Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Grzegorz Szpak
Nr albumu: 319400

Klasyfikacja wielowymiarowych szeregów czasowych przy ewoluujących pojęciach

Praca magisterska
na kierunku INFORMATYKA

Praca wykonana pod kierunkiem
dra Andrzeja Janusza
Instytut Informatyki

Grudzień 2016

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

W pracy przedstawiono sposoby wydajnej klasyfikacji szeregów czasowych dla danych pochodzących ze źródła o zmiennym rozkładzie. Opisane zostały metody ekstrakcji cech z wielowymiarowego szeregu czasowego. Autor opisuje także metody wyboru przestrzeni atrybutów odpornej na zmiany rozkładu źródła.

Słowa kluczowe

Eksploracja danych, wielowymiarowy szereg czasowy, ewoluujące pojęcia, dopasowanie dziedziny, ekstrakcja cech, selekcja cech, lasy losowe, regresja logistyczna, maszyna wektorów wspierających

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.4 Sztuczna inteligencja

Klasyfikacja tematyczna

D. Software
D.127. Blabalgorithms
D.127.6. Numerical blabalysis

Tytuł pracy w języku angielskim

Classification of multivariate time series in the presence of concept drift

Spis treści

Rozdział 1

Wprowadzenie

1.1. Uczenie z nadzorem a *concept drift*

Podstawowym problemem rozważanym w teorii uczenia maszynowego jest problem uczenia z nadzorem (ang. *supervised learning*). Niech dane będą:

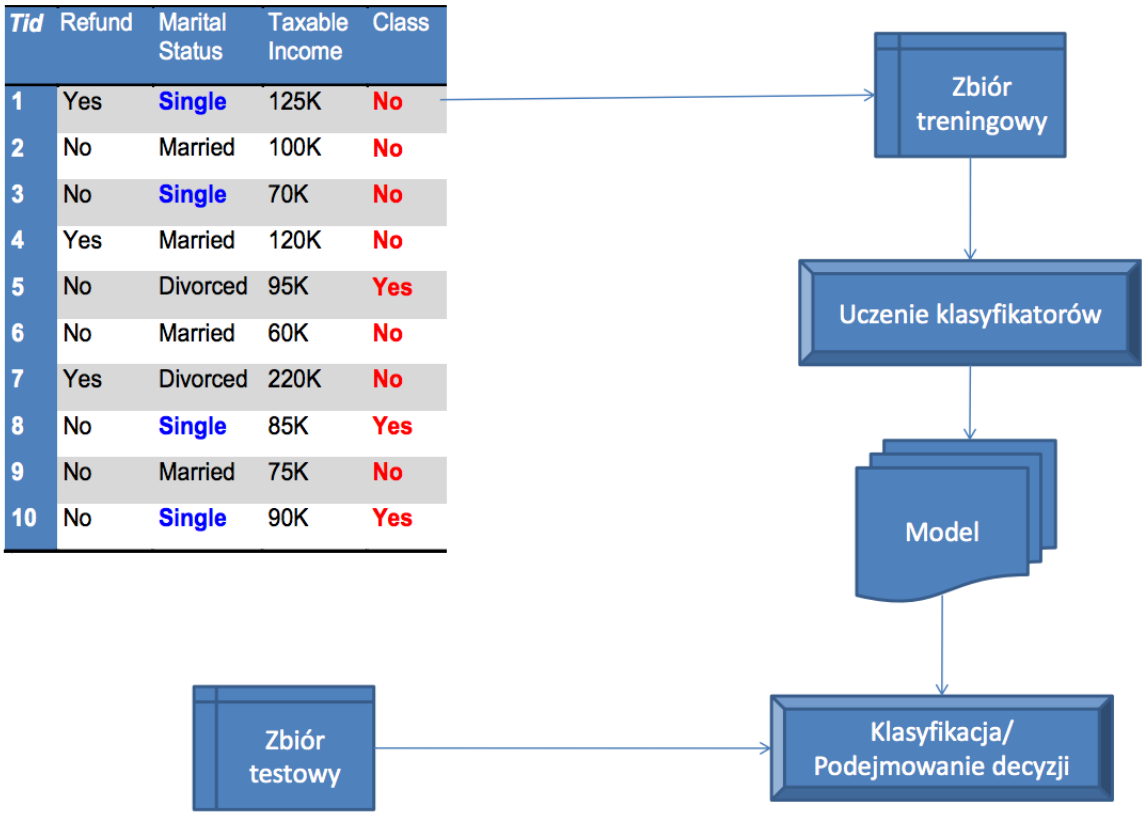
- zbiór X przykładów
- zbiór Y decyzji
- funkcja $f : X \to Y$
- Parę zbiorów $(X_{train} \subseteq X, Y_{train} \subseteq Y)$ instancji $x_1, x_2, \ldots, x_n$ oraz odpowiadających im decyzji $f(x_1), f(x_2), \ldots, f(x_n)$, zwaną *zbiorem treningowym*

Zadanie uczenia z nadzorem polega na wyznaczeniu na podstawie zbioru treningowego oraz przy użyciu pewnego algorytmu uczącego *klasyfikatora* takiej funkcji $c : X \to Y$ (zwanej *modelem*) będącej dobrą aproksymacją funkcji f. Jakość modelu c określa się, porównując jego wartości dla elementów skończonego *zbioru testowego* X_{test} z rzeczywistymi wartościami funkcji f dla tych elementów Y_{test}. Istotne jest przy tym założenie, że elementy zbiorów X_{train} oraz X_{test} losowane są ze zbioru X według tego samego rozkładu prawdopodobieństwa D.

Schemat rozwiązywania problemu uczenia z nadzorem przedstawia rys. 1.1.

Na przestrzeni ostatnich dziesięcioleci opracowanych zostało wiele algorytmów uczących wykazujących się dużą skutecznością w przeróżnych dziedzinach: od rozpoznawania obrazów, przez klasyfikację tekstów, rekomendację produktów, wykrywanie spamu, po przewidywanie zmian na giełdzie czy diagnostykę medyczną.

Sytuacja zmienia się diametralnie, gdy pominiemy założenie o równości rozkładów dla zbiorów treningowych i testowych. Problem ten nazywa się *ewolucją pojęć* (ang. *concept drift*) lub *dopasowaniem dziedziny* (ang. *domain adaptation*). Jest on szczególnie widoczny w zadaniach przetwarzania języka naturalnego (ang. *natural language processing*, w skrócie NLP). Rozpatrzmy dla przykładu problem rozpoznawania nazw własnych (ang. *named entity recognition*). Załóżmy, że klasyfikator uczony jest na podstawie danych encyklopedycznych oraz testowany na danych pochodzących z komunikatora internetowego. Obydwa zbiory, jakkolwiek powiązane, różnią się w znaczący sposób - przykładowo, szukanie wielkich liter może być bardzo pomocne w pierwszej dziedzinie, a nieść znacznie mniej informacji w wiadomościach z komunikatora.



Rysunek 1.1: Schemat uczenia z nadzorem

Stąd też właśnie w dziedzinie NLP powstało najwięcej metod mających rozwiązać problem ewoluujących pojęć. Przykładami takich metod są algorytm *structural correspondence learning* opisywany w [1] czy metoda odpowiedniego dopasowania przestrzeni parametrów zaproponowana przez Daume w [2].

Problem *domain adaptation* nie jest jednak często poruszany w przypadku klasyfikacji szeregów czasowych. W poniższej pracy autor przedstawia sposoby radzenia sobie z *concept drift* podczas klasyfikacji szeregów czasowych oraz wykonuje studium przypadku na wybranym zbiorze danych.

1.2. Formalizacja problemu

1.2.1. Paradygmaty uczenia się

Przyjmijmy definicje jak na początku sekcji 1.1. W zależności od rozkładów D_{train}, D_{test} oraz od dostępności zbiorów $Y_{train}, X_{test}, Y_{test}$, można (za [**?**]) zdefiniować inne paradygmaty uczenia.
I tak, jeśli zbiór Y_{train} jest nieznany w momencie tworzenia modelu, mamy do czynienia z *uczeniem bez nadzoru* (ang. *unsupervised learning*).
Gdy zbiór X_{test} nie jest znany podczas uczenia, mowa o *uczeniu indukcyjnym* (ang. *inductive learning*). W przeciwnym razie takie uczenie nazywa się *uczeniem transdukcyjnym* (ang.

transductive learning).
W powyższym przykładach istotne jest założenie, iż zbiory X_{train}, X_{test} pochodzą z tego samego rozkladu D. Odwrotna sytuacja rozpatrywana jest w paradygmacie *uczenia z przeniesieniem wiedzy* (ang. *transfer learning*). Przyjmuje się w nim, że dane są dwa różne rozkłady D^{source} i D^{target}. Model wyuczony na danych treningowych $X^{source}_{train}, Y^{source}_{train}$ wykorzystywany jest zatem do klasyfikacji zbioru testowego $X^{target}_{test}, Y^{target}_{test}$ pochodzących z rozkładu D^{target}. W poniższej pracy autor skupia się na problemie *dopasowania dziedziny*, który zakłada, że zbiór dostępnych klas Y jest ten sam dla D^{source} i D^{target}. Przeciwieństwem dopasowania dziedziny jest zadanie *uczenia wielozadaniowego* (ang. *multi-task learning*, więcej między innymi w [4]), gdzie zbiory X_{train}, X_{test} pochodzą z tego samego rozkładu, natomiast zbiory Y_{train}, Y_{test} są różne.

Powyższe rozważania podsumowuje tabela 1.1.

Tabela 1.1: Paradygmaty uczenia w teorii uczenia maszynowego. We wszystkich przypadkach zakładamy, że zbiór X_{train} jest dostępny podczas uczenia, podczas gdy zbiór Y_{test} nie jest znany.

Paradygmat	Y_{train} dostępny?	X_{test} dostępny?	Rozkład danych testowych
Indukcyjne uczenie bez nadzoru	Nie	Nie	D^{source}
Transdukcyjne uczenie bez nadzoru	Nie	Tak	D^{source}
Indukcyjne uczenie z nadzorem	Tak	Nie	D^{source}
Transdukcyjne uczenie z nadzorem	Tak	Tak	D^{source}
Indukcyjne uczenie bez nadzoru z przeniesieniem wiedzy	Nie	Nie	D^{target}
Transdukcyjne uczenie bez nadzoru z przeniesieniem wiedzy	Nie	Tak	D^{target}
Indukcyjne uczenie z nadzorem z przeniesieniem wiedzy	Tak	Nie	D^{target}
Transdukcyjne uczenie z nadzorem z przeniesieniem wiedzy	Tak	Tak	D^{target}

W poniższej pracy autor skupi się na problemie transdukcyjnego uczenia z nadzorem z przeniesieniem wiedzy. Przedstawione zostaną algorytmy, które wykorzystują dostępny zbiór X_{test} do znalezienia reprezentacji odpornej na zmiany rozkładu, co skutkować będzie zwiększoną jakością klasyfikacji w stosunku do standardowego podejścia opisanego w 1.1.

1.2.2. Ewolucja pojęć a *overfitting*

Mówiąc o problemie ewoluujących pojęć, należy wspomnieć o zagadnieniu przeuczenia (ang. *overfitting*). Polega on na zbudowaniu nadmiernie skomplikowanego modelu, co skutkuje słabą jego jakością.
Obydwa pojęcia mogą być mylone przy niewłaściwym sposobie walidacji modelu. Jeśli jakość klasyfikacji sprawdzana jest wyłącznie na zbiorach treningowym i testowym, zarówno *concept drift*, jak i *overfitting* dają podobne objawy - wysoki wynik na zbiorze treningowym oraz niski na zbiorze testowym. W przypadku przeuczenia jest to spowodowane nadmiernym dopasowaniem modelu do danych treningowych i jego niską zdolnością do uogólniania. Jeśli mamy do czynienia z ewoluującymi pojęciami, słaba jakość modelu jest spowodowana innym rozkładem dla zbioru testowego.
W rozróżnienu obydwu sytuacji pomagać może zastosowanie *walidacji krzyżowej* (ang. *cross-validation*) na zbiorze treningowym. Walidacja krzyżowa polega na podziale zbioru treningowego na n równolicznych części. Następnie budowane jest n modeli, przy czym $n-1$ części tworzy zbiór treningowy, natomiast pozostała część - zbiór testowy. Ostateczny wynik jest średnim wynikiem powstałych n modeli.
Nietrudno zauważyć, że w przypadku *concept drift* nie powinno się zauważyć znacznego spadku jakości modelu przy wykonaniu walidacji krzyżowej - w tym przypadku zbiór testowym pochodzi z tej samej dziedziny co treningowy. Inaczej będzie w przypadku przeuczenia - tu wynik walidacji krzyżowej będzie wyraźnie niższy niż wynik na zbiorze treningowym (tabela 1.2).

Tabela 1.2: *Concept drift* a *overfitting* - obniżony wynik modelu 1. przy walidacji krzyżowej świadczy o przeuczeniu, a nie występowaniu ewoluujących pojęć. W drugim przypadku model mimo dobrej umiejętności klasyfikacji elementów pochodzących z rozkładu D^{source}, cierpi na spadek jakości przy ewaluacji na zbiorze pochodzącym z rozkładu D^{target}.

	Wynik na X_{train}	Wynik CV	Wynik na X_{test}
Model 1	0.98	0.73	0.68
Model 2	0.97	0.92	0.76

Rozdział 2

Opis przeprowadzanego eksperymentu

2.1. Przedstawienie badanego zbioru danych

Dane, na których sprawdzana będzie jakość analizowanych algorytmów, pochodzą z konkursu *AAIA'15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene*[1] organizowanego przez Uniwersytet Warszawski oraz Szkołę Główną Służby Pożarniczej w Warszawie. Na potrzeby konkursu zebrano dane pochodzące z "inteligentnego kombinezonu", który monitoruje

[1] https://knowledgepit.fedcsis.org/contest/view.php?id=106

Bibliografia

[1] John Blitzer, Ryan McDonald, Fernando Pereira, *Domain adaptation with Structural Correspondence Learning*

[2] Hall Daume, *Frustratingly easy domain adaptation*

[3] Andrew Arnold, Ramesh Nallapati, William W. Cohen, *A comparative study of methods for transductive transfer learning*

[4] R. K. Ando, T. Zhang, *A framework for learning pre- dictive structures from multiple tasks and unlabeled data*

[Blar16] Elizjusz Blarbarucki, *O pewnych aspektach pewnych aspektów*, Astrolog Polski, Zeszyt 16, Warszawa 1916.