dr hab. Piotr Porwik
Uniwersystet Śląski, Katowice
Wydział Informatyki i Nauki o Materiałach
Zakład Systemów Komputerowych

Katowice 18.06.2016

OCENA ROZPRAWY DOKTORSKIEJ

"*Genetic and memetic algorithms for selection of training sets for support vector machines*"

mgra inż. Jakuba Nalepy

Ocena doboru tematu

Rozprawa doktorska opisuje zastosowania oraz modyfikacje algorytmów genetycznych i memetycznych w doborze zbiorów treningowych nadzorowanego klasyfikatora opartego na maszynie wektorów nośnych SVM (*Suppport Vector Machine*). Praca napisana została w języku angielskim. Tematyka pracy dobrze wpisuje się w bardzo obszerną oraz intensywnie rozwijaną dziedzinę sztucznej inteligencji - uczenia maszynowego. Na podstawie zgromadzonych doświadczeń (etykietowanych cech) maszyna uczy się w trybie predykcji rozpoznawania nowych, nieznanych danych w sposób możliwie najlepszy dla zadanego modelu uczenia. Klasycznym problemem klasyfikacji jest predykcja nieznanej etykiety na podstawie znanego wektora cech klasyfikowanego obiektu. Obserwując doniesienia światowej literatury można zauważyć, że klasyfikacja danych znajduje zastosowanie w wielu dziedzinach: w medycynie - jako element systemów wspomagania decyzji medycznych, biometrii - jako element biometrycznych systemów identyfikacji i weryfikacji osób, klasyfikacji danych masowych, czy w naukach technicznych, gdzie klasyfikatory stosowane są w diagnostyce technicznej. Zagadnienia klasyfikacji (klasyfikatory pojedyncze i zespoły klasyfikatorów) są stale anonsowane zarówno na konferencjach, jak i w artykułach publikowanych w takich periodykach jak Neurocomputing, Pattern Recognition, Applied Artificial Inteligence, Expert Systems with Applications i wielu innych - np. w książkach wydawanych także w Polsce. Obserwując publikowane tam prace naukowe zauważamy, że tendencja do podejmowania tych zagadnień utrzymuje się od lat, wliczając w to również rok 2016.

Powodem zainteresowania tematyką uczenia i ropoznawania maszynowego są możliwości analizy danych z różnych dziedzin, co z kolei wymusza (np. dla danych masowych) konieczność odpowiedniej selekcji lub redukcji cech obiektów dla skrócenia czasu analizy bez pogarszania jakości klasyfikacji. W swojej rozprawie Autor podjął próbę wpływania na liczbę danych treningowych klasyfikatora SVM za pomocą wybranych algorytmów

RAU
Biuro Dziekana
Wpłynęło dnia 22.06.2016
Nr 458 / zał. .....

genetycznych i memetycznych oraz ich parametrów. Tematyka rozprawy wpisuje się więc bardzo dobrze w nurt nowoczesnych i bardzo potrzebnych badań.

Charakterystyka rozprawy i jej zakres

Praca została przygotowana bardzo starannie i prezentuje zarówno wysoki poziom edytorski, jak i dobry poziom merytoryczny i naukowy. Lektura rozprawy wskazuje, że Autor włożył dużo pracy w przygotowanie tego opracowania.

Na stronie 4 Autor przedstawia tezy pracy doktorskiej, a na stronach 5 i 6 artykułuje podstawowe (7) oraz uzupełniające (2) cele, które zamierza osiągnąć. Od tego momentu cele te są konsekwentnie realizowane. Tezy pracy zostały przedstawione jasno i precyzyjnie. Lektura rozprawy, a wszczególności rozdział 6, gdzie przedstawione rozwiązania są walidowane, uzasadnia stwierdzenie, że tezy zostały poprawnie wykazane, a wytyczone cele zostały osiągnięte.

Praca podzielona została na kilka uzupełniających się części. W pierwszej, Autor opisuje strukturę pracy oraz przedstawia 5 współautorskich prac związanych bezpośrednio z realizowanym zadaniem. Do tej samej części (choć w różnych rozdziałach) zaliczyć można prezentację podstaw teoretycznych klasyfikatora pracującego w oparciu o maszynę wektorów nośnych oraz teorię algorytmów genetycznych i memetycznych. W drugiej części rozprawy zaprezentowane zostały szeroko eksperymenty walidacyjne klasyfikatora SVM prowadzone w oparciu o dokonane modyfikacje algorytmów ewolucyjnych.

W rozdziale 2, który ma charakter przeglądowy, przedstawione zostały podstawy teoretyczne metody wektorów nośnych (SVM), zarówno dla modelu liniowego jak i modelu nieliniowego z różnymi funkcjami jądrowymi. Ideą podejścia SVM w modelu liniowym jest wyznaczenie takiej hiperpłaszczyzny, aby była ona maksymalnie odległa od najbliższej obserwacji z próby uczącej. W większości przypadków klasy danych wejściowych nie są jednak liniowo separowalne, co oznacza, że nie da się dla nich skonstruować liniowej reguły klasyfikacyjnej.

W bardziej wymagającym modelu nieliniowym, zarejestrowane obserwacje transformuje się nieliniowo do przestrzeni wyższego wymiaru, a następnie stosuje w niej model liniowy. Podejście takie jest podstawą nieliniowych metod dyskryminacyjnych. Trzeba także zauważyć, że oprócz teoretyczych opisów SVM znanych z literatury już od 1971 roku (prace Vapnika), Autor przedstawił w tym rozdziale również własne przykłady obliczeń ilustrujące wpływ różnych funkcji jądrowych i ich parametrów na skuteczność klasyfikatora bazującego na wektorach nośnych (Tabela 2.2). Tutaj też została opisana neuro-rozmyta funkcja jądrowa opracowana przez Autora. Nie została jednak ona włączona do badań porównawczych (Tabela 2.3), gdyż doktorant uznał, że wykracza to poza zakres rozprawy.

W rozdziale 3 został przedstawiony problem trenowania klasyfikatora SVM na dużych zbiorach danych. Autor zauważa tutaj niebezpieczeństwa wynikające z niezrównoważenia danych. Tego typu dane (w szczególności dane medyczne) mogą niekorzystnie oddziały-

wać na ostateczną jakość klasyfikacji. Autor pominął w opisie tego zagadnienia ważny problem brakujących danych - co oznacza, że naprawdę duże zbiory danych mogą być nie tylko niezrównoważone (*imbalanced*), ale mogą być także niepełne (*missing values*), i to w różnych wymiarach: losowo lub nielosowo niepełne. Uniwersalne rozwiązania tego problemu z reguły zawodzą. Klasyfikatory trenowane na danych niezrównoważonych oraz niepełnych, np. z repozytoriów *UC Machine Learning*, wykazują często bardzo różne wyniki klasyfikacji. Praktyka pokazuje, że ten trudny problem wymaga rozwiązań dedykowanych dla konkretego typu danych. W tym samy rozdziale Autor przedstawia również możliwości przyspieszenia obliczeń poprzez ich zrównoleglenie lub redukcję danych, na których przeprowadza się trenowanie klasyfikatora.

Rozdział 4 zawiera informacje na temat zasad działania algorytmów ewolucyjnych, ze szczególnym uwzględnieniem algorytmów genetycznych i memetycznych.

W rozdziale 5, jednym z ważniejszych w rozprawie, Autor prezentuje 5 algorytmów ewolucyjnych, które zostały wykorzystane do selekcji danych treningowych klasyfikatora SVM. Są to modyfikacje 3 algorytmów genetycznych (GASVM, AGA oraz DAGA) i 2 algorytmów memetycznych (MASVM i PCA^{2}MA. Algorytm MASVM jest niekiedy myląco opisywany akronimem MA (np. na str. 59). Podstawowe zadania proponowanych algorytmów są dwojakiego rodzaju: mają dokonać heurystycznej selekcji danych treningowych do trenowania klasyfikatora SVM, a dynamiczne modyfikacje algorytmów mają zapewnić dobór parametrów sterujących procesem przeszukiwania przestrzeni rozwiązań. Proponowane przez Autora rozwiązania można zaliczyć do ewolucyjnego preprocessingu, który dzięki redukcji danych umożliwia w kolejnym etapie wybór wektorów nośnych klasyfikatora SVM, co w przypadku bez wstępnego przetwarzania danych byłoby niemożliwe lub bardzo utrudnione ze względu na rozmiar problemu. Tym sposobem zamiast klasycznego podejścia analizy wszystkich obserwacji (danych treningowych) i tworzenia z nich wektorów nośnych, Autor zaproponował losowy podział danych na mniejsze części i ich analizę metodami ewolucyjnymi. Ta technika wymaga jednak wielokrotnego strojenia algorytmu GASVM i powtarzania eksperymentu w celu znalezienia przynajmniej suboptymalnego rozwiązania, co w praktyce wiąże się z analizą wszystkich danych treningowych. Jest to więc rozwiązanie kosztowne obliczeniowo i należy je zaliczyć do metod *brute force*, czyli niezbyt atrakcyjnych. Autor wykazał jednak, że wyłonione w ten sposób wektory, będące potem wektorami treningowymi klasyfikatora SVM nie obniżają jakości klasyfikacji. Bardzo ciekawa byłaby jednak odpowiedź na pytanie jak wypadnie proponowne podejście w sytuacji, kiedy dane treningowe nie są dużymi zbiorami. W takim przypadku o ostatecznej postaci funkcji dyskryminującej klasyfikatora SVM decydują wyłącznie wektory nośne wyłonione bezpośrenio, bez pomocy algorytmów omawianych przez Autora. Jak wypadnie porównanie klasyfikatora SVM wspomaganego metodami genetycznymi i bez tego wspomagania? Czy liczba wygenerowanych wektorów treningowych będzie podobna? Czy w trakcie obrony Autor mógłby się ustosunkwać do tego zagadnienia?

Inne, ale mające podobny cel podejście, zaproponowano w algorytmach AGA i DAGA, które są modyfikowane dynamicznie w czasie ewolucji, co polepsza eksplorację przestrzeni rozwiązań oraz uwalnia od uciążliwego i kosztownego strojenia algorytmu. To rozwiązanie należy uznać za bardzo interesujące, gdyż daje wyraźnie lepsze wyniki w porównaniu z algorytmem GASVM. Jest to znana cecha algorytmów memetycznych. Przestrzeń rozwiązań uzyskiwana w algorytmach wymienionych powyżej, początkowo jest określana na podstawie losowo wybieranego pozdzbioru wektorów treningowych, co oznacza, że algorytmy te muszą być parametryzowane. W memetycznym algorytmie PCA^{2}MA dane wejściowe poddawane są najpierw klasteryzacji, aby potem na jej podstawie ustalić zestaw danych wejściowych algorytmu.

Celem wszystkich omawianych powyżej algorytmów, które są używane jako narzędzie wspomagające budowę klasyfikatora SVM jest wybranie odpowiedniego zbioru danych treningowych dla tego klasyfikatora. Skraca to przeważnie czas fazy trenowania klasyfikatora, gdyż z reguły liczba wektorów treningowych jest mała, a to z kolei wpływa na skrócenie czasu późniejszej klasyfikacji. Należy jednak spytać, czy uzyskiwane na tym etapie oszczędności czasowe nie będą niwelowane długim czasem działania algorytmów ewolucyjnych - szczególnie dla długich chromosomów?

Ocena przyjętych metod rozwiązania zagadnienia

Podstawowym zadaniem, które realizuje Autor w swojej rozprawie doktorkiej jest preprocessing danych wejściowych, z których wyłonić należy wektory treningowe klasyfikatora SVM. Zadanie to w różnym stopniu spełnia wymieniona już powyżej grupa algorytmów genetycznych i memetycznych. Zgodnie z zasadami nadzorowanego uczenia maszynowego, zbiór danych wejściowych należy podzielić na dwa rozłączne podzbiory: zbiór danych treningowych, na których klasyfikator jest uczony oraz zbiór danych, na których klasyfikator jest sprawdzany. Zbiór danych wejściowych jest wielokrotnie dzielony w procedurze k-krotnej walidacji krzyżowej. W swoim rozwiązaniu Autor proponuje ustalenie zbioru treningowego za pomocą algorytmów ewolucyjnych (AE). Algorytmy te podlegają w różnym stopniu autorskim modyfikacjom. Modyfikacje oceniane są dwojako: ocenie podlegają parametry AE (na podstawie uśrednionej krzywej dopasowania) oraz klasyfikator SVM (liczba wektorów nośnych, precyzja klasyfikatora oraz obszary pod krzywą ROC). Propozycje autorskie porównywane są następnie z rozwiązaniami znanymi z literatury.

Walidacja (z możliwością częściowej wizualizacji) proponowanych metod doboru zbiorów treningowych została wykonana przy użyciu różnych zbiorów danych - zbiorów sztucznie wygenerowanych, zbiorów benchmarkowych z repozytorium UCI, oraz zbioru zawierającego dane rzeczywiste. Pozwala to wiarygodnie ocenić proponowane rozwiązanie na tle innych. Dodatkowo, rozwiązania autorskie porównywane są między sobą parami za pomocą testu Wilcoxona, co pozwala na ocenę, czy różnice międzygrupowe dla grup algoryt-

mów są istotne ze statystycznego punktu widzenia. Z punktu widzenia oceny poprawności badań są to działania prawidłowe i wręcz niezbędne, lecz często ignorowane w innych pracach. Jest to dojrzała propozycja przeprowadzania i dokumentowania badań naukowych. Autor uznał jednak, że są to działania drugorzędne z czym nie mogę się zgodzić.

Rozwiązane autorskie wykazują istotną przewagę nad innymi metodami opisanymi w samej rozprawie. Eksperymenty zostały zaprojektowane wielowariantowo, co świadczy o dobrym przygotowaniu Autora do prowadzenia badań naukowych, a także o bardzo dobrej znajomości tematyki związanej z uczeniem maszynowym, statystyką i klasyfikacją danych.

W podsumowaniu tej części opinii należy uznać, że autorskie analizy danych algorytmami ewolucyjnymi są rozwiązaniem ciekawym i nowatorskim, dotąd nie stosowanym. Algorytmy genetyczne i memetyczne skutecznie wspierają proces uczenia klasyfikatora SVM, poprzez wstępną selekcję danych treningowych, nie obniżając jakości klasyfikacji SVM. Tym samym zaproponowane w rozprawie rozwiązania są przydatne w naukach technicznych zajmujących się systemami uczącymi, czyli systemami sztucznej inteligencji.

Analiza źródeł literaturowych

W rozprawie znajdują się odwołania do 227 prac, w tym 28 prac, gdzie doktorant jest współautorem artykułu. Niektóre z tych prac (21) są indeksowane w systemie WoS (Web of Science). System ten odnotowuje również indeks Hirscha $h = 5$, co jest wartością bardzo dobrą, znacząco przewyższającą przyjęte wymagania. Taką aktywność naukową doktoranta należy z dużym uznaniem podkreślić. Źródła literaturowe dotyczą w większości publikacji w czasopismach naukowych oraz materiałach konferencyjnych związanych z zagadnieniami klasyfikacji danych różnych aspektów i uczenia maszynowego. Zestaw pozycji literaturowych jest reprezentatywny dla eksplorowanej dziedziny naukowej i uwzględnia również doniesienia najnowsze, z roku 2016. Przywoływane w tekście cytowania, są odnotowywane w spisie literatury, gdzie Autor podaje nawet stronę, na której cytował daną pracę. Pozwala to w trakcie czytania tekstu pracy na łatwe sprawdzenie korelacji przwoływanej pracy z fragmentu prowadzonego wywodu. Z uznaniem stwierdzam, że zostało to zrobione wręcz wzorowo. Wnioski i polemiki wypływające z bardzo szerokiego i reprezentatywnego przeglądu literatury zostały sformułowane przekonywująco i należy je uznać za poprawne.

Ocena wyników przedstawionych w rozprawie

Uzyskane wyniki badań eksperymentalnych zostały zaprezentowane w solidnej, wielowariantowej analizie porównawczej, gdzie Autor konfrontował własne metody wyznaczania zbiorów treningowych klasyfikatora SVM z metodami znanymi z literatury. Część eksperymentalną, wraz analizą statystyczną należy zaliczyć do mocnych stron rozprawy. Analiza otrzymanych rezultatów wskazuje, że opisane w rozprawie metody oparte na al-

gorytmach ewolucyjnych przyczyniają sie do zmniejszenia liczby wektorów treningowych, co przyśpiesza fazę treningu klasyfikatora. Wykazano, że algorytmy genetyczne i memetyczne z adaptacyjnym doborem parametrów (wielkość osobników w populacji) pozwalają na uzyskanie mniejszej liczby wektorów podpierajacych w porównaniu z innymi technikami doboru tych wektorów. Należy również zauważyć, że oprócz badań mających wykazać postawioną tezę rozprawy, Autor wykonał oprogramowanie do wizualizacji wyników dla danych dwuwymiarowych, co ma wyraźny walor dydaktyczny i poznawczy. Na tle przedstawionej analizy porównanwczej, uzyskane rezultaty są nowe i zwiększają wiedzę o uczeniu maszynowym i klasyfikacji danych.

Wnioski końcowe

Podsumowując przedstawioną charakterystykę rozprawy doktorskiej stwierdzam, że doktorant, pan mgr inż. Jakub Nalepa zamieścił w niej samodzielnie uzyskane wyniki badań teoterycznych i eksperymetalnych popartych analizą statystyczną. Wykazał się znajomością programowania i zasad wizualizacji danych, co dodatkowo podnosi walory pracy doktorskiej. Do bardzo mocnych stron pracy zaliczam wartościową dyskusję proponowanych rozwiązań oraz szeroką interpretację uzyskanych wyników Autora rozprawy.

Pod względem merytorycznym praca z nadmiarem spełnia wymagania stawiane rozprawom doktorskim. Autor wykazał się znajomością metod uczenia maszynowego oraz dobrym opanowaniem warsztatu badawczego.

Wnoszę o dopuszczenie opiniowanej dysertacji do dalszych etapów przewodu doktorskiego, w tym do jej publicznej obrony.

Biorąc pod uwagę dojrzały charakter recenzowanej pracy oraz fakt, że Autor publikował 5 prac w czasopismach o dużej randze naukowej (lista A MNiSW), indeksowanych w systemie WoS, dla których współczynnik oddziaływania IF jest wyższy niż 0.5 **wnoszę o jej wyróżnienie,** co jest zgodne z Uchwałą Rady Wydziału Automatyki, Elektroniki i Informatyki Politechniki Śląskiej.

Następujące prace upoważniają do wystąpienia z wnioskiem o wyróżnienie pracy:

1. J. Nalepa, M. Kawulok, Adaptive memetic algorithm enhanced with data geometry analysis to select training data for SVMs. Neurocomputing, Vol. 185, pp.113-132, 2016 IF=2.083, liczba pkt. MNiSW: 30.

2. M. Kawulok. J. Kawulok, J. Nalepa, Spatial-based skin detection using discriminative skin-presence features, Pattern Recognition Letters, Vol. 41, pp. 3-13, 2014, IF=1,551, liczba pkt. MNiSW: 25.

3. J. Nalepa, M. Blocho, Adaptive memetic algorithm for minimizing distance in the vehicle routing problem with time windows, Soft Computing, Vol. 20 (6), pp 2309–2327, 2016, IF=1,271, liczba pkt. MNiSW: 25.

4. J. Nalepa, M. Blocho, Co-operation in the Parallel Memetic Algorithm, International Journal of Parallel Programming, Vol. 43, (5), pp 812–839, 2015, IF=0,491, liczba pkt. MNiSW: 15.

5. M. Kawulok, J. Kawulok, J. Nalepa, B. Smołka, Self-adaptive algorithm for segmenting skin regions, EURASIP Journal on Advances in Signal Processing, Vol. 2014, (170), pp. 1-22, 2014, IF=0,78, liczba pkt. MNiSW: 25.