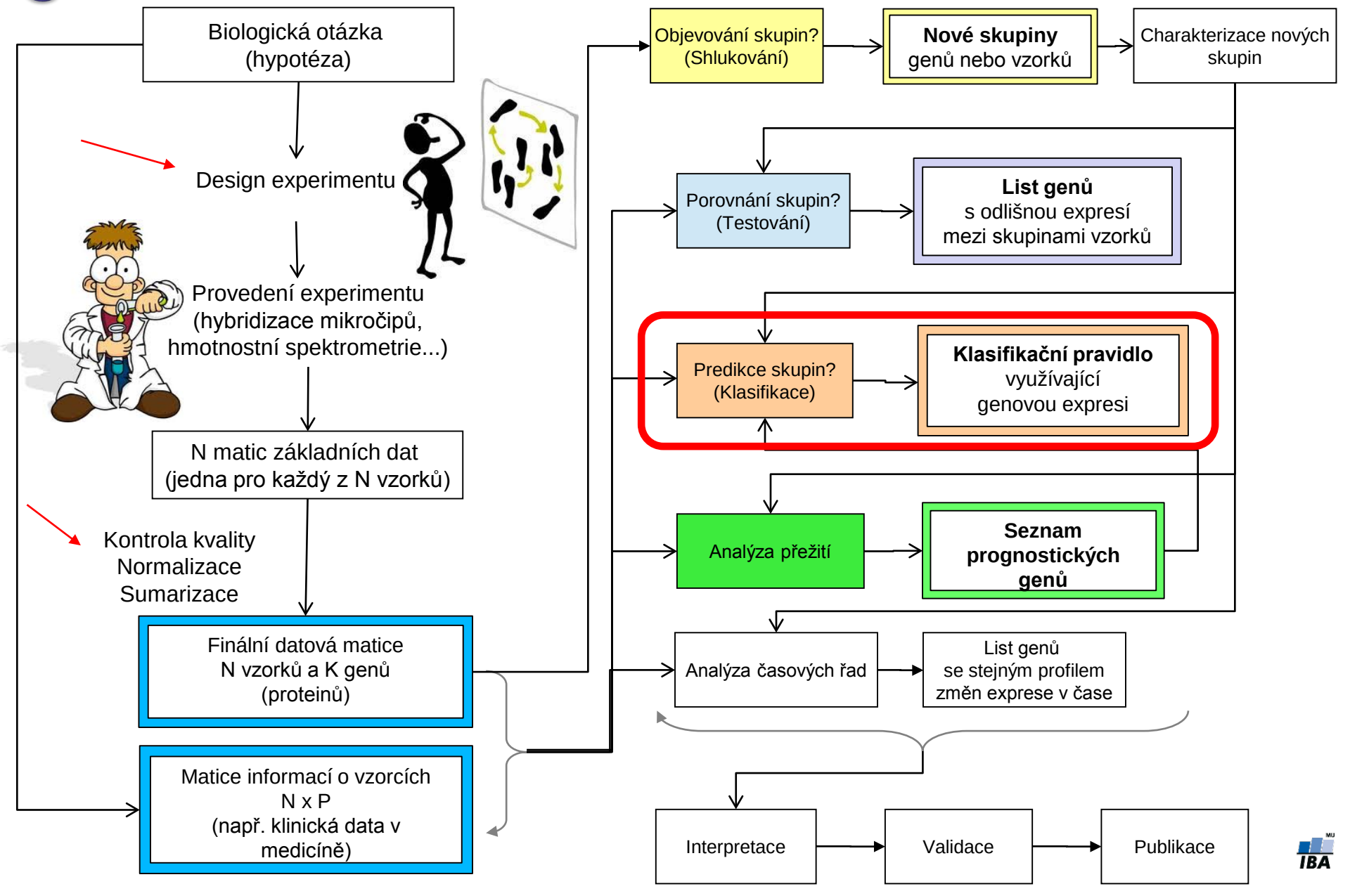
Předpovídání skupin (class prediction)

Společné schéma analýzy dat



Tradiční schéma analýzy

- **Učení s učitelem (supervised learning)**
  - V tomto případě zobecňujeme známou strukturu dat na nové data
  - **Porovnávání skupin (class comparison)**
    - hledáme rozdíly v expresi, počtu kopií genů nebo abundanci proteinů mezi již definovanými skupinami
  - **Předpovídání skupin (class prediction)**
    - na známých skupinách se snažíme vytvořit klasifikátor, který by dokázal zařadit nového pacienta do jedné ze skupin

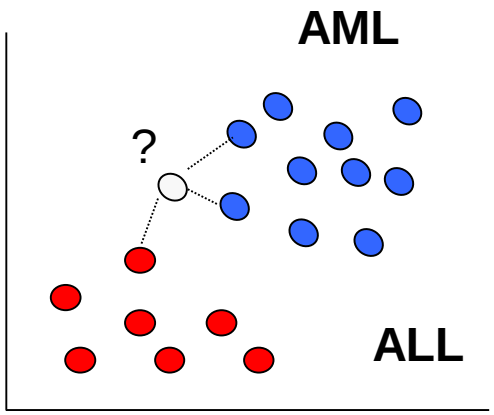
- **Učení bez učitele (unsupervised learning)**
  - V tomto případe struktura v datech není známá a musíme ji objevit
  - **Objevování skupin (class discovery)**
    - na základě informací o genech/proteinech hledáme nové skupiny
    - onemocnění X je velmi heterogenní a snažíme se identifikovat specifičtější podtypy, které by mohli být cílem cílené terapie

Společné znaky analýzy dat

- Velké množství proměnných
- Malé množství vzorek
- Proměnné jsou často korelované, s velmi komplexními vztahy
- Data obsahují množství šumu – biologická i technická variabilita

Předpovídání skupin

- V tomto typu analýzy se snažíme předpovědět příslušnost k jedné ze známých skupin na základě jejich genomického nebo proteomického profilu
- Například předpovídáme:
  - typ diagnózy
  - odpověď na terapii
  - přežití pacienta



- Cílem je vytvořit klasifikační pravidlo (soubor pravidel), které toto umožní
- Vytvoření klasifikátoru může sloužit jako nástroj pro selekci genů, které významně diskriminují mezi skupinami

- Shlukování s učitelem (supervised clustering)
- Regresní metody

Princip

1. **Výběr proměnných** pro klasifikaci
    - Vybíráme geny nebo proteiny, které se v klasifikátoru použijí

2. **Trénování**
    - Na trénovacích datech vytvoříme klasifikační pravidlo (klasifikátor, model)

3. **Testování**
    - Vytvořený klasifikátor se otestuje na testovacích datech
    - K odhadnutí výkonnosti (přesnosti) klasifikátoru a optimalizaci parametrů

Výběr proměnných I.

- Důvody k redukci dimenzionality dat:

- **Ze statistického hlediska**

  Eliminace tisíců nerelevantních genů významně ovlivní komplexitu vybraného klasifikátoru, stane se robustnější.

- **Z biologického hlediska**

  Výběr vhodných genů/proteinů silně korelovaných s danou skupinou pomůže pochopit mechanismus jejich působení.

- **Z praktického hlediska**

  Čím méně genů potřebujeme pro predikci, tím snadnější je uplatnení klasifikátoru v praxi.

Výběr proměnných II.

- U genomických a proteomických dat je výběr proměnných trochu problematický, protože geny jsou velmi korelované
  - Výběr jednoho reprezentanta je víceméně náhodný
  - Malé změny v trénovacích datech, případně aplikace jiného klasifikátoru může vyústit do úplně jiné selekce genů

  To je v pořádku, ale pozor na interpretaci!

- Při interpretaci je třeba brát na zřetel, že se jedná pouze o podskupinu genů
- Biologické závěry o podskupinách vzorek by měly být založené na studiu celé množiny významných genů

Příklad

ARTICLE

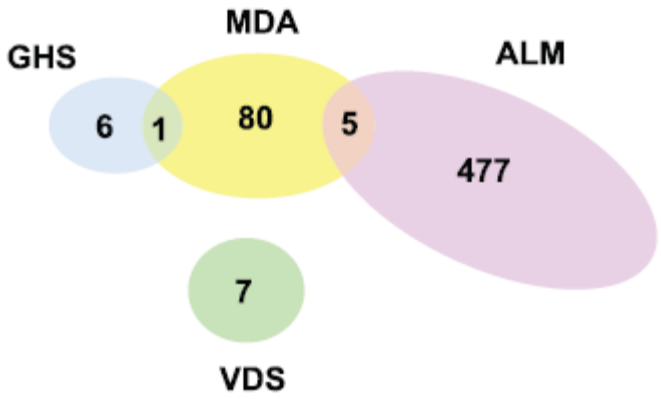
Test of Four Colon Cancer Risk-Scores in Formalin Fixed Paraffin Embedded Microarray Gene Expression Data

Antonio F. Di Narzo, Sabine Tejpar, Simona Rossi, Pu Yan, Vlad Popovici, Pratyaksha Wirapati, Eva Budinska, Tao Xie, Heather Estrella, Adam Pavlicek, Mao Mao, Eric Martin, Weinrich Scott, Fred T. Bosman, Arnaud Roth, Mauro Delorenzi

Manuscript received December 9, 2013; revised April 22, 2014; accepted July 2, 2014.

Table 1. Description of the four risk scores analyzed*

	Risk scores			
Abbreviation	**GHS**	**VDS**	**MDA**	**ALM**
Developer	Genomic Health	Veridex	MD Anderson	ALMAC diagnostics
Type of assay	Q-RT-PCR	microarray and Q-RT-PCR	microarray	microarray
Type of tissue	FFPE	fresh frozen and FFPE	fresh frozen	FFPE
Main publication	O'Connell et al. 2010.	Jiang et al. 2008.	Oh et al. 2011.	Kennedy et al. 2011.
Total number of features	7	7	114 (86 genes)	634 (482 genes)
Features used (genes)	7	6	85 (85 genes)	634 (identical platform)



Výběr proměnných III.

- Dva základní typy metod výběru proměnných:
  - **Filtrace**
    - Na základě diskriminační schopnosti jednotlivých proměnných (odlišně exprimované geny, prognostické geny,…)
  - **Wrapper metody**
    - Vybírají se přímo skupiny genů, na kterých se vybuduje klasifikátor, jehož výkon se následně otestuje
      - Forward sequential selection: geny jsou postupně vybrané na základě informace, kterou přispívají k diskriminaci
      - Backward selection začíná s celou množinou a postupně odstraňuje ty, které nepřispívají k diskriminaci (vzhledem k ostatním genům)
    - Tento přístup je výpočtově náročný, protože nemůžeme otestovat všechny možné podskupiny
    - Můžou být velmi nestabilní, výběr i-tého genu je velmi závislý na podmnožině už vybraných genů

Metody klasifikace vzorků I.

Black-box metody

- Často používají celý datový soubor použitý na trénování
- Obvykle nejsou jednoduše interpretovatelné

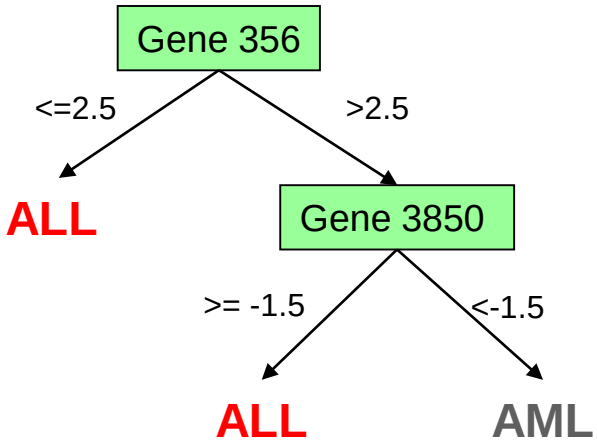
- K-nejbližších sousedů
- Support vector machine
- Neuronové sítě

Metody klasifikace vzorků II.

Metody vytvářející klasifikační pravidla

- Více intuitivní, jednoduše použitelné v praxi
- Dostáváme přímo skupinu důležitých parametrů, případně jasně interpretovatelné klasifikační pravidlo

- Regresní modely
- Bayesovský klasifikátor
- Fisherova diskriminační analýza
- Klasifikační stromy a lesy
- Top Scoring Pairs
- AdaBoost



Odhad výkonnosti

- Výkonnost každého klasifikátoru musí být otestovaná na jiném validačním souboru

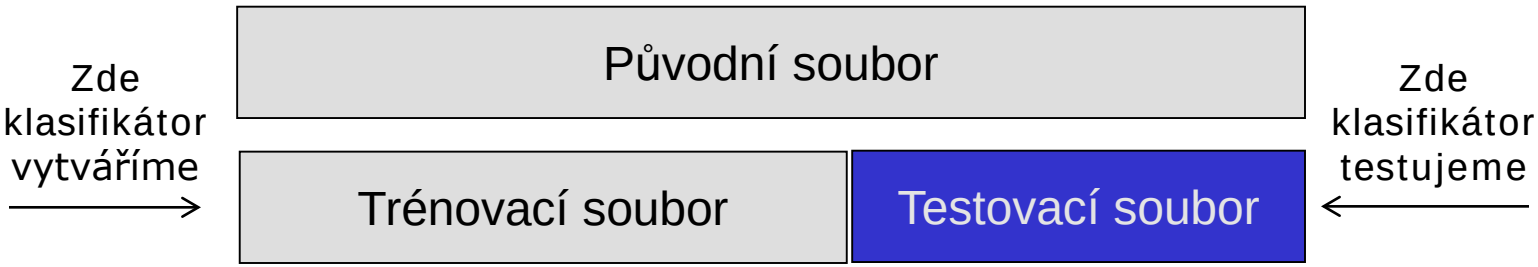
Proč odhadovat výkonnost klasifikátoru?

- Omezení trénovacím souborem
- Bez předpokladu o rozložení neexistuje žádný vzorec pro výpočet
- Často existuje jen jeden datový soubor pro trénování a testování klasifikátoru
- Odhad výkonnosti klasifikátoru na trénovacích datech je optimisticky zkreslený

Odhad výkonnosti

Základní myšlenka:

Převzorkováním rozdělit (opakovaně) datový soubor na trénovací a testovací, vytvořit klasifikátor na trénovacím souboru a změřit výkonnost klasifikátoru jen na datech, které nebyly použity pro jeho vytvoření.



UPOZORNĚNÍ: Všechny kroky, které závisí na převzorkování a které vedou k finálnímu **modelu musí být zopakované identicky na každém rozdělení na trénovací a testovací soubor**.

Patří sem například normalizace dat, výběr proměnných, trénovaní klasifikátoru, optimalizace parametrů,...

Odhad výkonnosti II.

- Každé dva trénovací soubory vytvořené z původního datového souoboru s pomocí převzorkování se do jisté míry překrývají -> vytvořené klasifikátory tedy nejsou úplně nezávislé
- Variabilita je obvykle podhodnocená

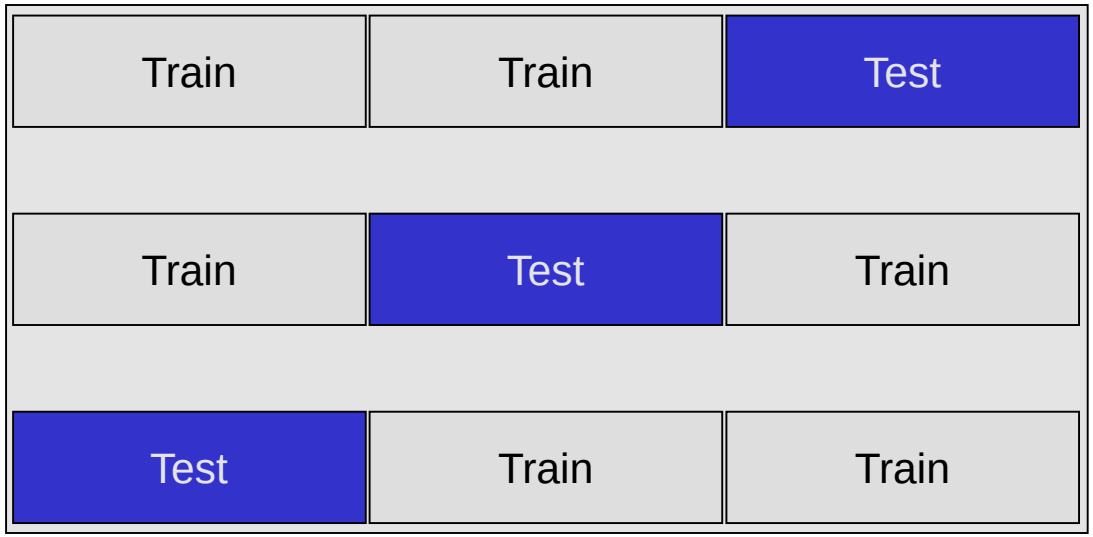
Převzorkovací metody

- Jednoduché **rozdělení na dva soubory**
- **k-násobná křížová validace** (k-fold cross validation)
  - Opakovaná k-násobná křížová validace
  - Monte-Carlo křížová validace
  - Leave-one-out křížová validace (n-násobná křížová validace, kde n je počet vzorků)
- **Bootstraping**

Křížová validace

- Oddělený trénovací a testovací soubor
- Náhodné rozdělení dat do *k* podmnožin
- Vytvoření klasifikátoru na k-1 množinách a otestování na zůstávající
- Každá podmnožina je jednou testovací
- Obvykle k=5 nebo k=10 (pokud se k=počtu vzorků, pak se jedná o leave-one-out odhad)
- Opakovaná křížová validace – ještě lepší odhad

k=3



Bootstrapping

- Vytvoření nového souboru vzorkováním s opakováním (rozdíl od křížové validace, kde je vzorek vždy jen jendou)
- Trénování klasifikátoru probíhá na nových datech a testuje se na vynechaných vzorcích
- Opakuje se B-krát
- Odhad chyby pomocí 0.632 pravidla

$$\bar{E} = 0.368E_0 + 0.632\frac{1}{B}\sum_{b=1}^{B}\hat{E}_b$$

Kde E0 je chyba na celém (původním) trénovacím souboru

Odhad výkonnosti III

- Zjistíme očekávanou výkonnost klasifikátoru na validačním souboru
- Můžeme identifikovat nejstabilnější proměnné (geny/proteiny)
- Které vzorky jsou stále špatné klasifikované (pokud takové jsou, naznačuje to odlehlé hodnoty)

Standardy pro mikročipy

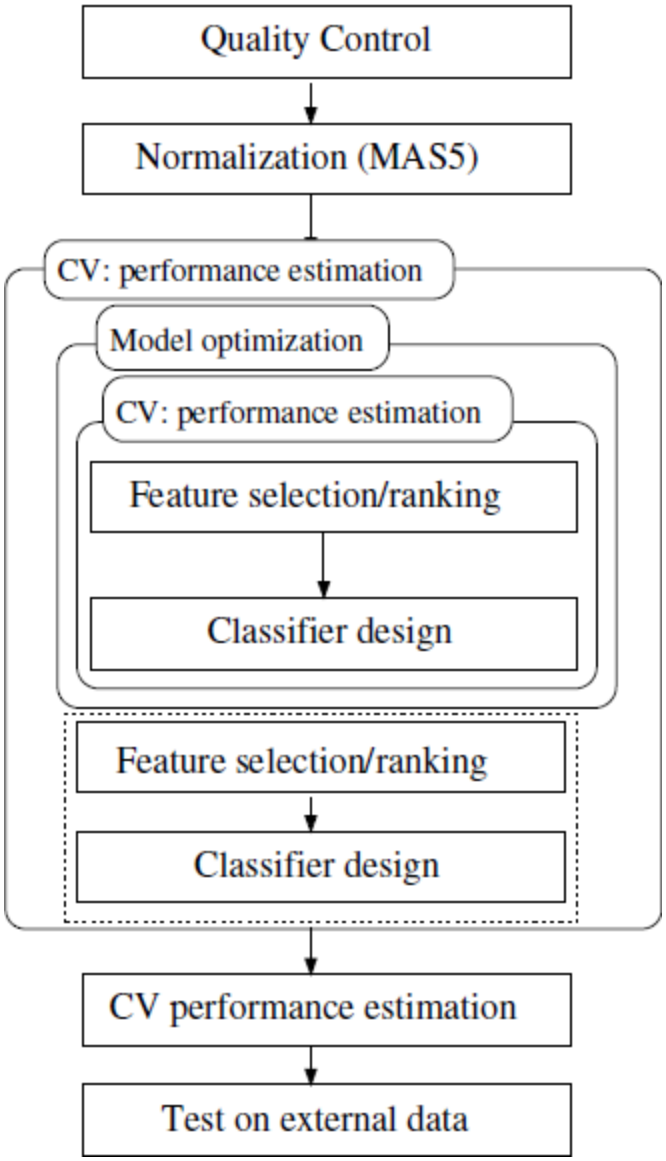
ARTICLES

nature biotechnology

The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models

MAQC Consortium*

Standardy pro mikročipy II



Validace

- **Vždy na nezávislém datovém souboru**
- Velmi důležitá pro otestování skutečné robustnosti klasifikátoru
- Absolutně nevyhnutelné v medicíně

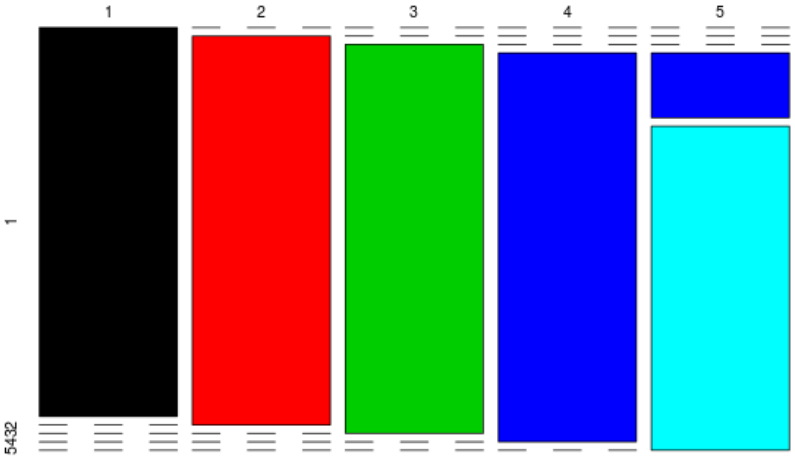
- Testovací soubor by měl splňovat následující vlastnosti:
  - Musí obsahovat parametry použité v klasifikátoru
  - Musí být známá příslušnost vzorků ke skupinám, které se klasifikátor snaží diskriminovat
  - Podobná struktura s ohledem na klinické a patologické parametry (např. Stejné rozložení věku, zastoupení pohlaví apd.)

Design experimentu je důležitý!

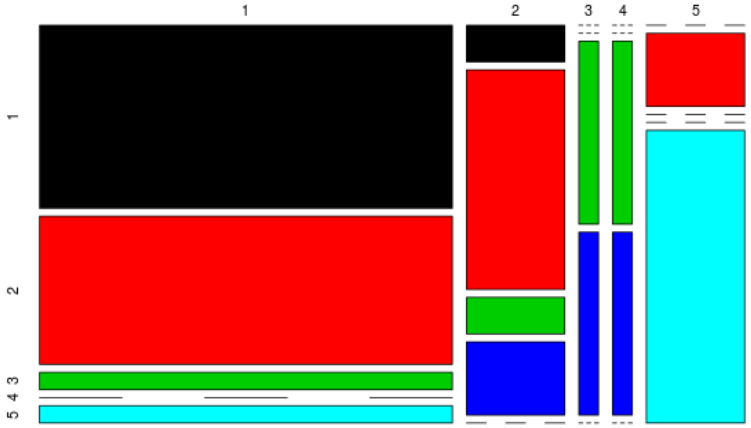
- **Myslete na dostatečně velký trénovací i testovací datový soubor!**

Příklad: 5 podtypů karcinomu prsu – 96 vzoriek (N1=48, N2=16, N3=8, N4=8, N5=16)

Trénovací soubor



Testovací soubor



Málo vzorků ve skupině, nemožnost tuningu, malá variabilita -> přetrénování => nefunguje na testovacím souboru.
Stačí jeden špatně klasifikovaný vzorek a výrazně se sníží výkonnost!

- **Datové soubory musí reprezentovat populaci, na které budete klasifikátor používat**

Shrnutí

- Je užitečné vybrat proměnné před aplikováním klasifikátoru
- Je lepší používat jednoduché klasifikátory
- Odhadujte výkonnost klasifikátoru a optimalizujte parametry na trénovacím souboru
- Vždy klasifikátor validujte na úplně jiném datovém souboru

Úloha [1 bod]

Pracujte s datovým souborem golub.Rdata, který naleznete v IS.

- X – matice genových expresí, kde v řádku jsou jednotlivý pacienti a ve sloupcích jednotlivé sondy
- Y – vektor určující, do které skupiny patří pacienti

1. Vhodnou metodou vyberte ty geny, které jsou odlišně exprimovány mezi skupinami.
2. Datový soubor náhodně rozdělte na trénovací (2/3 původního datového souboru) a testovací (1/3 původního datového souboru).
3. Na tyto geny aplikujte metodu k-nejbližších sousedů.
4. Porovnejte výsledky z klasifikátoru se známým zařazením pacientů do skupiny. Výsledky sumarizujte do čtyřpolní tabulky.

R skript obsahující komentáře k výsledkům zašlete do úterý 5.4.2016 na email (hanakova@recetox.muni.cz).