ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ 2012
Seria: ORGANIZACJA I ZARZĄDZANIE z. 63a Nr kol. 1891

Jakub BRYLAK, Jacek GRUBER, Ireneusz J. JÓŹWIAK
Politechnika Wrocławska
Instytut Informatyki

ZASTOSOWANIE EKSPLORACJI DANYCH DO WYKRYWANIA REKLAM W OBRAZACH NA STRONACH WWW

Streszczenie. W artykule przedstawiono sposób tworzenia klasyfikatorów, które mają posłużyć do wykrywania reklam w obrazach na stronach WWW. Opisano postać i rodzaj danych, jakie zostały wykorzystane do wytworzenia modeli. Omówiono wszystkie etapy tego procesu tworzenia modelu, na który składają się filtracja dostępnych danych, dobór odpowiednich zmiennych, ocena utworzonych klasyfikatorów.

APPLICATION OF DATA MINING IN DETECTING ADVERTISEMENTS IN IMAGES FROM WEB PAGES

Summary. This article describes how to create classifiers, which are used to detect advertising images on web pages. The character and type of data that were used to produce models were also discussed. Modeling process consists of filtering the data, selection of appropriate variables, the evaluation of classifiers created, has also been described.

1. Wprowadzenie

Serwisy internetowe często przepełnione są wszelkiego rodzaju reklamami. Zamieszczane reklamy są dla właścicieli serwisów źródłem znaczących zarobków. Natomiast dla użytkowników portali wygodne jest to, że często mogą korzystać z takich serwisów nieodpłatne. Niestety, wadą tego kompromisu jest to, że zamieszczane na portalach reklamy bywają uciążliwe. Istnieją narzędzia w postaci wtyczek do przeglądarek internetowych, które potrafią rozpoznawać elementy strony, będące reklamą. Użytkownik może te wtyczki aktywować lub ustawić, co skutkuje blokowaniem wyświetlania reklam. Rozpoznanie reklam na postawie danych o obrazie prezentowanym w przeglądarce, często jest ważnym zadaniem.

Istotne w tym przypadku jest utworzenie skutecznego klasyfikatora, który jedynie z danych wyłuskiwanych z treści prezentowanych użytkownikowi danego serwisu będzie rozpoznawać reklamy [3]. Badania omawiane w pracy przeprowadzono za pomocą narzędzi i metod dostępnych w pakiecie Statistica 10 PL [2], [4].

2. Dane wykorzystane do badań

Do badań wykorzystano zbiór danych zawierający informacje o obrazach, które były zamieszczane na stronach internetowych, a następnie przesyłane i prezentowane użytkownikowi – oczywiście były to zarówno reklamy, jak i inne obrazy. Nazwy zmiennych i ich zakresy w wybranym do badań pliku danych przedstawiono w tabeli 1.

Tabela 1

Nazwy zmiennych i ich zakresy w jednym z wziętych do badań plików danych

Definicje typów danych wziętych do badań	
ad, nonad | classes.	c.d. definicji z lewej kolumny
	ancurl*www.slake.com: 0,1.
height: continuous.	ancurl*cnet+cat: 0,1.
width: continuous.	ancurl*hydrogeologist: 0,1.
aratio: continuous.	ancurl*geoguide: 0,1.
local: 0,1.	ancurl*2fcrawler: 0,1.
| 457 features from url terms	ancurl*clawring+htm: 0,1.
url*images+buttons: 0,1.	ancurl*tkaine+kats: 0,1.
url*likesbooks.com: 0,1.	ancurl*labyrinth: 0,1.
url*www.slake.com: 0,1.	ancurl*clickthru+clickid: 0,1.
url*hydrogeologist: 0,1.	…
url*oso: 0,1.	| 111 features from alt terms
url*media: 0,1.	alt*your: 0,1.
…	alt*and: 0,1.
| 495 features from origurl terms	alt*top: 0,1.
origurl*labyrinth: 0,1.	alt*all: 0,1.
origurl*puc.edu: 0,1.	alt*email: 0,1.
origurl*charlie+charlie: 0,1.	…
origurl*hevern+psychref: 0,1.	| 19 features from caption terms
origurl*www.truluck.com: 0,1.	caption*and: 0,1.
…	caption*home+page: 0,1.
| 472 features from ancurl terms	caption*click+here: 0,1.

ancurl*search+direct: 0,1.	caption*the: 0,1.
ancurl*likesbooks.com: 0,1.	caption*pratchett: 0,1.
ancurl*mirror: 0,1.	…

Źródło: [5]

Pod uwagę wzięto dane o geometrii obrazów, czyli wysokość i szerokość. Wartości innych cech wziętych do badań, to informacje o adresie, pod jakim był przechowywany obraz, jeszcze inne wzięte pod uwagę dane dotyczyły fraz przechowywanych w atrybucie ‘alt’ oraz danych związanych z opisem obrazu. Trzy ze zmiennych występujących w zbiorze wziętym do badań były zmiennymi ilościowymi – wysokość, szerokość, kanał alfa. Reszta to zmienne jakościowe, przyjmujące dwie wartości. Zbiór danych składa się z 3279 wierszy. Każdy wiersz dotyczy jednego przypadku wyświetlanego obrazu. W omawianym zbiorze zawartych jest 2821 wierszy dotyczących obrazów niebędących reklamami oraz 458 wierszy dotyczących reklam. W tabeli 1 przedstawiono przykładowe nazwy zmiennych oraz ich zakresy podane w jednym z udostępnionych do badań plików [5].

3. Filtrowanie danych

Pierwszym etapem procesu, którego celem jest utworzenie klasyfikatora, jest wstępne przetworzenie oraz filtrowanie zbioru danych. Omawiany zbiór zawiera 28% niepełnych przypadków. W arkuszu dostępnych jest 3279 przypadków, stąd zdecydowano się owe 28% przypadków z całości procesu wyłączyć. Ilość pozostałych przypadków powinna być wystarczająca do budowy klasyfikatora. Po odrzuceniu przypadków niekompletnych, w zbiorze pozostaje 379 reklam, co stanowi około 11,5% liczności całego zbioru danych. Przy tak niezrównoważonej częstości grup, trudno jest zbudować poprawny model. Najprostszym rozwiązaniem tego problemu jest wylosowanie z danych próby o zrównoważonych licznościach. Do badań wykorzysta się 100% przypadków będących reklamami oraz wylosuje się około 19% przypadków obrazów niebędących reklamami. Taki zabieg da nam zbiór, w którym grupy będą miały bliskie sobie liczności (po około 50% przypadków na grupę). Po wykonaniu tych operacji zbiór zawiera 753 przypadków danych o obrazach. Ostatnią czynnością związaną ze wstępnym filtrowaniem danych jest odrzucenie zmiennych, które nie wnoszą ze sobą zmienności. Usunięto zmienne, których względne odchylenie standardowe wynosi 10^{-10}.

Po przeprowadzeniu omawianych czynności zbiór, który będzie wykorzystywany do dalszych etapów, zawiera 753 wierszy oraz 1332 zmienne.

4. Wybór zmiennych

Czynnością, która nastąpiła przed wyborem istotnych zmiennych, było wydzielenie zbioru testowego, który zawierał 20% przypadków. Często zdarza się, że dane zawierają zmienne, które przenoszą tę samą informację. W tym przypadku do badań wykorzystano zmienne, których współczynnik korelacji wynosi co najmniej 0,85. Okazało się, że dla tego zbioru danych zmienna nie została odrzucona.

Następny etap polega na usunięciu tych zmiennych, które nie mają wpływu na wielkość, którą chcemy przewidywać. W naszym przypadku trzeba usunąć zmienne niemające wpływu na sprawdzenie, czy dany obraz jest reklamą. W pakiecie Statistica 10 PL zaimplementowano metodę „Szybki dobór zmiennych", która potrafi samodzielnie wydzielić zmienne mające największe znaczenie. Zdecydowano się wydzielić 15 spośród 1331 zmiennych.

5. Ocena utworzonych modeli

Na potrzeby opisywanych badań zastosowano cztery spośród wielu dostępnych w tym programie metod budowy modeli klasyfikacyjnych do odkrywania wiedzy z danych. Należy nadmienić, że wszystkie cztery zastosowane metody mają także odmiany przeznaczone do modelowania regresyjnego i one również są zaimplementowane w programie Statistica 10 PL.

Modele Data Mining pozyskiwania wiedzy z danych są metodami analitycznymi tworzenia modeli opartych na danych, a nie na wiedzy analityka – jest to podejście nieparametryczne, eksploracyjne.

Istnieją cztery główne zadania realizowane przez modele Data Mining: pierwszym jest opis zależności regresyjnych – modele regresyjne realizują problemy regresyjne, odwzorowując wektory wejściowych zmiennych ciągłych i dyskretnych w pewną ilościową zmienną wyjściową (zmienną zależną). Drugim jest problem klasyfikacji wzorcowych, jako odwzorowanie zmiennych wejściowych ciągłych i dyskretnych na dyskretną zmienną jakościową dychotomizowaną, która jest zwykle dwustanowa – takie zadanie zostało postawione do realizacji w ramach omawianych w niniejszym artykule badań.

Trzecim jest problem klasyfikacji bezwzorcowej, sprowadzający się do badania i określenia struktury zbioru analizowanych obiektów i łączenia ich w klasy, czyli dokonywania segmentacji elementów zbioru – jedną z realizacji tego jest metoda analizy skupień. Czwartym jest prognozowanie szeregów czasowych – ta metoda jest podobna do problemu regresyjnego i polega na opisywaniu wartości, którą chcielibyśmy otrzymać w przyszłości na podstawie wartości obserwowanych w przeszłości.

Do badań wykorzystano następujące metody:

1. Drzewa C&RT do budowy modelu drzew klasyfikujących.
2. „Losowy las" (ang. „Random Forest"), będący metodą realizacji algorytmu Breaimana do przewidywania wartości klasyfikacyjnych.
3. Wzmacniane drzewa klasyfikacyjne (i regresyjne), stanowiące implementację metody stochastycznego, gradientowego wzmacniania drzew do modelowania klasyfikacyjnego (i regresyjnego). Do badań wykorzystano drzewa klasyfikujące.
4. Sieci neuronowe, zastosowane w badaniach do tworzenia modelu klasyfikacyjnego.

W procesie modelowania wykorzystywano w pełnym zakresie oferowane mechanizmy automatyzacji poszukiwania zarówno najbardziej odpowiedniej metody modelowania, jak i najlepszych parametrów modelu z zastosowaniem kontroli za pomocą odpowiednich miar ilościowych lub krzywych na specjalizowanych ich wykresach, obrazujących jakość modeli. Ta automatyzacja jest bardzo cenną funkcjonalnością programu Statistica, gdyż pozwala znacznie zredukować bardzo duże nakłady pracy, które badacz musiałby przeznaczyć na znalezienie odpowiedniego i prawie optymalnego modelu pozyskiwania wiedzy z danych. W sposób zautomatyzowany, w relatywnie krótkim czasie, można zbadać przydatność zwykle kliku oferowanych alternatywnie w ramach danej metody modeli, wybrać model najbardziej przydatny w sensie najmniejszych błędów klasyfikacji (lub regresji), a także niezwykle efektywnie i prawie optymalnie sparametryzować wybrane w procedurze automatycznej modele. Z łatwością można w ten sposób przeprowadzić automatyczne testowanie tysięcy sparametryzowanych modeli klasyfikacyjnych jednego, dwóch lub kilku wybranych automatycznie metod.

Alternatywą procedury automatycznej procedury doboru metod modelowania i doboru parametrów modeli jest uciążliwa i niezwykle pracochłonna procedura realizowana ręcznie przez badacza. Przykładowo, w automatycznym modelowaniu metodami sztucznych sieci neuronowych, w rozwiązywaniu problemu klasyfikacji za pomocą eksploracji danych, w przypadku modelowania metodą sztucznych sieci neuronowych, w procedurze automatycznej badane są dwie sztuczne sieci neuronowe: sieć MLP, czyli perceptron

wielowarstwowy z zastosowaniem liniowej funkcji agregującej, oraz sieć RBF, z agregacją radialną. Naturalnie obie te sztuczne sieci neuronowe realizują schemat uczenia z nauczycielem. Z łatwością można przetestować wiele tysięcy sparametryzowanych modeli zarówno MLP, jak i RBF, bez potrzeby wskazywania, którą z tych sieci – MLP czy RBF – należy wybrać. Program Statistica samodzielnie dokona wyboru typu modelu oraz doboru najlepszych, w sensie błędów, klasyfikacji parametrów modelu.

Tabela 2

Stopy błędów klasyfikacji dla zbiorów uczącego i testującego, otrzymane za pomocą badanych modeli

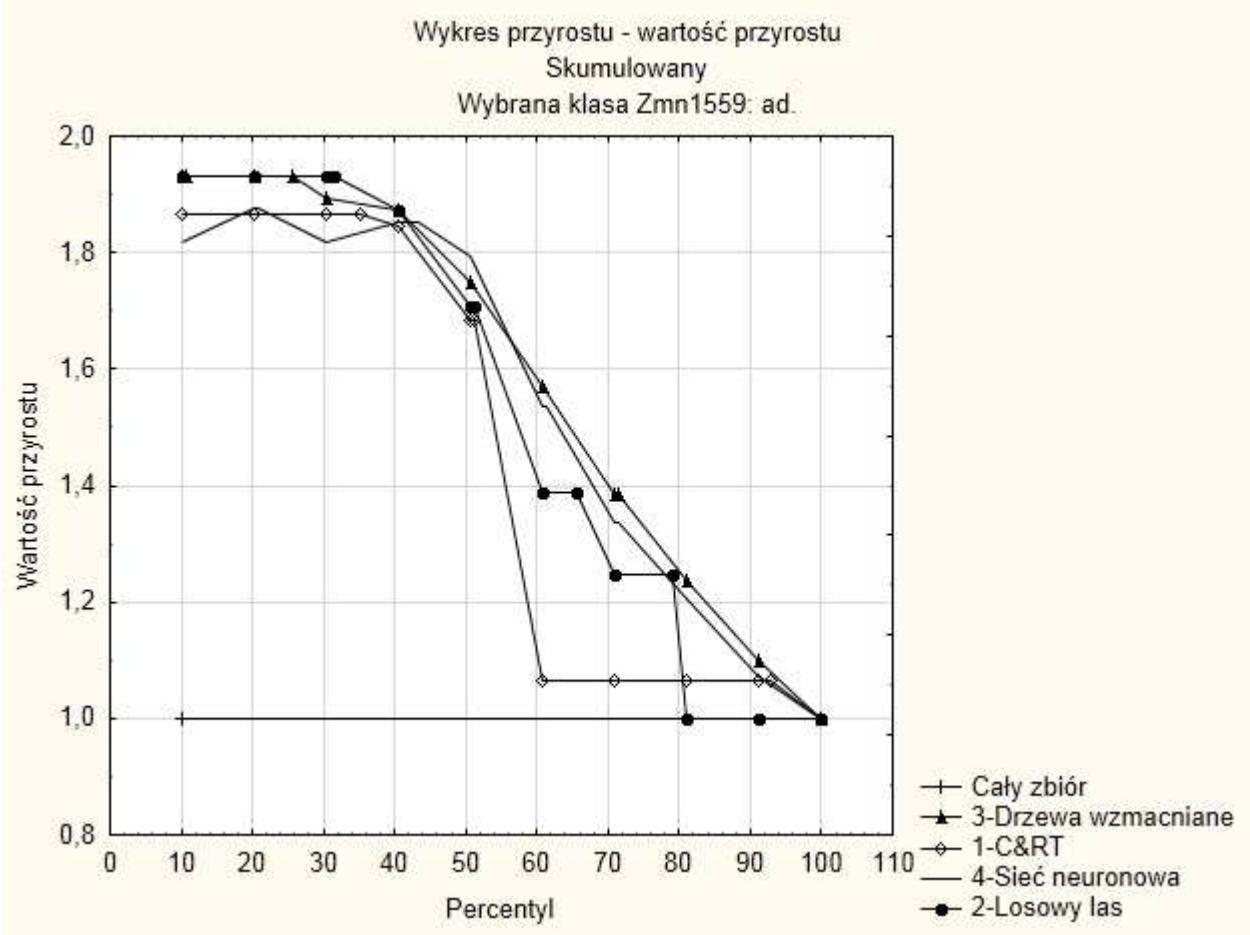
Model	Błąd (zbiór uczący) [%]	Błąd (zbiór testujący) [%]
Drzewa wzmacniane	2,39	10,71
C&RT	6,32	13,69
Sieć neuronowa	7,18	11,31
Losowy las	10,09	13,69

Źródło: Opracowanie własne.

Dla badanych modeli otrzymano wyniki przedstawione w tabeli 2. Obliczenia przeprowadzono za pomocą narzędzi i metod dostępnych w programie Statistica 10 PL, w grupie Data Mining.

Na podstawie błędu dla zbioru testującego można stwierdzić, że najlepszym modelem klasyfikującym jest model oparty na drzewach wzmacnianych. Stopy błędów przy zastosowaniu klasyfikatora opartego na drzewach wzmacnianych są najmniejsze zarówno dla zbioru uczącego, jak i dla zbioru testującego i wynoszą odpowiednio 2,39% i 10,71%. Przy wyborze modelu decydująca jest minimalna wartość stopy błędów dla zbioru testowego, z uwzględnieniem procentowego udziału tych przypadków w badanej próbie.

Do oceny modeli bardzo przydatny jest tzw. wykres przyrostu (*lift chart*), który pokazuje, o ile częściej niż w danych źródłowych przewidywana klasa występuje w próbie wskazanej przez dany model. Wartość wyznaczana jest na osi rzędnych, dla różnego stopnia pewności przewidywań modelu, wyznaczonych na osi odciętych, tzn. dla 10% przypadków, 20% przypadków itd. Uzyskana dla danego modelu krzywa powinna możliwie gładko spadać od wartości największej do 1, ponieważ gwałtowne skoki oznaczają, że model niezgodnie z rzeczywistością przewiduje szansę przynależności – tam gdzie szansa według modelu jest mniejsza, w rzeczywistości jest większa.



Rys. 1. Skumulowany wykres przyrostu dla czterech badanych modeli klasyfikacji [1]
Fig. 1. The cumulative growth chart for the four test models of classification [1]

Na rys. 1 przedstawiono wykres przyrostu, skumulowaną wartość przyrostu. Jak widać na wykresie na rys. 1, najlepiej zachowują się drzewa wzmacniane oraz model zbudowany na sieciach neuronowych. W szczególności dla percentyla 20%, równego procentowi próby wziętemu do badań jako próba testowa (zgodnie z wartościami 10,71% i 11,31% w trzeciej kolumnie tabeli 2). Krzywe wykreślone dla modeli wykorzystujących drzewa C&RT oraz algorytm „losowy las" opadają gwałtowniej oraz mają więcej gwałtownych załamań, więc zachowują się gorzej.

6. Podsumowanie

Wykrywanie reklam w obrazach na stronach internetowych jest trudnym zagadnieniem. Dzięki zastosowaniu technik data mining, problem ten jest jednak możliwy do rozwiązania z małą stopą błędu. Wyniki, jakie udało się osiągnąć, oscylują w zakresie 10,71% do 13,69%

błędu dla zbioru testującego. Najlepszym modelem wykorzystanym do rozwiązania problemu omawianego w tym artykule okazał się klasyfikator oparty na drzewach wzmacnianych.

Bibliografia

1. Aminian K.: Evaluation of Coalbed Methane Reservoirs. Petroleum & Natural Gas Engineering Department. West Virginia University, USA 2009.
2. Demski T.: Tworzenie i stosowanie modelu data mining za pomocą przepisów Statistica Data Miner na przykładzie wykrywania nadużyć. Statsoft, Kraków 2009.
3. Gibiec M.: Soft Computing tools for machine diagnosing. Journal of Theoretical and Applied Mechanics 2008, Vol. 42, No. 3, p. 483-501.
4. Gibiec M.: Zastosowana data mining w systemie monitorowania pracy kombajnów górniczych. Statsoft, Kraków 2008.
5. Internet Advertisements Data Set. http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements, 22/05/2012.

Abstract

The article describes how to create classifiers, which are used to detect advertising images from web pages. The data used to create models were discussed as well. The modeling uses the characteristics of images that can be read with HTML tags and URL patterns, for example, the width and height of the images. All stages of this process are described. This process involves filtering is based on the data, selection of appropriate variables, evaluation of classifiers created, such as enhanced trees, neural networks, C & RT regression trees and Random Forest.