Wrocław 2023.07.20

dr hab. inż. Michał Przewoźniczek, prof. uczelni
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska

Recenzja rozprawy doktorskiej
Ensembles of Support Vector Machines with evolutionarily optimized hyperparmaters and training sets
autorstwa Pana
mgr inż. Wojciecha Dudzika

Promotor: dr hab. inż. Michał Kawulok, prof. PŚ
Promotor pomocniczy: dr hab. inż. Jakub Nalepa, prof. PŚ

1. Cel naukowy rozprawy

Praca doktorska Pana mgra inż. Wojciecha Dudzika dotyczy optymalizacji ewolucyjnej w zagadnieniach dotyczących poprawy działania tzw. *Support Vector Machines (SVM)*. Autor stawia w pracy dwie następujące tezy:

1. Jednoczesna optymalizacja zestawu treningowego (ang. *training set*), zestawu cech i hiperparametrów SVM poprawia czas treningu i klasyfikacji w porównaniu do innych aktualnych metod zaproponowanych do tego celu bez wpływu na jakość klasyfikacji.
2. Budowa zestawów SVM (ang. *SVM ensembles*) z wykorzystaniem algorytmów ewolucyjnych zapewnia poprawę jakości klasyfikacji w porównaniu do innych uznanych w dziedzinie metod przeznaczonych do tego celu.

Ze względu na skalę trudności postawionego zadania, Autor jako punkt wyjścia do swoich badań obiera Algorytmy Ewolucyjne. Biorąc pod uwagę aktualny stan wiedzy w dziedzinie, której dotyczy praca, wybór taki należy uznać za prawidłowy. Należy jednak zauważyć, że w takiej sytuacji, formalne dowiedzenie lub obalenie tez postawionych przez Autora może znajdować się poza możliwościami dzisiejszej nauki. Ta obserwacja ma istotny wpływ na charakter pracy. Autor stara się zaproponować mechanizmy optymalizacji, które mają służyć konkretnym celom, a następnie eksperymentalnie weryfikuje ich skuteczność. Takie postępowanie jest zgodne ze stanem badań w dziedzinie optymalizacji ewolucyjnej nie może jednak być uznane za formalny dowód. Dlatego, w mojej opinii, obie hipotezy należy traktować jako cele pracy, co jest zgodne z obowiązującymi przepisami dotyczącymi oceny rozpraw doktorskich. Weryfikacja osiągnięcia (bądź nie) obranego celu nie musi być formalna, musi jednak opierać się na przekonujących wynikach badań. Dlatego zamieszczona w pracy eksperymentalna weryfikacja propozycji Autora wykorzystująca narzędzia analizy statystycznej jest w mojej opinii prawidłowa i wystarczająca do uznania, że Autor osiągnął cele swojej pracy.

Poprawa jakości działania SVM jest celem aktualnym i ważnym zarówno z punktu widzenia teorii, jak i praktyki.

2. Oryginalny wkład Autora

Pierwszy cel rozprawy został zrealizowany głównie poprzez zaproponowanie metody *Simultaneously-Evolved SVMs* (SE-SVM). Ta metoda optymalizuje jednocześnie hiperparametry SVM, wybór zbioru treningowego, oraz wybór cech. Moim zdaniem interesujące z punktu widzenia optymalizacji wydaje się skonstruowanie metody z użyciem następujących mechanizmów. Po pierwsze, lista wybranych próbek i cech jest dynamiczna, a jej długość zmienia się w zależności od analizy wyników uzyskanych przez metodę we wcześniejszych iteracjach. Takie dynamiczne podejście do kodowania rozwiązań nie jest często spotykane w literaturze dotyczącej optymalizacji ewolucyjnej, głównie ze względu na trudności związane z definiowaniem operatorów genetycznych. Po drugie, SE-SVM wykorzystuje mechanizmy zaproponowane wcześniej (w tym przez innych badaczy) m.in. do inicjacji populacji. Analiza wyników przeprowadzonych eksperymentów pozwala stwierdzić, że jakość klasyfikatorów wytrenowanych za pomocą SE-SVM była wyższa niż w przypadku metod konkurencyjnych.

Drugi cel rozprawy został zrealizowany poprzez propozycję metod *Cascades of Evolutionarily optimized SVMs* (CE-SVM) i *Ensembles of Cascades of Evolutionarily optimized SVMs* (ECE-SVM). Obie metody są dedykowane do trenowania zestawów SVM. Obie wykorzystują również wiedzę dziedzinową polegającą na intuicji, że przestrzeń podlegająca klasyfikacji może zostać podzielona na regiony homogeniczne (łatwe do klasyfikacji i jednorodne) oraz heterogeniczne (trudne do klasyfikacji i niejednorodne). Główną ideą zaproponowanych klasyfikatorów jest przedstawienie klasyfikatora w postaci hierarchicznej, gdzie każdy węzeł jest odpowiedzialny za klasyfikację innej części przestrzeni. Każdy taki węzeł jest oddzielnym klasyfikatorem trenowanym przez metodę SE-SVM (przy czym optymalizacji podlegają hiperparametry SVM i zestaw danych treningowych, w tym miejscu Autor nie wspomina o optymalizacji wyboru cech).

Badania nad metodą CE-SVM wykazały pewne jej wady. Na przykład, w niektórych przypadkach metoda była zatrzymywana przedwcześnie. Dlatego Autor zaproponował metodę ECE-SVM, która jest ulepszoną wersją metody CE-SVM. W ramach metody ECE-SVM Autor wprowadza szereg usprawnień. Między innymi rezygnuje z uwzględniania niektórych jąder dla trenowanych SVM, ogranicza rozmiar zbioru walidacyjnego, oraz pozwala na to, aby węzły w tej samej kaskadzie posiadały różne rodzaje jąder. Ten ostatni efekt jest osiągana poprzez wprowadzenie dwóch równolegle ewoluowanych populacji. Wreszcie do poprawy jakości klasyfikacji zostaje wykorzystany mechanizm *Extra Tree Classifier* (Autor wykorzystuje tutaj mechanizm zaproponowany przez innych badaczy). W części badawczej pracy, przekonująco wykazano przewagę ECE-SVM nad konkurencyjnymi metodami.

Podsumowując badania przedstawione w ocenianej rozprawie doktorskiej, można powiedzieć, że polegają one na następujących elementach:

1. Analizie istniejących metod i technik i wykryciu ich słabości.
2. Zrozumieniu istoty słabych i mocnych stron istniejących propozycji.
3. Obserwacji i zrozumieniu charakterystyki zbiorów danych mających podlegać klasyfikacji.
4. Zaproponowaniu podstawowych intuicji mających na celu poprawę jakości wyników.
5. Zaproponowaniu metody opartej o własne intuicje i istniejące rozwiązania.
6. Weryfikacji i poprawie zaproponowanych własnych optymalizatorów poprzez wprowadzenie nowych intuicji bądź mechanizmów.

Powyższy schemat jest typowy dla wysokiej jakości badań w dziedzinie Obliczeń Ewolucyjnych. Wymaga on nie tylko dogłębnego zrozumienia dotyczącego wiedzy dziedzinowej, która dotyczy optymalizowanego

problemu (w tym przypadku jest to problematyka dotycząca SVM), ale również typowych cech zbiorów danych, które podlegają klasyfikacji. Dla osiągnięcia wyników wysokiej jakości (co w mojej ocenie zostało osiągnięte przez Autora), niezbędne jest odpowiednie uogólnienie takiej wiedzy, w przeciwnym przypadku opracowana metoda będzie dawać dobre jakościowo wyniki wyłącznie dla wąskiej grupy przypadków testowych, które są do siebie bardzo podobne. Jako propozycje kluczowych podstawowych intuicji (krok nr 4) wskazałbym mechanizmy związane z dynamicznym kodowaniem w metodzie SE-SVM, oraz wykorzystanie wiedzy o homo- i heterogenicznych regionach w metodzie CE-SVM. Wszystkie zaproponowane przez autora metody (SE-SVM, CE-SVM i ECE-SVM) są hybrydami pomysłów Autora i istniejących w literaturze rozwiązań (krok 5). Wreszcie, metoda ECE-SVM eliminująca wady swojej poprzedniczki (CE-SVM) poprzez wprowadzenie nowych mechanizmów usuwających zaobserwowane wady jest typową implementacją kroku nr 6. Cały opisany powyżej proces wymaga dogłębnej wiedzy dziedzinowej, twórczego podejścia oraz wysokich umiejętności technicznych polegających na integracji wszystkich mechanizmów.

W związku z powyższym można stwierdzić, że oceniana praca doktorska przekonująco wykazuje, że postawione w niej cele zostały osiągnięte. Oznacza to również, że Autor przyczynił się do rozwiązania problemu ważnego w nauce i posiadającego istotne znaczenie praktyczne.

3. Uwagi krytyczne i polemiczne

Uwagi zawarte w tej części recenzji dzielą się na dwie grupy. Pierwszą stanowią uwagi redakcyjno-techniczne, które, choć krytyczne, nie wpływają ani na ocenę pracy. Druga grupa uwag to uwagi w kwestii możliwych alternatywnych technik optymalizacji, które Autor mógłby (choć nie musiał) uwzględnić.

W ocenianej pracy doktorskiej można znaleźć zdanie „This means that SE-SVM is a pareto-optimal method which proves this thesis". Dotyczy ono jakości działania metody SE-SVM w odniesieniu do metody ALMA. Nie mogę zgodzić się z następującymi elementami tego zdania. **Po pierwsze**, moim zdaniem, Autor nie wie i nie może udowodnić, że metoda SE-SVM zwraca rozwiązania z optymalnego frontu Pareto. Takie rozwiązania nie są dominowane przez żadne inne. Aby udowodnić, że SE-SVM faktycznie zwraca takie rozwiązania dla rozpatrywanego problemu należałoby sprawdzić całą przestrzeń rozwiązań. **Po drugie**, sformułowanie „pareto-optimal" w odniesieniu do metody optymalizacji jest dla mnie zaskakujące. Optymalny może być konkretny front Pareto, a nie metoda optymalizacji. Być może Autorowi chodziło o to, że metoda SE-SVM zawsze zwraca optymalny front Pareto? Takie stwierdzenie byłoby jednak jeszcze trudniejsze do obrony, ponieważ wymagałoby dowodu uwzględniającego minimalny nakład obliczeniowy potrzebny do znalezienia takiego frontu oraz, że jest on kompletny (zwracana są wszystkie rozwiązania, które pozwalają na zdefiniowanie optymalnego frontu Pareto). **Po trzecie**, charakter ocenianej rozprawy doktorskiej jest eksperymentalny. Autor nie może więc niczego udowodnić. Może jedynie pokazać (i robi to!), że zaproponowana przez niego metoda, dla danej klasy problemów, w ramach przeprowadzonych w określonym środowisku badań daje wyniki lepsze niż metody konkurencyjne. W mojej opinii, niefortunne sformułowanie Autora wynika przede wszystkim z (nieprawidłowego moim zdaniem) przedstawienia celów pracy jako hipotez badawczych, co omówiłem w pierwszej części recenzji.

Podobna jak powyżej uwaga dotyczy występujących w pracy niefortunnych w mojej opinii stwierdzeń takich jak „Although as proved in the conference paper [40] the method provides great classification performance". Moim zdaniem w pracy [40] **niczego** nie udowodniono. Zamiast tego **pokazano** (przekonująco i prawidłowo), że dana metoda może zapewnić wysoką jakość klasyfikacji. W związku z tym

zdanie powinno brzmieć raczej w następujący sposób (razem z poprawkami stylistycznymi): „Although as *SHOWN* in [40], the method provides *HIGH/OUTSTANDING* classification performance" lub „Although as *SHOWN* in [40], the method provides classification performance *OF SIGNIFICANTLY HIGHER QUALITY THAN OTHER METHODS*".

W swoich badaniach Autor skupia się na wykorzystaniu stosunkowo prostych algorytmów ewolucyjnych. Tymczasem, dziedzina ta intensywnie się rozwija. Na przykład, w ramach prowadzonych badań możliwe byłoby wykorzystanie inspiracji pochodzących z badań nad tzw. Szarą Skrzynką (ang. *Gray-box optimization*). Jedną z technik optymalizacji Szarej Skrzynki są tzw. częściowe obliczania jakości rozwiązania (ang. *partial evaluations*). Taki zabieg polega na tym, że posiadając ocenione rozwiązanie wprowadzamy do niego niewielką modyfikację (w sensie genotypu). Aby ocenić zmodyfikowane rozwiązanie wystarczy policzyć powtórnie tylko część składowych optymalizowanej funkcji co jest znacznie tańsze obliczeniowo. Częściowe obliczanie jakości rozwiązania pozwala na przykład na znaczące (np. o kilka rzędów wielkości) obniżenie kosztów optymalizacji genotypu algorytmem zachłannym. Taka intuicja nie przekłada się bezpośrednio na problemy optymalizacji rozważane w ocenianej rozprawie. Jednak posiadając odpowiednią kontrolę nad kodem odpowiedzialnym za trenowanie i dobór klasyfikatorów można by zapamiętywać cząstkowe wyniki treningu już ocenionych rozwiązań i taniej oceniać nowe, nieznacznie zmodyfikowane (np. poprzez zmianę jednego z wybranych przypadków testowych) rozwiązania.

W badaniach dotyczących optymalizacji ewolucyjnej istotnym elementem jest znaczenie wartości genu i jego pozycji. Na przykład, optymalizując problem składający się z pięciu zmiennych wiemy, że wartość na pierwszej pozycji genotypu zawsze odpowiada pierwszej zmiennej, druga drugiej itd. Tymczasem, kodowanie rozwiązań zaproponowane w metodzie SE-SVM łamie powiązanie pomiędzy pozycją a znaczeniem genu. Na przykład, jeżeli w genotypie zapisuję 5 wybranych próbek o numerach 1,4,5,16 i 78, to taki zapis może wystąpić w dowolnej kolejności. W takiej sytuacji powiązanie pomiędzy pozycją genu w genotypie a jego znaczeniem zostaje zerwane. W badaniach Autora zabrakło mi analizy skutków tego zjawiska. Powyższa uwaga jest raczej jest głównie sugestią dotyczącą ewentualnej kontynuacji badań, ponieważ zastosowanie kodowania takiego jak zaproponowane w metodzie SE-SVM jest dobrze uzasadnione (skrócenie długości genotypu).

Inną kwestią związaną z powyższym wątkiem jest fakt, że współczesne osiągnięcia dotyczące rozwoju metod ewolucyjnych dają skuteczne narzędzia pozwalające radzić sobie z problemami zakodowanymi z wykorzystaniem długich genotypów. Są to m.in. techniki *linkage learning* (w pracach dotyczących przestrzeni ciągłych nazywane również *decomposition strategies*), które wykrywają powiązania pomiędzy genami. Taka wiedza pozwala na konstruowanie optymalizatorów o nieosiągalnej wcześniej skuteczności. Być może warto wykorzystać te narzędzia również w badaniach dotyczących optymalizacji SVM.

4. Ocena dorobku naukowego Autora

Zgodnie z informacjami podanymi przez portal Google Scholar (w dniu 2023.07.26), Pan mgr inż. Wojciech Dudzik jest współautorem 19 prac, które były cytowane 1615 razy. Pierwsza praca ukazała się w 2017 roku. Zdecydowanie najczęściej cytowana praca (1456 cytowań), to praca zbiorowa autorstwa kilkuset współautorów. Pozostałe prace, gdzie wkład autorski Pana Wojciecha Dudzika wydaje się być znacznie większy, osiągnęły około 200 cytowań. W mojej opinii, uwzględniając dziedzinę nauki, na tym etapie kariery jest to wynik dobry lub bardzo dobry.

Kilkanaście prac, w których Pan Wojciech Dudzik **nie jest pierwszym autorem**, to referaty konferencyjne i publikacje w czasopismach. Wedle mojej wiedzy, prace w tej grupie zostały opublikowane w czasopismach spoza Listy Filadelfijskiej. Natomiast wśród publikacji na konferencjach co najmniej cztery z nich zostały opublikowane na konferencjach, które w rankingu Core należą do kategorii B (*Asian Conference on Intelligent Information and Database Systems*, *International Conference on Pattern Recognition*, *Artificial Intelligence in Medicine*, *IEEE International Conference on Image Processing*). W mojej ocenie, takie osiągnięcia uzasadniają stwierdzenie, że Pan Wojciech Dudzik jest cennym członkiem zespołów badawczych i może wnieść istotny wkład do dziedzin, którymi się zajmuje.

Wedle mojej wiedzy, Pan Wojciech Dudzik jest pierwszym autorem pięciu opublikowanych prac. Dwie z nich to referaty konferencyjne spoza rankingu Core. Dwie to prace typu *poster* opublikowane na konferencji GECCO (ranking Core A, konferencja powszechnie uważana za najlepszą lub jedną z najlepszych na świecie w dziedzinie obliczeń ewolucyjnych). Obie prace na konferencji GECCO zostały wydane w ramach obszaru tematycznego *Evolutionary Machine Learning*. Prace typu *poster* są zdecydowanie najniższe rangą na konferencji GECCO (wyżej stoją tzw. *workshop papers* i oczywiście prace typu *full paper*), jednak taka publikacja jest zauważalnym osiągnięciem na każdym etapie kariery. W przypadku pracy doktorskiej Pana Wojciecha Dudzika jest to istotna przesłanka, że prezentuje on wartościowe pomysły dotyczące optymalizacji zagadnień z obszaru Uczenia Maszynowego.

Piątą pracą, gdzie Pan Wojciech Dudzik jest pierwszym autorem, jest publikacja w czasopiśmie *Knowledge-Based Systems* (IF w roku publikacji: 8.139). Moim zdaniem jest to zdecydowanie najważniejsze osiągnięcie publikacyjne w Jego dotychczasowej karierze. Prace w czasopismach tej rangi uzasadniają stwierdzenie, że Autor jest zdolny do samodzielnego prowadzenia badań na wysokim poziomie. Ponadto, tego typu publikacje pokazują, że propozycje Autora są ważne w uprawianej przez niego dziedzinie.

Moim zdaniem rozmiar i jakość dorobku Pana Wojciecha Dudzika jest odpowiednia dla tego etapu kariery naukowej. Jednocześnie uważam, że taki dorobek nie stanowi przesłanki do wnioskowania o wyróżnienie rozprawy doktorskiej. W mojej opinii dorobek stanowiący podstawę do wyróżnienia powinien być opublikowany w miejscach powszechnie uznawanych za elitarne, czyli najlepsze w danej dziedzinie. W kontekście ocenianej rozprawy doktorskiej mogłoby to być 2-3 publikacje typu *full paper* na konferencjach GECCO, ICML lub NeurIPS. Innym wybitnym osiągnięciem mogłaby być publikacja w czasopiśmie uznawanym za najlepsze w dziedzinie. Na przykład w *IEEE Transactions on Evolutionary Computation*.

5. Podsumowanie

Wykorzystując analizę i argumentację przedstawioną powyżej, uzasadnione są stwierdzenia, że:

- Autor posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka Techniczna i Telekomunikacja
- Autor jest zdolny do samodzielnego prowadzenia pracy naukowej
- Rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego

Tym samym spełnione są wymogi ustawowe i zwyczajowe stawiane pracom doktorskim.