ČVUT

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Úloha C – Klasifikace diabetu Pima Indians

Marek Darsa

Výběr klasifikátorů

- Vybrány dva klasifikátory:
  1. Logistická regrese – jednoduchá a interpretovatelná.
  2. Rozhodovací strom – snadná vizualizace rozhodování
- Kritéria: jednoduchost, interpretovatelnost

Trénování a přesnost modelu

- Data rozdělena na 70 % trénovací / 30 % testovací (náhodně, bez fixního seed)
- Logistická regrese (nevyvážená):
  - Přesnost: 78,42 % (min: 71,19 %, max: 85,59 %)
- Logistická regrese (ROSE):
  - Přesnost: 71,90 % (min: 61,45 %, max: 79,52 %)
- Rozhodovací strom (nevyvážený):
  - Přesnost: 70,06 % (min: 55,42 %, max: 83,13 %)
- Rozhodovací strom (ROSE):
  - Přesnost: 69,82 % (min: 59,04 %, max: 83,13 %)

Důležité příznaky a zjednodušení

- Logistická regrese ukázala důležité příznaky:
  - Glucose, bmi, diabetes (v rodině), age
- Nevýznamné:
  - Pregnant, triceps, diastolic, insulin
- Lze zvažovat snížení počtu měření o nevýznamné parametry

Úspěšnost při nasazení v praxi

- Pro zhodnocení stability a očekávané přesnosti klasifikátoru byla provedena simulace 100 opakovaných náhodných rozdělení dat
- Tímto způsobem jsme získali realističtější odhad výkonnosti modelu při nasazení na nové pacienty
- Minimální přesnost: 71,19%
- Průměrná přesnost: 78,42%
- Maximální přesnost: 85,59%

Pravděpodobnost onemocnění

- Celková prevalence diabetu v datech : 130 / 392 ≈ 33,2%
- Vyvážené modely více reflektují rizikové skupiny

Vliv nevyváženosti dat

- Nezvyvážená data:
  - model preferuje většinovou třídu (zdravé pacientky)
  - více FN → přehlédnutých nemocných
- Po vyvážení pomocí ROSE:
  - více FP, ale méně FN
  - model lépe detekuje nemocné (vyšší citlivost)
- Vyvážení zvyšuje citlivost modelu, ale často snižuje celkovou přesnost.

Falešné předpovědi - diskuze

- Logistická regrese (nevyv.):
- - FP (neg → pos): 8
- - FN (pos → neg): 14
- Po vyvážení:
- - FP: 18, FN: 8 → vyšší citlivost, nižší přesnost
- Trade-off mezi falešnými poplachy a přehlédnutím

Závěr

- Logistická regrese je vhodnější pro tento úkol
- Vyvážení dat zvyšuje citlivost modelu ale snižuje přesnost