dr hab. inż. Janusz Starczewski, Prof. PCz
Instytut Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
Al. Armii Krajowej 36, Częstochowa

Częstochowa, 02.05.2019

Recenzja

Przedmiotem niniejszej recenzji jest rozprawa doktorska pana magistra **Marcina Pietrzykowskiego** pod tytułem „***Lokalne uczenie algorytmów regresyjnych metodą mini-modeli***" napisana pod kierownictwem promotora Pana profesora dr. hab. inż. Andrzeja Piegata i promotora pomocniczego Pana dr. inż. Marcina Plucińskiego na Wydziale Informatyki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie w 2019 roku. Recenzja została sporządzona w odpowiedzi na pismo Dziekana Wydziału Informatyki Pana dr. hab. inż. Jerzego Pejasia z dnia 24.01.2019 r.

1 Przedmiot i problematyka rozprawy

Przedłożona do oceny rozprawa dotyczy algorytmów aproksymacji i modelowania danych bazujących na próbkach. Proponowane jest tu modelowanie lokalne z wykorzystaniem tak zwanych mini-modeli ograniczających obliczenia do lokalnej domeny zamiast modelowania globalnego w całej dziedzinie problemowej. Koncepcja mini-modeli została wprowadzona przez profesora Andrzeja Piegata jako alternatywa dla metody *k*-Najbliższych Sąsiadów. W ujęciu lokalnym podejście wykorzystuje w procesie uczenia metody statystyczne takie, jak np. regresja liniowa. Niedostatkiem oryginalnej metody jest jej niska efektywność w zadaniach więcej niż trójwymiarowych.

Celem rozprawy jest udoskonalenie metody uczenia mini-modeli na bazie próbek lokalnych w przestrzeni wielowymiarowej. W toku zaprezentowanych badań Autor weryfikuje hipotezę badawczą mówiącą, że "*udoskonalenie metody uczenia mini-modeli umożliwi wykazanie co najmniej części przewidywanych zalet tej metody względem geometrycznych metod modelowania bazujących na próbkach*". Zatem Autor dokładnie sprecyzował problem naukowy i cel badawczy oraz w toku całej rozprawy skrupulatnie wypełnił ciąg czynności badawczych skutkujący rozwiązaniem postawionych problemów.

2 Analiza merytoryczna treści rozprawy i największe osiągnięcia naukowo-badawcze

Proponowane mini-modele posiadają nie tylko dobre właściwości interpolacyjne, ale również ekstrapolacyjne w przypadku korzystnego dobrania domen. Toteż w tym kontekście proponowane są do wykorzystania nieskomplikowane funkcje, częstokroć liniowe, co istotnie upraszcza obliczenia. W odniesieniu do referencyjnej metody k-Najbliższych Sąsiadów k-NN, mini-modele obliczając odległość w kontekście najbliższego sąsiedztwa wykorzystują dodatkowo informacje o nachyleniu płaszczyzny funkcyjnej trendu.

Głównym osiągnięciem pracy jest rozwinięcie metody mini-modeli do szeregu jej wersji działających na danych wysoko-wymiarowych. Do szczegółowych osiągnięć autora należy zaliczyć:

- modyfikacje metody mini-modeli do wersji funkcjonującej w przestrzeniach wielowymiarowych,
- wykorzystanie sferycznego układu współrzędnych przestrzeniach wielowymiarowych przy określaniu domeny mini-modeli,
- nowatorski heurystyczny algorytm uczenia mini-modeli skonstruowany na bryłach wielowymiarowych,
- opracowanie lokalnej metody redukcji wymiarowości,
- wykorzystanie algorytmów grupowania danych do dekompozycji przestrzeni wejść na domeny mini-modeli.

Proponowany heurystyczny algorytm uczenia składa się z dwóch etapów: identyfikacji adekwatnego lokalnego otoczenia punktu zapytania oraz nastrojenia wybranej metody modelowania lokalnego. Określenie optymalnego otoczenia punktu zapytania definiuje obszar domeny mini-modelu i polega na heurystycznej adaptacji generatorów ścian bryły stanowiącej domenę, tj. rotacji bryły i zmiany promieni generatorów. Natomiast jako metody matematycznego modelowania lokalnego wykorzystywane są metody klasycznej regresji liniowej, autorskiej heurystyki liniowej, heurystyki nieliniowej i aproksymacji wielomianowej, aczkolwiek w miejscu tym może zostać także zastosowana bardziej złożona metoda modelowania, jak np. jednokierunkowa sztuczna sieć neuronowa, czy system wnioskowania rozmytego.

Opracowane rozwiązania należą do klasy metod wykorzystujących podobieństwo, nazywanych w rozprawie metodami bazującymi na próbkach, stąd słusznie w badaniach porównawczych w głównej mierze odniesiono się do algorytmów z zakresu tzw. *Memory-Based Methods:* tj. k-NN, Regresyjnej Maszyny Wektorów Nośnych (RSVM), gaussowskiej regresji jądrowej (*Gaussian kernel regression*) sieci o radialnych funkcjach bazowych (RBF) i sieci typu *General Regression Neural Network* (GRNN). Badania przeprowadzono z wykorzystaniem własnych implementacji oraz funkcji bibliotecznych w środowisku Matlab.

W grupie mini-modeli dwuwymiarowych (rozdz. 4.1) przeprowadzone zostały badania mające na celu wizualizację i interpretację zachowania się metody w kontekście przestrzeni wielowymiarowych. Potwierdzona została skuteczność elaborowanej metody w kontekście uczenia

lokalnego. Nie we wszystkich przypadkach (np. liczba lat edukacji w modelowaniu wysokości wynagrodzenia – Tab. 4.9, liczba godzin koncentracji dwutlenku azotu – Tab. 4.8, czy liczba cylindrów w modelowaniu spalania benzyny – Tab. 4.10) proponowana metoda przyniosła najlepsze rezultaty dokładności działania. W większości jednak w grupie metod obarczonych większym niż minimalny osiągnięty, poziom skuteczności dla mini-modeli był (poza jednym przypadkiem – Tab. 4.10) porównywalny z osiągami najlepiej działającej metody. Należało się też spodziewać, co zostało wykazane w badaniach, że metody, które mają lokalnie założoną minimalną liczbę punktów modelują z mniejszym błędem, zapewne z racji lepszego reprezentowania domen przez dane. Kwestie wątpliwe są następujące:

- Co oznacza i jak należy interpretować (np. kolumnach Tab. 4.9) wskazanie na lokalną minimalną liczbę punktów równą zeru? Czy w tych miejscach podanie wartości uśrednionej nie byłoby miarodajne?
- Czy wartości parametrów minimalnej ilości próbek w domenie mini-modelu nie powinny być równe dla różnych metod modelowania lokalnego w danym eksperymencie badawczym (np. w Tab. 4.6)? Czy wartość ta nie powinna korespondować do liczby sąsiadów w metodzie k-NN (np. w Tab. 4.7, 4.8, 4.10)?

Podsumowując wyniki analizy otrzymanych rezultatów Autor słusznie zauważa, że nie można dobrać optymalnej metody dla wszystkich zbiorów danych testowanych w ramach badań. Tym samym proponowana metoda znajduje się w grupie metod opartych na próbkach optymalnych w sensie Pareto, jeśli zadania można potraktować jako kryteria.

W podrozdziale 4.1.2. dotyczącym badania tzw. luk informacyjnych, tj. w przypadku braku części danych (10%, 20% albo nawet 30%) wyniki posiadają hazardowy rozrzut przy testowaniu danych z luki informacyjnej, czego należało się spodziewać w przypadku tak wysokich procentowości brakujących danych oraz braku porównania z metodami wykorzystującymi zbiory interwałowe albo też zbiory rozmyto-przybliżone, które wykazują naturalną właściwość modelowania danych posiadających ubytki. Zastanawiająca jest przewyższająca skuteczność mini-modeli z lokalnie obraną minimalną liczbą punktów. W tym miejscu należy zadać pytania:

- Czy ten fakt nie wynika z metodologicznie błędnie przyjętego założenia, że pomimo nawet 30% brakujących danych, nadal każda domena jest obsadzona minimalną liczbą danych uczących? Czy w przypadku braku prawie co trzeciej danej nie powinno się zdarzać tak, że niektóre domeny są puste, jeśli przyjęliśmy dowolność lokalnego parametru m?

W drugiej grupie badań znalazły się mini-modele oparte o bryły wielowymiarowe w układzie sferycznym. Proponowane metody odznaczają się najmniejszym błędem tylko w dwóch zadaniach (Concret Compressive Strength oraz Yacht Hydrodynamics) spośród zaprezentowanych w tym zestawieniu sześciu. Najlepszymi metodami okazały się sieci RBF i GRNN oraz gaussowska regresja jądrowa w jednym przypadku. Jednakże proponowana wersja metody mini-modeli wyposażona jest w zdolność identyfikowania przypadków, w których algorytm nie określił domeny mini-modelu zgodnie z przyjętymi kryteriami. Jest to podejście szczególnie przydatne, gdy domena jest nazbyt obszerna i powoduje błędy numeryczne przy krańcach domeny lub też w całym obszarze w przypadku niedostatecznej liczby danych uczących. Należy zgodzić się z Autorem, iż „metoda jest w stanie konkurować z innymi metodami na próbkach" oraz że opracowanie efektywnej modyfikacji metody wymaga dalszych badań eksperymentalnych.

W odniesieniu do mini-modeli opartych o bryły wielowymiarowe (krótki rozdz. 4.3) potwierdzona została korzyść wyposażania metody w lokalną redukcję wielowymiarowości względem pozostałych metod opartych o mini-modele. Opracowanie skutecznej metody względem RBF, GRNN i regresji jądrowych wymaga wciąż nakładu pracy. Nie jasna jest tu konkluzja Autora,

- czy zawsze zredukowane zmienne są nieistotne i tym samym podwyższają błąd modelu jako swego rodzaju „szum informacyjny" (str. 149). Z opisu wynika, że opracowana metoda nie została wyposażona w sprzężenie zwrotne określające istotność odrzucanych danych.

Badania metody mini-modeli opartych o algorytmy grupowania danych wskazują na możliwości dalszego rozwoju, które tkwią w połączeniu modelowania lokalnego z identyfikowaniem domen lokalnych poprzez klasteryzację. Obiecujące, tj. w dwóch zadaniach najlepsze, dała w wyniku swego działania metoda mini-modeli bazująca na algorytmie *k*-średnich w wykorzystaniem wyjścia jako atrybutu wejściowego dla miary euklidesowej (Boston Housing i Servo – Tab. 4.25-4.26). Na bardzo wstępnym etapie są badania nad połączeniem modelowania lokalnego z metodą rozmytą *c*-średnich. Przełomowych rezultatów nie przyniosło ani podejście z wyostrzonymi ani z rozmytymi granicami klastrów, choć użycie *c*-średnich wydaje się być bardziej elastyczne i powinno dać lepsze rezultaty.

Jako możliwość szerszego zastosowania w rozdziale 4.5 zaprezentowana została adaptacja metody mini-modeli do zadań klasyfikacji danych. Wyniki proponowanej adaptacji choć tylko w zadaniu klasyfikacji kwiatów Irysa są najlepsze, nie odbiegają znacząco od czołowych standardowych metod klasyfikowania danych.

Należy zauważyć, że największą wadą opracowanej grupy metod jest złożoność obliczeniowa zależna od liczby próbek uczących i naturalnie metody te nie mogą być wprost zastosowane do analizy dużych zbiorów danych tzw. Big Data. W podsumowaniu Autor trafnie identyfikuje konsekwentny kierunek badań nad wykorzystaniem algorytmu kD-drzew w celu zmniejszenia złożoności obliczeniowej dla mini-modeli.

Kolejną naturalną drogą dalszego rozwoju wskazywaną przez Autora jest optymalizacja domeny mini-modelu. Również pewną poprawę działania mogą przynieść inne niecentroidalne algorytmy klasteryzacji, jak na przykład przytaczany DBSCAN. Jednak należy się spodziewać, że pomyślność tego podejścia będzie silnie uzależniona od rodzaju zadania regresji czy aproksymacji danych. Wreszcie w przypadku danych obarczonych znaczącą niedokładnością pomiarową zgodnie z propozycją Autora wykorzystanie metod logiki rozmytej i matematyki interwałowej powinno przynieść istotną poprawę skuteczności działania, zdaniem recenzenta zwłaszcza używając do tego celu zbiorów rozmytych typu-2 lub też podejścia posybilistycznego z całym dobrodziejstwem rozwijanych w ostatnich latach metod matematyki interwałowej oraz algebry zbiorów rozmytych typu-2. Natomiast w przypadku gdyby Autor miał na myśli niepewność pomiarową jako parametr rozkładu prawdopodobieństwa błędu, rozwiązań upatrywać należy w teorii Dempstera-Shafera lub innych metodach probabilistycznych.

3 Analiza struktury

Rozprawa obejmuje 196 stron i złożona jest z Wprowadzenia, pięciu numerowanych rozdziałów, rozdziału dodatkowego, spisów rysunków i tablic, Bibliografii liczącej 97 pozycji oraz Skorowidza. Rozprawa zawiera 34 rysunki i 48 tablic.

Dwa pierwsze rozdziały prezentują w wyczerpujący sposób informacje podstawowe niezbędne dla określenia autorskich elementów pracy. Rozdział pierwszy prezentuje w głównej mierze algorytmy oparte na próbkach oraz algorytmy grupowania danych wykorzystywane do modyfikacji metody mini-modeli w dalszej części pracy. Rozdział drugi traktuje o operacjach przekształcających domeny w postaci brył wielowymiarowych.

Najistotniejszą część rozprawy z punktu widzenia oryginalnego wkładu naukowego obejmują rozdziały 3, 4 i 5, w których kolejno poruszanymi obszarami są:

- autorskie warianty metody mini-modeli,
- badania eksperymentalne (podrozdziały 4.3 i 4.4 prezentują wyniki badań w znacznym skrócie w odniesieniu do wyczerpująco opisanych badaniach w poprzedzających podrozdziałach),
- podsumowanie wyników badań i wskazanie na dalsze kierunki rozwoju proponowanych metod.

4 Wątpliwości redakcyjne

- Na str. vii pojawia się pojęcie objętości jednostkowej hipersfery, podczas gdy objętością jest miara Lebesgue'a obszaru ograniczanego przez hipersferę.
- Na str. 60 pojawia się skrót myślowy wymagający wyjaśnienia z podaniem warunków takiego twierdzenia: *„Nie ma potrzeby używania aproksymacji wyższego rzędu, gdyż doprowadziłoby to jedynie do zwiększenia złożoności obliczeniowej algorytmu."*
- Tablica 4.2 nie została w tekście dostatecznie objaśniona. Przy pierwszym czytaniu rodziło się pytanie, w jakim celu wprowadzone są w wierszach metody, dla których nie przeprowadzono eksperymentów, dopiero w analizie wyników oczywistym staje się, że "l. eksperymentów" oznacza liczbę eksperymentów zakończonych sukcesem.

5 Wnioski końcowe

Proponowane w rozprawie metody prezentują ważny z punktu widzenia konstrukcji metod opartych na próbkach i autorski wkład do dyscypliny informatyka, a także przyczynek do rozwoju metod lokalnych analizy dużych zbiorów danych, czy też danych niepewnych, niedokładnych lub niekompletnych. W świetle przedstawionej rozprawy oraz wyników badań Autora

zaprezentowanych w czasopismach, publikacjach konferencyjnych i w zeszytach naukowych należy stwierdzić, że badania zostały zaprezentowane w sposób bardzo dokładny i szczegółowy. Przedstawione uwagi mają wyłącznie charakter redakcyjny, a jakość zaproponowanych rozwiązań algorytmicznych jest w mojej ocenia bardzo wysoka.

Stwierdzam, że przedstawiona do oceny rozprawa doktorska pana mgr. inż. **Marcina Pietrzykowskiego** pod tytułem ***„Lokalne uczenie algorytmów regresyjnych metodą mini-modeli"*** spełnia w sposób wyczerpujący wszystkie wymagania stawiane rozprawom doktorskim przez obowiązującą *Ustawę z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Art. 13 ust 1, z późn. zm., na mocy przepisów przejściowych)* odnośnie stopnia doktora nauk technicznych. Wnioskuję do Komisji **o dopuszczenie rozprawy doktorskiej do publicznej obrony a ponadto uwzględniając szczególność osiągnięć naukowych popartych publikacjami w czasopismach naukowych zgłaszam wniosek o wyróżnienie rozprawy.**

Janusz Starczewski