Eksploracja danych w genotypowych bazach danych

Przemysław Klęsk
pklesk@wi.zut.edu.pl

Zakład Sztucznej Inteligencji
Wydział Informatyki, ZUT

1. „Sodowrażliwość”

- Cel badań — sprawdzenie możliwości rozpoznawania sodowrażliwości na podstawie czynników genetycznych.
- Istnieją osoby, które po spożyciu sodu (m.in. sól kuchenna, benzoesan sodu) doświadczają skoków ciśnienia tętniczego. Pojawia się to także u osób, które nie chorują nominalnie na nadciśnienie.
- Pomorska Akademia Medyczna (obecnie: PUM) zebrała zbiór danych na grupie 106 osób (bez i z sodowrażliwością).
- Eksperyment trwał 3 tygodnie: tydzień na diecie bezsolnej, tydzień na diecie solnej, tydzień na diecie mieszanej.
- Zbiór danych zawiera: 106 przykładów i 24 atrybuty (wśród atrybutów wejściowych jest 19 genetycznych i 4 inne).

(!) Własność: Zakład Biochemii Klinicznej i Molekularnej, Pomorski Uniwersytet Medyczny w Szczecinie (prof. A. Ciechanowicz).

1. „Sodowrażliwość” (c.d.)

Atrybuty

1. płeć — $\{F, M\}$,
2. wiek — $\{< 39.5, \geqslant 39.5\}$,
3. BMI — Body Mass Index $\{< 23.45, \geqslant 23.45\}$,
4. NT — nadciśnienie tętnicze $\{0, 1\}$,
5. dSS — wskaźnik sodowrażliwości (atrybut decyzyjny) $\{< 8, \geqslant 8\}$,
6. PROK — $\{?, AA, AB, AH, AI, AK, AQ, AR, BB, BH, BI, BK, HI, HK, HQ, IK\}$,
7. GSL — $\{?, CC, CT, TT\}$,
8. BE16 — $\{AA, AG, GG\}$,
9. BE27 — $\{CC, GC, GG\}$,
10. BE1 — $\{?, CC, CG, GG\}$,
11. G3NB — $\{CC, CT, TT\}$,
12. ACE — $\{DD, ID, II\}$,
13. HPA — $\{WM, WW\}$,
14. SYAL — $\{CC, TC, TT\}$,
15. ESC — $\{CC, CG, GG\}$,

1. „Sodowrażliwość” (c.d.)

Atrybuty

16. ADD — $\{GG, GT, TT\}$,
17. AT1R — $\{AA, AC, CC\}$,
18. ATG — $\{AA, AG, GG\}$,
19. KAL1 — $\{CC, GC, GG\}$,
20. KAL3 — $\{GA, GG\}$,
21. KAL4 — $\{AA, AG, GG\}$,
22. KAL5 — $\{AA, AC, CC\}$,
23. eNOS — $\{GG, GT, TT\}$.

Atrybut decyzyjny

dSS — wartość $\geqslant 8$ — osoba sodowrażliwa; wartość < 8 — osoba niesodowrażliwa.

2. „Inseminacja krów mlecznych”

- Cel badań — sprawdzenie możliwości przewidywania klasy zabiegu inseminacyjnego u krów mlecznych.
- Zbiór danych zawiera: 409 przykładów i 13 atrybutów.
- Dane zebrane na podstawie dokumentacji gospodarstw wiejskich.
- Wiadomo, że do czynników wpływających na łatwość zacieleń u krów należą m.in. wiek krowy, jej kondycja, wydajność mleka, sezon zabiegu itd.
- Pytanie: czy istnieje wpływ udziału genów HF na łatwość zabiegu inseminacyjnego?

2. „Inseminacja krów mlecznych” (c.d.)

Atrybuty wejściowe

1. MIES — miesiąc inseminacji,
2. LAK — kolejna laktacja,
3. WYD — średnia wydajność,
4. LAK_W — wydajność laktacyjna,
5. GOS – średnia wydajność mleka w gospodarstwie,
6. TL – średnia wydajność tłuszczu w gospodarstwie,
7. TL% — średnia zawartość tłuszczu w gospodarstwie,
8. BL% — średnia zawartość białka w gospodarstwie,
9. HF — procent genów HF (bydło holsztyno-fryzjerskie),
10. POROD — numer porodu,
11. KOND — kondycja,
12. OCEN — ocena inseminacyjna.

Atrybut decyzyjny

INS — klasa zabiegu inseminacyjnego; A — zacielenie łatwe (następowało po 1, 2 zabiegach inseminacyjnych), B — zacielenie trudne (zacielenie po 3-11 zabiegach).

3. „Wydajność krów mlecznych”

- Cel badań — sprawdzenie możliwości przewidywania wydajności krów mlecznych (zadanie estymacji regresji).
- Zbiór danych zawiera: 188 przykładów i 11 atrybutów.
- Dane zebrane na podstawie dokumentacji gospodarstw wiejskich.
- Atrybut ALU dotyczy genotypu albuminy {LL, LV}.

(!) Własność: Zakład Biostatystyki, Zachodniopomorski Uniwersytet Technologiczny w Szczecinie (prof. W. Grzesiak).

3. „Wydajność krów mlecznych” (c.d.)

Atrybuty wejściowe

1. LAK — kolejna laktacja,
2. RAHF — udział genów HF,
3. WYD_MAT — wydajność matki,
4. ROHF – udział genów HF ojca,
5. WYD_OJCA — wydajność ojca oceniona na podstawie córek,
6. SEZON — sezon wycielenia {1 — jesienno zimowy, 2 — wiosenno-letni},
7. WIEK — wiek krowy,
8. MLEKO (atrybut wyjściowy) — wydajność mleczna krowy,
9. ALU — genotyp albuminy,
10. WYD — wydajność mleka w gospodarstwie.

Atrybut wyjściowy (regresja)

MLEKO — wydajność mleczna krowy; zakres: 2692 ÷ 9467, średnia: 5521.44, odchylenie std.: 1347.03.

4. „Standaryzowana masa ciała”

- Cel badań — sprawdzenie możliwości przewidywania standaryzowanej masy ciała (zadanie estymacji regresji).
- Zbiór danych zawiera: 183 przykładów i 11 atrybutów.
- Dane zebrane na podstawie dokumentacji gospodarstw wiejskich.
- Atrybut genotypowy HOR_WZR dotyczy hormonu wzrostu $\{1 - AA, 2 - AB, 3 - BB\}$.

4. „Standaryzowana masa ciała” (c.d.)

Atrybuty wejściowe

1. WIEK_KR — wiek krowy,
2. MIES_WY — miesiąc wycielenia,
3. OMW_4 — okres międzywycieleniowy,
4. MAS_KR_P — masa krowy po wycieleniu,
5. PLEC4 — płeć cielęcia {j — jałówka, b — byczek},
6. HOR_WZR — genotyp hormonu wzrostu $\{1 - AA, 2 - AB, 3 - BB\}$,
7. MASA_CIE — masa cielęcia,
8. MASA_CIEL — masa cielęcia 2,
9. PRZYR_sR — przyrost masy ciała,
10. sR_MASA — średnia masa ciała,
11. MASA_STD_ (atrybut wyjściowy) — standaryzowana masa ciała krowy.

Atrybut wyjściowy (regresja)

MASA_STD_ — standaryzowana masa ciała krowy; zakres: $184 \div 349$, średnia: 266.68, odchylenie std.: 32.83.

5. „Mastitis”

- Cel badań — sprawdzenie możliwości przewidywania liczby komórek somatycznych w mleku (lub klasy liczby komórek: dobra / zła).
- Nadmierny poziom liczby komórek somatycznych w mleku może świadczyć o chorobach wymion (stan zapalny — *mastitis*) i/lub nieprawidłowym dojeniu krów.
- W mleku klasy extra dopuszczalny poziom komórek somatycznych: 400 tysięcy. W mleku klasy A dopuszczalny poziom komórek somatycznych: 500 tysięcy.
- Zbiór danych zawiera: 596 przykładów i 18 atrybutów.
- Dwie wersje zadania: klasyfikacja i estymacja regresji.
- Dane zebrane na podstawie dokumentacji gospodarstw wiejskich.

5. „Mastitis” (c.d.)

Atrybuty wejściowe

1. KOL — kolejność rekordu,
2. NR_KROWY — numer krowy,
3. %HF — udział genów HF,
4. LAK — kolejna laktacja,
5. D_URODZ_ — data urodzenia,
6. D_WYC_ — data wycielenia,
7. DNI — liczba dni doju,
8. WIEK — wiek krowy,
9. TD — wydajność mleka w próbnym udoju,
10. TL — wydajność tłuszczu,
11. TL% — zawartość tłuszczu,
12. BL — wydajność białka,
13. BL% — zawartość białka,

5. „Mastitis” (c.d.)

Atrybuty wejściowe

14. %_LAKTOZY — zawartość procentowa laktozy w mleku,
15. SM — zawartość suchej masy w mleku,
16. MOCZNIK — zawartość mocznika.
17. LKS lub LKS_ (atrybut wyjściowy) — liczba komórek somatycznych w mleku (regresja); lub klasa liczby komórek {A — dobra, B — zła} (klasyfikacja).

Atrybut wyjściowy (regresja)

LKS — liczba komórek somatycznych w mleku (regresja); zakres: $6 \div 12944$, średnia: 650.61, odchylenie std.: 1249.49.
LKS_ — klasa liczby komórek {A — dobra, B — zła} (klasyfikacja).

6. „Przyrosty”

- Cel badań — sprawdzenie możliwości przewidywania klasy przyrostu masy ciała cieląt. W szczególności sprawdzenie wpływu genotypów miostatyny, leptyny oraz białka prionowego.
- Podział na klasy dokonany na podstawie średniej masy (poniżej, średnia lub powyżej).
- Zbiór danych zawiera: 261 przykładów i 18 atrybutów.
- Dane zebrane na podstawie dokumentacji gospodarstw wiejskich.

6. „Przyrosty” (c.d.)

Atrybuty wejściowe

1. RASA — kod rasy krowy {CHLH — mieszaniec Charolais-Heroford, CHL — Charolais, CHLS — Charolais-Simmental},
2. MR — masa ciała urodzeniowa,
3. WD2 — masa ciała,
4. MR2 — masa ciała przy odsadzeniu,
5. MS_210 — masa ciała w 210 dniu życia,
6. PS_210 — przyrosty w 210 dniu życia,
7. MLECZNOS — mleczność krowy matki,
8. SEZON — sezon wycielenia {ZIMA, LATO},
9. M_C__PO — masa ciała po wycieleniu,
10. OMW — okres międzywycieleniowy,

6. „Przyrosty” (c.d.)

Atrybuty wejściowe

11. GDF8 — genotyp miostatyny $\{AA, AB, BB\}$,
12. LEP — genotyp leptyny $\{AA, AB, BB\}$,
13. PRNP — genotyp białka prionowego $\{AA, AB, BB\}$,
14. INSDEL12 — genotyp insercyjno-delecyjny białka PRNP $\{ins12, insdel12, del12\}$,
15. INSDEL23 — genotyp insercyjno-delecyjny białka PRNP $\{ins23, insdel23, del23\}$,
16. HAP — haplotyp (kombinacja poprzednich genotypów),
17. HAP1 — haplotyp kombinowany,

Atrybut wyjściowy (klasyfikacja)

PS_KL — klasa przyrostu {A — dobry, B — zły}.

Klasyfikacja i regresja

- naiwny klasyfikator Bayesa,
- drzewa decyzyjne CART,
- regresja liniowa i wielomianowa (klasyczne najmniejsze kwadraty LSQ),
- regresja liniowa i wielomianowa z regularyzacją L_1 (*ridge regression*),
- regresja liniowa i wielomianowa z regularyzacją L_2 (*lasso regression*),
- regresja logistyczna.

Meta-klasyfikatory

- techniki: bagging, boosting, stacking,
- AdaBoost + decision stump,
- RealBoost,
- ResponseBinningBoost.

Indukcja reguł

- wyczerpujące wyszukiwanie reguł decyzyjnych,
- wykrywanie reguł Pareto-optymalnych,
- klasyfikatory regułowe,
- reguły asocjacyjne — algortym A priori.

Literatura

1. D. Hand, H. Mannila, P. Smyth, *Eksploracja danych*. WNT, Warszawa, 2005.
2. J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się*. WNT, Warszawa, 2005.
3. P. Cichosz, *Systemy uczące się*. WNT, 2007.
4. W. J. Ewens, G. R. Grant, *Statistical Methods in Bionformatics: An Introduction*, Springer, 2010, 2
5. A. D. Baxevanis, B. F. F. Quellette, *Bioinformatyka. Podrecznik do analizy genów i białek*, PWN, 2005