Strojno učenje

Selekcija (probir) varijabli
(en. Feature selection)

Tomislav Šmuc

Literatura

- Guyon & Elisseeff “An Introduction to Variable and Feature Selection”, J Mach Learn Res, 3 (2003), 1157-1182
- Feature Extraction, Foundations and Applications, I. Guyon et al, Eds. Springer, 2006.
- WEKA – Preprocessing algorithms

Svrha

- Smanjenje količine podataka (kod problema sa vrlo velikim brojem primjera i brojem varijabli/atributa: 1000-100000+)
- Smanjiti broj varijabli
- Odabrati samo najvažnije varijable/atribute
- Prednosti:
    - Bolji modeli (poboljšanje točnosti)
    - Brže - do modela; brži modeli (u eksploataciji)
    - Interpretabilniji modeli

	a_1	a_2	...	a_i	a_j	a_{n-1}	a_n
P_1							
P_2							
...							
P_m							

Važne/relevantne varijable

Svrha - realni problemi - primjeri

Klasifikacija teksta

– Varijable ~ 10^5 riječi, parovi riječi (?)
– Tipična praksa: koristi sve riječi - prepusti metodama odabira varijabli da riješe nekorisni višak varijabli
– Treniranje sa svim varijablama (riječima) je preskupo
– Prisutnost irelevantnih varijabli može negativno utjecati na generalizaciju

Klasifikacija tumora na osnovu ekspresije gena [Xing, Jordan, Karp ’01]

– 72 pacijenta (primjeri)
– 7130 varijabli (nivoi ekspresije različitih gena)

Dijagnoza bolesti

– Varijable su rezultat (skupih) laboratorijskih testova
– koji bi test trebali napraviti na pacijentu?

Ugrađeni sustavi (Embedded systems) sa limitiranim resursima

– klasifikator mora biti kompaktan:
  - npr. prepoznavanje glasa s mobitela
– Predikcija na CPU (4KB ograničenje!)

Neki osnovni pojmovi

Individualna irelevantnost varijable $\boldsymbol{X_i}$ (V^{-i}- podskup bez varijable X_i)

$$P(X_i, Y|\boldsymbol{V}^{-i}) = P(X_i|\boldsymbol{V}^{-i})P(Y|\boldsymbol{V}^{-i}).$$

Dostatni podskup varijabli ***V*** *(u odnosu na kompletan skup* ***X****)*

$$P(Y|\boldsymbol{V}) = P(Y|\boldsymbol{X}).$$

Gornji izrazi su u praksi nerealni (znak jednakosti)

- u stvarnosti radi se obično o približno zadovoljenim tvrdnjama (razlika < ε), t.j.:
  - Vjerojatno približno irelevantnim varijablama
  - Minimalnom približno dostatnom podskupu varijabli ***V***

Odabir metode

Odabir metode selekcije varijabli ovisi:

- skupu varijabli i ciljnoj varijabli (binarne, kategoričke, kontinuirane)
- samom problemu (kakve su zavisnosti između varijabli/ciljne varijable, linearne/ne-linearne)
- Količini dostupnih podataka (odnosu broja primjera naspram broja varijabli, točnosti podataka - šum)

Osnovna podjela

- **Pojedinačno rangiranje**
  –> nezavisna ocjena svake varijable pojedinačno

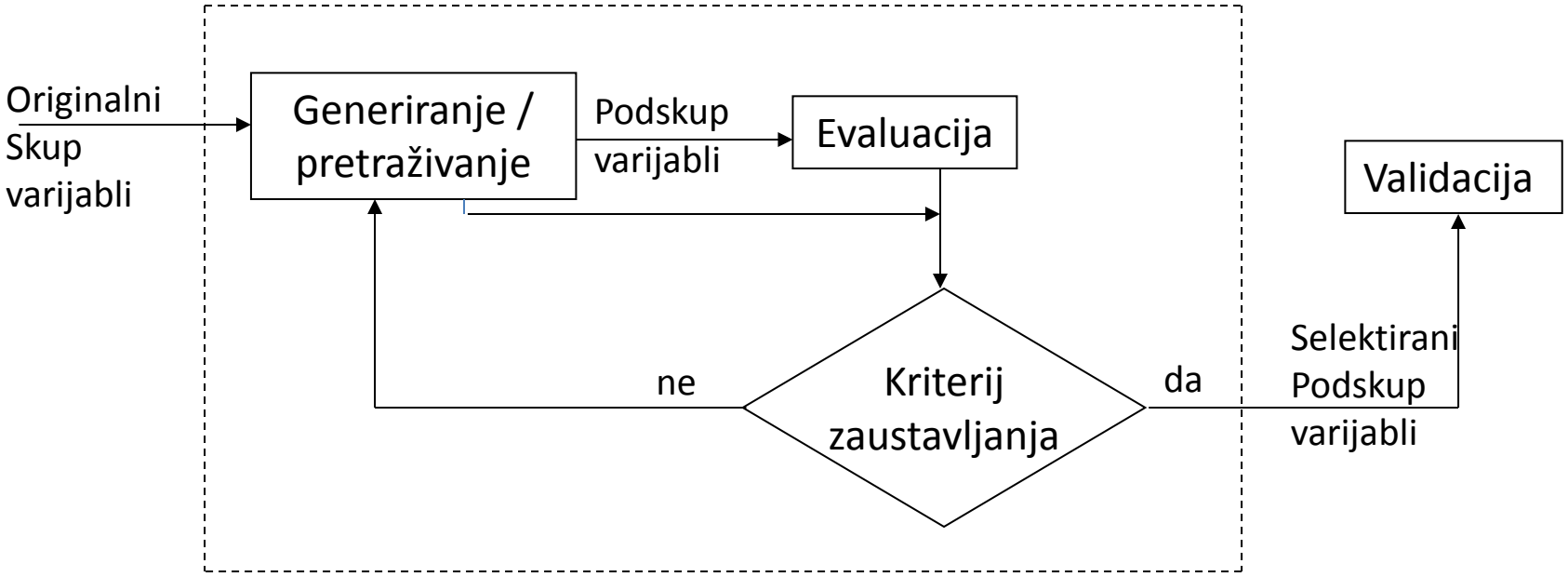
- **Grupno (multivarijantno) rangiranje**
  –> uzima u obzir istovremeno skup varijabli

- **Filter metode**
  –> rangiranje varijabli ili skupa varijabli na bazi indeksa(relevantnosti), nezavisno od algoritma za učenje(klasifikatora)

- **Metode "omotača" (wrapper) metode**
  -> koristi se klasifikator da bi se odredila vrijednost varijabli ili skupa varijabli

- **Ugrađene metode (embedded) ili algoritmi**
  -> istovremeno zajednički se uči i model i selekcija varijabli
  (primjer stabla odlučivanja !)

Generalna shema procesa odabira varijabli



Generiranje/pretraživanje - odabir podskupa (varijable)
Evaluacija - izračunati relevantnost podskupa varijabli.
Kriterij zaustavljanja - odrediti da li je podskup relevantan.
Validacija - nezav. verificiranje odabranog podskupa

Generiranje/pretraživanje:

- odabir podskupa ili varijable za evaluaciju
  - Početak: prazan skup, sve varijable, sluč. gen. podskup.
  - Inkrement : dodavanje, uklanjanje, dod/ukl varijabli
- kategorizacija načina generiranja/pretraživanja

  - Iscrpne (exhaustive, complete)
  - Heurističke
  - Slučajne

Proces: Generiranje

Iscrpno (complete/exhaustive)

- Ispitivanje svih kombinacija podskupova varijabli
  - {a1,a2,a3} => { {a1},{a2},{a3},{a1,a2},{a1,a3},{a2,a3},{a1,a2,a3} }
- Prostor pretraživanja - $O(2^m)$, m - # varijabli
- Optimalni podskup je dohvatljiv
- Nedopustiva složenost za m>>

Heuristički pristup

- Selekcija po određenom principu
  - izbacivanje varijabli
  - kandidati = { {f1,f2,f3}, {f2,f3}, {f3} }
- inkrementalno generiranje podskupova
- Prostor pretraživanja je drastično manji
- Neki od relevantnih podskupova varijabli mogu biti preskočeni !

Slučajno generiranje

- Slučajni odabir varijable (Probabilistički pristup)
- Optimalni podskup zavisi od broja pokušaja (~ ovisi o resursima)

Proces: Evaluacija

Evaluacija

- Odredi važnost generiranog podskupa varijabli za klasifikacijski problem

  R_v = J(podskupa varijabli)

  if (R_v > best_R_v) => best_R_v = R_v

- Osnovni tipovi evaluacijskih funkcija.
  - udaljenost

    (euklidska udaljenost, Manh. Udaljenost, sl.)
  - informacijske mjere

    (entropija, Infogain - porast informacije, sl.)
  - zavisnost između varijable i ciljne varijable

    (Pearsonov korelacijski koeficijent)
  - konzistentnost podskupa

    (minimalni konzistentan broj varijabli)

  } Filter metode

  - pogreška klasifikatora

  } "wrapper"

Proces: Evaluacija

Udaljenost

- $z^2 = x^2 + y^2$
- Selektirati one varijable koje podupiru “bliskost” primjera iste klase
- Primjeri iste klase morali bi biti međusobno bliži u smislu udaljenosti nego primjeri različitih klasa

Informacijske mjere

- Entropija – mjera sadržaja informacije
- Info-gain varijable : (kao kod stabla odlučivanja)
  IG(V) = I(p,n) - E(V)
  IG(V) = prije grananja – suma po svim čvorovima poslije grananja
- Odaberi A <- if IG(A) > IG(B).

Proces: Evaluacija

Mjere zavisnosti

- Korelacija između varijable i ciljne varijable

$$C(j) = \frac{|\sum_{i=1}^{m}(x_{i,j} - \bar{x}_j)(y_i - \bar{y})|}{\sqrt{\sum_{i=1}^{m}(x_{i,j} - \bar{x}_j)^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} ,$$

- Zavisnost između prediktorskih varijabli = nivo redundancije
  - ako je neka varijabla zavisna o drugoj, tada je i redundantna

Konzistentnost

- Nekonzistentni primjeri - ako imaju iste vrijednosti varijabli - različite klase

Primjer	a1	a2	klasa
P1	a	b	C1
P2	a	b	C2

- Odaberi {f1,f2}
  => ako u trening setu nema primjera kao u gornjoj tablici
- min-feature = traži se najmanji podskup koji je konzistentan

Proces: Evaluacija

Klasifikacijska pogreška

- Samo u “wrapper” metodi

  evaluacija = pogreška_klasifikatora(podskup varijabli)
  if (error_rate < predefinirani threshold) select the feature subset

- selekcija varijabli – nije generalna (zavisi o konkretnom klasifikatoru), ali poboljšava točnost konkretnog klasifikatora na danom problemu
- Računalno skup pristup

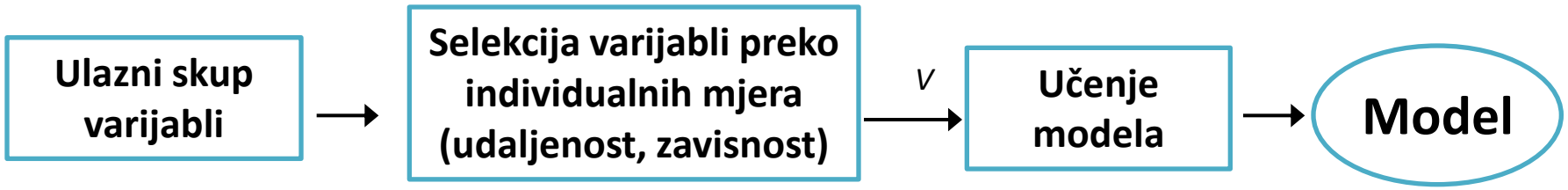
Proces: Evaluacija

metoda	Primjenjivost/ općenitost	Vremenska složenost	Točnost
Udaljenost	Da	Mala	*
Informacijske mjere	Da	Mala	*
Zavisnost	Da	Mala	*
Konzistentnost	Da	Srednja	*
Klasifikacijska točnost	ne	Velika	Visoka

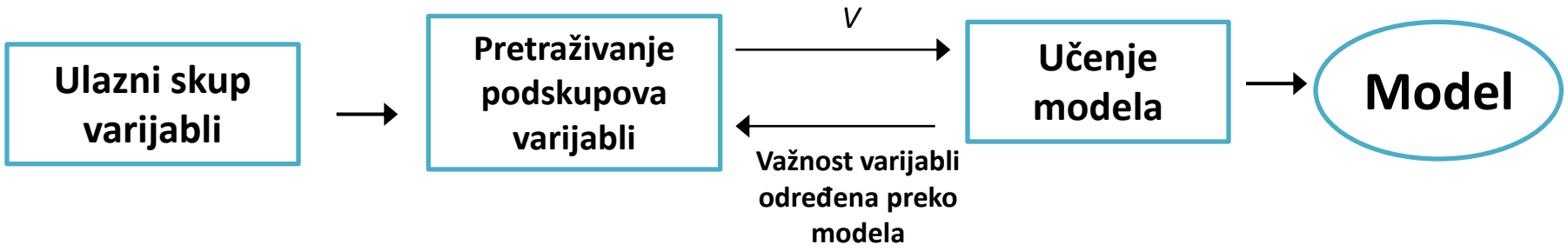
Primjenjivost – koliko su rezultati generalni (primjenjivi) kod različitih klasifikatora
Točnost – koliko je točan konačni (klasifikacijski) model
(*) Točnost ovisi od konkretne kombinacije metode selekcije i algoritma za klasifikaciju

Filter, “wrapper” i ugrađene metode

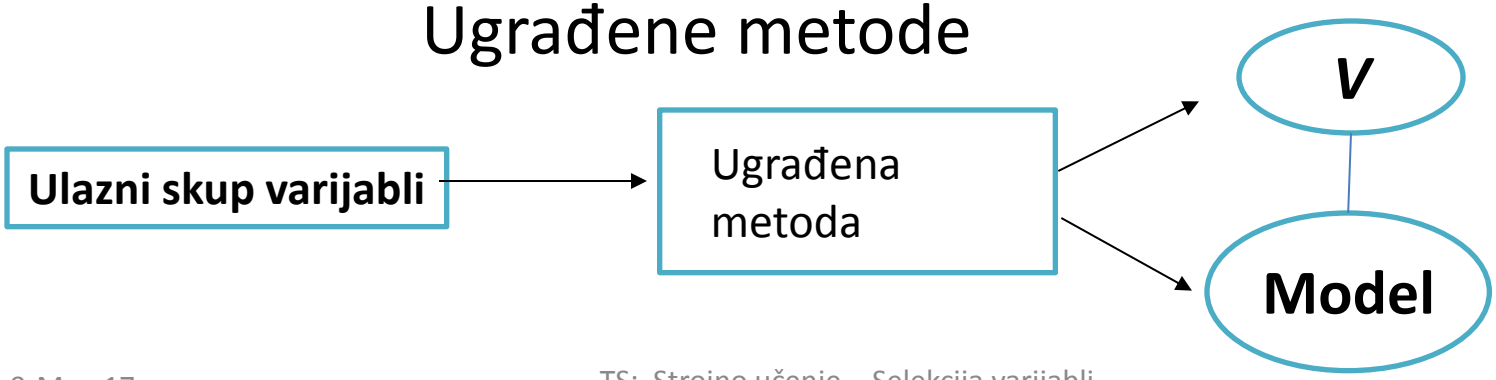
Filter



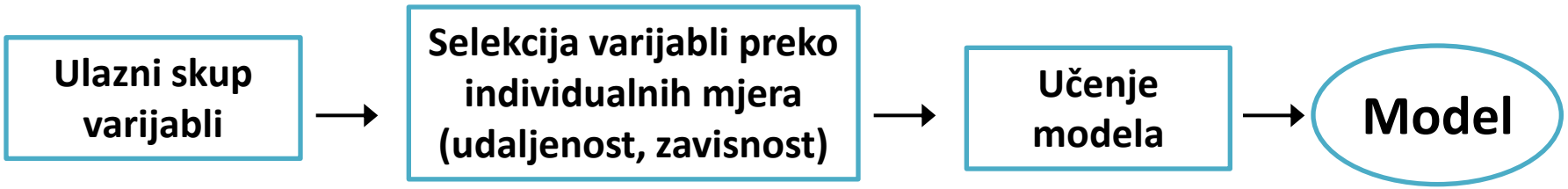
“Wrapper”



Ugrađene metode



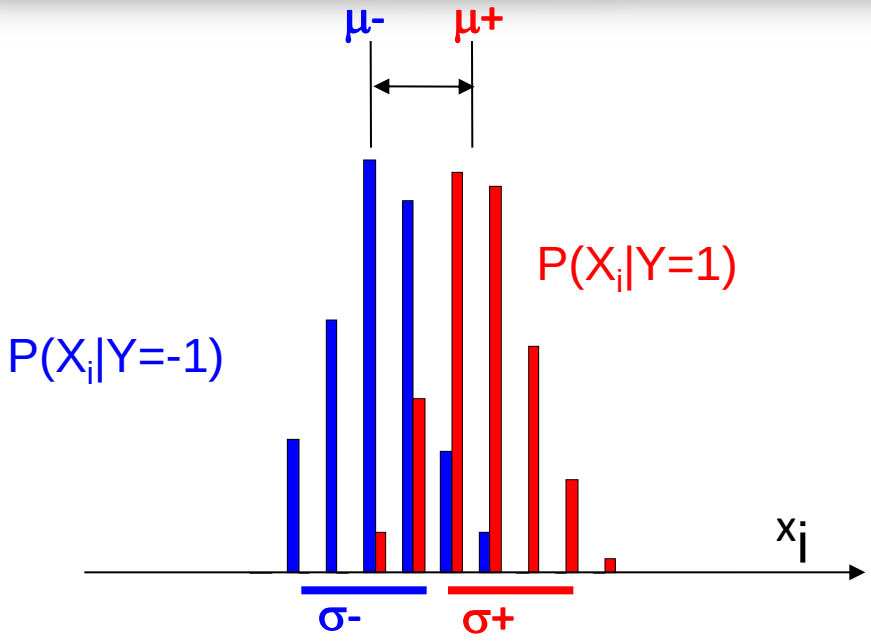
Filter



- Varijable se evaluiraju individualno a najboljih *v* se selektira i kasnije koristi u učenju modela

- Evaluacijske mjere: korelacija, uzajamna informacija, t-statistika, p-vrijednost, itd.

Univarijantne metode selekcije: primjer



- Normalna distr. klasa, σ^2 nepoznato - procjena iz podataka kao σ^2_p.
- Nulta hipoteza H_0: $\mu+ = \mu-$
- **T - test**:

$$t = (\mu^+ - \mu^-)/(\sigma_p \sqrt{1/m^+ + 1/m^-}) \propto Student(m^+ + m^- - 2d.f.)$$

Problemi Univarijantnog filter pristupa

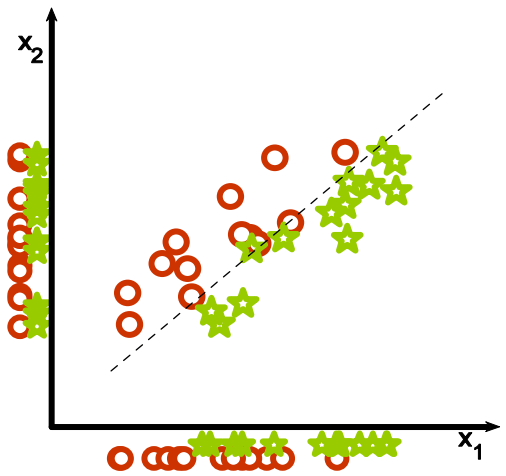
- Redundancija odabranih varijabli => varijable su odabrane nezavisno, ne kontrolira se donose li dodatnu informaciju u skup

- Interakcije između varijabli ne mogu se eksplicitno uključiti u određivanje podskupa varijabli. Individualno nevažne varijable, mogu biti važne u interakciji!

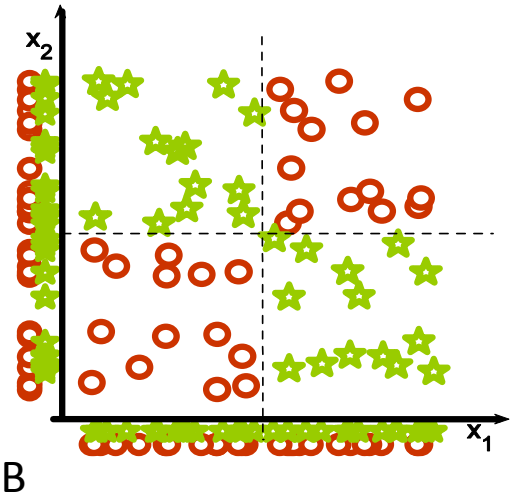
- Zanemarena je važnost (specifičnost) klasifikacijskog algoritma: neke filter metode su prikladne za određene klasifikatore, a za neke nisu.

Problemi univarijantnog filter pristupa

Guyon-Elisseeff, JMLR 2004



A



B

Prediktivna snaga varijabli - kada se promatraju zajedno

A) x_2 – irelevantna varijabla sama za sebe; relevantna u kombinaciji s x_1

B) Dvije varijable koje su individualno irelevantne - postaju relevantne u kombinaciji

Multivarijantni filter pristup

Relief algoritam [generiranje=heurističko, evaluacija=udaljenost].

- Osnove

  - svaka varijabla dobiva kumulativno težinu koja se određuje preko primjera iz trening skupa

  - varijable sa težinom iznad zadane vrijednosti *T* se selektiraju u odabrani skup varijabli

- Određivanje težine varijabli

  - princip => primjeri koji pripadaju istoj klasi trebali bi biti bliže negoli primjeri različite klase

  - bliski-pogodak (near-hit) primjer = najbliži primjer iste klase

  - bliski-promašaj (near-miss) primjer = najbliži primjer suprotne klase

  - update mehanizam za težine => $W = W - d(X,nearhit)^2 + d(X,nearmiss)^2$

Relief algoritam

1. odabrani_podskup = {}

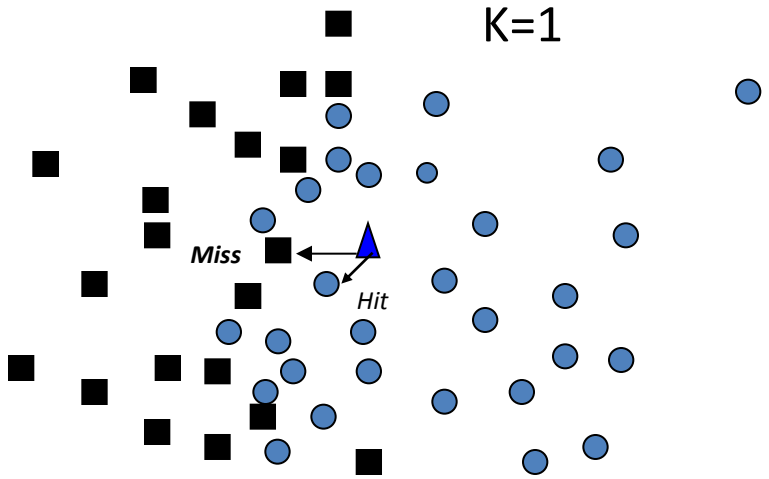
2. inicijaliziraj težine varijabli $w_i = 0$ (i=1,M)

3. za i = 1 to N % N - # primjera
   uzmi jedan primjer X iz trening skupa D.
   pronađi near-hit H = primjer iz D za kojeg je d(X,H) udaljenost najmanja & X.class=H.class
   pronađi near-miss M = primjer iz D za kojeg je d(X,M) udaljenost najmanja & X.class<>M.class
   osvježi težine svih varijabli:
   $w_i = w_i - d(X,H)^2 /N + d(X,M)^2 /N$

4. za j = 1 to M (npr. 2)
   if $w_j >= T$, dodaj v_j u odabrani skup varijabli

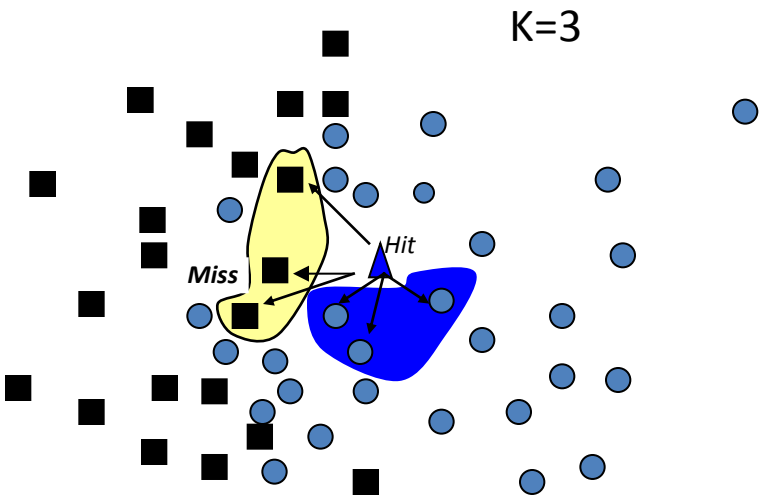
Multivarijantni filter pristup

Relief algoritam Kira & Rendell, 1992



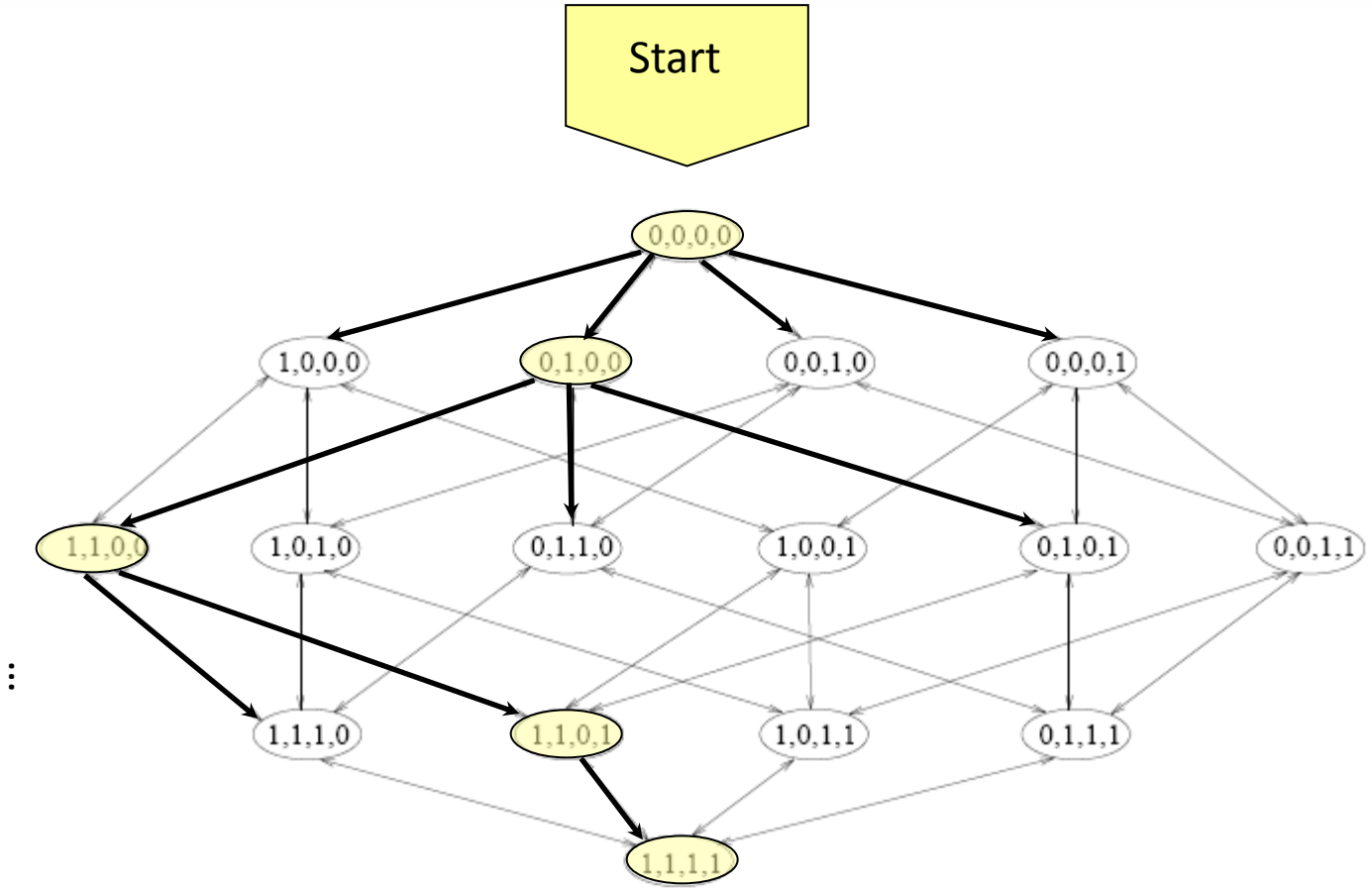
The Relief algorithm works by randomly sampling an instance and locating its nearest neighbour from the same and opposite class. The values of the features of the nearest neighbours are compared to the sampled instance and used to update the relevance scores for each feature.

Multivarijantni filter pristup



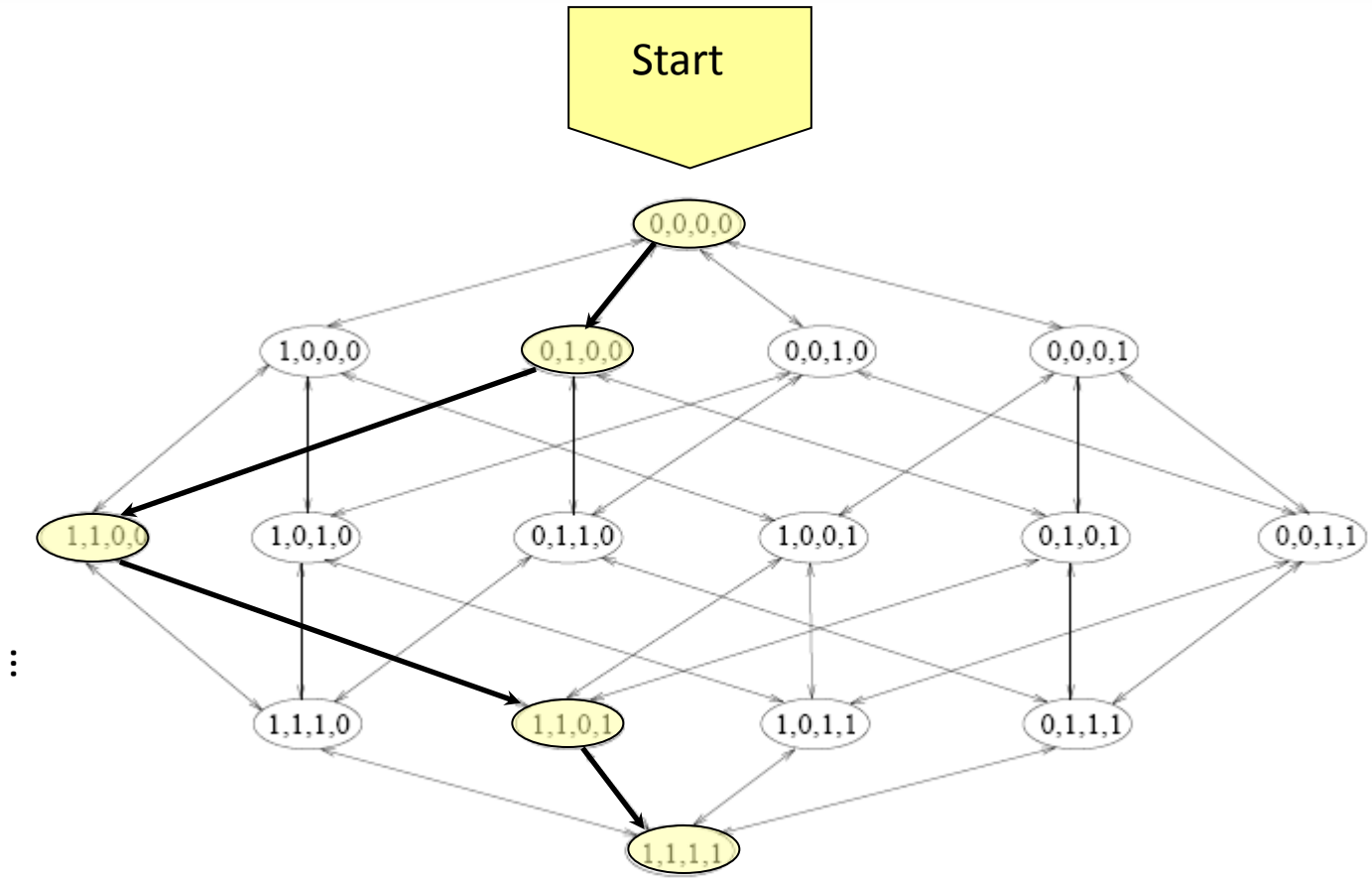
ReliefF algoritam
(više-klasni problemi, otporan na šum)
Robnik-Sikonja and I. Kononenko, 2003

Pretraživanje podskupova– wrapper pristup



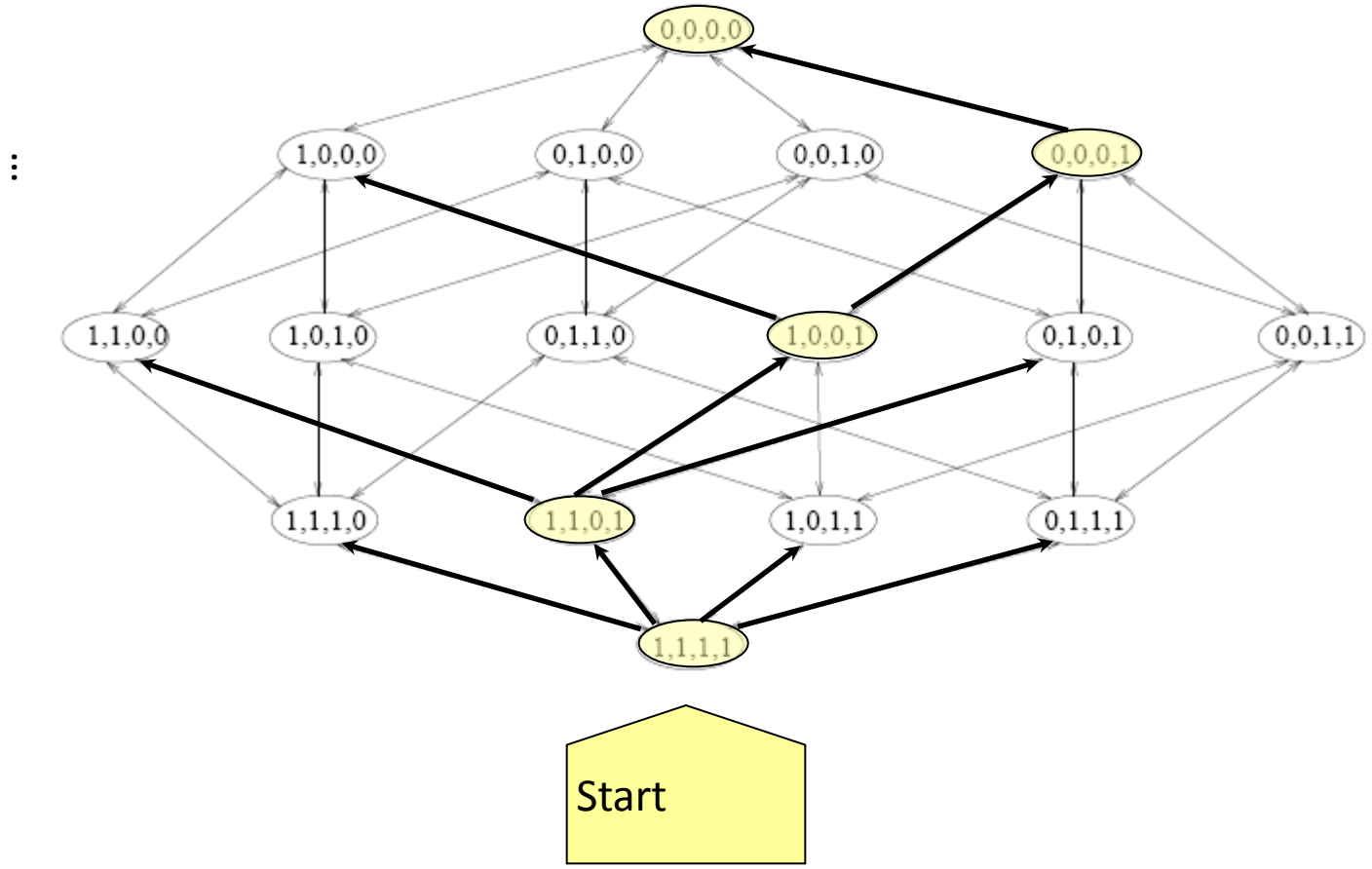
Sekvencijalno pretraživanje unaprijed (SFS: Sequential Forward Selection)

Pretraživanje podskupova - ugrađeni (embedded) pristup



Vođeno pretraživanje: ne razmatraju se alternativni putevi
Primjer.: stabla odlučivanja !

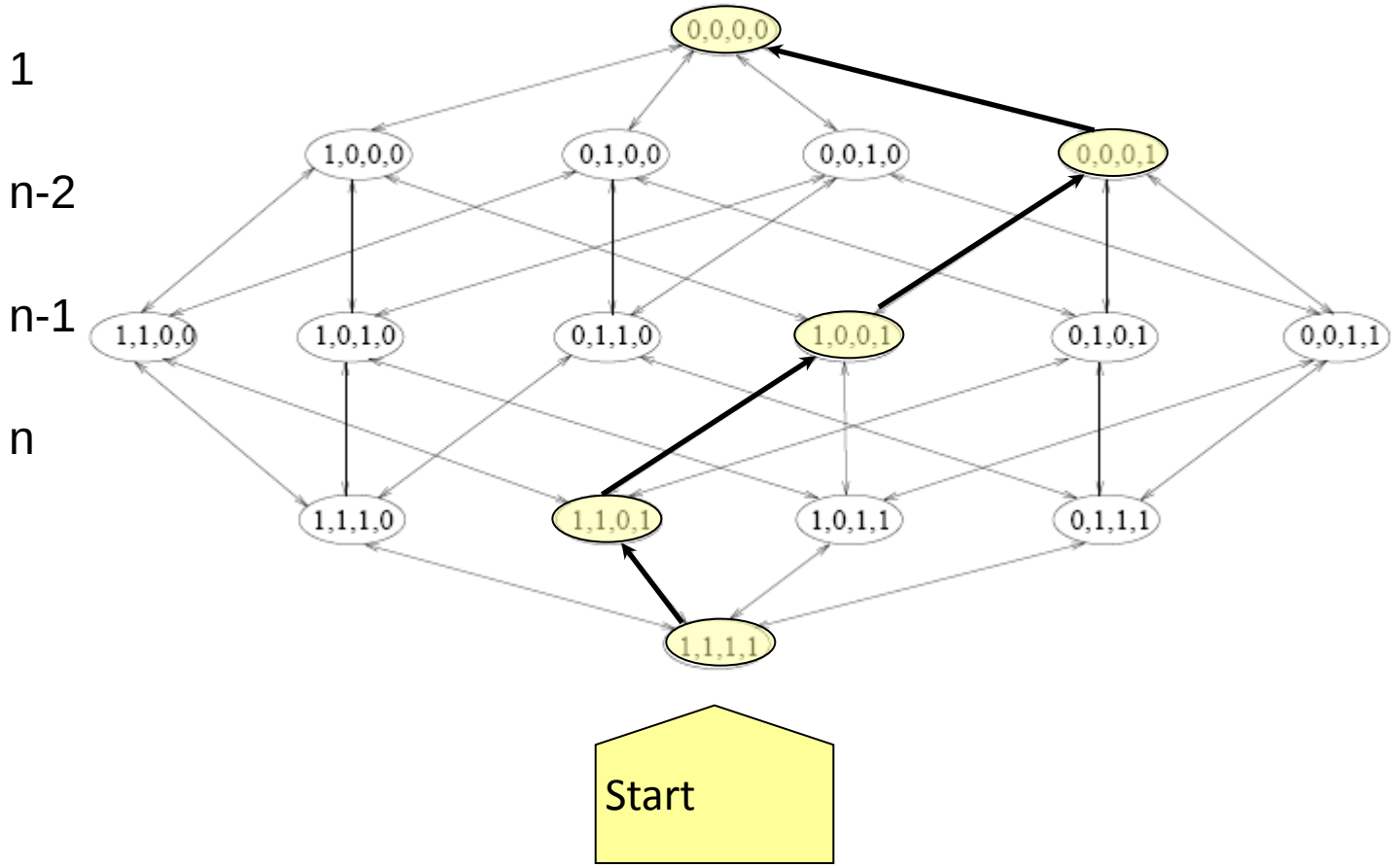
Pretraživanje podskupova – wrapper pristup



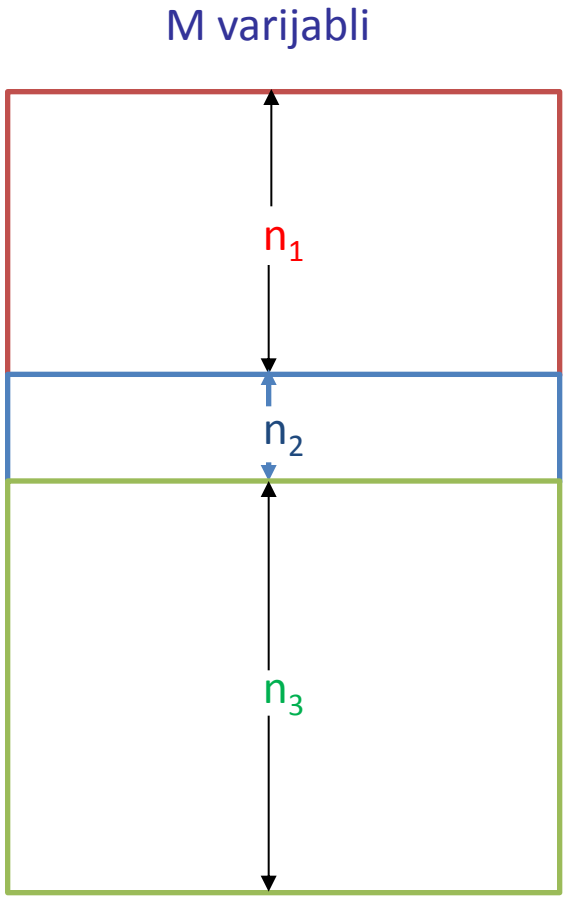
Sekvencijalno peretraživanje unatrag (SBS - Sequential Backward Selection)

Pretraživanje podskupova – embedded(ugrađeni) pristup

Vođeno pretraživanje: ne uzimaju se u obzir alternativni putevi
Primjer: “rekurzivna eliminacija varijabli bazirana na težinama (SVM)”
RFE-SVM.



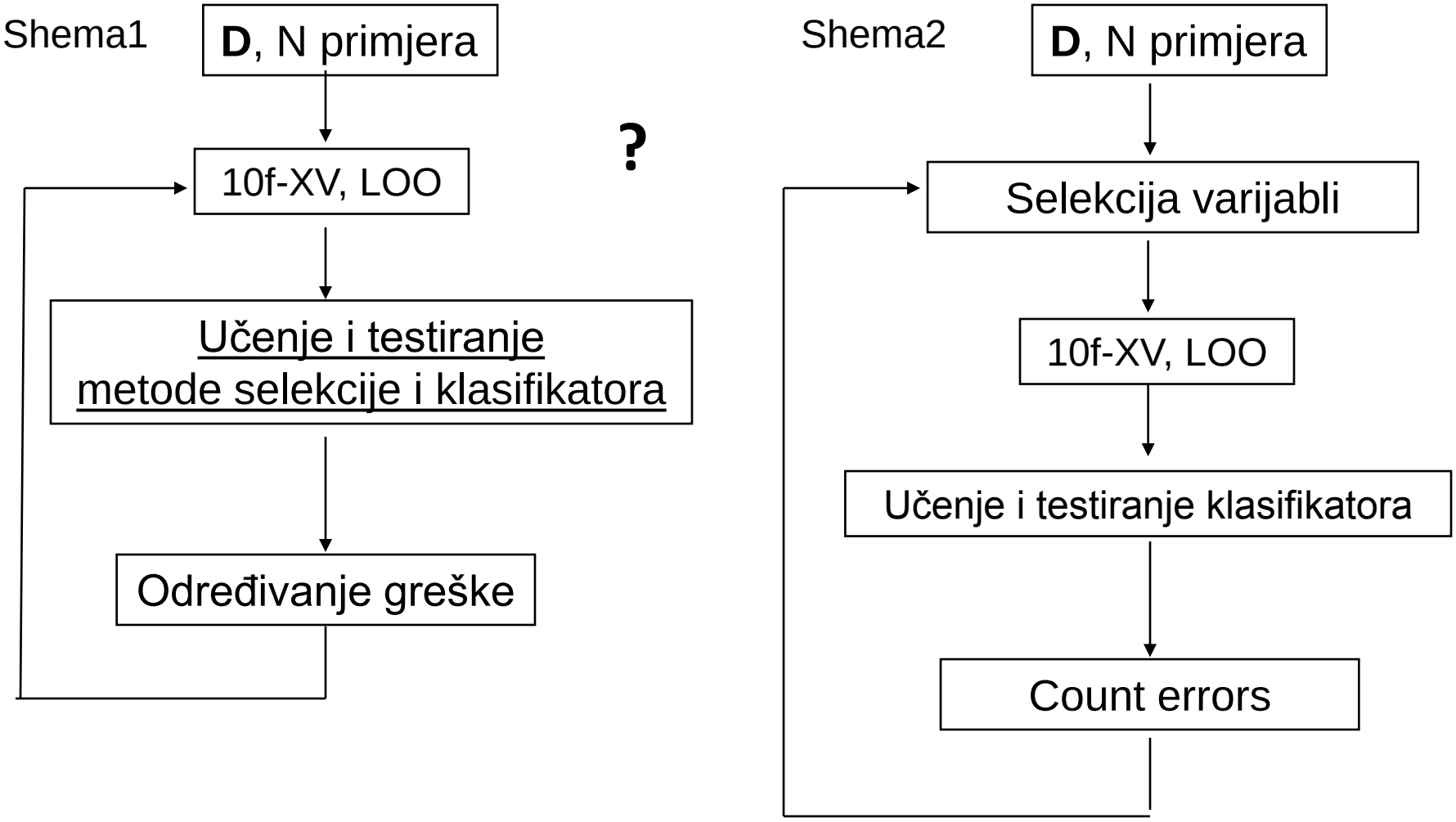
Procjena greške: XV - dvije sheme



Podjela originalnog skupa na 3 dijela:
n1-trening; n2-validacijski, i n3- test set

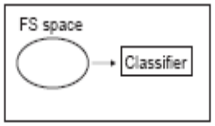
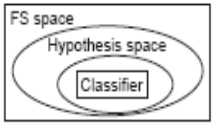
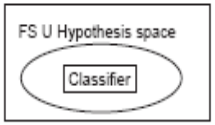
1) Za svaki podskup varijabli, uči model na n1.
2) Odaberi podskup ***V***, koji je najbolji na n2
   - Ponoviti i usrednjiti (XV?)
3) Testiraj na n3.

Procjena greške: XV - dvije sheme

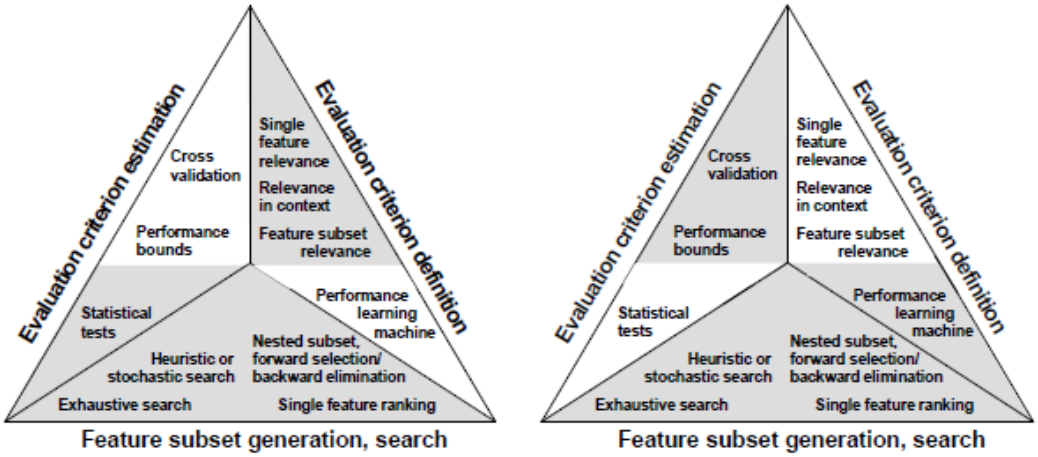


Taksonimija(e) metoda selekcije varijabli

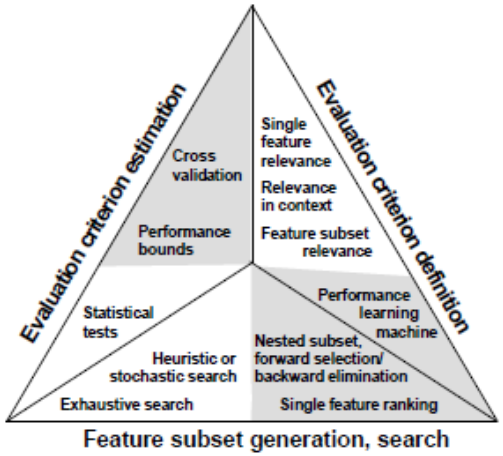
Table 1. A taxonomy of feature selection techniques. For each feature selection type, we highlight a set of characteristics which can guide the choice for a technique suited to the goals and resources of practitioners in the field.

	Model search	Advantages		Disadvantages	Examples
Filter		Univariate	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	Chi-square Euclidean distance t-test Information gain, Gain ratio [6]
		Multivariate	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation based feature selection (CFS) [45] Markov blanket filter (MBF) [62] Fast correlation based feature selection (FCBF) [136]
Wrapper		Deterministic	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) [60] Sequential backward elimination (SBE) [60] Plus q take-away r [33] Beam search [106]
		Randomized	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing [110] Genetic algorithms [50] Estimation of distribution algorithms [52]
Embedded		Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies		Classifier dependent selection	Decision trees Weighted naive Bayes [28] Feature selection using the weight vector of SVM [44, 125]

Taksonimija(e) metoda selekcije varijabli



(a) Filters (b) Wrappers



(c) Embedded methods

Fig. 1. *The three principal approaches of feature selection.* The shades show the components used by the three approaches: filters, wrappers and embedded methods.