Raport

Projekt: Filtrowanie reklam

Autor: Mateusz Bieliński

1. Wstęp i cel badania

Celem niniejszego badania było opracowanie modelu predykcyjnego będącego w stanie skutecznie klasyfikować strony internetowe pod względem obecności reklam. Model analizuje cechy takie jak rozmiary, adresy URL oraz atrybuty alt i caption elementów HTML oraz adresy URL stron internetowych. Na ich podstawie jest w stanie przewidzieć czy strona zawiera reklamę przypisując klasę ad lub nonad.

2. Opis danych i ich cechy

Zbiór danych składa się z 3 plików: ad.data, ad.DOCUMENTATION i ad.names. Ad.data zawiera rekordy z danymi, w ad.DOCUMENTATION znajduje się dokumentacja danych, a plik ad.names zawiera nazwy atrybutów i zakresy wartości. Zbiór danych liczy 3279 rekordów i składa się z 1558 atrybutów i jednej klasy decyzyjnej (ad/nonad). Wartości klasy decyzyjnej określają czy strona zawiera reklamę czy nie. Zbiór dzieli się na 2820 rekordy nonad i 459 rekordów ad. Wśród atrybutów można wymienić:

- height – wysokość elementu, typ integer (oraz ?),
- width – szerokość elementu, typ integer (oraz ?),
- aratio – stosunek wysokości do szerokości, typ real (oraz ?),
- local – wskazuje czy element jest lokalny, typ integer (wartości 0/1)
- atrybuty url* - wskazują czy adres URL strony zawiera ciąg znaków zawarty w nazwie atrybutu, typ integer (wartości 0/1),
- atrybuty origurl* - wskazują czy adres URL, z którego element został pobrany zawiera ciąg znaków zawarty w nazwie atrybutu, typ integer (wartości 0/1),
- atrybuty ancurl* - wskazują czy atrybut href w elemencie anchor (<a>) zawiera ciąg znaków zawarty w nazwie atrybutu, typ integer (wartości 0/1),
- atrybuty alt* - wskazują czy atrybut alt w elementach HTML zawiera ciąg znaków zawarty w nazwie atrybutu, typ integer (wartości 0/1),

- atrybuty caption* - wskazują czy atrybut caption w elementach HTML zawiera ciąg znaków zawarty w nazwie atrybutu, typ integer (wartości 0/1).

3. Metodologia i rozwiązanie

Pierwszym krokiem było przygotowanie zbioru danych. W tym celu zaimplementowano program w języku Python do połączenia plików ad.data i ad.names w jeden plik CSV. Najpierw przetworzono plik ad.names, odrzucając zbędne wiersze i przekształcając pozostałe w jeden wiersz będący nagłówkiem. Następnie wczytano wiersze z pliku ad.data, dołączono je do nagłówka i zapisano całość w nowym pliku CSV.

Do stworzenia modelu wykorzystano program AI Studio (wcześniej RapidMiner). W programie do trenowania i klasyfikacji został wybrany model drzewa decyzyjnego (operator Decision Tree).

4. Metoda oceniania jakości modelu

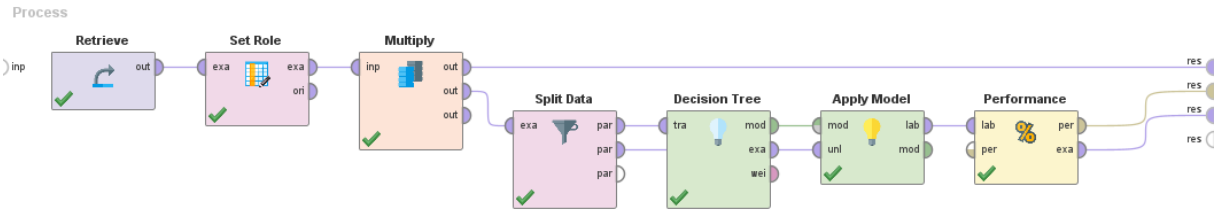
Do oceny modelu została wykorzystana miara Accuracy (dokładność). Jest to miara określająca, jak dobrze model przewiduje klasy na podstawie dostarczonych danych i jest stosunkiem liczby poprawnych predykcji do całkowitej liczby predykcji.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

W kontekście niniejszego projektu będzie to stosunek poprawnych klasyfikacji obecności i braków obecności reklam na stronach internetowych do wszystkich klasyfikacji.

5. Wyniki eksperymentalne

a) Test klasyfikatora na danych oryginalnych



W operatorze Set Role ustawiono atrybut class na label. W operatorze Split Data podzielono zbiór na 70% danych treningowych i 30% danych testowych. W operatorze Decision Tree pozostawiono domyślne parametry, w tym criterion: accuracy.

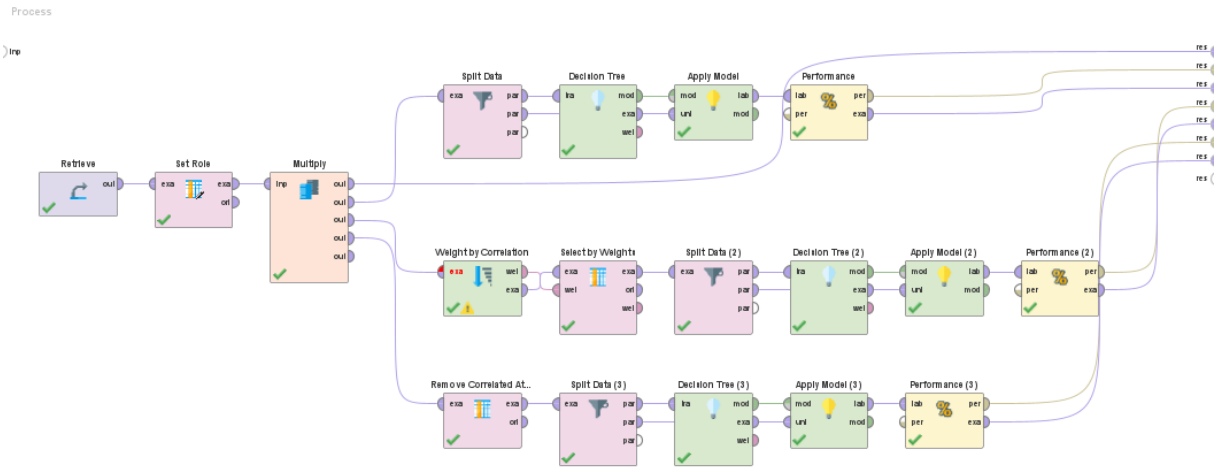
Wyniki:

accuracy: 96.04%

	true ad.	true nonad.	class precision
pred. ad.	108	9	92.31%
pred. nonad.	30	837	96.54%
class recall	78.26%	98.94%	

b) Test klasyfikatora na danych po selekcji

Selekcja cech została przeprowadzona na dwa sposoby. Pierwszy z nich to użycie operatorów Weight by Correlation i Select by Weights (selekcja atrybutów o największej korelacji z wynikiem), drugi to użycie operatora Remove Correlated Attributes (usunięcie skorelowanych ze sobą atrybutów).



Wyniki dla Weight by Correlation i Select by Weights:

Parametr weight w Select by Weights ustawiono na 0.2. W wyniku selekcji zostało 70 atrybutów.

accuracy: 96.04%

	true ad.	true nonad.	class precision
pred. ad.	108	9	92.31%
pred. nonad.	30	837	96.54%
class recall	78.26%	98.94%	

Wobec takich samych wyników jak w przypadku braku selekcji postanowiono zmienić weight na 0.3. Przy takiej wartości zostało 27 atrybutów.

accuracy: 95.12%

	true ad.	true nonad.	class precision
pred. ad.	99	9	91.67%
pred. nonad.	39	837	95.55%
class recall	71.74%	98.94%	

Wyniki dla Remove Correlated Attributes:

Correlation w operatorze Remove Correlated Attributes pozostawiono na 0.95.

accuracy: 95.93%

	true ad.	true nonad.	class precision
pred. ad.	106	8	92.98%
pred. nonad.	32	838	96.32%
class recall	76.81%	99.05%	

6. Podsumowanie i wnioski

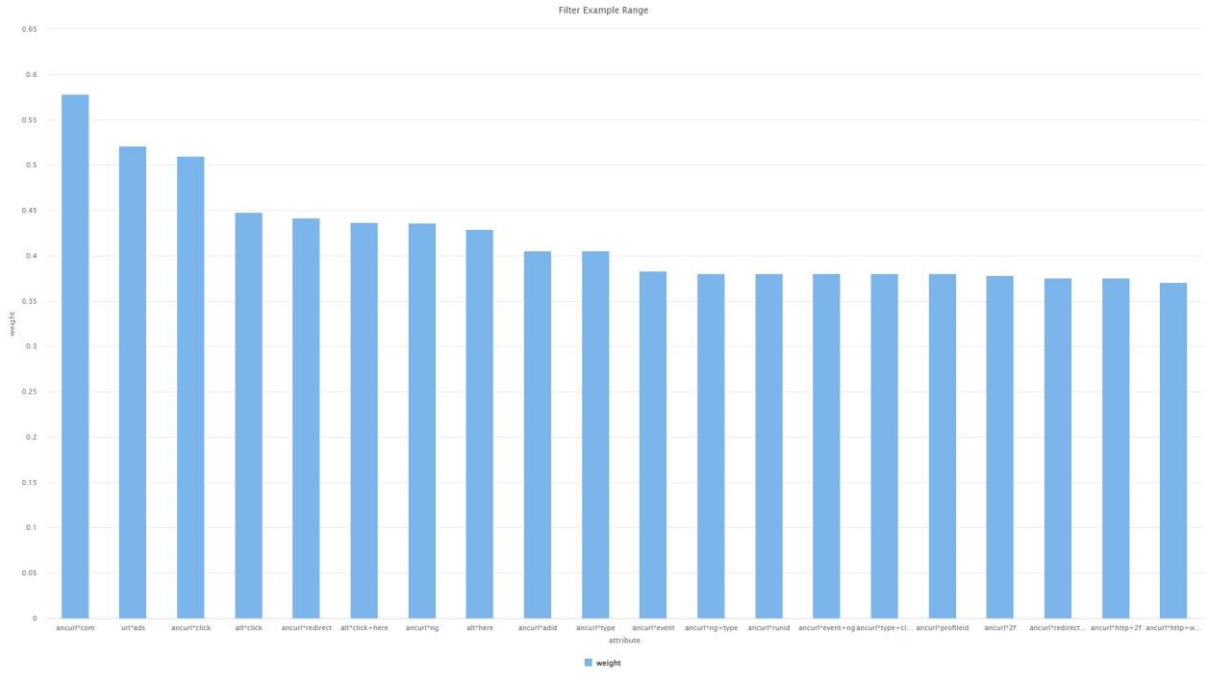
Program AI Studio pozwolił na zrealizowanie postawionego zadania. Jest to dobre narzędzie do przeprowadzania analizy danych i stosowania klasyfikatorów. Klasyfikacje przeprowadzono dla 4 różnych konfiguracji.

	Brak selekcji	Weight by Correlation i Select by Weights (weight 0.2)	Weight by Correlation i Select by Weights (weight 0.3)	Remove Correlated Attributes
Atrybuty	1558	70	27	662
Accuracy	96.04	96.04	95.12	95.93

Z powyższej tabeli wynika, że dokładność klasyfikatora nie jest większa po selekcji cech. Natomiast jest taka sama mimo usunięcia znacznej ilości atrybutów w przypadku selekcji przez „Weight by Correlation i Select by Weights (weight 0.2)", gdzie ilość atrybutów zmniejszyła się z 1558 do 70. W przypadku pozostałych konfiguracji z selekcją cech wartość dokładności spadła nieznaczne wobec usunięcia znacznej ilości atrybutów. Pokazuje to, że większość cech jest nieistotna w klasyfikacji.

Warto zauważyć, że „Weight by Correlation i Select by Weights (weight 0.2)" ma lepszą dokładność względem „Remove Correlated Attributes" mimo, że druga metoda posiada 592 atrybutów więcej. Wskazuje to, że istotny jest sposób selekcji i wybranie odpowiednich cech.

Poniższy wykres przedstawia 20 cech z największymi wagami:



Na powyższym wykresie można zauważyć, że wśród cech z największymi wagami znajdują się ciągi znaków w adresach URL elementów <a>, np. „com”, „click”, „redirect”. Ponadto wśród największych wag znajdują się między innymi adresy URL stron zawierające „ads” i atrybuty alt „click”.

Moim zdaniem dokładności modelu od 95.1% do 96.04% to dobry wynik i świadczy o tym, że klasyfikator jest skuteczny w rozpoznawaniu stron z reklamami.