UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matija Polajnar

Vedenje algoritma FreeViz v visokorazsežnostnih prostorih

DIPLOMSKO DELO
NA INTERDISCIPLINARNEM UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Janez Demšar

Ljubljana, 2009

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: 00013/2009
Datum: 01.09.2009

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

Kandidat: **MATIJA POLAJNAR**

Naslov: **VEDENJE ALGORITMA FREEVIZ V VISOKORAZSEŽNOSTNIH PROSTORIH**

BEHAVIOUR OF FREEVIZ IN HIGH-DIMENSIONAL SPACES

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

FreeViz je vizualizacijska metoda, ki za dane podatke, opisane z vrednostmi atributov in razredom, poišče dvodimenzionalno projekcijo, ki čim jasneje loči med primeri iz različnih razredov. Takšne projekcije lahko uporabljamo za vizualizacijo podatkov in za klasifikacijo.

Raziščite, kako se metoda vede v visokorazsežnostnih prostorih. S poskusi na resničnih domenah, kot so podatki iz genetskih mrež, in primerno konstruiranih umetnih domenah ugotovite, ali metoda v takšnih primerih deluje in od česa je odvisna kvaliteta dobljenih projekcij.

Mentor:

doc. dr. Janez Demšar

Fakulteta za računalništvo in informatiko
UNIVERZA V LJUBLJANI
2

Dekan Fakultete za računalništvo in informatiko:

prof. dr. Franc Solina

Dekan Fakultete za matematiko in fiziko:

akad. prof. dr. Franc Forstnerič

UNIVERZA V LJUBLJANI
Fakulteta za matematiko in fiziko
6

Namesto te strani **vstavite** original izdane teme diplomskega dela s podpisom mentorja in dekana ter žigom fakultete, ki ga diplomant dvigne v študentskem referatu, preden odda izdelek v vezavo!

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani Matija Polajnar

z vpisno številko 63050208

sem avtor diplomskega dela z naslovom:

Vedenje algoritma FreeViz v visokorazsežnostnih prostorih

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom
  doc. dr. Janeza Demšarja,
- 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 8. 9. 2009 Podpis avtorja:

Zahvala

Najbolj sem hvaležen svojemu mentorju doc. dr. Janezu Demšarju, ki je neutrudno spremljal in usmerjal moje delo. Brez njega to delo ne bi nastalo v tako omejenem času, če sploh.

Zahvalo dolgujem tudi Tomažu Curku in Tomažu Gregorcu, ki sta mi dovolila in omogočila uporabo računske moči zadostnega števila računalnikov, da sem lahko pravočasno izvedel vse poskuse.

Za pomoč na področju medicinskega izrazoslovja se zahvaljujem Evi Kočevar.

Vsem bližnjim se opravičujem za asocialnost v času trdega dela in se jim zahvaljujem za razumevanje in podporo.

To delo posvečam
svoji izbranki, domačim in rodni grudi
v zahvalo za
potrpežljivost, vzgojo in žulje.

Kazalo

Seznam uporabljenih kratic in simbolov

AUC	*Area Under Curve* – ploščina pod krivuljo
CA	*Classification Accuracy* – klasifikacijska točnost
***k*NN**	*k Nearest Neighbours* – k najbližjih sosedov
LOESS	*Locally Weighted Scatterplot Smoothing* – lokalno uteženo glajenje razsevnega diagrama
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machine* – metoda podpornih vektorjev
XOR	*Exclusive OR* – ekskluzivni ALI

Povzetek

FreeViz je metoda za lokalno optimizacijo linearnih projekcij na področju *odkrivanja zakonitosti v podatkih* (angl. *data mining*). V diplomskem delu obravnavamo uspešnost metode na podatkih s področja genetike. Ti podatki imajo tipično bistveno več atributov kot učnih primerov, zato najprej z uporabo linearne algebre pokažemo, kakšne težave zaradi te lastnosti pričakujemo. V nadaljevanju se posvečamo iskanju lastnosti množic podatkov, ki pozitivno ali negativno vplivajo na delovanje FreeViza.

Ker je metoda namenjena iskanju dobrih (informativnih) vizualizacij, njeno uspešnost merimo s kvaliteto klasifikacije klasifikatorja k najbližjih sosedov (angl. *k Nearest Neighbours, kNN*) na dobljeni projekciji.

Predstavljeni rezultati kažejo, da FreeViz na nekaterih domenah s področja genetike res ni uspešen, na eni izmed uporabljenih množic podatkov pa je dosegel relativno dobre rezultate.

Z namenom iskanja lastnosti množic, ki vplivajo na delovanje metode, smo le-to preizkusili na sintetičnih množicah podatkov. Rezultati kažejo na zanemarljiv vpliv razmerja med številom atributov in številom učnih primerov, slabo delovanje v primeru majhnega števila pomembnih atributov in izboljšanje delovanja ob prisotnosti medsebojno koreliranih atributov.

Opazovali smo tudi poti projekcij atributov med optimizacijo projekcije in nismo odkrili pravila za ločevanje pomembnih in redundantnih atributov. S FreeVizom dobljena projekcija sicer pri velikem številu učnih primerov redundantne atribute projicira bistveno bližje koordinatnemu izhodišču, če pa je primerov manj kot atributov, se ta razlika ne pojavi. V tem primeru sicer atributov po pomembnosti ne loči niti naivni Bayesov klasifikator.

Ključne besede:

vizualizacija, linearna projekcija, FreeViz, genetika, redundanca, korelacija, pomembnost atributa

Abstract

FreeViz is a data mining method for local optimization of linear projections. In this thesis we cover method's performance in field of genetics. Genetic datasets usually contain much more attributes than instances; we first use linear algebra to predict method's problems on such data. Then we try to find properties of datasets that influence the performance of FreeViz.

The goal of the analysed method is to find good (informative) visualizations, so we estimate its performance by measuring quality of a kNN (k Nearest Neighbours) classifier on the projections it yields.

The results confirm FreeViz's poor performance on genetic data, but it nevertheless proved successful on one of the used dataset.

In pursue of dataset properties that influence method's performance, we generated synthetic datasets. Results show that the ratio between attribute count and instance count has negligible influence. On the other hand, FreeViz's quality is degraded when most of the attributes are redundant and improved when there are mutually correlated attributes.

We have also observed the paths that attribute projections make during optimization, but found no rule to distinguish redundant attributes from the rest. In case there is a large number of instances, FreeViz yields a projection that maps redundant attributes closer to the origin. That is not the case when there are more attributes than instances. However, in that case not even a nomogram for a naive Bayesian classifier can distinguish between informative and redundant attributes.

Key words:

visualization, linear projection, FreeViz, genetics, redundancy, correlation, attribute importance

Poglavje 1

Uvod

Vizualizacija podatkov je pomemben del odkrivanja zakonitosti v podatkih. Dobra vizualizacija omogoča interpretacijo podatkov in odkrivanje zakonitosti v podatkih.

Vsak učni primer (angl. *learning instance*) opisuje množica atributov, zato učno množico lahko opišemo kot prostor visoke dimenzije (vsaka dimenzija je en atribut). Pri vizualizacijah smo omejeni z dvema ali včasih tremi dimenzijami.

Če se omejimo na dve dimenziji, mora metoda vizualizacije uporabiti projekcijo iz prostora atributov v dvodimenzionalni prostor. **FreeViz** je nadzorovana (angl. *supervised*) hevristična metoda za optimizacijo separacije razredov pri linearni projekciji v dvodimenzionalni prostor. Z drugimi besedami, učne primere želi linearno projicirati tako, da bodo primeri istega razreda čim bolj skupaj, primeri različnih razredov pa čim bolj oddaljeni. [1]

Težava se pojavi pri visokodimenzionalnih podatkih (tj. ko imamo opravka z veliko atributi), saj je pogosto možno najti linearno projekcijo, ki je popolnoma prilagojena učnim podatkom. Pojav imenujemo prekomerno prileganje (angl. *overfitting*) in je običajno nezaželen, saj pomeni izgubo splošnosti in prilagajanje šumu v podatkih.

Namen tega diplomskega dela je raziskati delovanje FreeViza na podatkih s področja genetike ter vpliv

- dimenzionalnosti podatkov,
- redundantnih atributov (takih, ki ne vplivajo na razred) in
- koreliranih atributov

na kvaliteto FreeViza. Konkretno bo s prečnim preverjanjem ocenjen klasifikator, ki učne podatke projicira v ravnino, projekcijo optimizira s postopkom FreeViz in nato nove primere klasificira glede na k najbližjih sosedov projekcije novega primera.

Prav tako je namen odkriti, če se med optimizacijo projekcije redundatnih atributov premikajo na matematično opisljivo drugačen način kot projekcije pomembnih (neredundantnih) atributov.

Poglavje 2

Teorija

2.1 Linearne projekcije

Naj bo E matrika učnih primerov dimenzije $n \times m$, kjer vrstice predstavljajo n primerov, stolpci pa m atributov. Diskretni atributi so pretvorjeni v zvezne binarne atribute. Element $e_{i,j}$ matrike E je torej vrednost j-tega atributa pri i-tem primeru, e_i pa vodoraven vektor, ki predstavlja i-ti primer.

Linearno projekcijo vrstic te matrike v dvodimenzionalni prostor lahko zapišemo z matriko projekcije A velikosti $m \times 2$. Ta matrika vsebuje projekcije baznih vektorjev (tj. atributov) v 2D prostor. Projekcija i-tega atributa je torej A_i, deli pa se na komponenti $A_{i,\mathrm{X}}$ in $A_{i,\mathrm{Y}}$.

$$E' = E \cdot A \tag{2.1}$$

E' v enačbi 2.1 je slika učnih primerov po projekciji; dimenzija E' je $n \times 2$ – n primerov v 2D prostoru. Sliko primera e_i v 2D prostoru označimo $e'_i = e_i \cdot A$; sestavljena je iz komponent $e'_{i,\mathrm{X}}$ in $e'_{i,\mathrm{Y}}$.

Vsak učni primer pripada določenemu razredu. Razred i-tega primera označimo R_i.

2.2 FreeViz

FreeViz izhaja iz poljubne projekcijske matrike A in jo spreminja tako, da z gradientno metodo zmanjšuje *potencialno energijo* fizikalnega sistema delcev. Projekcije učnih primerov nastopajo kot delci v ravnini, ki se med seboj privlačijo, če pripadajo istemu razredu, oziroma odbijajo, če pripadajo različnim

razredom. Nižja potencialna energija sistema tako ustreza boljši separaciji razredov.

Silo delca i s položajem e'_i na delec j s položajem e'_j označimo $F_{i\to j}$.

Metoda je neodvisna od izbire načina delovanja sil, je pa smiselno, da

- odbojna sila med delci različnih razredov pada z razdaljo (kot v fiziki elektromagnetna sila), saj zadosti oddaljenih primerov različnih razredov ne želimo nadalje oddaljevati,
- privlačna sila med istorazrednimi delci raste z razdaljo (kot v fiziki močna jedrska sila), saj ni potrebe po zbliževanju tistih primerov istega razreda, ki so si že blizu.

V osnovni izvedbi FreeViz uporablja model sil, pri katerem se delca z razdaljo r privlačita s silo velikosti r, če pripadata istemu razredu, sicer pa odbijata s silo velikosti $\frac{1}{r}$ (enačbi 2.2 in 2.3).

$$F_{i\to j} = \begin{bmatrix} F_{i\to j,\mathrm{x}} \\ F_{i\to j,\mathrm{y}} \end{bmatrix} = \begin{cases} e'_i - e'_j & R_i = R_j \\ (e'_j - e'_i)\cdot r^{-2} & R_i \neq R_j \end{cases} \tag{2.2}$$

$$|F_{i\to j}| = \sqrt{F^2_{i\to j,\mathrm{x}} + F^2_{i\to j,\mathrm{y}}} = \begin{cases} r & R_i = R_j \\ \frac{1}{r} & R_i \neq R_j \end{cases} \tag{2.3}$$

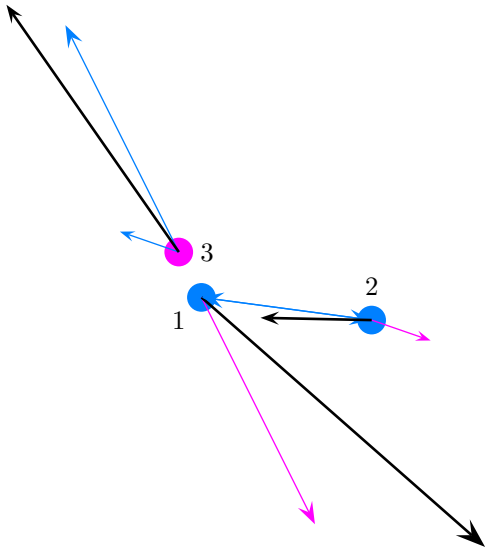
Rezultanta sil F_j na j-ti delec je seveda vsota vseh sil, ki delujejo nanj (enačba 2.4). Komponenti označimo $F_{j,\mathrm{x}}$ in $F_{j,\mathrm{y}}$.

$$F_j = \sum_{i=1}^{n} F_{i\to j} \tag{2.4}$$

Primer sil pri treh delcih (tj. treh projekcijah učnih primerov) je prikazan na sliki 2.1.

Projekcije primerov je seveda možno spremeniti le, če spremenimo projekcije atributov, zato FreeViz izračuna gradient potencialne energije W kot funkcije projekcijske matrike A, tj. $\frac{dW}{dA}$. Sprememba potencialne energije je enaka vloženemu delu, ki je v fiziki definirano kot skalarni produkt sile in odmika (spremembe položaja). Da premaknemo delec j za vektor de'_j, moramo v to vložiti $-F_j \cdot de'_j$ dela. Spremembo potencialne energije torej izračunamo po enačbi 2.5.

$$dW = -\sum_{j=1}^{n} F_j \cdot de'_j \tag{2.5}$$



Slika 2.1: Primer sil na delce pri FreeViz optimizaciji. Z oštevilčenimi krogi so prikazane projekcije učnih primerov. Dve različni barvi ustrezata dvema razredoma. Barva puščice ustreza barvi delca, ki je silo povzročil, črne puščice pa so rezultante (vsote sil na posamezne delce).
Vidimo, da se delca 1 in 3 močno odbijata, saj je razdalja med njima majhna. Sili med 2 in 3 sta šibkejši zaradi večje razdalje.
Delca 1 in 2 pripadata istemu razredu, zato se privlačita s silo, katere velikost je enaka razdalji med njima.

Gradient potencialne energije kot funkcije projekcijske matrike je podan z enačbo 2.6. Za vsako komponento $A_{i,k}$ matrike A ga je možno izračunati: enačba 2.7 podaja gradient po x koordinati i-tega atributa, analogna enačba pa velja za koordinato y.

$$\frac{dW}{dA} = -\sum_{j=1}^{n} F_j \cdot \frac{de'_j}{dA} \tag{2.6}$$

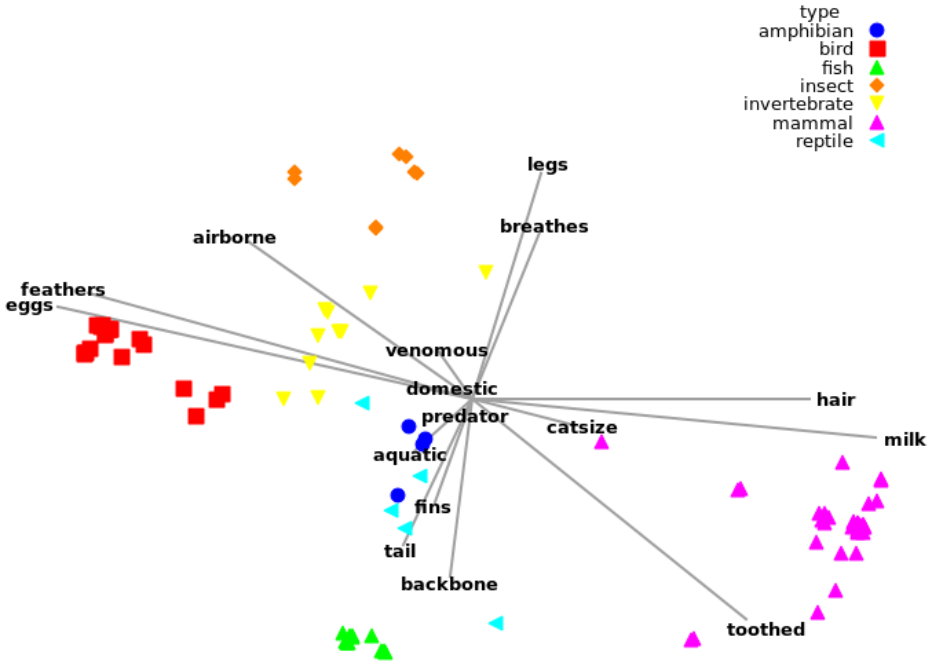
$$\frac{dW}{dA_{i,\mathrm{X}}} = -\sum_{j=1}^{n} F_j \cdot \frac{de'_j}{dA_{i,\mathrm{X}}} = -\sum_{j=1}^{n} F_{j,\mathrm{X}} \cdot e_{j,i} \tag{2.7}$$

FreeViz nato na vsakem koraku optimizacije matriko A spremeni v smeri negiranega gradienta.

Za preprečitev eksplozije ali implozije po vsakem koraku optimizacije FreeViz projekcijo centrira (tako, da je vsota vseh krajevnih vektorjev do projekcij primerov enaka 0) in normalizira (tako, da ima najdaljši krajevni vektor dolžino 1).

Optimizacija se ustavi, ko se projekcijska matrika v 50 iteracijah optimizacije ne spremeni več bistveno.

Na sliki 2.2 je primer s FreeVizom optimizirane projekcije podatkov uvrstitve živali v živalske vrste. [2]



Slika 2.2: S FreeVizom optimizirana projekcija podatkov uvrstitve živali v živalske vrste. Sive črte označujejo projekcije atributov. Vsi atributi so binarni, le *legs* ima vrednosti med 0 in 6. Barve in oblike primerov ustrezajo njihovim razredom.
Število atributov: 16. Število učnih primerov: 101. Število razredov: 7.

2.2.1 Klasifikacija s FreeVizom

S FreeVizom je možno tudi klasificirati nove primere.

Učno množico projiciramo v ravnino in nad projekcijo poženemo optimizacijo FreeViz. Dobljena projekcijska matrika in projekcije vseh učnih primerov sestavljajo naš klasifikacijski model.

Nov primer klasificiramo tako, da ga s projekcijsko matriko projiciramo v ravnino. V sliki projekcije nato najdemo k najbližjih sosedov (angl. *k Nearest Neighbours, kNN*), torej najbližjih projekcij učnih primerov, in primer uvrstimo v večinski razred.

2.2.2 Omejitve FreeViza

Če ima matrika E toliko vrstic kot stolpcev ($m = n$), torej če število atributov dosega število učnih primerov, in če matrika ni singularna, kar zaradi šuma v podatkih običajno ni, lahko za poljuben E' po enačbi 2.8 izračunamo projekcijsko matriko. Enačbo smo dobili tako, da smo enačbo 2.1 z leve množili z E^{-1}.

$$A = E^{-1} \cdot E' \tag{2.8}$$

Za E' lahko zahtevamo poljubno sliko. FreeViz bo pogosto uspel najti sliko z največjo separacijo razredov: vsi primeri istega razreda so v eni točki, različni razredi pa razporejeni po krožnici.

V primeru, ko je atributov celo več kot učnih primerov (E ima več stolpcev kot vrstic; $m > n$), je število različnih matrik A, ki za poljubno sliko E' ustrezajo enačbi linearne projekcije 2.1, neskočno. Sistem linearnih enačb, ki ga določa ta matrična enačba, je namreč poddoločen: enačb je n, vektorskih neznank (vrstic matrike A) pa $m > n$.

To dejstvo ima neželene posledice. Na podlagi razporeditve atributov po ravnini namreč lahko običajno delamo hipoteze glede korelacije med vrednostmi različnih atributov. Če k temu dodamo še položaje večjih gruč primerov istega razreda, je iz slike možno razbrati tudi vpliv vrednosti atributov na razred primera. Ker pa obstaja mnogo različnih projekcij z istim rezultatom, to morda pomeni, da položaji projekcij atributov v ravnini v resnici niso informativni, ampak je bila separacija razredov dosežena po naključju zaradi prekomernega prileganja učni množici. Naše hipoteze, dobljene na podlagi take projekcije, bodo torej nesmiselne.

Da FreeViz v situaciji z več atributi kot primeri uspe najti eno izmed projekcij, ki za dano učno množico da optimalno separacijo razredov, je prikazano na sliki 2.3. Prikazana je s FreeVizom optimizirana projekcija genetskih podatkov iz podatkovne množice *DLBCL* (vsebina domene je opisana v podpoglavju

3.1.1). Število atributov je bistveno večje od števila učnih primerov, saj je z atributi popisana aktivnost večjega števila genov v tkivih obravnavanih oseb.

Slika 2.3: S FreeVizom optimizirana projekcija genetskih podatkov, kjer je atributov bistveno več kot učnih primerov. Opazimo morebitno prekomerno prileganje podatkom.
Projekcije atributov na sliki niso prikazane, saj je število atributov preveliko. Število atributov: 7700. Število učnih primerov: 77. Število razredov: 2.

Poglavje 3

Poskusi

Za vpogled v delovanje FreeViza v nekaterih težjih okoliščinah, kot so domene z več atributi kot učnimi primeri, domene z redundantnimi atributi in domene s koreliranimi atributi, smo izvedli več eksperimentov:

- preverjanje delovanja na štirih domenah z genetskimi podatki,
- preverjanje delovanja na več vrstah umetnih (sintetičnih) množic,
- primerjavo projekcij redundantnih in pomembnih (neredundantnih) atributov med in po optimizaciji.

To poglavje je namenjeno opisom in rezultatom eksperimentov.

3.1 Podatki s področja genetike

Ker obstaja možnost, da se FreeViz na domenah z več atributi kot učnimi primeri prekomerno prilega učnim podatkom, smo uspešnost klasifikatorja FreeViz (ta je opisan v poglavju 2.2.1) z 10-kratnim prečnim preverjanjem najprej ocenili na podatkih s področja genetike. Le-ti so tipičen primer neumetne domene z več atributi kot učnimi primeri. Z atributi je popisana aktivnost večjega števila genov v rakastem tkivu, razred pa opisuje vrsto raka, za katero pacient boleha. Atributov je veliko, ker ustrezajo popisanim genom, učnih primerov pa je malo, saj je pridobivanje podatkov drago.

Zaradi majhnega števila učnih primerov lahko izbor podmnožic za prečno preverjanje opazno vpliva na izmerjeno kvaliteto klasifikacije, zato je bil vsak poskus izveden 10-krat, kot končni rezultati pa so podane mediane izmerjenih vrednosti.

3.1.1 Opis podatkov

Razpolagali smo s tremi množicami genetskih podatkov in ustvarili derivat ene izmed njih:

DLBCL [3] Difuzni velikocelični limfom B (angl. *Diffuse Large B-cell Lymphomas, DLBCL*) in folikularni limfom (angl. *Follicular Lymphomas, FL*) sta dve vrsti raka s sprva različnimi kliničnimi in morfološkimi značilnostmi. Pogosto se sčasoma spremenita tako, da ju po značilnostih ni možno več ločiti, vseeno pa zahtevata različni zdravljenji. V tej množici podatkov so za namen ločevanja teh dveh oblik raka zbrani genetski podatki.

SRBCT [4] Množica podatkov združuje genetske podatke 83 pacientov z namenom diagnostike ene od štirih vrst raka, ki zaradi podobne histološke slike spadajo v skupino malih okroglih modroceličnih tumorjev (angl. *Small Round Blue Cell Tumors, SRBCTs*).

SRBCT-2 Množica je derivat SRBCT, pri katerem smo ohranili le učne primere tistih dveh razredov, ki sta imela najvišji delež učnih primerov. Tako je problem postal dvorazreden.

lung [5] Popis aktivnosti genov 203 pacientov in klasifikacija v 5 razredov – štiri vrste pljučnega raka ter zdrava pljuča.

Število atributov in primerov, število razredov ter deleži učnih primerov, ki pripadajo posameznim razredom, so za vse štiri množice opisani v tabeli 3.1.

domena	**atributov**	**primerov**	**št. in distribucija razredov**
DLBCL	7.070	77	2: 75%, 25%
SRBCT	2.308	83	4: 35%, 13%, 22%, 30%
SRBCT-2	2.308	54	2: 54%, 46%
lung	12.600	203	5: 69%, 10%, 10%, 8%, 3%

Tabela 3.1: Podatki o uporabljenih genetskih množicah.

Za mero kvalitete klasifikatorja sta uporabljeni ploščina pod ROC krivuljo

(AUC) in klasifikacijska točnost (CA).

Za primerjavo smo na podatkih poleg FreeViza preizkusili še naivni Bayesov klasifikator in metodo podpornih vektorjev (angl. *Support Vector Machine, SVM*). Naivni Bayes in SVM sta bila uporabljena s privzetimi nastavitvami paketa Orange. To pomeni, da je Bayes ocenjeval apriorne verjetnosti razredov z relativno frekvenco, verjetnosti razredov pri vrednostih atributov pa z metodo *LOESS* (ker so atributi zvezni). SVM je za jedro uporabljal radialne bazne funkcije s parametrom $\gamma = \frac{1}{n}$ (kjer je n število učnih primerov), parameter ν pa je bil nastavljen na 0,5.

3.1.2 Rezultati

Rezultati so predstavljeni v tabelah 3.2 in 3.3.

domena	**AUC:**	naivni Bayes	SVM	**FreeViz**
DLBCL		0,74	0,49	**0,81**
SRBCT		0,98	0,95	**0,60**
SRBCT-2		1,00	1,00	**0,50**
lung		0,80	0,99	**0,64**

Tabela 3.2: Primerjava vrednosti AUC (ploščine pod ROC krivuljo) za različne klasifikatorje na genetskih domenah.

domena	**CA:**	naivni Bayes	SVM	**FreeViz**
DLBCL		0,82	0,75	**0,70**
SRBCT		0,93	0,80	**0,39**
SRBCT-2		0,98	1,00	**0,54**
lung		0,62	0,93	**0,66**

Tabela 3.3: Primerjava klasifikacijske točnosti (angl. *Classification Accuracy, CA*) za različne klasifikatorje na genetskih domenah.

Pri domeni DLBCL je FreeVizova kvaliteta primerljiva s kvaliteto ostalih dveh klasifikatorjev. Na podlagi tega lahko postavimo hipotezo, da neugodno razmerje med številom atributov in učnih primerov ni dejavnik, ki bi primerjalno znižal kvaliteto klasifikacije s FreeVizom.

Pri preostalih treh domenah (od katerih dve primere ločujeta v več kot dva razreda) je FreeViz dosegal primerjalno slabše rezultate. Domena *SRBCT-2* je

bila konstruirana ravno z namenom ugotoviti, če je večje število razredov tista lastnost domen, ki FreeVizu povzroča težave. Dejstvo, da je pri tej množici FreeViz dosegel še slabšo vrednost AUC kot pri *SRBCT*, to idejo negira. Klasifikacijska točnost med množicama z različnim številom razredov seveda ni primerljiva.

3.2 Umetni podatki

Da bi bilo vpliv lastnosti domen na FreeViz lažje razumeti, smo ga testirali tudi na umetnih podatkih. Namen je bil raziskati vplive različnih lastnosti domen, zato smo pripravili različne *vrste* umetnih množic, kot je opisano v tabeli 3.4.

Obravnavana lastnost domene	**Lastnosti umetnih množic**
razmerje med številom atributov in številom učnih primerov vrsta množice: **A**	**št. primerov:** med 60 in 1000 **št. atributov:** med 10 in 1000 **št. redundantnih atr.:** 0 **št. koreliranih atr.:** 0
razmerje med pomembnimi in redundantnimi atributi vrsta množice: **B**	**št. primerov:** 200 **št. atributov:** med 10 in 1000 **št. redundantnih atr.:** med 10 in št. atr. **št. koreliranih atr.:** 0
korelacije med atributi vrsta množice: **C**	**št. primerov:** 200 **št. atributov:** med 10 in 1000 **št. redundantnih atr.:** 0 **št. koreliranih atr.:** med 10 in št. atr.

Tabela 3.4: Lastnosti treh vrst umetnih množic.

Iz tabele je razvidno, da se pri vsaki vrsti množic dva parametra spreminjata. Generiranih je bilo veliko različnih množic s parametri na danem obsegu, da je bilo možno oceniti občutljivost kvalitete FreeViza na spremembo posameznih parametrov.

Za vsako generirano množico velja naslednje:

- Vsi atributi so zvezni z zalogo vrednosti $[-1, 1]$.
- Razred je diskretna spremenljivka z vrednostjo $\times$ ali $\circ$.

- Vsem učnim primerom znotraj ene množice je bil razred določen na enak način – naključno, pri čemer so vrednosti atributov vplivale na verjetnost posameznega razreda.

Za vsako vrsto smo generirali množice za tri *ciljne koncepte* (način vpliva atributov na razred). Ti koncepti so opisani v nadaljevanju.

3.2.1 Ciljni koncepti

Pri sledečem opisu ciljnih konceptov so uporabljene naslednje oznake:

- m je število atributov in m_r je število redundantnih atributov,
- a_i je vrednost i-tega atributa pri obravnavanem primeru,
- R je razred obravnavanega primera.

Ciljni koncept *črta* Za posamezno generirano množico podatkov je bil najprej enakomerno naključno izbran vektor koeficientov $\vec{c} \in [-1, 1]^{m-m_r}$ in normaliziran tako, da je bila vsota pozitivnih elementov manjša ali enaka 1, vsota negativnih elementov večja ali enaka -1, za vsaj eno od teh dveh zvez pa je veljala enakost.

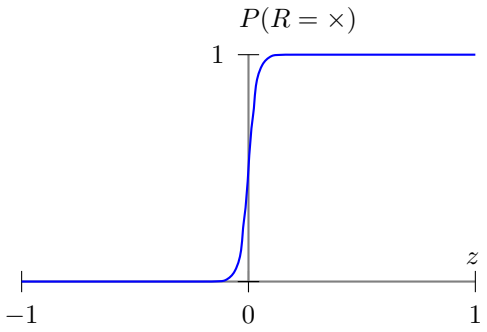
Nato je bil za vsak generiran primer izračunan skalar z kot linearna kombinacija vrednosti atributov s koeficienti iz c (enačba 3.1), verjetnost, da je primer iz razreda $\times$, pa je nato dana z enačbo 3.2.

$$z = z_1 = \sum_{i=1}^{m-m_r} c_i a_i \tag{3.1}$$

$$P(R = \times) = \frac{1}{1 + e^{-50 \cdot z}} \tag{3.2}$$

Zaradi prej omenjene normalizacije vektorja koeficientov je $z \in [-1, 1]$. Enačba 3.2 je za to zalogo vrednosti z upodobljena na grafu 3.1.

Ciljni koncept *krog* Število atributov in število redundantnih atributov sta bili sodi. Vsak par atributov je tako nastopal kot vektor v ravnini. Na enak način kot pri konceptu *črta* je bil izbran vektor koeficientov, le da je bil za polovico krajši: $\vec{c} \in [-1, 1]^{\frac{1}{2}(m-m_r)}$.



Slika 3.1: Verjetnost, da je razred primera enak ×, v odvisnosti od vrednosti z.

Za vsak primer je bil izračunan vektor w kot linearna kombinacija vektorjev (parov atributov) s koeficienti iz c (enačba 3.3).

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad w_1 = \sum_{i=1}^{\frac{1}{2}(m-m_r)} c_i a_{2i-1}, \quad w_2 = \sum_{i=1}^{\frac{1}{2}(m-m_r)} c_i a_{2i} \tag{3.3}$$

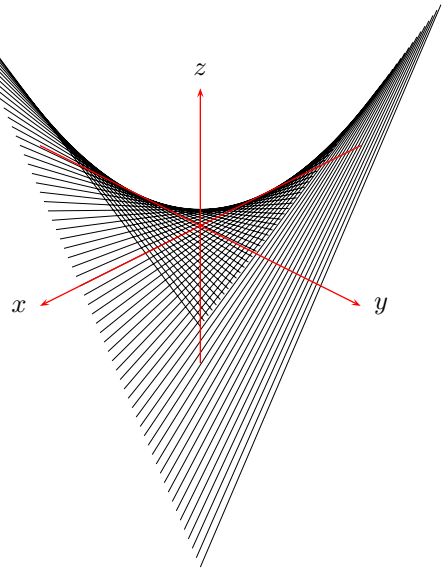
Verjetnost, da je primer iz razreda ×, je zopet dana z enačbo 3.2, pri čemer je z za $\frac{\sqrt{2}}{2}$ manjši od dolžine vektorja $\vec{w}$ (enačba 3.4).

$$z = z_2 = |\vec{w}| - \frac{\sqrt{2}}{2} \tag{3.4}$$

Zaradi normalizacije vektorja $\vec{c}$ velja $w_1 \in [-1, 1]$ in $w_2 \in [-1, 1]$, zato $z \in [0 - \frac{\sqrt{2}}{2}, \sqrt{2} - \frac{\sqrt{2}}{2}] = [-0.71, 0.71]$.

Ciljni koncept *XOR* Postopek generiranja množice je enak kot pri konceptu *krog*, le z je definiran drugače. Ima obliko take sedlaste funkcije, da je njena vrednost na koordinatnih oseh enaka 0, nato pa v 1. in 3. kvadrantu pada, v ostalih dveh pa raste. Konkretno gre za sedelno funkcijo $z = w_1^2 - w_2^2$, rotirano za 45° v smeri, nasprotni od smeri urinega kazalca (enačba 3.5, graf na sliki 3.2).

$$z = z_3 = \left(\cos\left(\frac{\pi}{4}\right) w_1 - \sin\left(\frac{\pi}{4}\right) w_2\right)^2 - \left(\sin\left(\frac{\pi}{4}\right) w_1 + \cos\left(\frac{\pi}{4}\right) w_2\right)^2 \tag{3.5}$$



Slika 3.2: Sedelna funkcija, rotirana za 45° v smeri, nasprotni od smeri urinega kazalca. Na oseh ima vrednost 0, nato pa v lihih kvadrantih pada in v sodih kvadrantih raste.

Na sliki 3.3 je v obliki razsevnega diagrama vizualiziran po en primer množice za vsak koncept.

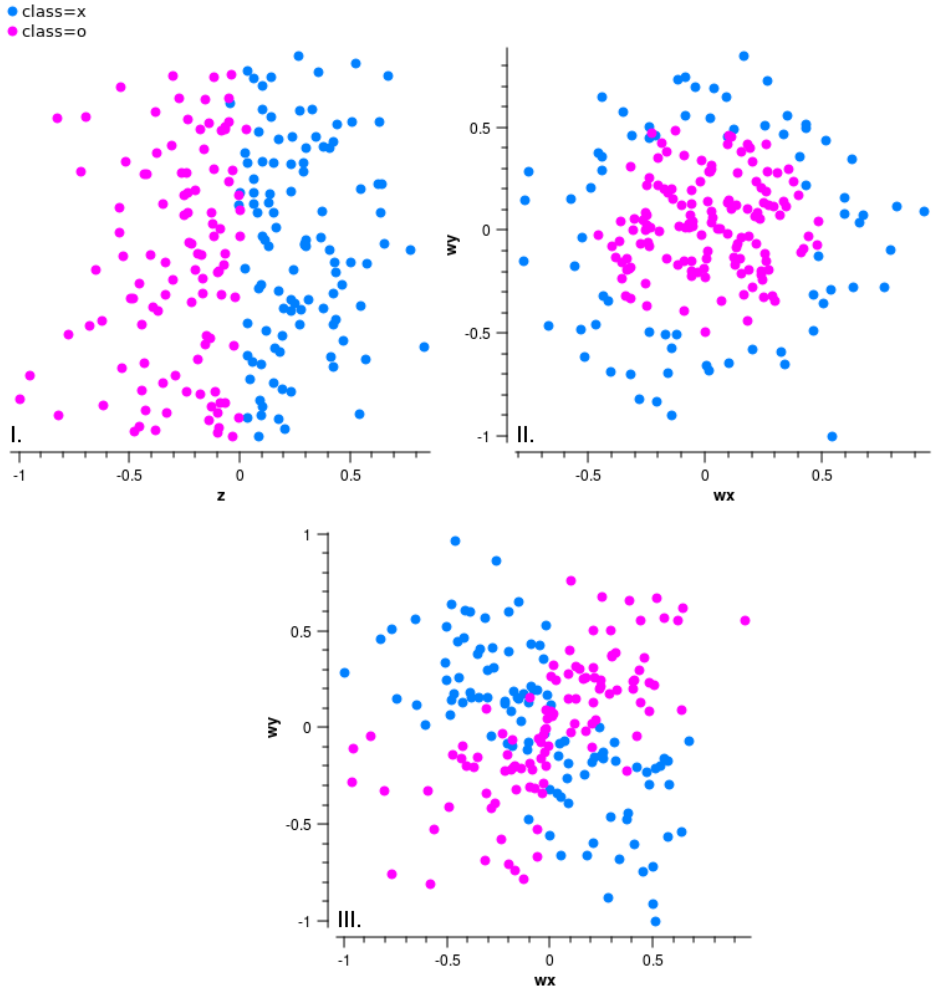
3.2.2 Določanje vrednosti atributov

Vrednosti atributov so bile za množice vrst A in B (ki ne predvidevata korelacije med atributi) izbrane neodvisno enakomerno naključno na intervalu $[-1, 1]$.

Pri množicah vrste C je bil uporabljen postopek za generiranje vektorjev, katerih korelacijo določa vnaprej izbrana kovariančna matrika. [6] Postopek predvideva:

- izbor poljubne kovariančne matrike[1] Σ,

[1]Kovariančna matrika ima tri pomembne lastnosti: na diagonali ima nenegativne elemente, saj vsebujejo varianco posameznih komponent vektorjev, je simetrična in pozitivno semi-definitna. Velja tudi obratno: vsaka matrika s temi lastnostmi je kovariančna matrika.



Slika 3.3: Primeri generiranih umetnih množic podatkov.
Pri vizualizaciji koncepta *črta* (levo zgoraj) so primeri v navpični smeri razpršeni naključno. Na vodoravni osi je skalar z, ki je pri tem konceptu definiran kot linearna kombinacija atributov.
Pri vizualizacijah preostalih dveh konceptov (*krog* in *XOR*) je vodoravna os wx oziroma w_1 in navpična os wy oziroma w_2. Vektor $\vec{w}$ s komponentama w_1 in w_2 je linearna kombinacija atributov.

- izračun razcepa Choleskega[2] kovariančne matrike (rezultat je spodnjetrikotna matrika L),
- generiranje poljubnega števila vektorjev, vsakega na sledeč način:
  - generiranje vektorja x_i z neodvisno naključno izbranimi normalno porazdeljenimi komponentami,
  - množenje vektorja z matriko L: $y_i = x_i \cdot L$.

Če imajo vektorji x_i porazdelitev $N(0, I_m)$, kjer je I_m identična matrika velikosti $m \times m$, so vektorji y_i porazdeljeni po $N(0, L \cdot I_m \cdot L^{\mathrm{T}}) = N(0, \Sigma)$.

Pri generiranju množic vrste C je potrebno generirati primere (vektorje) tako, da bo v korelaciji le točno določeno število atributov (komponent), ostali pa bodo neodvisni. Označimo število vseh atributov z m, število koreliranih atributov pa z m_k. Kovariančna matrika Σ velikosti $m \times m$ mora vsebovati $m - m_k$ takih stolpcev, ki imajo neničelen (pozitiven) element le na diagonali matrike. Zaradi simetričnosti matrike bo ta lastnost veljala tudi za vrstice.

Ker v splošnem ni lahko zagotoviti, da bo velika simetrična naključna matrika pozitivno semi-definitna, namesto kovariančne matrike Σ generiramo kar spodnjetrikotno matriko L, ki jo bomo nato uporabili v zvezi $\Sigma = L \cdot L^{\mathrm{T}}$. Če ta matrika v prvih $m - m_k$ stolpcih vsebuje neničelen element le na diagonali (primer daje enačba 3.6; $\times$ označuje neničelen element), bo to veljalo tudi za matriko Σ (enačba 3.7). Sledi formalen dokaz.

$$L = \begin{bmatrix} \times & 0 & 0 & 0 & 0 & 0 \\ 0 & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times \end{bmatrix} \tag{3.6}$$

$$\Sigma = L \cdot L^{\mathrm{T}} = \begin{bmatrix} \times & 0 & 0 & 0 & 0 & 0 \\ 0 & \times & 0 & 0 & 0 & 0 \\ 0 & 0 & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \end{bmatrix} \tag{3.7}$$

[2]Razcep Choleskega za matriko Σ obstaja če in samo če je ta simetrična pozitivno semi-definitna. Rezultat razcepa je spodnjetrikotna matrika L, da velja $\Sigma = L \cdot L^{\mathrm{T}}$, kjer $^{\mathrm{T}}$ označuje operacijo transponiranja.

Trditev. *Naj bo L spodnjetrikotna matrika velikosti $m \times m$, ki ima v prvih u stolpcih neničelne elemente le na diagonali, torej:*

$$\forall i, j : (i \neq j \wedge (i \leq u \vee j \leq u)) \Rightarrow L_{i,j} = 0$$

Potem ima tudi matrika $\Sigma = L \cdot L^{\mathrm{T}}$ v prvih u stolpcih in prvih u vrsticah neničelne elemente le na diagonali, torej:

$$\forall i, j : (i \neq j \wedge (i \leq u \vee j \leq u)) \Rightarrow \Sigma_{i,j} = 0$$

Dokaz. Poglejmo, kako se izračuna $\Sigma_{i,j}$ pri pogojih $i \neq j \wedge (i \leq u \vee j \leq u)$:

$$\Sigma_{i,j} = \sum_{k=1}^{m} L_{i,k} \cdot L_{k,j}$$

Zaradi simetrije lahko brez škode za splošnost predpostavimo $i \leq u$ in $j > i$. Potem je člen $L_{i,k}$ neničelen le pri $k = i$. Drugi člen je takrat enak $L_{i,j} = 0$. □

3.2.3 Rezultati

Rezultati so predstavljeni z grafi, ki imajo na oseh spreminjajoča se parametra (število atributov in še en parameter, odvisno od vrste množice), vrednost v določeni točki pa je ponazorjena s svetlostjo točke. Angleški izraz za take grafe je *heat map*.

Na grafih prikazana vrednost je ploščina pod ROC krivuljo (angl. *Area Under Curve, AUC*), ki jo je FreeVizov klasifikator dosegel pri 10-kratnem prečnem preverjanju nad umetno množico. En graf prikazuje podatke za eno vrsto in koncept množice, parametre (npr. število atributov) posamezne množice pa določa položaj točke na grafu. Tako je z grafov možno razbrati občutljivost kvalitete FreeViza na spreminjajoče se parametre. Za primerjavo je ob vsakem grafu narisan še ekvivalenten graf za klasifikator SVM.

Da se izniči vpliv izbora podmnožic za prečno preverjanje, je bil vsak eksperiment ponovljen 35-krat, podani rezultati pa so mediane izmerjenih vrednosti. Tako visoko število ponovitev je bilo izbrano na podlagi predpostavke, da se FreeVizova kvaliteta v odvisnosti od parametrov spreminja *gladko*, pri nižjem številu ponovitev pa je bilo v rezultatih opaziti veliko nihanj.

Črna točka na grafu predstavlja vrednost 1.0 (največjo možno vrednost AUC), bela pa vrednost manjšo ali enako 0.5. Vmesne vrednosti so predstavljene z ustreznimi odtenki sive.

Levi spodnji kvadratek na grafih za podatke vrste A ustreza točki (60, 10), kar pomeni da je imela množica 60 učnih primerov z 10 atributi. Na grafih

za podatke vrst B in C spodnji levi kvadratek ustreza točki (10, 10). Premik za en kvadratek pomeni spremembo parametra (npr. števila atributov) za 50. Zaradi računske zahtevnosti poskusov jih žal ni bilo možno izvesti z manjšim korakom.

Vrsta množic A

Grafi na slikah 3.4, 3.5 in 3.6 prikazujejo ploščino pod ROC krivuljo (AUC) za FreeViz in SVM pri množicah vrste A (brez redundantnih atributov, brez korelacij med atributi) z različnima številoma atributov in učnih primerov.

Na grafih je pri fiksnem številu atributov opaziti rahlo poslabšanje kvalitete FreeViza pri zmanjševanju števila učnih primerov. Glede na to, da je ta trend pri SVM še bolj izrazit (razen pri konceptu *krog*, ki se ga ni uspel naučiti), najbrž kakovostnega modela pri prenizkem številu primerov sploh ni možno sestaviti, FreeViz pa je delo opravil dobro. Ob tem moramo opozoriti, da smo SVM uporabljali s privzetimi nastavitvami. Korektnejša primerjava bi zahtevala iskanje optimalnih nastavitev, kar pa bi zahtevalo več časa, kot smo ga imeli na razpolago.

Vrsta množic B

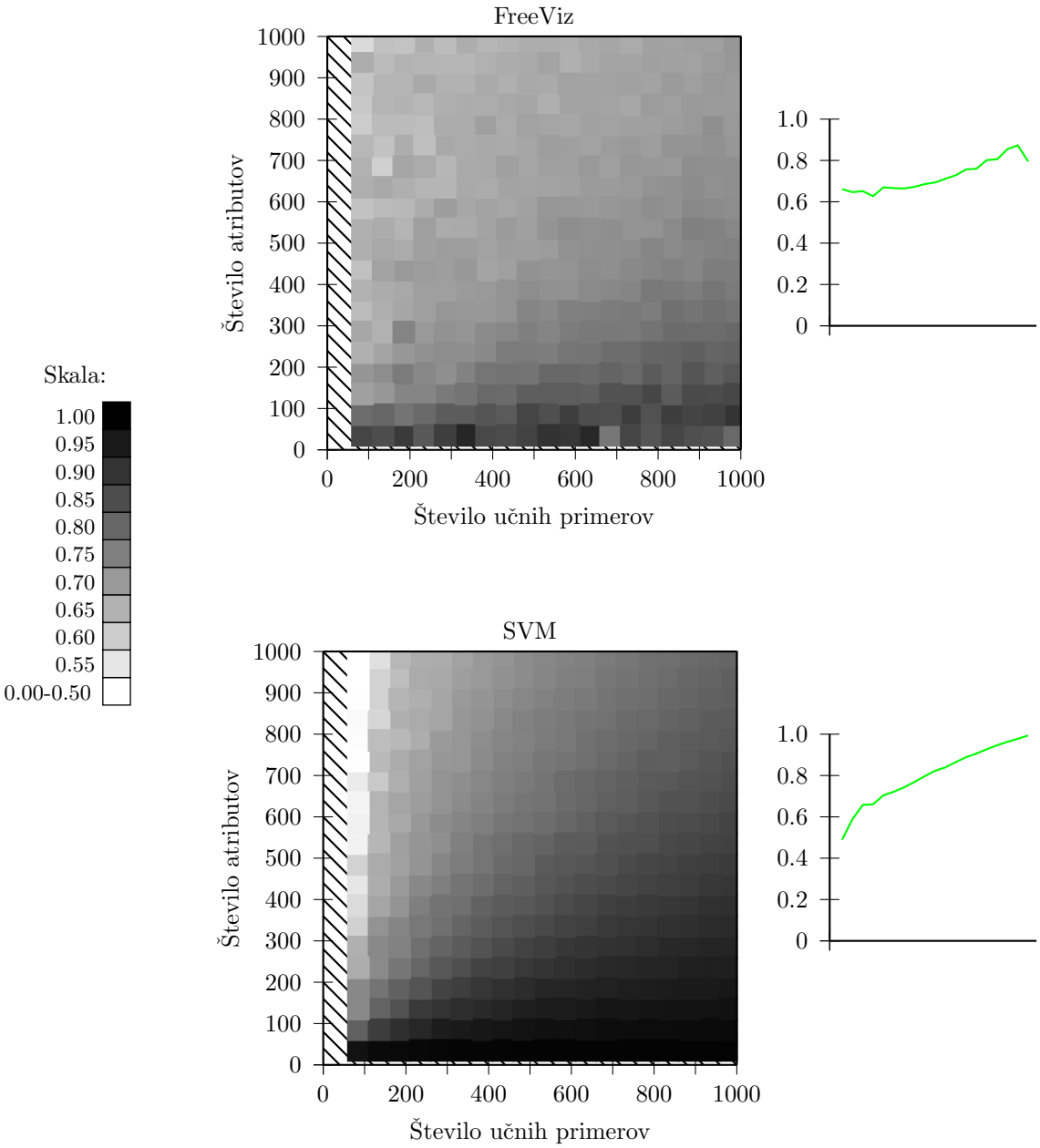
Grafi na slikah 3.7, 3.8 in 3.9 prikazujejo ploščino pod ROC krivuljo (AUC) za FreeViz in SVM pri množicah vrste B (z redundantnimi atributi, brez korelacij med atributi) z različnima številoma vseh in pomembnih (neredundantnih) atributov.

Pri podatkih za težje koncepte (*krog* in *XOR*) spremembe AUC zaradi nizke vrednosti na celotnem območju niso izrazite. Pri konceptu *črta* sta izraziti značilnosti le slaba kvaliteta FreeViza pri zelo nizkem številu pomembnih atributov (10) in splošno slabšanje kvalitete obeh klasifikatorjev z večanjem števila vseh atributov. Prva značilnost bi lahko povzročala težave pri genetskih podatkih, saj imamo v množici popisanih genov pogosto le nekaj takih, ki vplivajo na iskano lastnost oseb.

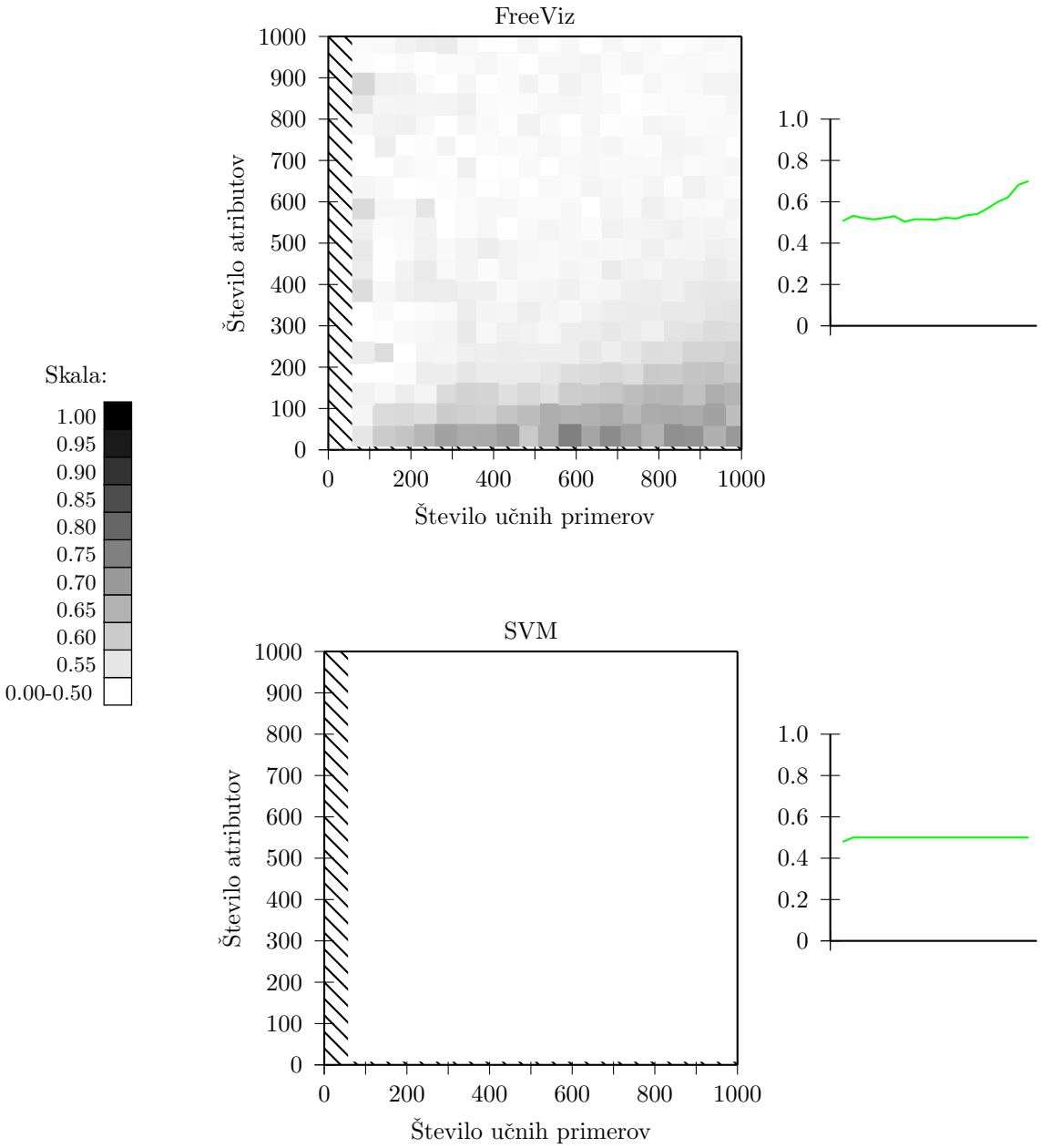
Vrsta množic C

Grafi na slikah 3.10, 3.11 in 3.12 prikazujejo ploščino pod ROC krivuljo (AUC) za FreeViz in SVM pri množicah vrste C (brez redundantnih atributov, s korelacijami med nekaterimi atributi) z različnima številoma vseh in koreliranih atributov.

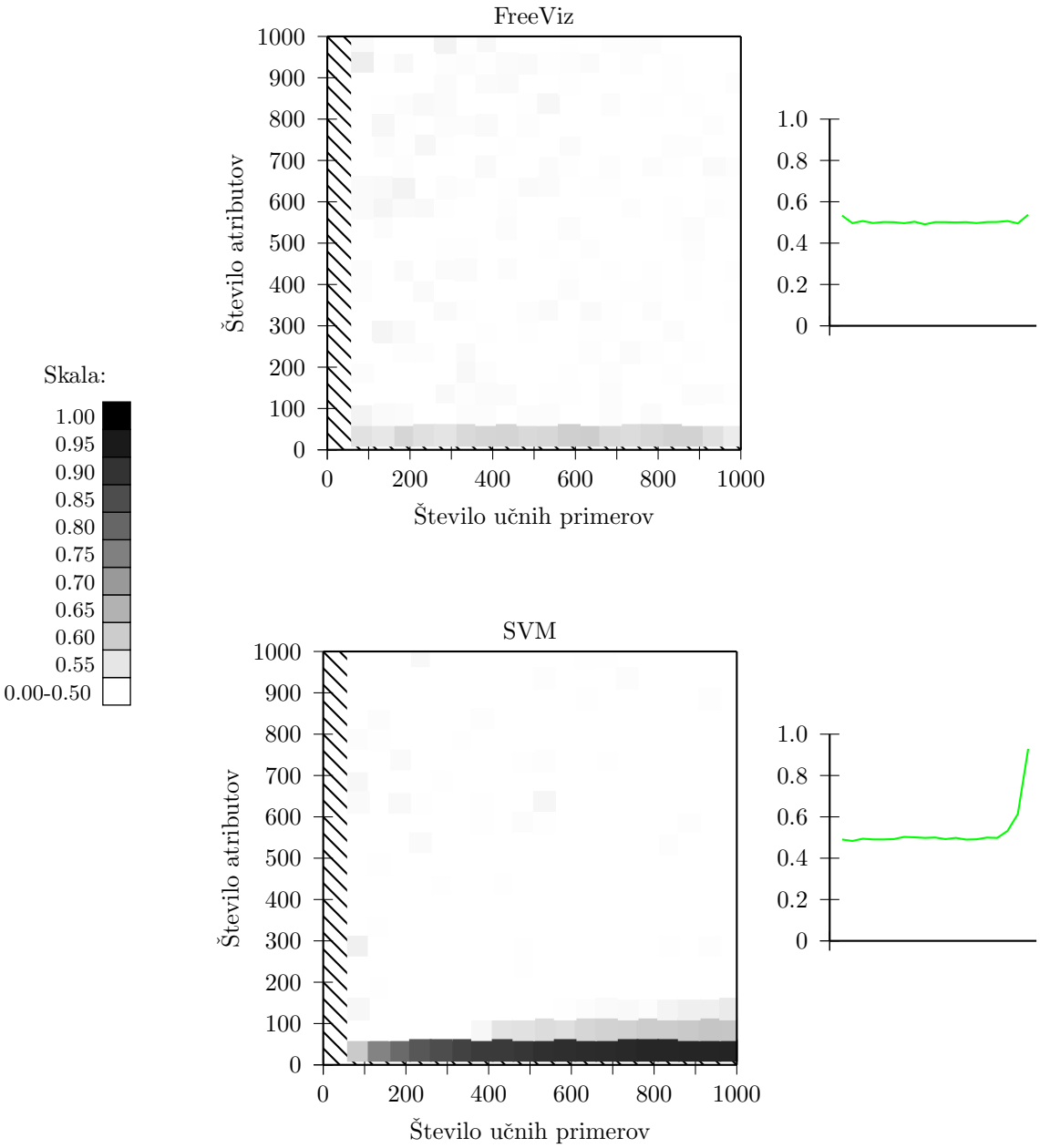
Korelacije med atributi učenje olajšajo; to je razvidno iz:



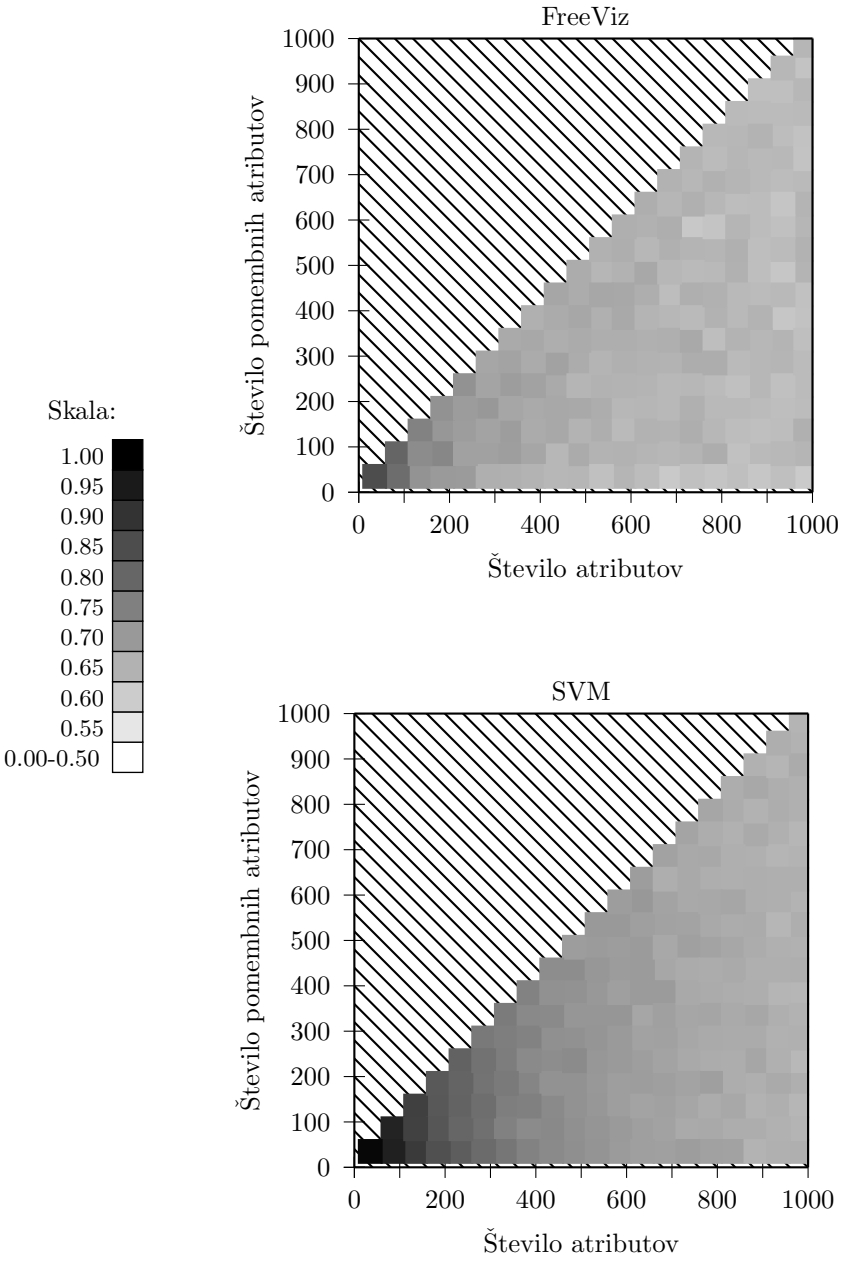
Slika 3.4: AUC za FreeViz in SVM pri umetnih podatkih **vrste A** (brez redundantnih atributov) za koncept ***črta***. Grafa na desni prikazujeta vrednosti AUC po drugi diagonali (od 60 primerov z 910 atributi do 960 primerov z 10 atributi).



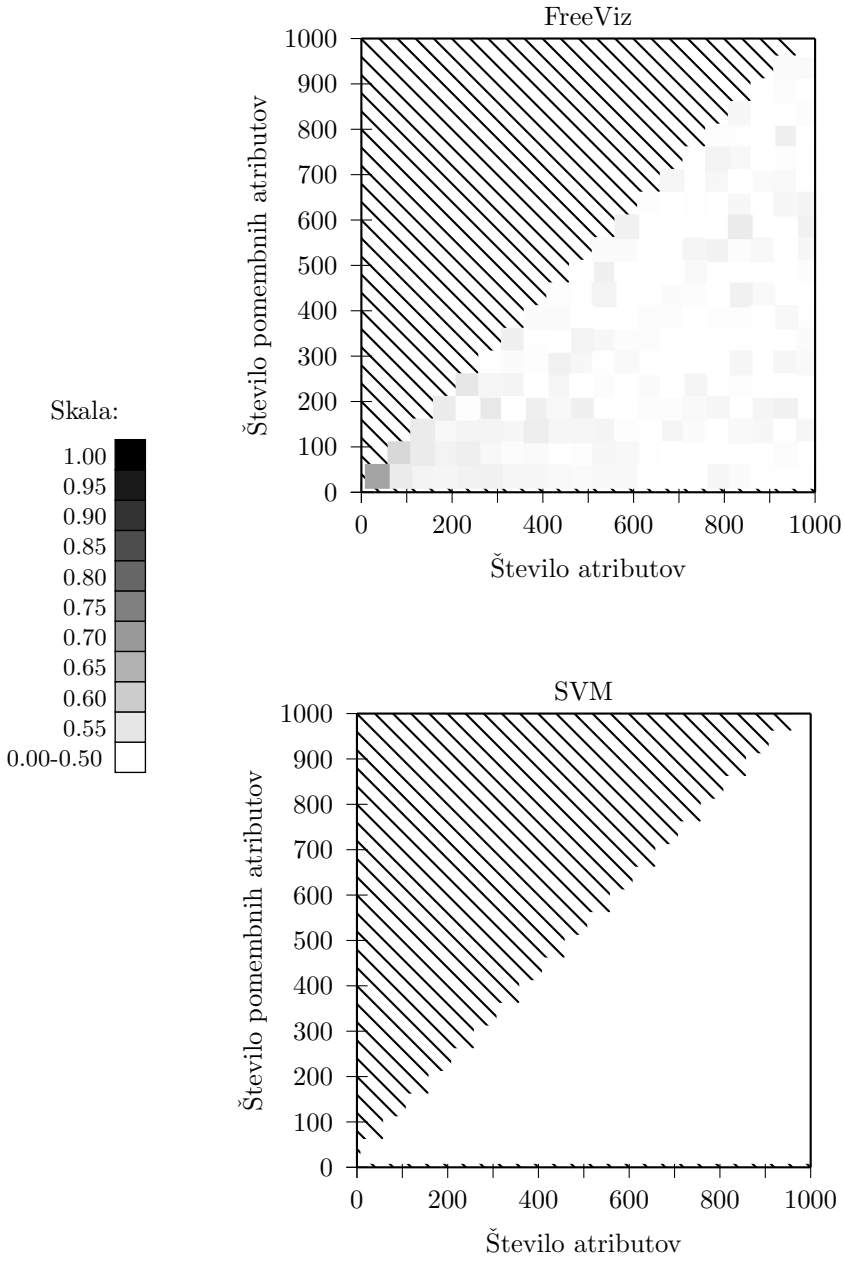
Slika 3.5: AUC za FreeViz in SVM pri umetnih podatkih **vrste A** (brez redundantnih atributov) za koncept ***krog***. Grafa na desni prikazujeta vrednosti AUC po drugi diagonali (od 60 primerov z 910 atributi do 960 primerov z 10 atributi).



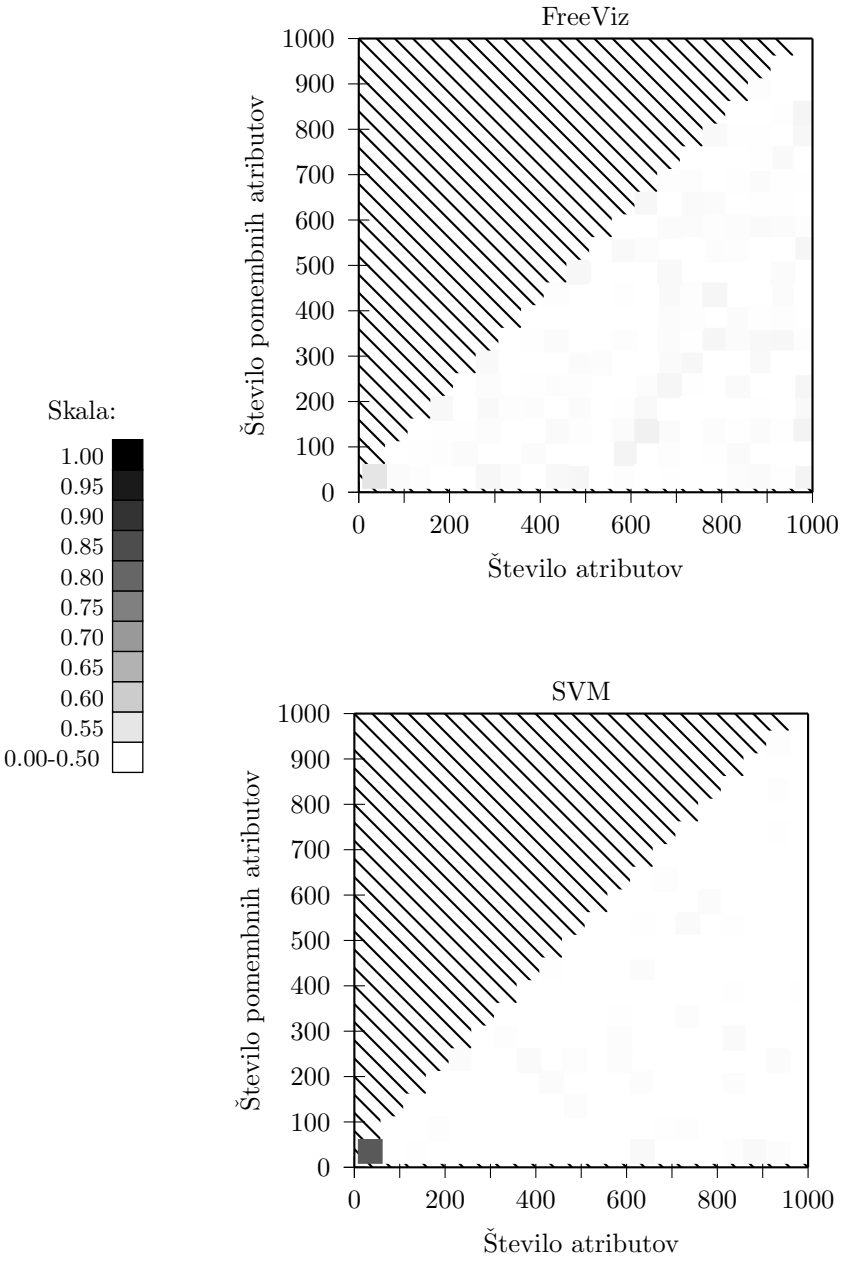
Slika 3.6: AUC za FreeViz in SVM pri umetnih podatkih **vrste A** (brez redundantnih atributov) za koncept ***XOR***. Grafa na desni prikazujeta vrednosti AUC po drugi diagonali (od 60 primerov z 910 atributi do 960 primerov z 10 atributi).



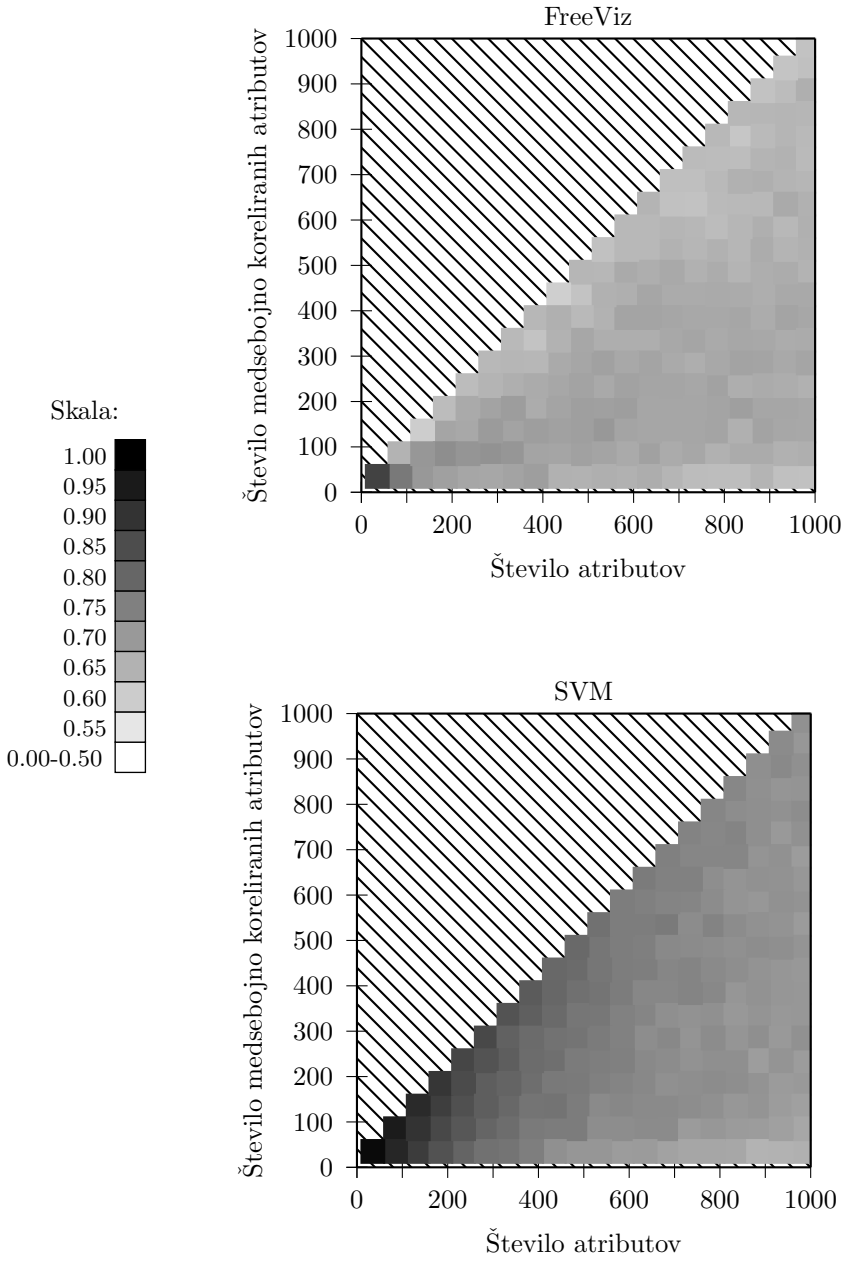
Slika 3.7: AUC za FreeViz in SVM pri umetnih podatkih **vrste B** (z redundantnimi atributi) za koncept ***črta***.



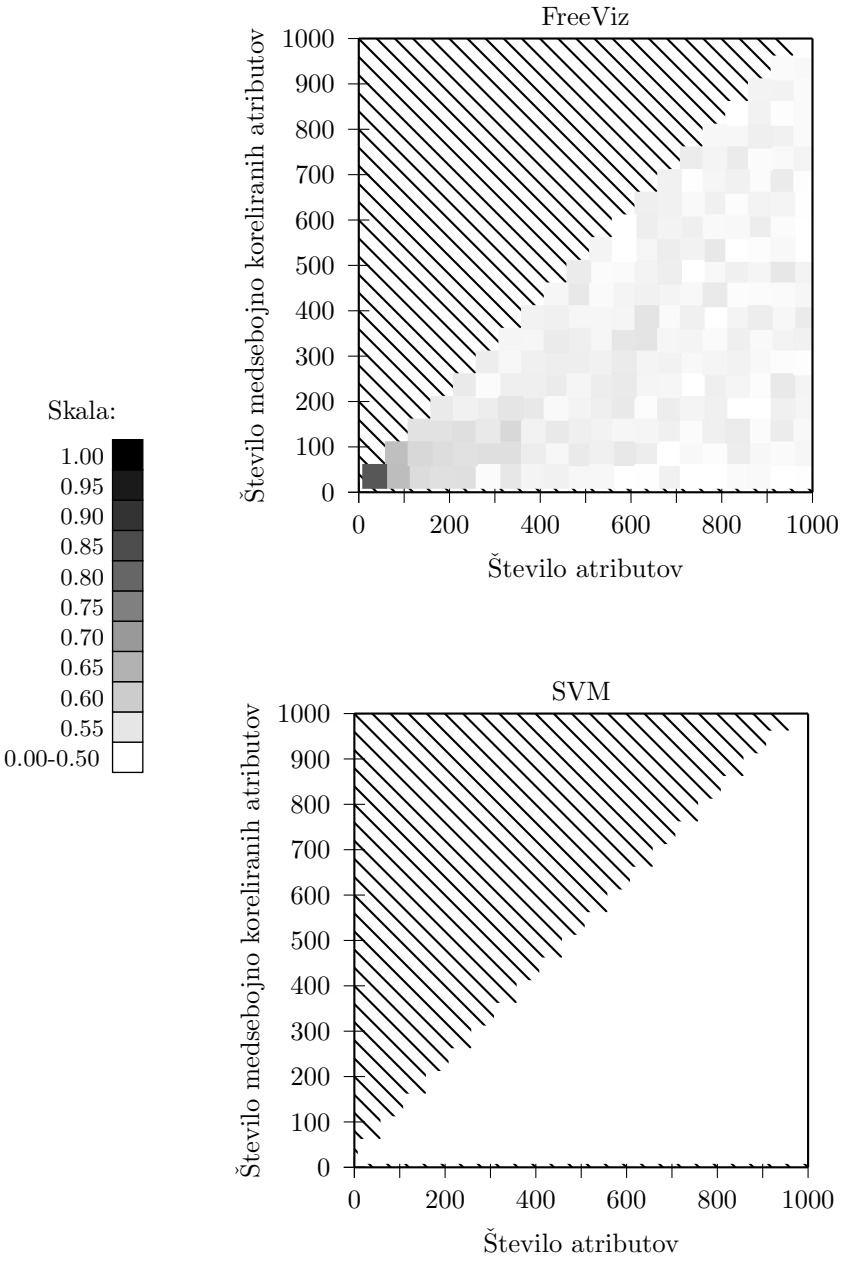
Slika 3.8: AUC za FreeViz in SVM pri umetnih podatkih **vrste B** (z redundantnimi atributi) za koncept ***krog***.



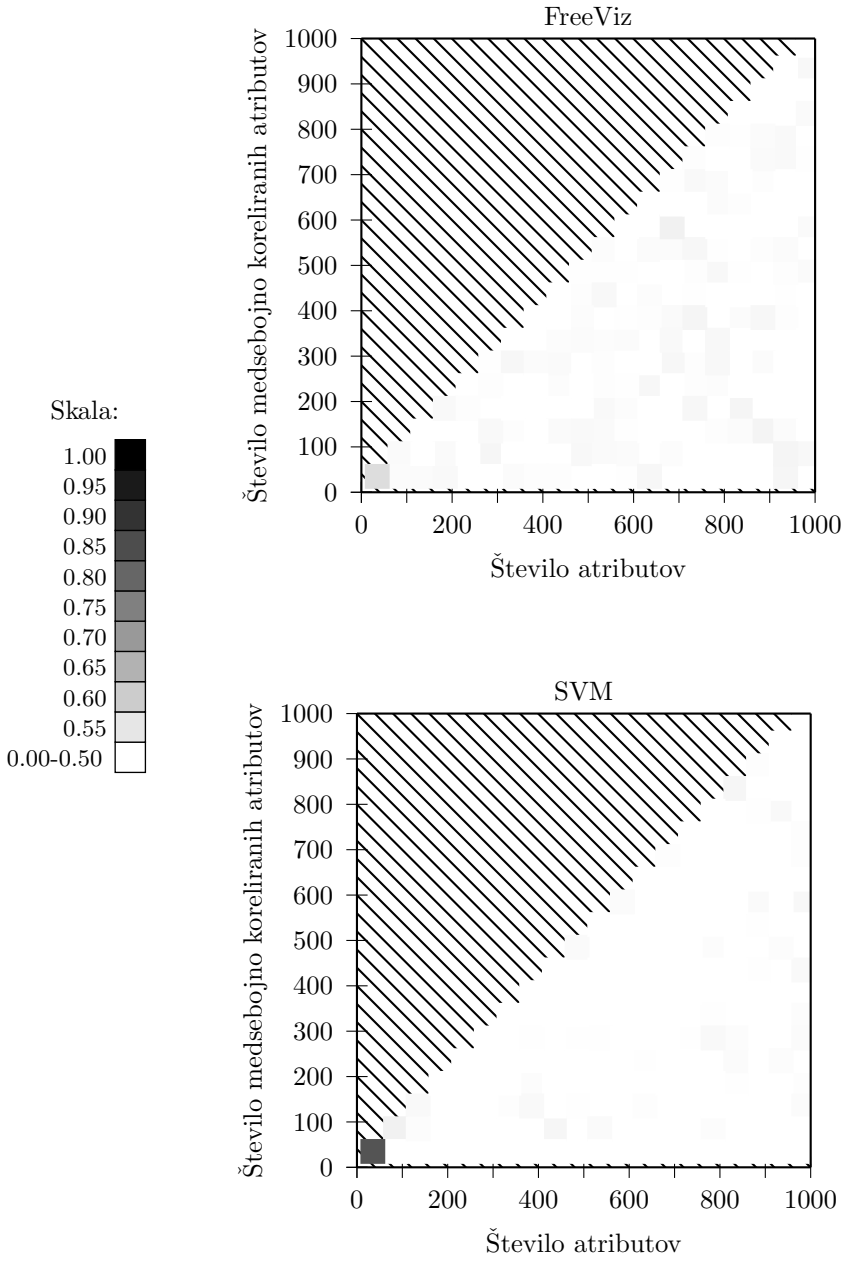
Slika 3.9: AUC za FreeViz in SVM pri umetnih podatkih **vrste B** (z redundantnimi atributi) za koncept ***XOR***.



Slika 3.10: AUC za FreeViz in SVM pri umetnih podatkih **vrste C** (s koreliranimi atributi) za koncept ***črta***.



Slika 3.11: AUC za FreeViz in SVM pri umetnih podatkih **vrste C** (s koreliranimi atributi) za koncept ***krog***.



Slika 3.12: AUC za FreeViz in SVM pri umetnih podatkih **vrste C** (s koreliranimi atributi) za koncept ***XOR***.

- splošnega povečevanja vrednosti AUC pri obeh klasifikatorjih, če pri fiksnem številu atributov povečujemo število koreliranih atributov,
- pri težjih konceptih (*krog* in *XOR*) je pri nizkem številu atributov (10) korelacija zvišala kvaliteto obeh klasifikatorjev; primerjava je možna z grafi za množice vrste A pri 10 (nekoreliranih) atributih in 200 učnih primerih.

Rezultati za koncept *črta* (in nekoliko manj izrazito tudi pri konceptu *krog*) kažejo, da FreeVizova kvaliteta pade, če je medsebojno koreliranih atributov zelo veliko (tj. enako ali skoraj enako številu vseh atributov).

3.3 Projekcije redundantnih atributov med in po optimizaciji

Pri tem eksperimentu smo preverjali hipotezo, da je možno meriti pomembnost (*kvaliteto*) atributa glede na:

- gibanje projekcije atributa med koraki gradientne FreeViz optimizacije,
- oddaljenost projekcije atributa od središča (koordinatnega izhodišča) po končani optimizaciji.

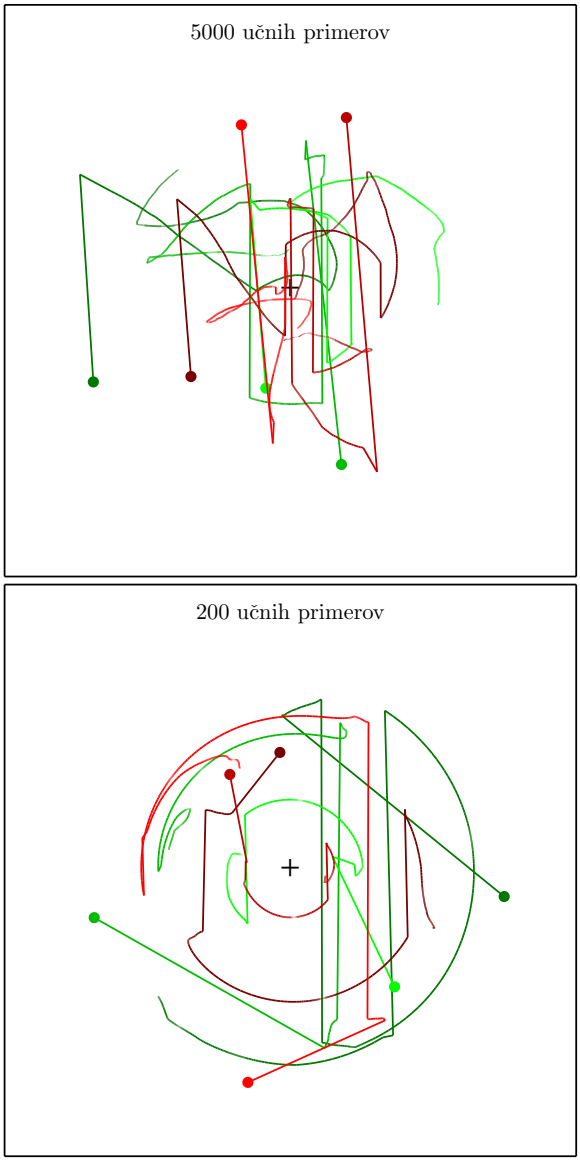
Najprej smo generirali množici s 300 atributi, od tega 50 pomembnimi. Ena množica je vsebovala 200, druga pa 5000 učnih primerov. Na obeh smo pognali FreeViz optimizacijo linearne projekcije in za nekaj naključnih pomembnih in redundantnih atributov narisal poti, ki so jih njihove projekcije opravile med optimizacijo. Rezultat je prikazan na sliki 3.13.

Medtem ko se pri množici s 5000 učnimi primeri projekcije redundantnih atributov gibljejo mnogo bolj v okolici koordinatnega izhodišča (torej območja manjšega vpliva na projekcijo primerov) kot projekcije pomembnih atributov, te razlike pri množici z malo (200) učnimi primeri ni. To lepo kaže na FreeVizovo prekomerno prileganje šumu v podatkih.

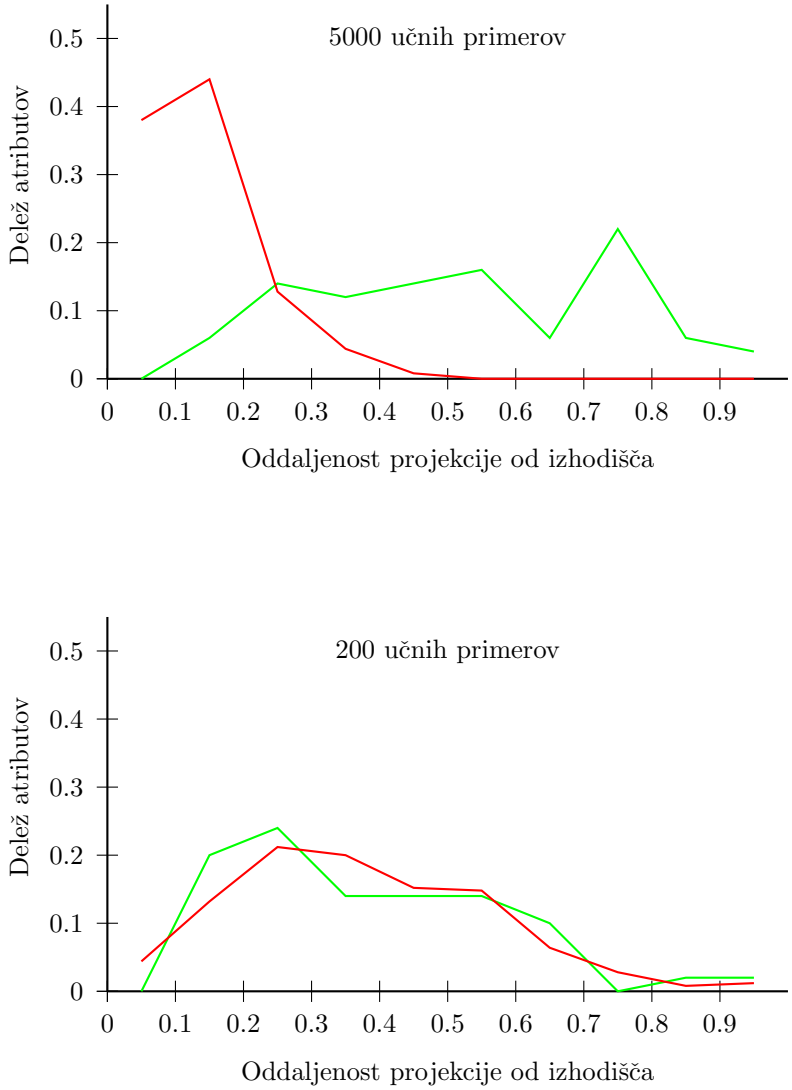
Še bolj nazorno ta pojav prikazujeta histograma na sliki 3.14. Prikazujeta delež redundantnih (rdeče) in delež pomembnih (zeleno) atributov, katerih projekcije se po končani optimizaciji nahajajo na posamezni oddaljenosti od koordinatnega izhodišča.

Pri množici z veliko učnimi primeri je bila večina redundantnih atributov projicirana blizu koordinatnega izhodišča. Pri majhni učni množici sta histograma za pomembne in redundantne atribute skoraj enaka. FreeViz torej s

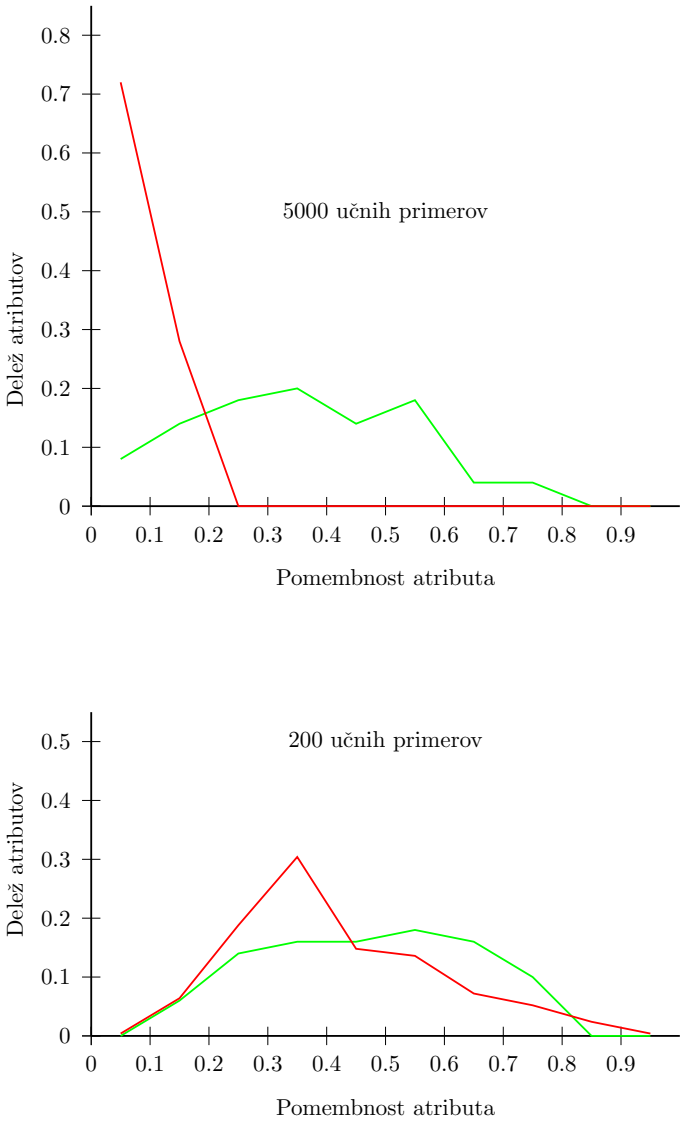
premalo učnimi primeri ni uspel ločiti pomembnih atributov od redundantnih, a enako velja za naivni Bayesov klasifikator: slika 3.15 prikazuje histograma deležev atributov glede na njihovo *pomembnost* (tj. absolutno razliko med največjo in najmanjšo verjetnostjo enega od razredov preko vseh vrednosti ocenjevanega atributa).



Slika 3.13: Primeri poti projekcij atributov med optimizacijo. Zgornja slika ustreza množici s 5000, spodnja pa množici z 200 učnimi primeri. Zelene črte (3) predstavljajo poti pomembnih (neredundantnih) atributov, rdeče (3) pa poti redundantnih atributov. Začetni položaji projekcij (pred optimizacijo) so bili nastavljeni naključno in so označeni s pikami na začetku črt. Koordinatno izhodišče je označeno s črnim križcem.



Slika 3.14: Histograma deležev redundantnih (rdeče) in pomembnih (zeleno) atributov po oddaljenosti projekcije od koordinatnega izhodišča po končani FreeViz optimizaciji. Zgornji histogram ustreza množici s 5000, spodnji pa množici z 200 učnimi primeri.



Slika 3.15: Histograma deležev redundantnih (rdeče) in pomembnih (zeleno) atributov po pomembnosti za naivni Bayesov klasifikator. Zgornji histogram ustreza množici s 5000, spodnji pa množici z 200 učnimi primeri. Pomembnost atributa je definirana kot razlika med največjo in najmanjšo verjetnostjo enega od razredov, če je znana le vrednost tega atributa. Ker so atributi zvezni, naivni Bayes uporablja metodo LOESS za določitev verjetnosti.

Poglavje 4

Diskusija in zaključek

FreeViz je algoritem za optimizacijo dvodimenzionalne projekcije večdimenzionalnih podatkov. Namen projiciranja je izris podatkov v razsevnem diagramu. Projekcijo pa je mogoče v kombinaciji z metodo najbližjih sosedov uporabljati tudi za klasifikacijo novih primerov. Takšen klasifikator je mogoče uporabljati tudi za ocenjevanje kvalitete projekcije. Ker je FreeViz tehnika lokalne optimizacije, lahko vrne lokalno optimalno a nesmiselno projekcijo; po pravilnosti klasifikacije lahko ločimo med smiselnimi in nesmiselnimi projekcijami. [7]

Lastnosti FreeViza kot algoritma za učenje klasifikacijskih modelov še niso dobro raziskane. Za preprostejše domene, v katerih je število primerov bistveno manjše od števila atributov, je Demšar že izvedel poskuse, ki so pokazali, da je njegova kvaliteta primerljiva s kvaliteto drugih algoritmov učenja[1]. V tem delu smo se osredotočili na njegovo delovanje pri podatkovnih zbirkah iz genetike ter zbirkah, kjer razredi niso linearno ločljivi, prisotni pa so tudi redundantni in korelirani atributi. Tudi pri izboru dimenzionalnosti za sintetične množice smo se zgledovali po tipičnih razsežnostih podatkovnih zbirk iz genetike, konkretno, podatkov dobljenih z mikročipi.

Že v okviru seminarske naloge pri predmetu *umetna inteligenca 2* smo želeli izboljšati delovanje FreeViza v primeru, ko je atributov bistveno več kot učnih primerov. Takrat se je izkazalo, da FreeViz že brez prilagoditev zadovoljivo deluje. V tem delu to trditev negiramo: delovanje FreeViza je bilo po naključju zadovoljivo ravno na učni množici, ki smo jo pri seminarski nalogi uporabljali.

Poskusili smo raziskati, v čem se ta množica razlikuje od ostalih. Najbolj očitna razlika je bila delitev primerov v dva razreda, medtem ko so ostale množice razlikovale več razredov. Poskus z zmanjševanjem števila razredov množici z več razredi je pokazal, da število razredov ni tisti dejavnik, ki pov-

[1] osebna komunikacija

zroči zmanjšanje uspešnosti FreeViza.

Za temeljitejšo raziskavo obnašanja FreeViza nad podatki z več atributi kot primeri smo sistematično generirali večje število sintetičnih množic podatkov. Rezultati niso posebej obetavni. FreeViz je uspešno reševal preprost (linearno ločljiv) koncept, še posebej pri nizkem številu atributov. Nizko število učnih primerov mu je povzročalo celo manj težav kot klasifikatorju SVM. V nasprotju s slednjim žal FreeViz ni deloval bolje niti pri bistvenem povečanju števila učnih primerov.

Na sintetičnih množicah podatkov se je izkazalo tudi, da je FreeViz uspešnejši, če so atributi medsebojno korelirani, in da dosega slabo kvaliteto, če na razred vpliva le manjše število atributov. Slednja lastnost bi lahko bila dejavnik, ki povzroča slabo delovanje na podatkih s področja genetike.

Ker rezultati niso pojasnili dejstva, da FreeViz uspešno deluje v domeni *DLBCL*, smo predpostavili, da se odgovor morda skriva v uporabljenem postopku lokalne optimizacije, v katerem se pri množicah, na katerih je FreeViz uspešen, projekcije relevantnih atributov umestijo na ustrezna mesta, projekcije redundantnih pa se zaradi sprotne normalizacije projekcije počasi bližajo koordinatnemu izhodišču. Na sintetični množici podatkov je lažje določiti pomembnost atributov, zato smo poti projekcij atributov med optimizacijo projekcije spremljali na preprosti generirani domeni, na kateri je FreeViz dosegal podobno uspešnost kot na *DLBCL*. Predpostavka se je izkazala za napačno: poti projekcij pomembnih in redundantnih atributov se med optimizacijo ne da ločiti, redundantni atributi pa so projicirani bližje koordinatnega izhodišča šele pri mnogo večjem številu učnih primerov. Pomembnosti atributov na majhni učni množici ni uspel določiti niti naivni Bayes, zato sklepamo, da tako delovanje ni indikator slabosti FreeViza.

V diplomski nalogi nismo dokončno odgovorili na nobeno od zastavljenih vprašanj. Potrdili smo, da FreeViz občasno deluje tudi, ko nam intuicija pravi, da ne more, npr. v domeni *DLBCL*. Kateri dejavniki v tem konkretnem primeru in v splošnem vplivajo na uspešnost, nismo uspeli ugotoviti.

Kljub temu v diplomo vloženo delo ni bilo zaman. Opravljeni poskusi so bili izvedeni korektno. Čeprav nam ne dajejo dokončnih odgovorov, predstavljajo prvi korak pri raziskovanju vedenja FreeViza v zanimivem in aktualnem kontekstu podatkov s področja genetike.

Slike

Tabele

Literatura

[1] J. Demšar, G. Leban in B. Zupan, "Freeviz - an intelligent multivariate visualization approach to explorative analysis of biomedical data," *Journal of Biomedical Informatics*, zv. 40, št. 6, str. 661–671, 2007.

[2] R. Forsyth, "UCI machine learning repository," 1990. [Na spletu]. Dosegljivo na: http://www.ics.uci.edu/~mlearn/MLRepository.html

[3] M. Shipp *et al.*, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nat Med*, zv. 8, št. 1, str. 68–74, 2001.

[4] J. Khan *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat Med*, zv. 7, št. 6, str. 673–679, 2001.

[5] A. Bhattacharjee *et al.*, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc Natl Acad Sci USA*, zv. 98, št. 24, str. 13 790–13 795, 2001.

[6] P. D. Scott in E. Wilkins, "Evaluating data mining procedures: techniques for generating artificial data sets," *Information & Software Technology*, zv. 41, št. 9, str. 579–587, 1999.

[7] G. Leban, I. Bratko, U. Petrovič, T. Curk in B. Zupan, "Vizrank: finding informative data projections in functional genomics by machine learning," *Bioinformatics*, zv. 21, št. 3, str. 413–414, 2005.