Sztuczna inteligencja w informatyce biomedycznej

Projekt 01: Analiza danych medycznych z wykorzystaniem tradycyjnych metod uczenia maszynowego

Szymon Wilk

28.02 i 7.03.2024

Spis treści

1 Wprowadzenie 2

2 Zadania 3

3 Literatura 5

1 Wprowadzenie

Cele pierwszego cyklu zajęć laboratoryjnych są następujące:

1. Zapoznanie się z rzeczywistymi zbiorami danych opisującymi wybrane problemy decyzyjne rozważane w izbie przyjęć szpitala dziecięcego (triaż[1] wybranych typów ostrego bólu u dzieci)
2. Przygotowanie środowiska eksperymentalnego (np. *Colab*) i przeprowadzenie analizy wskazanych udostępnionych zbiorów medycznych (zebranych w izbie przyjęć szpitala pediatrycznego) w celu budowy modeli decyzyjnych wspomagających *triaż*.

Dostępne są 4 rzeczywiste zbiory danych pozyskane podczas badań prospektywnych i retrospektywnych w izbie przyjęć szpitala pediatrycznego. Opisują one pacjentów z wybranymi typami ostrego bólu (są to problemy spotykane albo najczęściej, ale wskazane przez współpracujących ekspertów medycznych jako najbardziej wyzywające) i zawierają informację o poprawnym wstępnym postępowaniu.
Krótka charakterystyka zbioru dostępna jest w tabeli 1. Sekwencja klas przedstawia porządek klas z uwagi na ich istotność kliniczną – klasa najbardziej istotna jest oznaczona za pomocą „‼” (w przypadku astmy istotne są dwie klasy z uwagi na konieczność szybkiego podania leków sterydowych). Oryginalne zbiory zawierają zarówno atrybuty jakościowe, jak i ilościowe (numeryczne), przy czym dla tych ostatnich zdefiniowano przedziały dyskretyzacji na podstawie wiedzy eksperckiej. Wszystkie te zbiory są niezrównoważone – liczba obiektów w najważniejszej klasie jest zazwyczaj najmniejsza.

[1] Triaż (ang. *triage*) to określenie rodzaju pomocy oraz jej pilności dla danego pacjenta. Po triażu następuje diagnoza.

Tabela 1. Charakterystyka rozważanych zbiorów danych

Zbiór	# obiektów	# cech	# klas	Sekwencja klas	Uwagi
ap_pro	457	13	3	discharge → observation → consult!!	Ból brzucha (*abdominal pain*) Dane prospektywne, w kolumnie *Observer* dostępna decyzja lekarza
sp_retro	470	29	3	discharge → clinic → consult!!	Ból moszny (*scrotal pain*) Dane retrospektywne
hp_retro	413	24	3	discharge → xlab → lab_xray_bscan!!	Ból biodra (*hip pain*) Dane retrospektywne
ae_retro	427	48	3	short → long!! →admit!!	Atak astmy (*asthma exacerbation*) Dane retrospektywne Skupiamy się na decyzji podejmowanej po godzinie pobytu (decision = 60)

Warto wyjaśnić, że wszystkie cztery zbiory są scharakteryzowane za pomocą ograniczonego zbioru cech – odpowiadają one badaniom wykonywanym zaraz po rejestracji pacjenta w izbie przyjęć (stąd np. brak badań obrazowych). Opracowane na podstawie danych modele miały stanowić konkurencję dla często stosowanych w takich sytuacjach systemów punktowych (np. MANTRELS dla bólu brzucha, czy PRAM dla astmy).

2 Zadania

Zbuduj i oceń *możliwie najlepszy* model decyzyjny (klasyfikator) dla każdego zbioru danych. Zalecane jest wykonanie tego zadania w grupach 4-osobowych (ewentualnie mniejszych), w których jedna osoba zajmie się analizą wybranego zbioru danych. Do analizy wykorzystaj dostępne biblioteki implementujące tradycyjne techniki uczenia maszynowego oraz techniki wstępnego przetwarzania danych (`scikit-learn`, `imbalanced-learn`, `multi-imbalanced` [1]) – mogą się one okazać przydatne z uwagi na niezrównoważenie danych. Wykonując analizę zastosuj się do poniższych sugestii:

1. Dokonaj binaryzacji klasy decyzyjnej, gdzie *klasa pozytywna* obejmuje klasę lub klasy istotne (!!), natomiast klasa negatywna pozostałe klasy.

2. Zbuduj klasyfikator bazowy (proste rozwiązanie wykorzystujące typowe techniki, np. regresję logistyczną) oraz klasyfikator docelowy dopasowany do rozważanego problemu decyzyjnego. Klasyfikator docelowy może obejmować bardziej złożony *pipeline* z krokami odpowiedzialnymi za wstępne przetwarzanie danych uczących.
3. Do oceny klasyfikatora użyj 5-krotnej warstwowej walidacji krzyżowej (*stratified cross validation*) powtórzonej 3-krotnie (proponuję ustawić ziarno losowania na 42☺).
4. Podczas każdej iteracji wykonaj następujące czynności:
    a. Dokonaj oceny zaproponowanych klasyfikatorów (bazowy i docelowy) na zbiorze testowym wykorzystując miary AUPRC (*area under the precision-recall curve*) oraz AUROC (*area under the ROC curve*). AUPRC jest lepiej dopasowana do danych niezrównoważonych [2], AUROC stanowi jej uzupełnienie.
    b. Dla każdego z klasyfikatorów wyznacz na podstawie zbioru uczącego progi średniego i wysokiego ryzyka (*medium- and high-risk thresholds*) [3]:
        i. Próg *medium-risk* = próg, dla którego *sensitivity* >= 99%,
        ii. Próg *high-risk* = próg, dla którego *specificity* >= 90%.
    c. Zastosuj progi do klasyfikacji danych testowych wykorzystując ciągłą odpowiedź klasyfikatora (*response*):
        i. Jeśli *response* < *medium-risk*, wówczas decyzja = *negative*
        ii. Jeśli *response >= high-risk*, wówczas decyzja = *positive*
        iii. Domyślnie brak odpowiedzi („szara strefa” – umiarkowane ryzyko)
    d. Wyznacz miary *false-negative rate* (FNR) oraz *false-positive rate* (FPR) dla przykładów testowych zaklasyfikowanych odpowiednio jako *negative* i *positive*.
5. Po wykonaniu wszystkich iteracji (3*5-fcv) wyznacz uśrednione wielkości
    a. AUPRC i AUROC
    b. FNR i TNR
    c. Odsetek przykładów uczących zaklasyfikowanych jako *negative* i *positive* przy użyciu progów *medium-risk* oraz *high-risk*

Po wykonaniu analizy przygotuj krótki raport (~2 strony) przedstawiającą zaproponowane rozwiązanie dla danego problemu wraz z uzasadnieniem zastosowanego rozwiązania oraz

uzyskane wyniki. Prezentacje wykonane przez poszczególne osoby z grupy powinny zostać połączone tak, aby można było je przekazać łącznie.

Zadanie wykonaj do 21.03 i wyślij końcową prezentację przez eKursy.

3 Literatura

1. Multi-imbalance: open source Python toolbox for multi-class imbalance classification. http://www.cs.put.poznan.pl/mlango/publications/multiimbalance/
2. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015 Mar 4;10(3):e0118432. doi: 10.1371/journal.pone.0118432. PMID: 25738806; PMCID: PMC4349800.
3. Than MP, Pickering JW, Sandoval Y, Shah ASV, Tsanas A, Apple FS, Blankenberg S, Cullen L, Mueller C, Neumann JT, Twerenbold R, Westermann D, Beshiri A, Mills NL; MI3 Collaborative. Machine Learning to Predict the Likelihood of Acute Myocardial Infarction. Circulation. 2019 Sep 10;140(11):899-909. doi: 10.1161/CIRCULATIONAHA.119.041980. Epub 2019 Aug 16. PMID: 31416346; PMCID: PMC6749969.