

Prof. dr hab. inż. Eulalia Szmidt
Instytut Badań Systemowych
Polskiej Akademii Nauk
ul. Newelska 6
01-447 Warszawa
E-mail: szmidt@ibspan.waw.pl

Warszawa, 30.04.2019r.

**Recenzja rozprawy doktorskiej mgr Łukasza Gadomera
pt. „Rozmyte lasy losowe oparte na modelach klastrowych drzew
decyzyjnych w zadaniach klasyfikacji”**

Niniejsza recenzja została przygotowana na prośbę Rady Wydziału Informatyki Politechniki Białostockiej wyrażoną w piśmie z dn. 11 lutego 2019r.

Recenzowana rozprawa doktorska dotyczy ciągle aktualnego, ważnego i interesującego zagadnienia badawczego, zarówno w kontekście teoretycznym jak i aplikacyjnym, mianowicie, problemu klasyfikacji i regresji. Autor rozprawy zaproponował nowy klasyfikator wykorzystując teorię logiki rozmytej do zagadnienia drzew decyzyjnych. Mówiąc dokładniej, zastosował znane z literatury, klastrowo-rozmyte drzewa decyzyjne oraz klastrowo-kontekstowe rozmyte drzewa decyzyjne do opracowania autorskiej metody rozmytych lasów losowych, w których wykorzystane są wymienione drzewa decyzyjne.

Praca liczy 186 stron, składa się 8 rozdziałów, wykazu literatury obejmującego 117 pozycji (w tym kilku publikacji Doktoranta), oraz odrębnych spisów: oznaczeń, tabel, rysunków oraz algorytmów.

We Wstępie (Rozdział 1) jest dobrze wyjaśniony rozważany problem. Autor podkreśla potrzebę poszukiwania coraz lepszych metod klasyfikacji i regresji, sygnalizuje najbardziej znane podejścia pokazując, gdzie wpisuje się problem,

którym zajmuje się w rozprawie. Po jasno sformułowanym celu rozprawy i hipotezach badawczych, przedstawiony jest układ pracy.

W Rozdziale 2 pracy doktorskiej przedstawiono podstawowe pojęcia i zagadnienia związane z różnymi rodzajami rozmytych drzew decyzyjnych, lasami losowymi, rozmytymi lasami losowymi i przykładowymi metodami grupowania rozmytego, w szczególności grupowania rozmytego uwzględniającego kontekst danych. Zagadnienia te są ściśle związane z badaniami prowadzonymi przez Autora rozprawy, są opisane w sposób przejrzysty, pozwalający stwierdzić, że Autor zna odpowiednią literaturę.

W Rozdziale 3 przedstawiono oryginalny rezultat pracy doktorskiej, czyli zaproponowano nowy, rozmyty-klastrowy las losowy. Rozdział rozpoczyna opis drzew będących elementami składowymi nowego klasyfikatora. Są to dwa rodzaje drzew znanych z literatury, mianowicie, klastrowo-rozmyte drzewa decyzyjne oraz klastrowo-kontekstowe rozmyte drzewa decyzyjne. Wymienione drzewa, zastępując w lesie losowym tradycyjnie stosowane rozmyte drzewa decyzyjne, dają w efekcie dwie podstawowe wersje nowego klasyfikatora. Obydwie wersje klasyfikatora mogą być wykorzystane zarówno w procesach klasyfikacji jak i regresji. Przejrzystym opisom klasyfikatorów towarzyszą uproszczone zapisy algorytmów, dzięki czemu łatwo jest śledzić procesy tworzenia, uczenia oraz przeprowadzania procesów klasyfikacji i regresji danych. Autor odsyła czytelnika do zapoznani się z rezultatami uzyskanymi przez zaproponowane klasyfikatory do Rozdziału 7 oraz do dwóch artykułów konferencyjnych własnego autorstwa w *LNCS* (wydawnictwo Springer, *Lecture Notes in Computer Science*).

W Rozdziale 4 przedstawiono modyfikacje zaproponowanych algorytmów o możliwość ważonego podejmowania decyzji. Celem była chęć poprawy jakości klasyfikacji dzięki wykorzystaniu całej wiedzy gromadzonej we wszystkich węzłach w drzewie oraz wszystkich drzewach w lesie, poprzez

przypisanie większych wag przy podejmowaniu decyzji niektórym z nich. Wykorzystano w tym celu tzw. uporządkowane operatory średniej ważonej OWA (ang. *Ordered Weighted Averaging*). Autor, swoim dobrym zwyczajem, w sposób przejrzysty zaczyna od zapoznania czytelnika z podstawowymi pojęciami, ilustrując je przykładem i zarysowując szeroki obszar zastosowań operatorów OWA. Następnie, proponuje trzy autorskie sposoby obliczania wag z wykorzystaniem operatorów OWA w opracowanym klasyfikatorze: lokalne OWA, globalne OWA dla drzewa oraz globalne OWA dla lasu. W kolejnych podrozdziałach przedstawione są dwie kolejne autorskie metody ważenia decyzji, mianowicie, ważenie proporcjonalne na podstawie obiektów ze zbioru uczącego oraz ważenie proporcjonalne na podstawie indeksu wydajności drzewa. Wszystkie przedstawione metody są opisane dobrze i dodatkowo zilustrowane uproszczonymi zapisami algorytmów, co jeszcze bardziej ułatwia czytanie tekstu. Wyniki zaproponowanych rozwiązań przedstawiono w Rozdziale 7, a ponadto opublikowano w czasopiśmie *Soft Computing* oraz jako rozdział w *LNCS* (Springer).

Rozdział 5 poświęcony jest nieodzownemu tematowi przy omawianiu drzew, czyli ich przycinaniu. Szczegółowo przedstawionych jest pięć autorskich wariantów metod przycinania drzew wchodzących w skład rozmytego-klastrowego lasu losowego z klastrowymi-rozmytymi drzewami decyzyjnymi oraz trzy dalsze metody dla klasyfikatora z klastrowo-kontekstowymi drzewami decyzyjnymi. Każda z wymienionych metod została dokładnie opisana z uwzględnieniem ich zalet, wad oraz algorytmów działania.

W Rozdziale 6 rozważane są możliwości przyspieszenia działania zaproponowanych klasyfikatorów poprzez zrównoleglenie obliczeń. Omówiono równoległe obliczenia przetestowane w zaproponowanych w rozprawie klasyfikatorach oraz przedstawiono inne możliwości zrównoleglania obliczeń w ich przypadku.

Rozdział 7 zawiera sprawdzenie hipotez badawczych rozprawy poprzez eksperymenty numeryczne, pozwalające na ocenę skuteczności działania wszystkich zaproponowanych rozwiązań. Na podstawie dokładnych opisów można stwierdzić, że Autor rozprawy rygorystycznie dba o zachowanie tych samych warunków (np. te same podziały zbioru danych) we wszystkich przeprowadzonych eksperymentach. Implementacja zaproponowanego klasyfikatora, rozwiązania dotyczące zastosowania operatorów OWA, metody przycinania i zrównoleglania obliczeń zostały wykonane za pomocą kodu maszynowego napisanego przez Doktoranta. Klasyfikatory są porównywane na podstawie procentowej dokładności i odchylenia standardowego świadczącego o ich „stabilności”. Istotność otrzymanych wyników sprawdzano przy użyciu odpowiednich testów statystycznych. W efekcie przeprowadzonych badań na 10 referencyjnych zbiorach danych (repozytorium UCI) udało się potwierdzić, że opracowany klasyfikator osiąga znacznie lepsze rezultaty niż drzewo C4.5 będące bazą porównań. Jak należało się spodziewać, również pojedyncze drzewa wchodzące w skład nowego klasyfikatora, osiągają gorsze rezultaty. Pokazano również, że zastosowanie wag zwiększa jakość klasyfikacji. W wielu przypadkach analogicznie działa wykorzystanie losowości. Przycinanie drzew oraz zrównoleglanie obliczeń powodowało skrócenie czasu obliczeń,

Rozdział 8 stanowi umiejętne podsumowanie treści rozprawy i uzyskanych w niej wyników.

Praca napisana jest w sposób przejrzysty, co zostało już podkreślone przy omawianiu zawartości rozdziałów. Autor proponuje nie tylko własne rozwiązania, ale daje także przegląd innych podejść znanych z literatury.

W części eksperymentalnej podkreślona jest dbałość o warunki wykonywania eksperymentów. Między innymi podane jest, że do eksperymentów wybrane zostały takie bazy danych, w których nie ma niezbalansowanych klas. Jest to ważne założenie przy porównywaniu klasyfikatorów, z czego Doktorant

doskonale zdaje sobie sprawę. Jednak w rzeczywistości, niektóre z baz danych mają jednak klasy niezbalansowane. Przykładem może być *Pima Indian Diabetes* – baza danych zawierająca 768 obiektów, z których 500 należy do klasy 1, a 268 do klasy 2. Kolejnym przykładem jest zbiór danych *Climate Model Simulation Crashes* zawierający 540 obiektów, z których 494 należy do klasy 1, a 46, a więc tylko 8,5%, do klasy 2. W takich przypadkach zazwyczaj podaje się oprócz dokładności, jeszcze inne miary oceny klasyfikatorów, np. miary utworzone na podstawie macierzy krosvalidacji albo/i krzywe ROC. W rozprawie wykorzystana jest też baza danych *Dermatology* zawierająca sześć klas zawierających odpowiednio: 112, 61, 72, 49, 52 i 20 elementów, a w opisie bazy danych mamy podane dwie klasy, które prawdopodobnie powstały w ten sposób, że połączono jakoś wymienione oryginalne klasy, otrzymując w efekcie tylko dwie, które zostały uwzględnione w obliczeniach. Szkoda, że zabrakło wzmianki na ten temat. Na usprawiedliwienie Autora trzeba zauważyć, że odciął się od problematyki niezbalansowanych klas, a każdy z porównywanych klasyfikatorów staje przed tym samym wyzwaniem, którego wyniki zostały ocenione testami istotności.

Pewien niedosyt budzi fakt, że zaproponowane klasyfikatory porównano jedynie z algorytmem C4.5 oraz pojedynczymi drzewami wchodzącymi w skład lasu losowego, będącego przedmiotem rozprawy. Trzeba jednak podkreślić, że Autor wyjaśnił motyw – chodziło mianowicie o zachowanie dokładnie tych samych warunków eksperymentów, co wiązało się z napisaniem kodu maszynowego, a trudno wymagać od Doktoranta tworzenia własnego oprogramowania do kilku innych metod klasyfikacji.

Na uwagę zasługuje staranna strona edycyjna pracy. Do drobnych usterek można zaliczyć:

- str 29 – nad (2.12) jest: „może odbywać się zgodnie następująco [47]:” zamiast np.: może odbywać się zgodnie z następującym równaniem [47]:

- str 35 – równanie (2.24) jest dokładnie takie samo jak (2.4); wystarczyło powołać się na poprzednie równanie;

- str 39 – równanie (3.2) jest dokładnie takie samo jak (2.6);

str – 138 – „W większości przypadków większość wariantów klasyfikatora...”

Powtórzenia wzorów można jednak wytłumaczyć chęcią Autora do jak najbardziej przejrzystego przedstawiania zagadnień.

Wymienione, przykładowe usterki nie podważają jednak opinii o starannym zredagowaniu pracy.

Na zakończenie warto zauważyć, że nie jest łatwo zmierzyć się z klasycznym problemem i uzyskać nowe, interesujące wyniki. Doktorant dobrze wywiązał się z postawionego zadania, czyli zaproponował, a następnie udoskonalił metodę, która może być stosowana do zadań klasyfikacji i regresji. Zweryfikował również pozytywnie wszystkie hipotezy.

Można więc stwierdzić, że przedstawiona do recenzji rozprawa stanowi oryginalne rozwiązanie problemu naukowego i wykazuje erudycję Autora. W związku z powyższym stwierdzam, że w mojej opinii, praca spełnia wymagania stawiane rozprawom doktorskim przez stosowne przepisy, wnoszę zatem o jej przyjęcie i dopuszczenie Pana mgr Łukasza Gadoamera do publicznej obrony.

