UNIWERSYTET W BIAŁYMSTOKU
INSTYTUT INFORMATYKI

Mateusz JABŁOŃSKI

Efektywne metody selekcji cech dla predykcji punktów końcowych pacjentów z rakiem pęcherza moczowego

Promotor:
dr Aneta POLEWKO-KLIM

BIAŁYSTOK 2022

Spis treści

Wstęp

Według World Cancer Research Fund International, rak pęcherza moczowego jest jednym z najczęściej występujących nowotworów na świecie. Zajmuje 10 miejsce w rankingu najczęstrzej zachorowalności na raka. Jest to choroba diagnozowana przeważnie u osób powyżej 55 roku życia w wysokorozwiniętych krajach południowej i zachodniej Europie jak i w Północnej Ameryce. U mężczyzn zachorowalność na tego typu nowotwór jest 4 krotnie wyższa niżu kobiet. Najczęstrzymi czynnikami które zwiększają ryzyko zachorowania na raka pęcherza moczowego to: palenie papierosów, narażenie na działanie niektórych substancji chemicznych(takich jak aminy aromatyczne, wielopierścieniowe węglowodory aromatyczne węglowodory i chlorowane węglowodory oraz alkohol), dieta bogata w czerwone mięso oraz genetyczne obciążenie[**medsci8010015**]. Rak pechęrza moczowego możemy podzielić w zależności od wyników klinicznych i możliwością terapii. Rak pęcherza moczowego nieinwazyjny (NMIBC) i rak pęcherza moczowego inwazyjny (MIBC). MIBC są agresywnymi nowotworami, charakteryzującymi się pięcioletnim przeżyciem poniżej 50% [**10.1145/3136625**] .

W 15% przypadkach pierwotnie agresywny typ nowotworu (MIBC) jest wykrywany jako nieagresywny(NMIBC) który ulega pogorszeniu się do agresywnego(MIBC) [**CHEN2018214**] .

Celem niniejszej pracy jest projekt i implementacja algorytmu który pozwoli nam na predykcje punktów końcowych pacjentów z rakiem pęcherza moczowego z wykorzystaniem algorytmu nadzorowanego uczenia maszynowego. Testy efektywności stworzonego modelu zostaną przeprowadzone na danych "Tennis Major Tournament Match Statistics Data Set".

W pierwszej części mojej pracy przedstawię opis wybranych algorytmów selekcji cech oraz klasyfikatory, miary oceny jakości modelu oraz metody walidacji modelu.

W drugiej części mojej pracy przedstawie opis danych na których zbudowany będzie model uczenia maszynowego, opis język programowania w którymm model będzie napisany a także ocena skutczności modelu dla różnych slekecji cech. Implementacja podanych wyżej algorytmów zostanie przeprowadzona przy użyciu języka python.

Rozdział 1

Wybrane metody selekcji cech

Selekcja cech jest to strategia przetwarzania wysokowymiarowych danych do rozmaitych problemów takich jak analiza danych i uczenie maszynowe. Celem selekcji cech jest przede wszystkim budowa wysoce skutecznego modelu klasyfikatora [**10.1145/3136625**] , poprzez znalezienie jak najlepszego zestawu atrybutów(predyktorów) które to mają wpływ na skuteczność modelu.

1.1 Filtry

ReliefF

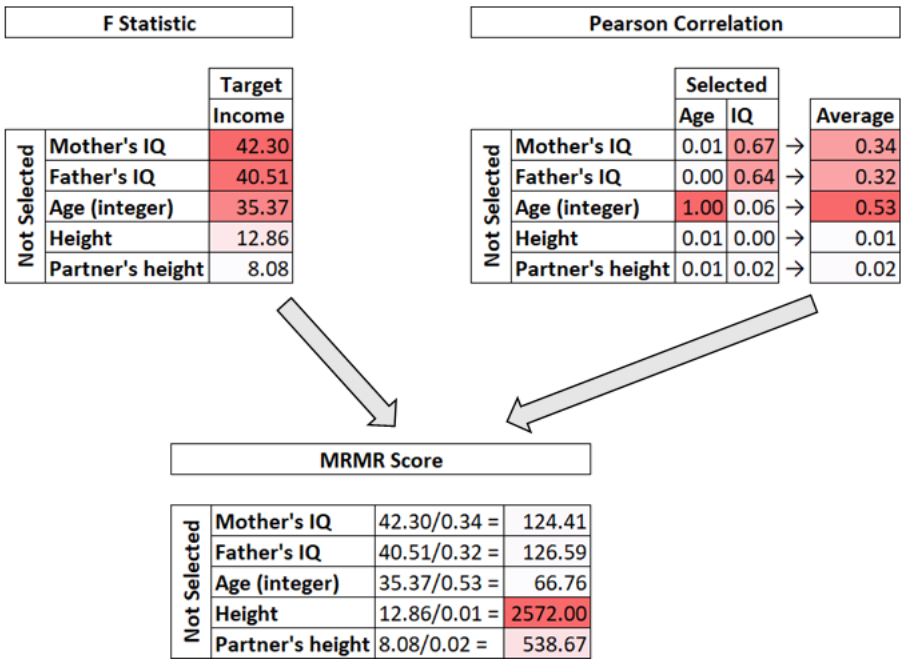
Główną koncepcją ReliefF polega na ocenie jakości cech poprzez ich zdolność do odróżniania poszczególnych przypadków z jednej klasy od innych w lokalnym sąsiedztwie, tzn. najlepsze cechy to te, które w większym stopniu przyczyniają się do zwiększenia dystansu pomiędzy różnymi instancjami klasowymi, natomiast w mniejszym stopniu przyczyniają się do zwiększenia dystansu pomiędzy instancjami tej samej klasy. ReliefF, jak wspomniano powyżej, jest rozszerzeniem oryginalnej metody Relief, która jest w stanie pracować z wieloklasowymi i niekompletnymi zbiorami danych.[**Palma-Mendoza**].

mRMR

mRMR (ang. minimum Redundancy - Maximum Relevance) jest metodą selekcji cech która ma preferencje do wybierania predyktorów o wysokiej zależnością z klasą i niskiej zależności między sobą. Jedną z możliwości opisania trafności jest użycie F-statistic, jest to wartość, którą dostajemy poprzez zastosowanie testu ANOVA lub analizy regrasji, aby dowiedzieć się czy średnia między dwiema populacjami różnią się znacząco. Natomiast do obliczenia redundacji możemy użyć współczynnik korelacji liniowej Pearson. Jest on wyrażony wzorem:

$$\mathcal{R}(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$$

gdzie X_i jest i-tym współczynnikiem wektora cech, var(X_i) - wariancją, a cov(X_i,Y)-kowariancją Następnie cechy są wybierane jedna po drugiej poprzez zastosowanie wyszukiwania zachłannego w celi maksymalizacji celu.



źródło: https://towardsdatascience.com/mrmr-explained-exactly-how-you-wished-someone-explained-to-you-9cf4ed27458b

Jak możemy zauważyć na powyższym przykładzie, następnym predyktorem który będzie najbardziej istotny jest ten z największym wynikiem czyli wzrost.

Mann Whitney U test

Mann Whitney U Test również znany jako "Wilcoxon Rank Sum Test" jest to nieparametryczny test statystyczny który służy do porwównywania różnic między dwiema niezależnymi grupami, gdy zmienna zależna jest porządkowa lub ciągła. Przykładem użycie testu może być, zrozumienie czy wynagrodzenie wyrażone w skali ciągłej, różniłoby się w zależności od poziomu wykształcenia, które to posiada dwie grupy: "szkoła średnia" i "uniwersytet". U test jest traktowany jako nieparametryczna alternatywa dla "Student's t-test", chociaż nie zawsze tak jest. Zaletą tego testu i przewagą nad "Student's t-test" jest fakt że można wyciągnąć różne wnioski na podstawie danych w stosunku od przyjetych założeń dotyczące ich rozkładu. Test U Manna-Whitneya wyrażamy wzorem:

$$U = R_{min(k)} - \frac{n_k(n_k + 1)}{2}$$

gdzie:

U - wynik testu U Manna-Whitneya

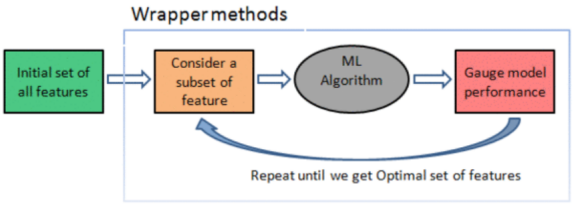
$R_{min(k)}$ - suma rang dla grupy, w której suma jest mniejsza

n_k - liczba obserwacji w grupie z mniejszą sumą rang

1.2 Wrappery

W metodach wrapperowych przebieg wyboru cech bazuje na danym algorytmie uczenia maszynowego, który staramy się zaadaptować na danym zbiorze danych. Metody oceniają szereg modeli za pomocą procedur które dodają i/lub usuwają cechy by znaleźć optymalną kombinacje która zwiększa wydajność modelu. Procedury te bazują na podejściu techniki (algorytmu) zachłannego wyszukiwania (ang. greedy algorithms). Algorytm zachłanny polega na podejmowaniu decyzji która w danym momencie wydaje się być najkorzystniejsza.

Schemat przepływu - metody Wrapper:



źródło:https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python/

Istnieją trzy kierunki przeprowadzenia procedur:

- Selekcja w przód (ang. Forward selection)
- Eliminacja wsteczna (ang. Backward elimination)
- Selekcja krokowa (ang. Step-wise selection)

Selekcja w przód

W selekcji w przód startujemy z modelem pustym po czym zaczynamy dopasowywać model z każdym pojedynczym predyktorem po kolei i wybieramy atrybut z minimalną wartością p (prawdopodobieństwem testowym). Po wybraniu pierwszej cechy dopasowujemy model z dwoma cechami, próbując kombinacji pierwszej cechy z resztą predyktorów. Ponownie wybieramy atrybut z minimalną wartością p. Kontynuujemy ten proces do momentu uzyskania zestawu wybranych cech z wartością p o poszczególnych predyktorów mniejszą od poziomu istotności.

Eliminacja wsteczna

W eliminacji wstecznej zaczynamy od pełnego modelu po czym zaczynamy usuwać nieistotne cechy z najwyższą wartością p (poziom istotności), proces ten powtarza się do momentu aż uzyskamy ostateczny zestaw istotnych cech.

Selekcja krokowa

Eliminacja ta jest podobna do selekcji w przód, różnicą jest mechanizm, który podczas dodawania nowej cechy sprawdza również istotność już dodanych wcześniej predyktorów i jeśli natrafi na jakiś mało znaczący atrybut, usuwa go poprzez eliminacje wsteczną. Stąd selekcja krokowa to połączenie dwóch poprzednich eliminacji.

1.3 Metody wbudowane

W metodach wbudowanych algorytmy selekcji cech są ujednolicone z algorytem uczenia maszynowego. Metody te zawierają w sobie cechy filtrów a także wraperów. Algorytmy uczące które posiadają własne metody selekcji cech wykonują w tym samym momencie selekcje i klasyfikacje. Najbardziej popularną techniką wbudowną są algorytmy drzewiaste takie jak **RandomForest, ExtraTree** i tak dalej. Algorytmy drzewiaste wybierają predyktor w każdym kroku rekurencyjnym procesu wzrostu drzewa i dzielą zbiór próbek na mniejsze podzbiory. Im więcej węzłów posiada "dzieci"w podzbiorze w tej samej klasie, tym bardziej istotne są predyktory. Inne metody wbudowane to **LASSO z regularyzacją L1 i Ridge z regularyzacją L2** do konstruowania modelu liniowego. Te dwie metody zmniejszą wiele cech do wartości zera lub blisku zera.

/TODO

Rozdział 2

Klasyfikacja nadzorowana

2.1 Budowa klasyfikatora

2.2 Metody walidacji modelu

2.3 Metryki oceny skuteczności klasyfikatora modelu

2.4 Wybrane algorytmy klasyfikacji danych molekularnych

Rozdział 3

Opis zbiorów danych

Rozdział 4

Projekt i implementacja klasyfikatora punktów końcowych pacjentów z rakiem pęcherza moczowego z selekcją cech

Rozdział 5

Wybrane wyniki budowy i oceny skuteczności opracowanego modelu dla różnych metod selekcji cech

5.1 Porównanie zmiennych informacyjnych

5.2 Skuteczność klasyfikatora dla różnych metod selekcji cech

5.3 Wydajność czasowa i pamięciowa opracowanych algorytmów

Spis rysunków

Spis tabel