijskofizikalni lastnosti. Tabela 7.3 za vsako kemijskofizikalno lastnost prikazuje relativni koren srednje kvadratne napake najboljših modelov pri uporabi posamezne metode predprocesiranja spektrov. Vsako polje v določeni vrstici je obarvano tako, da zelena barva predstavlja najboljši model, rdeča pa najslabšega.

	Brez	A	KM	MSC	SNV	SGD	OSC	A + MSC	A + SNV	A + SGD	KM + MSC	KM + SNV	KM + SGD
KFL 1	39,83	38,12	35,66	37,90	37,64	33,15	29,81	38,13	36,50	32,15	36,39	36,42	33,01
KFL 2	51,10	50,67	56,95	51,40	51,46	51,13	51,33	51,43	51,49	50,59	53,05	53,60	56,81
KFL 3	65,81	65,03	66,79	65,73	65,78	65,69	66,36	65,41	65,45	65,04	65,23	65,32	66,93
KFL 4	32,77	32,44	40,01	29,83	29,99	31,34	30,69	29,73	29,95	31,42	30,65	30,85	31,84
KFL 5	46,11	46,05	47,28	45,24	45,26	45,36	43,98	45,35	45,33	44,64	45,12	45,06	45,20
KFL 6	48,95	48,74	54,96	48,87	48,66	48,93	49,39	48,87	48,72	49,02	50,48	50,40	56,03
KFL 7	78,59	76,88	75,28	77,98	77,24	78,55	77,95	77,93	77,15	77,47	76,35	75,87	75,53
KFL 8	47,82	48,12	52,47	46,54	46,51	48,46	48,19	46,50	46,19	48,49	46,21	46,33	47,35
KFL 9	48,00	48,04	50,66	46,91	46,92	48,08	47,34	47,00	47,01	47,78	47,58	47,34	50,09
KFL 10	88,03	88,65	91,59	86,41	86,68	85,79	86,92	86,41	86,84	85,20	86,73	86,75	86,04
KFL 11	57,68	56,87	62,17	54,71	54,55	57,74	58,09	54,74	54,62	56,71	55,33	55,30	60,89
KFL 12	67,67	66,69	67,13	66,21	66,25	68,60	68,70	66,27	66,31	67,65	65,38	65,34	68,15
KFL 13	36,57	36,04	41,21	33,71	33,38	33,90	34,05	33,70	33,68	34,81	33,93	34,07	36,66
KFL 14	63,38	61,57	60,85	62,42	62,58	60,07	63,01	62,56	62,74	59,54	62,76	62,71	61,19

Tabele 7.3: Rezultati najboljših modelov posamezne kemijskofizikalne lastnosti glede na uporabljeno metodo predprocesiranje spektrov.

Opazimo lahko, da je vedno dobro uporabiti vsaj eno metodo predprocesiranja, saj v prvem stolpcu ne zasledimo zelene barve. Drugače pa vidimo, da je pri skoraj vseh kemijskofizikalnih lastnostih najboljša drugačna metoda predprocesiranja. Zanimivo je, da je pri KFL 7 najboljša Kubelka-Munkova transformacija, čeprav je pri ostalih

kemijskofizikalnih lastnostih pogostokrat najslabša. Iz tega lahko sklepamo, da je pri iskanju najboljšega modela najbolje preizkusiti čim več metod predprocesiranja.

Sedaj pa si poglejmo še, katere metode strojnega učenja so najboljše pri posameznih kemijskofizikalnih lastnostih. Tabela 7.4 prikazuje rezultate metod strojnega učenja na originalnih podatkih, tabela 7.5 pa prikazuje najboljše rezultate metod strojnega učenja na spektrih, na katerih je bila uporablja ena od metod predprocesiranja. Vrstice so obarvane na enak način kot v tabeli 7.3.

	SLR	LR	PR	IBK	LWL	SMO	M5P	REP	PCR	PLSR	MP	RBFN
KFL 1	88,69	39,83	42,12	72,76	40,18	52,61	70,53	79,57	39,93	40,11	43,63	80,62
KFL 2	94,49	51,10	54,87	83,81	52,92	64,70	84,34	93,78	51,60	51,70	64,67	85,60
KFL 3	96,24	65,81	68,49	90,67	65,92	68,89	92,05	94,54	65,94	66,08	79,10	92,70
KFL 4	77,90	32,89	33,81	54,73	33,25	35,43	56,28	65,32	32,77	32,92	41,24	73,50
KFL 5	79,43	46,33	47,27	60,69	46,11	47,50	56,03	63,68	46,49	46,48	47,15	68,61
KFL 6	96,46	49,35	50,67	82,86	49,78	58,28	83,10	90,97	49,24	48,95	58,46	85,31
KFL 7	94,83	78,59	82,43	93,81	78,81	90,05	91,05	94,43	79,21	79,03	86,36	93,71
KFL 8	82,89	47,82	48,28	67,84	48,47	51,91	68,74	77,39	48,32	48,25	57,79	72,41
KFL 9	98,64	48,03	49,17	71,80	48,00	48,96	75,81	83,37	48,22	48,24	50,57	86,98
KFL 10	100,41	88,77	89,76	97,98	88,03	90,30	97,90	99,93	90,19	90,01	103,73	99,66
KFL 11	96,89	58,26	60,98	92,65	63,33	66,56	92,86	96,54	58,66	57,68	76,16	92,90
KFL 12	94,61	67,67	70,45	88,80	68,48	70,01	91,69	95,59	68,50	68,37	82,64	90,95
KFL 13	82,26	36,57	37,84	57,42	36,57	37,51	56,07	71,22	36,89	36,94	38,08	78,30
KFL 14	97,49	63,38	67,18	91,91	66,97	83,44	90,96	96,83	64,48	64,16	86,38	93,29

Tabele 7.4: Rezultati najboljših modelov posamezne kemijskofizikalne lastnosti pri neuporabi nobene metode predprocesiranja spektrov.

	SLR	LR	PR	IBK	LWL	SMO	M5P	REP	PCR	PLSR	MP	RBFN
KFL 1	71,80	35,66	36,70	29,81	35,46	35,88	32,15	39,80	35,76	35,60	37,23	50,67
KFL 2	76,00	50,59	53,34	56,49	51,40	51,86	53,00	67,61	51,13	51,07	61,58	63,30
KFL 3	86,27	65,03	66,17	71,43	65,33	66,31	68,28	85,03	65,15	65,35	71,81	79,62
KFL 4	43,50	29,95	30,53	31,25	30,59	30,27	29,73	34,77	29,99	30,21	31,85	32,48
KFL 5	57,17	45,10	45,34	43,98	45,90	46,11	44,83	48,97	45,06	45,18	48,17	45,61
KFL 6	68,51	48,66	50,44	54,13	49,64	49,61	52,41	60,74	48,92	48,74	55,38	55,79
KFL 7	89,71	75,48	79,50	77,95	75,87	77,10	79,65	87,23	75,28	75,49	85,76	84,65
KFL 8	67,63	46,19	47,28	52,83	46,21	47,17	48,49	60,28	46,71	46,68	47,35	55,52
KFL 9	62,70	46,91	47,20	48,50	47,67	47,21	48,94	53,61	47,05	47,08	50,17	47,34
KFL 10	97,26	88,52	91,31	86,92	88,65	85,20	88,31	97,36	90,35	89,87	96,41	93,10
KFL 11	87,99	55,77	58,06	64,25	54,55	58,99	58,06	74,13	56,37	55,61	58,73	78,89
KFL 12	84,36	65,34	68,39	68,71	65,35	69,40	71,27	81,23	66,48	66,43	75,47	78,70
KFL 13	48,66	33,68	33,62	34,05	33,70	33,94	33,38	37,25	34,02	33,91	34,22	36,56
KFL 14	84,85	60,85	62,07	64,43	61,36	64,48	66,91	76,23	62,61	61,74	59,54	78,15

Tabele 7.5: Rezultati najboljših modelov posamezne kemijskofizikalne lastnosti pri uporabi ene od metod predprocesiranja spektrov.

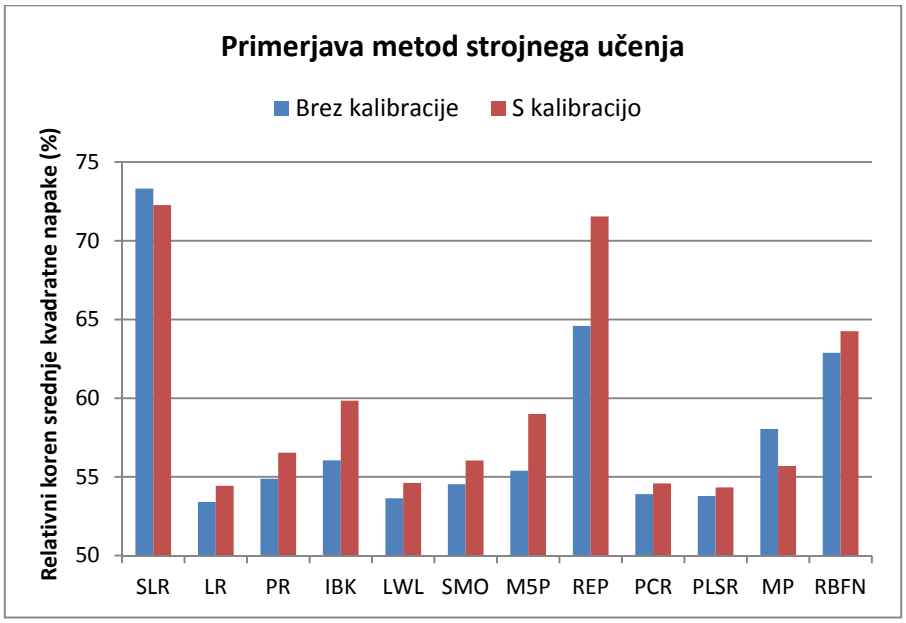
Vidimo lahko, da so metode strojnega učenja predvsem pri uporabi originalnih spektrov bolj konsistentne kot metode predprocesiranja, saj so stolpci v tabeli bolj enakomerno obarvani. Za napovedovanje kemijskofizikalnih lastnosti papirja z uporabo originalnih spektrov je najbolje vzeti linearno regresijo (LR), linearno lokalno uteženo regresijo (LWL) ali pa eno od dveh metod (PCR in PLSR), ki temeljita na podatkovni kompresiji.

Pri uporabi spektrov, na katerih je bila izvedena ena od metod predprocesiranja, pa je

poleg že naštetih metod smiselno preizkusiti tudi metodo k-najbližjih sosedov (IBK), metodo podpornih vektorjev (SMO) in drevesa modelov (M5P). Nikakor pa ni dobro uporabiti enostavno linearno regresijo (SLR), regresijska drevesa (REP) ali mrežo radialnih baznih funkcij (RBFN).

7.7 Rezultati kalibracije modelov

Kakšni so rezultati metod strojnega učenja po uporabi naknadne kalibracije, ponazarja naslednji graf (glej sliko 7.22). Na grafu sta za vsako metodo strojnega učenja prikazana povprečna rezultata najboljših modelov pred in po uporabi kalibracije, pri čemer je povprečni rezultat izračunan kot povprečje rezultatov najboljšega modela posamezne kemijskofizikalne lastnosti. Pri izbiri najboljšega modela se je upoštevalo vse modele, ne glede na to, katera metoda predprocesiranja je bila uporabljena.

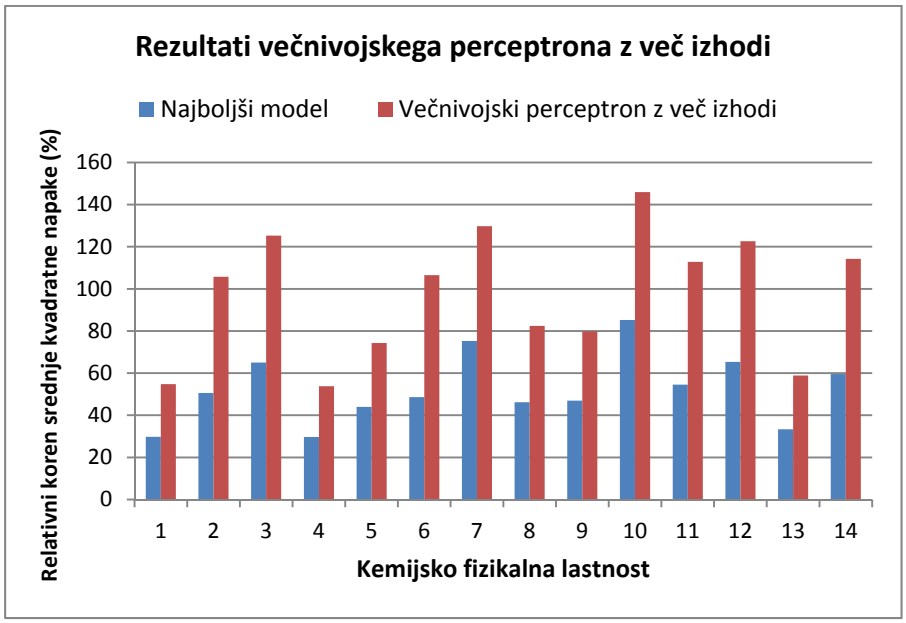


Slika 7.22: Primerjava rezultatov modelov pred in po uporabi kalibracije.

Iz grafa je razvidno, da kalibracija ne pripomore k izboljšanju napovedi posameznih metod strojnega učenja. Edini izjemi sta enostavna linearna regresija in večnivojski perceptron. Čeprav se pri enostavni linearni regresiji napoved nekoliko izboljša, je ta napoved še vedno najslabša od napovedi vseh drugih metod strojnega učenja in je zato v praksi neuporabna. Izboljšanje napovedi večnivojskega perceptrona pa je verjetno posledica naključja. Za nevronske mreže je namreč značilno, da je njihova napoved zelo odvisna od same strukture mreže in števila iteracij v učnem procesu. Ker je učenje večnivojskega perceptrona zahtevalo preveč časa, je bila preizkušena samo ena struktura umetne nevronske mreže. Če bi preizkusili večje število mrež in bi uporabili tudi različno število prehodov preko učnih primerov, bi dobili verjetno drugačne rezultate.

7.8 Rezultati večnivojskega perceptrona z več izhodi

Preizkušena je bila tudi večnivojska umetna nevronska mreža z več izhodnimi nevroni, za vsako kemijskofizikalno lastnost enega. Ta umetna nevronska mreža je torej hkrati izračunala napovedi vseh kemijskofizikalnih lastnosti. Ker umetna nevronska mreža za učenje zahteva veliko časa, je bila preizkušena le ena struktura umetne nevronske mreže. Pri učenju je bilo uporabljenih 3000 prehodov preko vseh učnih primerov. Učenje mreže se je izvajalo na originalnih spektrih, kar pomeni, da ni bilo uporabljenega nobenega predprocesiranja spektrov.

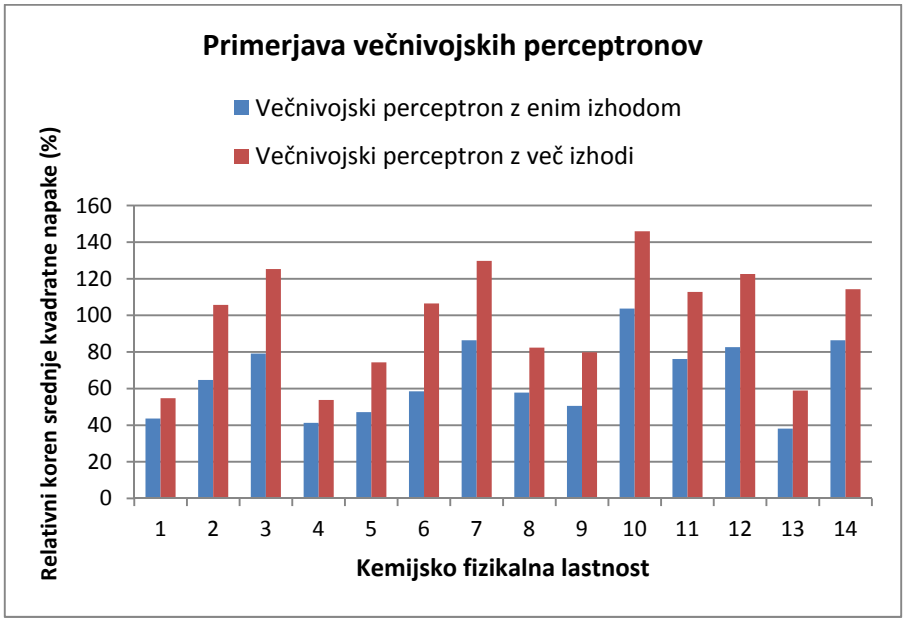


Slika 7.23: Rezultati najboljšega modela in rezultati večnivojskega perceptrona z več izhodi za posamezno kemijskofizikalno lastnost.

Graf (glej sliko 7.23) prikazuje rezultate večnivojskega perceptrona z več izhodi in rezultate najboljših modelov, ki smo jih dobili do sedaj. Iz grafa je razvidno, da umetna nevronska mreža pri napovedovanju kemijskofizikalnih lastnosti daje veliko slabše rezultate kot metode, ki smo jih preizkusili do sedaj, in je za napovedovanje kemijskofizikalnih lastnosti neuporabna. To je verjetno posledica tega, da je bila preizkušena le ena struktura mreže pri točno določenem številu ponovitev učnih primerov. Za bolj zanesljive rezultate bi bilo treba preizkusiti veliko več različnih umetnih nevronskih mrež in ugotoviti, pri koliko prehodih preko učnih primerov daje mreža najboljše rezultate. Za to pa je potrebno veliko več časa.

Primerjajmo še rezultate večnivojskega perceptrona z več izhodi z rezultati večnivojskih perceptronov z enim izhodom. Rezultati so prikazani na sliki 7.24. Izkaže se, da umetna nevronska mreža z več izhodi veliko slabše napoveduje kemijskofizikalne lastnosti kot umetna nevronska mreža z enim izhodom. Vrednost RRSE mreže z več izhodi je vsaj za 10 odstotkov večja od vrednosti RRSE mreže z enim izhodom. Iz tega lahko sklepamo, da posamezni izhodi mreže v več izhodi negativno vplivajo na sosednje izhode. Za učenje mreže z več izhodi bi potrebovali večje število skritih nivojev in večje število nevronov v skritih nivojih, da bi mrežo z več izhodi naučiti naučili vsaj tako dobro napovedovati kot mrežo z enim izhodom.

Ker umetna nevronska mreža z več izhodi napoveduje vse kemijskofizikalne lastnosti hkrati, bi pričakovali, da bodo napovedi boljše, saj lahko mreža pri napovedovanju ene kemijskofizikalne lastnosti uporabi tudi informacijo, ki jo vsebujejo druge kemijskofizikalne lastnosti. Ker pa se to v tem primeru ni zgodilo, potem to pomeni, da bodisi kemijskofizikalne lastnosti med seboj niso povezane bodisi napovedi te mreže zaradi prej omenjenih razlogov niso točne.



Slika 7.24: Primerjava rezultatov večnivojskega perceptrona z več izhodi z rezultati večnivojskih perceptronov z enim izhodom

Poglavje 8

Zaključek

V diplomskem delu smo spoznali spektroskopijo NIR in uporabo le te pri napovedovanju kemijskofizikalnih lastnosti papirja. Ugotovili smo, da je napovedovanje kemijskofizikalnih lastnosti papirja linearen problem.

Preizkušenih je bilo več metod predprocesiranja spektrov. Ugotovili smo, da je pri napovedovanju lastnosti papirja vedno dobro uporabiti vsaj eno od metod predprocesiranja spektrov. Najboljše rezultate pa vračata metodi Standard Normal Variate in multiplikativna korekcija razpršenosti.

Preizkušenih je bilo tudi več metod strojnega učenja. Najbolj se obnesejo metode, ki gradijo linearne modele. To so predvsem linearna regresija, regresija glavnih komponent in regresija delnih najmanjših kvadratov.

Ugotovili smo, da je iz spektrov možno napovedovati vse kemijskofizikalne lastnosti papirja, ki so bile izmerjene, saj je pri vseh relativni koren srednje kvadratne napake napovedi manjši od 100 odstotkov. To pomeni, da je v spektrih vsebovana informacija o prav vseh izmerjenih kemijskofizikalnih lastnostih.

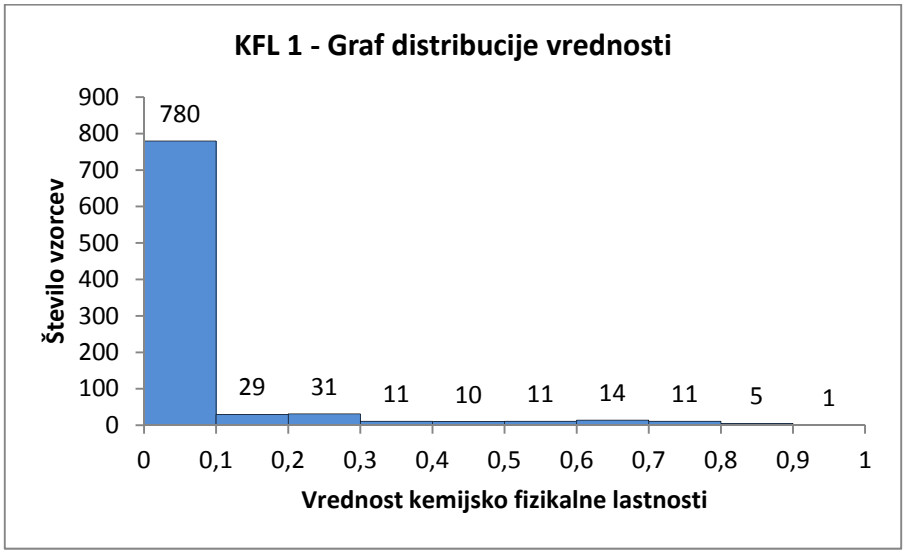
V nadaljnjem raziskovanju bi bilo dobro preizkusiti še kakšno drugo metodo strojnega učenja. Predvsem pa bi bilo dobro pogledati, ali se da z umetnimi nevronskimi mrežami izboljšati rezultate. Umetne nevronske mreže imajo namreč velik potencial, vendar je za preizkušanje različnih mrež potrebno zelo veliko časa.

Ker podatkovna množica vsebuje zelo raznolike vzorce papirja, bi bilo pametno, da bi vzorce papirja najprej razporedili glede na njihovo sestavo v nekaj podmnožic. Potem pa bi poskusili napovedovati kemijskofizikalne lastnosti papirja za vsako podmnožico posebej. Tako bi verjetno dobili boljše rezultate, saj bi bili vzorci v posameznih podmnožicah med seboj bolj podobni.

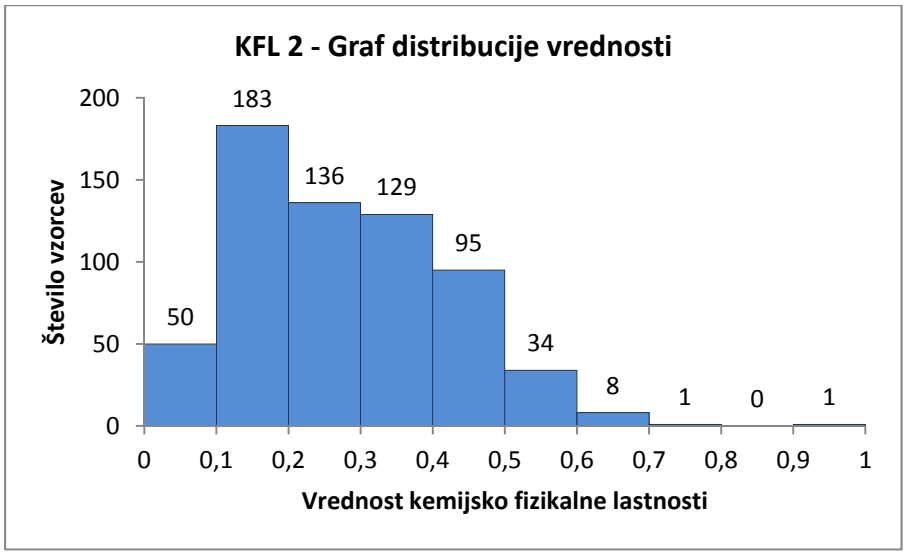
Dodatek A

Grafi porazdelitev vrednosti kemijskofizikalnih lastnosti

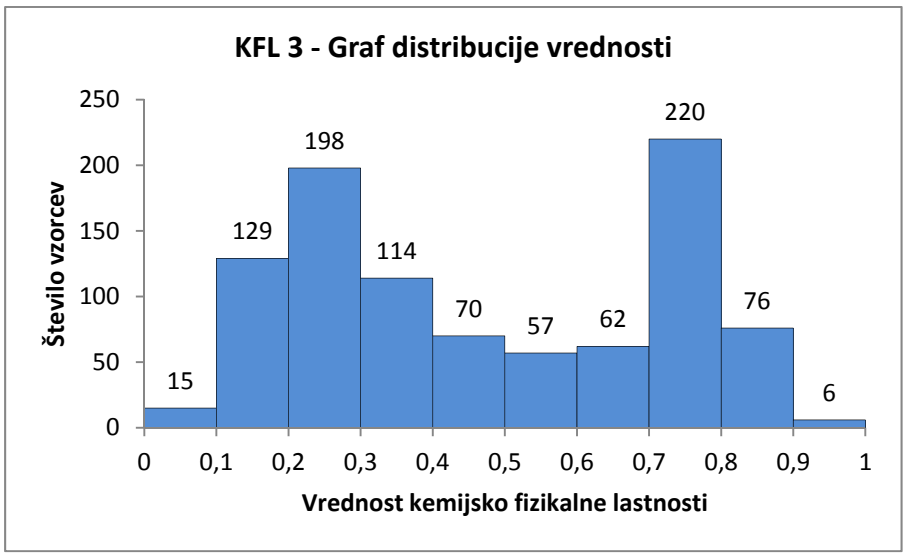
Tukaj so zbrani grafični prikazi porazdelitev vrednosti posameznih kemijskofizikalnih lastnosti papirja.



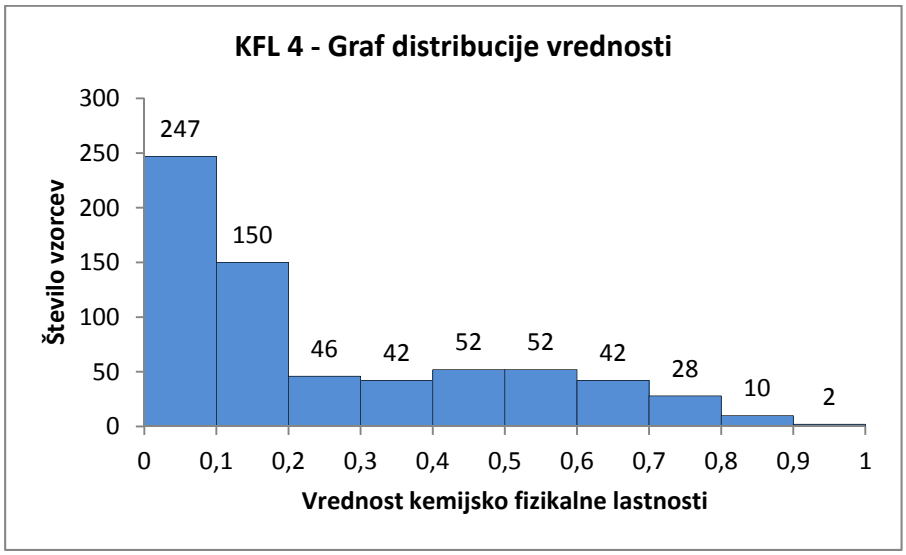
Slika A.1: Porazdelitev vrednosti kemijskofizikalne lastnosti 1.



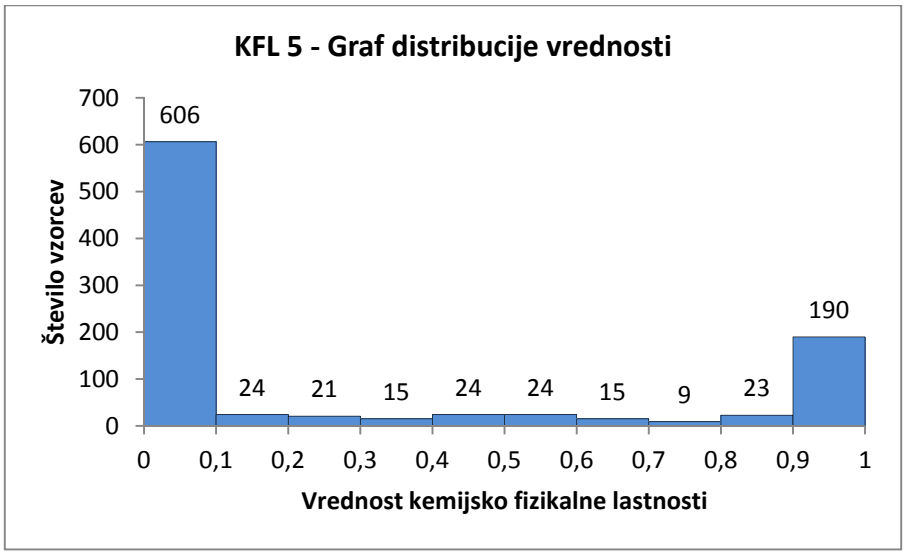
Slika A.2: Porazdelitev vrednosti kemijskofizikalne lastnosti 2.



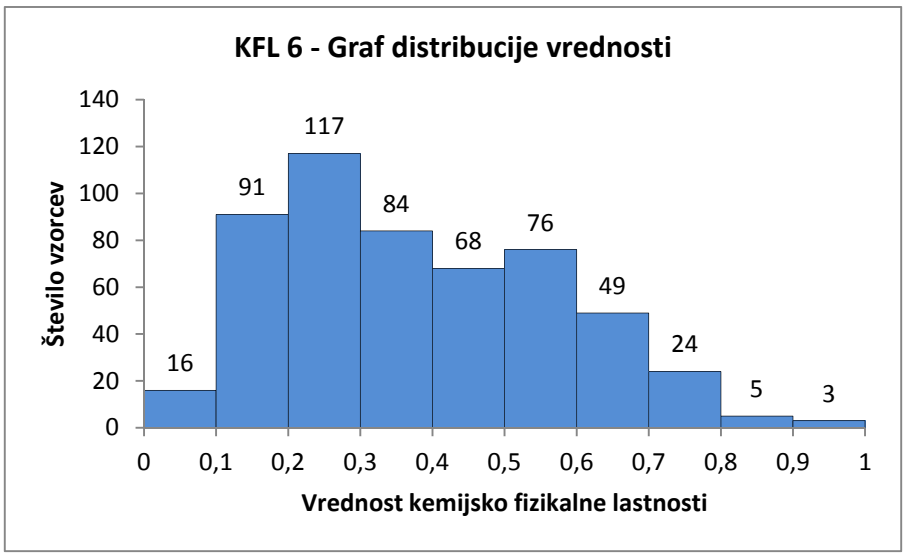
Slika A.3: Porazdelitev vrednosti kemijskofizikalne lastnosti 3.



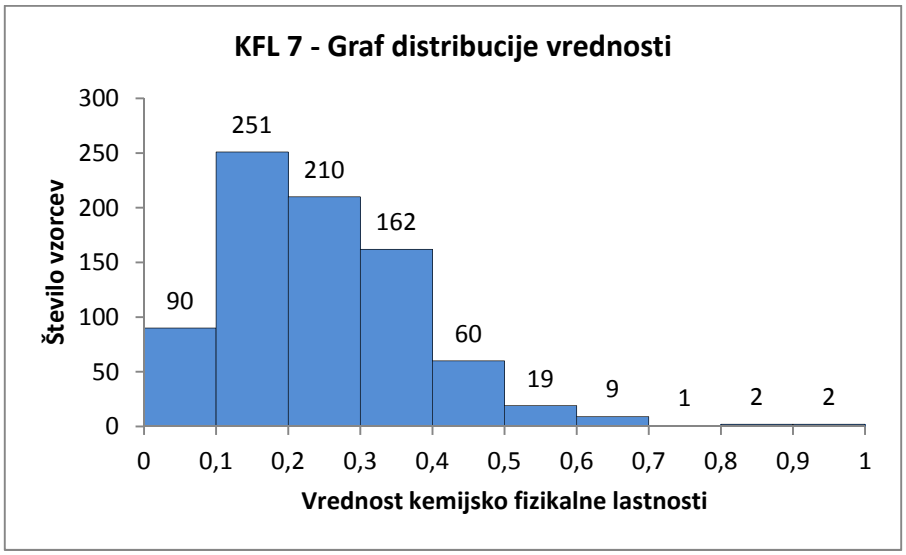
Slika A.4: Porazdelitev vrednosti kemijskofizikalne lastnosti 4.



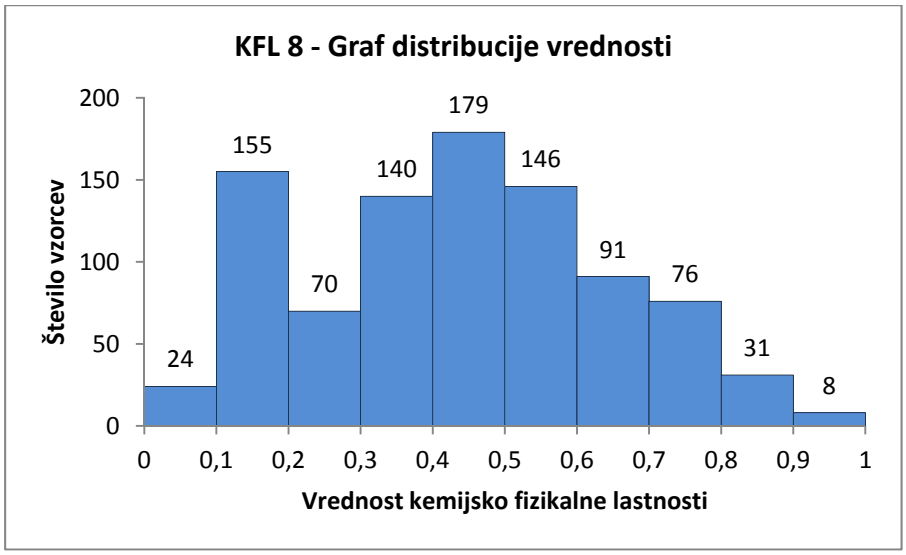
Slika A.5: Porazdelitev vrednosti kemijskofizikalne lastnosti 5.



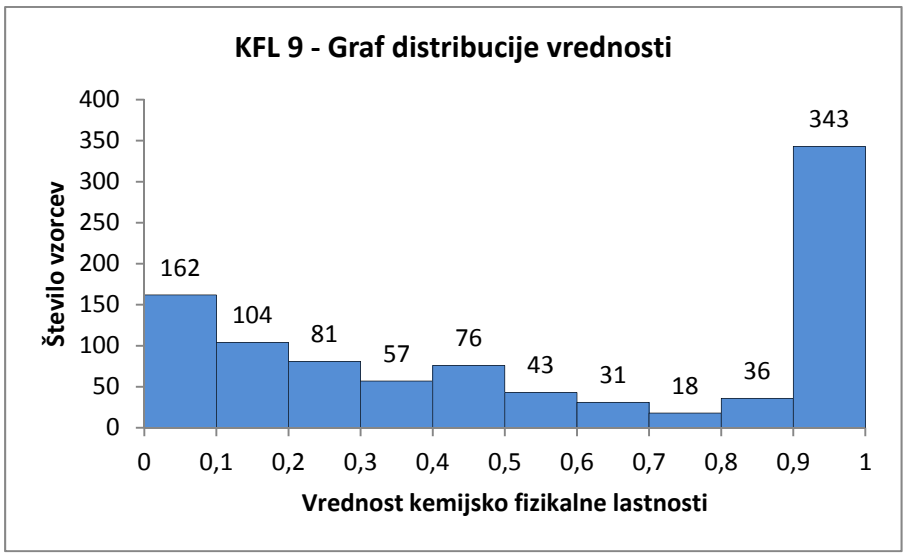
Slika A.6: Porazdelitev vrednosti kemijskofizikalne lastnosti 6.



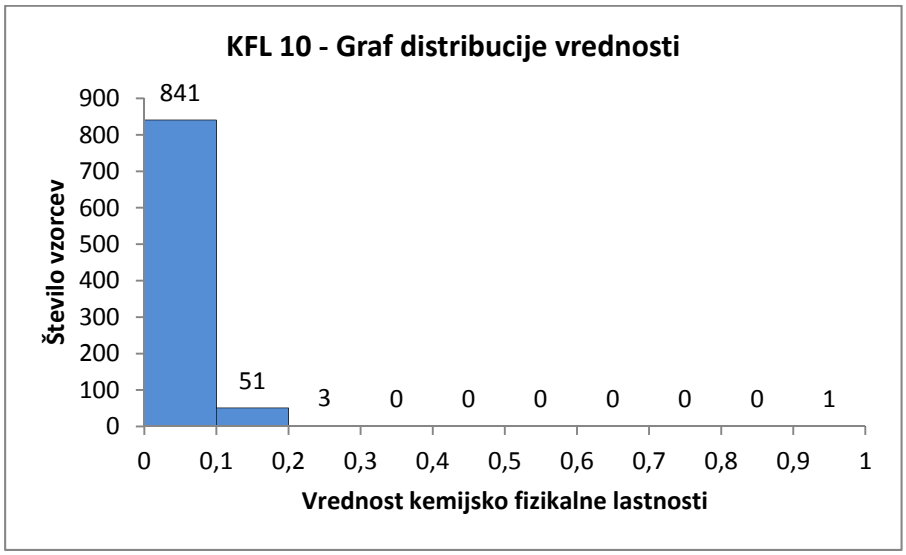
Slika A.7: Porazdelitev vrednosti kemijskofizikalne lastnosti 7.



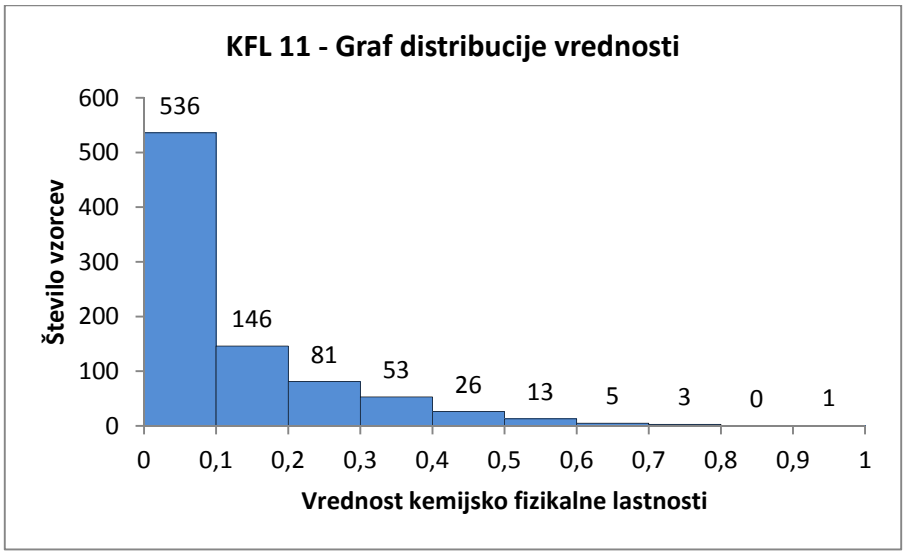
Slika A.8: Porazdelitev vrednosti kemijskofizikalne lastnosti 8.



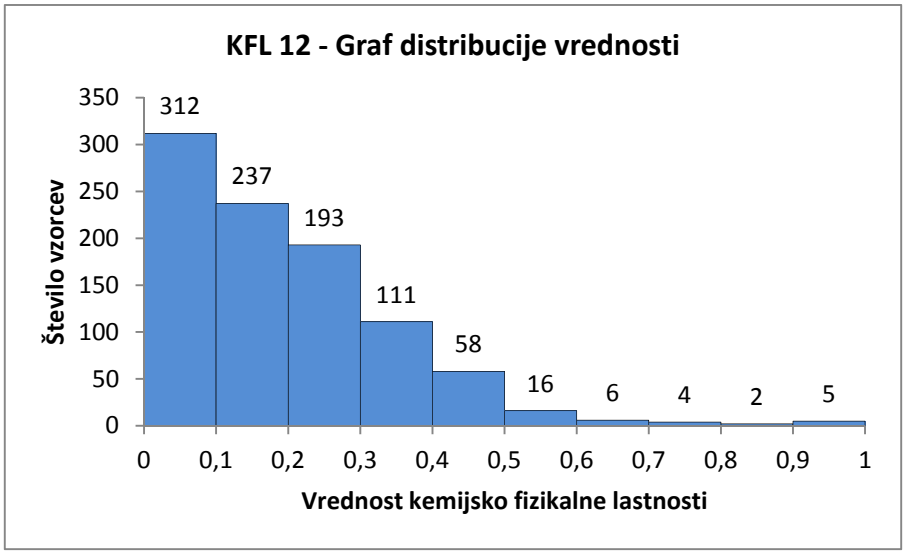
Slika A.9: Porazdelitev vrednosti kemijskofizikalne lastnosti 9.



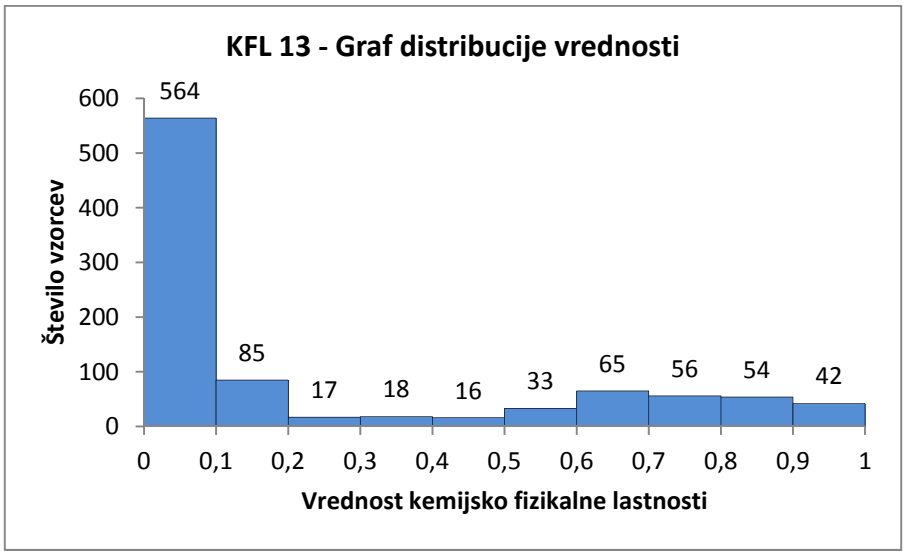
Slika A.10: Porazdelitev vrednosti kemijskofizikalne lastnosti 10.



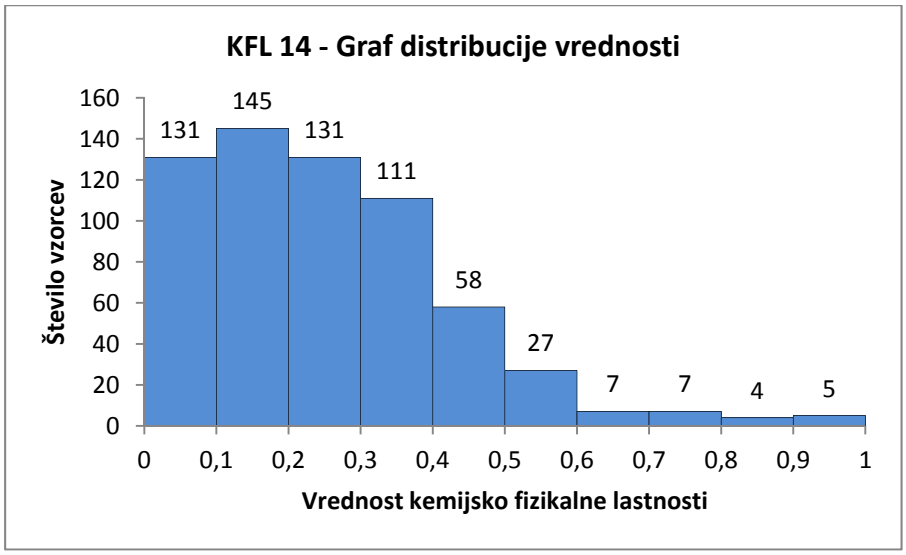
Slika A.11: Porazdelitev vrednosti kemijskofizikalne lastnosti 11.



Slika A.12: Porazdelitev vrednosti kemijskofizikalne lastnosti 12.



Slika A.13: Porazdelitev vrednosti kemijskofizikalne lastnosti 13.



Slika A.14: Porazdelitev vrednosti kemijskofizikalne lastnosti 14.

Literatura

[1] *C. G. Atkenson, A. W. Moore, S. Schaal, "Locally Weighted Learning", Artificial Intelligence Review, št. 11, str. 11-73, 1997.*

[2] *D. A. Burns, E. W. Ciurczak, Handbook of Near-Infrared Analysis, New York, 2001.*

[3] *M. Carlin, "Radial Basis Function Networks and Nonlinear Data Modelling", Proceedings of Neuro-Nimes, str. 623-633, 1992.*

[4] *R. N. Feudale, H. Tan, S. D. Brown, "Piecewise orthogonal signal correction", Chemometrics and Intelligent Laboratory Systems, št. 63, str. 129-138, 2002.*

[5] *R. N. Feudale, H. Tan, S. D. Brown, "Improved Piecewise Orthogonal Signal Correction Algorithm", Applied Spectroscopy, št. 10, zv. 57, str. 1201-1206, 2003.*

[6] *U. Henniges, M. Schwanninger, A. Potthast, "Non-Destructive Determination of Cellulose Functional Groups and Molecular Weight in Pulp Sheets and Historical Papers by NIR-PLS-R" v zborniku Durability of paper and writing 2, 2nd International Symposium and Workshops, Ljubljana, julij 2008, str. 46-47.*

[7] *I. Kononenko, Strojno učenje, Ljubljana: Fakulteta za računalništvo in informatiko, 2005.*

[8] *R. Kreslin, I. Kononenko, "A general method for calibrating regression models" v zborniku konference Information Society 2011, IJS, oktober 2011.*

[9] *D. Lichtblau, M. Strlič, T. Trafela, J. Kolar, M. Anders, "Determination of mechanical properties of historical paper based on NIR spectroscopy and chemometrics – a new instrument", Applied Physics A, št. 92, str. 191-195, 2008.*

[10] *T. Naes, T. Isaksson, T. Fearn, T. Davies, A User-Friendly Guide to Multivariate Calibration and Classification, Chichester: NIR Publications, 2002.*

[11] *H. Martens, T. Naes, Multivariate Calibration, Chichester: John Wiley & Sons, 1991.*

[12] H. W. Siesler, Y. Ozaki, S. Kawata, H. M. Heise, Near-Infrared Spectroscopy: Principles, Instruments, Applications, Weinheim: Wiley-VCH, 2002.

[13] M. Strlič, J. Kolar, D. Lichtblau, "The SurveNIR project – a dedicated near infrared instrument for paper characterization", Museum Microclimates, National Museum of Denmark, str. 81-84, 2007.

[14] T. Trafela, M. Strlič, J.kolar, D. A. Lichtblau, M. Anders, D. Pucko Mencigar, B. Pihlar, "Nondestructive Analysis and Dating of Historical Paper Based on IR Spectroscopy and Chemometric Data Evaluation", Anal. Chem., št. 79, str. 6319-6323, 2007.

[15] I. H. Witten, Data Mining: Practical machine learning tools and techniques, 3. izdaja, Burlington: Morgan Kaufmann, 2011.

[16] Y. Wang, I. H. Witten. "Pace regression", Technical Report 99/12, Department of Computer Science, University of Waikato, 1999.

[17] J. Workmann, L. Weyer, Practical Guide to Interpretive Near-Infrared Spectroscopy, Boca Raton: CRC Press - Taylor & Francis Group, 2008.

[18] (2011) SurveNIR Project. Dostopno na: http://www.science4heritage.org/survenir/

[19] (2011) J. Trygg, "Every thing you need to know about Orthogonal Signal Correction (OSC) filters – and how they can improve interpretation of your data", Homepage of Chemometrics, Editorial March 2002. Dostopno na: http://www.chemometrics.se/editorial/may2002.pdf

[20] (2011) U. Halici, Artificial Neural Networks, pogl. 9. Dostopno na: http://www.eee.metu.edu.tr/~halici/courses/543LectureNotes/lecturenotes-pdf/ch9.pdf

[21] (2011) Eclipse. Dostopno na: http://www.eclipse.org/

[22] (2011) Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Dostopno na: http://www.cs.waikato.ac.nz/ml/weka/