ZACHODNIOPOMORSKI UNIWERSYTET TECHNOLOGICZNY W SZCZECINIE

Wydział Informatyki

STRESZCZENIE ROZPRAWY DOKTORSKIEJ

mgr inż. Marcin Pietrzykowski

Lokalne uczenie algorytmów regresyjnych metodą mini-modeli

Promotor
prof. dr hab. inż. Andrzej Piegat

Promotor Pomocniczy
dr inż. Marcin Pluciński

Recenzenci
prof. dr hab. Leon Bobrowski
dr hab. inż. Janusz Starczewski, Prof. PCz

Szczecin 2019

Spis treści

1 Aktualność problemu

Ludzkość na przestrzeni swojej historii nieustannie tworzyła nowe dane. Wraz z nastaniem ery komputerów, a później z coraz powszechniejszym dostępem do Internetu i podłączaniem do niego coraz większej liczby urządzeń, przyrost ten stał się coraz gwałtowniejszy. Szacuje się, że w roku 2012 ludzkość tworzyła 2.5 miliarda nowych bajtów danych każdego dnia. 80% tych danych była nie ustrukturyzowana i zaliczały się do nich: filmy, zapisy audio, dane z różnego rodzaju czujników oraz dane z mediów społecznościowych[1]. Ludzie oprócz tworzenia danych zawsze starali się je analizować. Robili to w różnym celu: od zrozumienia praw rządzących światem, do zdobycia informacji w celu osiągnięcia różnego rodzaju korzyści. Z czasem zaczęły powstawać różnego rodzaju metody i algorytmy pozwalające na przetwarzanie danych. Jedną z pierwszych metod używanych do tego celu jest metoda najmniejszych kwadratów, będąca w użyciu od początku XIX wieku kiedy to została sformułowana[2].

Wraz z rozwojem nauki, powstaniem komputerów, a później wraz ze wzrostem ich mocy obliczeniowej powstał szereg różnego rodzaju algorytmów przetwarzania danych, w tym szereg metod sztucznej inteligencji. Dzięki tym algorytmom, dane mogą być m. in. analizowane, aproksymowane, klasyfikowane, grupowane. Na potrzeby niniejszej pracy skupiono się jedynie nad algorytmami służącymi do aproksymowania lub modelowania danych. Jedną z pierwszych metod mających niebagatelny wpływ na naukę była metoda najmniejszych kwadratów. Innymi ważnymi metodami z punktu widzenia niniejszego opracowania jest metoda najbliższych sąsiadów opracowana w latach 50-tych XX wieku przez E. Fix oraz J. Hodgesa [8]. W latach 60-tych XX wieku powstała podstawowa wersja Maszyny Wektorów Nośnych opracowana przez V. Vapnika [37]. W latach 80-tych ubiegłego wieku powstała metoda LOWESS [4] oraz sieci neuronowe typu RBF [2]. Lata 90-te ubiegłego wieku zaowocowały powstaniem sieci neuronowych typu GRNN[36] oraz opracowano Regresyjną Maszynę Wektorów Nośnych [7].

Krótki rys historyczny skupia się na metodach należących do grupy algorytmów bazujących na próbkach i obejmuje jedynie najważniejsze pozycje. Algorytmy te nie wykonują generalizacji, a jedynie porównują nową próbkę z danymi uzyskanymi w procesie uczenia. Metoda mini-modeli ze względu na swój sposób działania również wpisuje się do tego zestawienia. Algorytmy

[1]Dane korporacji IBM, `www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf`, dostęp: 24 lipca 2018

[2]Metoda została opublikowana przez Legendre'a w 1805 roku. W 1809 roku Carl F. Gauss opublikował swoją metodę obliczania pozycji ciał niebieskich, deklarując w artykule, że był w posiadaniu metody od 1795 roku.

tego typu charakteryzowane są przez szereg cech pozytywnych oraz negatywnych i z tego właśnie względu to one będą stanowiły grupę porównawczą dla badanej metody, o której traktuje ta praca.

Podstawowa koncepcja metody mini-modeli została opracowana przez prof. Andrzeja Piegata. Motywacją do opracowania metody była chęć ulepszenia algorytmu najbliższych sąsiadów. W swoim zamyśle metoda mini-modeli ma być metodą modelowania ogólnego i może zostać zastosowana dla danych różnorakiego pochodzenia. Zbiory danych numerycznych mogą pochodzić z różnego rodzaju systemów, na przykład: fizycznych, ekonomicznych czy też biologicznych. Jednym z przykładów modelowania może być określenie wielkości stopy bezrobocia przy znanym poziomie innych czynników ekonomicznych. Innym klasycznym przykładem, z którego prawie wszyscy na co dzień korzystamy, będzie modelowanie temperatury powietrza na danym obszarze. Kolejnym przykładem, tym razem dotykającym mediów społecznościowych, a mających duże znaczenie dla biznesu, będzie określenie wieku lub wielkości zarobków użytkownika na podstawie polubionych stron i wydarzeń na Facebooku w celu lepszego dostosowania do niego reklam. Wszystkie metody modelowania danych w sensie matematycznym starają się szacować postać funkcji opisującej dane wyjściowe na podstawie danych wejściowych. Dzieje się tak dlatego, że bardzo często nie znamy równania rzeczywistej funkcji rządzącej danym zjawiskiem, która opisuje zależności występujące pomiędzy danymi. Zaawansowane algorytmy sztucznej inteligencji w połączeniu z dużą mocą obliczeniową współczesnych komputerów są w stanie wykryć takie zależności nawet w danych wielowymiarowych.

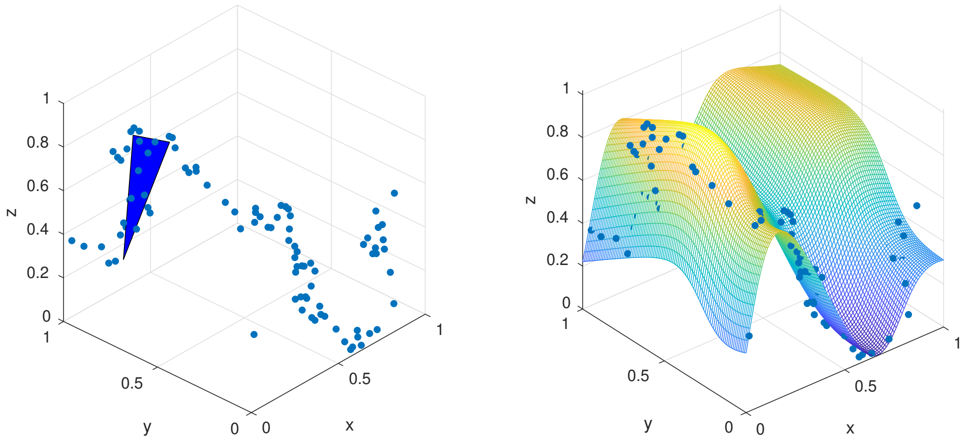
2 Przedmiot badań

Praca podejmuje temat lokalnego uczenia algorytmów regresyjnych metodą mini-modeli. Główna zasada działania prezentowanej metody opiera się na założeniu, że w procesie modelowania danych bardzo często jesteśmy zainteresowani odpowiedzią na jedno konkretne zapytanie. Motywacją do opracowania metody była chęć ulepszenia algorytmu najbliższych sąsiadów. W odróżnieniu od znanych i powszechnie stosowanych metod modelowania globalnego takich jak: sieci neuronowe, sieci neurorozmyte, aproksymacja wielomianowa, które dokonują aproksymacji w całej domenie systemu, metoda mini-modeli (metoda MM) nie tworzy modelu globalnego, jeżeli nie jest on wymagany. Główna zasada działania prezentowanej metody opiera się na założeniu, że w procesie modelowania danych bardzo często jesteśmy zainteresowani odpowiedzią na jedno tylko konkretne zapytanie. Wiedza na temat wartości zmiennej zależnej w całej modelowanej dziedzinie nie zawsze jest potrzebna. Dlatego metoda MM stara się zidentyfikować funkcję matematyczną opisującą zależność pomiędzy zmiennymi wejściowymi i wyjściowymi operując jedynie na danych znajdujących się w lokalnym otoczeniu punktu zapytania. Dla przykładowego zapytania: "Jaka wysoka będzie stopa bezrobocia w sytuacji, gdy podaż pieniądza będzie wynosić 950 mld zł, stopa inflacji 10%, liczba ludności 37.5 mln osób, a minimalne wynagrodzenie będzie na poziomie 1500 zł?", odpowiedź wymaga jedynie danych znajdujących się w otoczeniu punktu zapytania tj. danych, dla których podaż pieniądza wynosi około 950 mld zł, stopa inflacja jest na poziomie około 10%, liczba ludności wynosi około 37.5 mln osób, a minimalne wynagrodzenie wynosi około 1500 zł.

Metoda mini-modeli u swych podstaw wychodzi z założenia, że aby otrzymać odpowiedź na konkretne zapytanie nie trzeba budować globalnego modelu w całej dostępnej dziedzinie. Jest to odmienne od sposobu w jaki działają dla przykładu sieci neuronowe, które na podstawie próbek uczących starają się zbudować model mający zastosowanie w całej dziedzinie problemu. Zamiast tego wystarczy jedynie lokalny, prosty model zbudowany na danych znajdujących się w pewnym otoczeniu punktu zapytania. Funkcja matematyczna i w tym przypadku również jest estymowana, jednak nie obejmuje całej obszernej dostępnej dziedziny, a jedynie jej wycinek.

Na wstępie należy uwypuklić zasadniczą różnicę pomiędzy mini-modelem, a modelem globalnym. Mini-model operuje tylko i wyłącznie na punktach znajdujących się w granicach jego zasięgu, które są ściśle określone i nie obejmują całej modelowanej dziedziny tak jak w przypadku modeli globalnych. Mini-model liniowy tworzony jest na podstawie próbek należących do ściśle określonego ciągłego obszaru przestrzeni wejść zwanego dalej otoczeniem punktu zapytania, obszarem mini-modelu lub domeną mini-modelu. Co prawda w ekstremalnych przypadkach mini-model liniowy może rozszerzyć otoczenie punktu zapytania na całą dostępną dziedzinę, ale jest to sytuacja mogąca mieć miejsce jedynie podczas modelowania danych, w których zachodzi ścisła zależność liniowa, co jest raczej rzadkością. Model globalny operuje i musi uczyć się na wszystkich dostępnych próbkach. Różnice pomiędzy nimi ilustruje Rysunek 1.

Mini-model w swym założeniu pracuje przy użyciu prostych narzędzi takich jak na przykład regresja liniowa. Oczywiście istnieje możliwość tworzenia mini-modeli pracujących w oparciu o bardziej złożone metody. Preferowanym podejściem jest jednak odnajdywanie optymalnej domeny (podzbioru przestrzeni wejść), w której regresja liniowa przynosi najlepsze rezultaty, zamiast na przykład zwiększania złożoności modelu.



(a) Mini-model 3D (b) Model globalny 3D

Rysunek 1: Porównanie mini-modelu i modelu globalnego. *Źródło: opracowanie własne.*

3 Metody naukowe stosowane podczas wykonania badań

Hipoteza badawcza została uzasadniona empirycznie poprzez przedstawienie wyników eksperymentów oraz ich analizę. Metodyka badań była zgodna z ogólną procedurą wykrywania zależności pomiędzy danymi i budowania modeli na ich podstawie. Procedura ta polega na: postawieniu problemu, sformułowaniu hipotezy, przygotowaniu eksperymentu, pozyskaniu danych pomiarowych i ich wstępnym przetworzeniu, estymacji modelu, interpretacji wyników i formułowania wniosków [6].

Empiryczne udowodnienie hipotezy pracy wymagało przeprowadzania wielu eksperymentów, które zostały przeprowadzone na drodze symulacji komputerowej na zbiorach danych zaczerpniętych z repozytoriów internetowych. Badania nad metodą mini-modeli zostały przeprowadzone przy użyciu autorskiego oprogramowania komputerowego. Badana metoda jest oryginalnym pomysłem i nie jest szerzej znana w literaturze naukowej, a co za tym idzie nie posiadała ona istniejącej implementacji w postaci programu komputerowego. Uzyskane wyniki zostaną opracowane i porównane z wynikami metod konkurencyjnych, w szczególności z metodami bazującymi na próbkach.

4 Główny cel rozprawy

Głównym celem rozprawy jest:

> **udoskonalenie metody uczenia mini-modeli na bazie próbek lokalnych w przestrzeni wielowymiarowej**.

Hipoteza badawcza mówi, że:

> **udoskonalenie metody uczenia mini-modeli umożliwi wykazanie co najmniej części przewidywanych zalet tej metody względem geometrycznych metod modelowania bazujących na próbkach.**

5 Zadania do rozwiązania

W trakcie prac nad rozprawą konieczne stało się rozwiązanie następujących zadań:

- Opracowanie wersji mini-modeli działających w przestrzeni wielowymiarowej ($> 3D$).
- Konstruowanie domeny-mini modelu jako bryły geometrycznej w przestrzeni wielowymiarowej.
- Opracowanie algorytmów modyfikacji położenia bryły geometrycznej w przestrzeni wielowymiarowej.
- Opracowanie algorytmów działania i uczeni mini-modeli w oparciu o algorytmy klasteryzacji.
- Opracowanie algorytmów lokalnej redukcji wymiarowości.
- Przebadanie działania mini-modeli z danymi zawierającymi "luki informacyjne".

6 Wartość teoretyczna

Przeprowadzone badania wykonane na potrzeby rozprawy miały charakter badań podstawowych. W trakcie pisania pracy opisano oraz przebadano różne wersje metody mini-modeli. Opracowano mini-modele działające w przestrzeni wielowymiarowej, bazujące na sferycznym układzie współrzędnych. Zagadnienie to wymagało opracowania algorytmu ucznia mini-modeli oraz algorytmów konstruowania i modyfikowania położenia bryły geometrycznej w przestrzeni wielowymiarowej. Ponadto opracowano mini-modele bazujące na algorytmach klasteryzacji danych oraz mini-modele używające lokalnej redukcji wymiarowości.

Biorąc pod uwagę szereg parametrów definiujących sposób ucznia mini-modeli (m. in. różne bryły definiujące domenę mini-modelu, różne metody grupowania danych czy różne metryki) można przyjąć, że metoda mini-modeli została łącznie przebadana w blisko czterdziestu wariantach.

7 Wartość praktyczna

Opracowana w rozprawie metoda mini-modeli pozwala na jej praktyczne zastosowanie w zadaniu modelowania danych wielowymiarowych. W szczególności należy podkreślić tu następujące wartości metody:

- Dokładność metody jest na poziomie metod, z którymi była porównywana. Zgodnie z zasadą *No free lunch theorem*, nie jest to jednak metoda idealna oraz nie istnieje jedna wybrana wersja metody osiągająca optymalne rezultaty dla wszystkich zbiorów danych.
- Wersja metody bazująca na bryłach wielowymiarowych jest w stanie wykryć sytuacje, w której algorytm nie odnalazł domeny mini-modelu spełniającej nałożone na nią kryteria. Odpowiedź numeryczna modelu z reguły jest obarczona dużym błędem w sytuacji, w której domena mini-modelu jest zbyt rozciągnięta lub zawiera niewłaściwą liczbę próbek uczących.
- Mini-modele posiadają dobre właściwości ekstrapolacyjne. Metoda w obszarach wykraczających poza obszar interpolacyjny będzie zachowywała linię trendu wygenerowaną na podstawie danych leżących najbliżej punktu zapytania. Zachowywanie linii trendu ma bardzo duże znaczenie dla danych z luką informacyjną.
- Można dokonywać lokalnej redukcji wymiarowości co jest ważne w przypadku problemów wysoko-wymiarowych. Zmienne nieistotne mogą podwyższać błąd modelu działając jako szum informacyjny.
- Przewaga mini-modeli względem k-NN polega na wykorzystywaniu nie tylko odległości do najbliższego sąsiada, ale także informacji o nachyleniu powierzchni funkcyjnej.
- Mini-modele mogą być prostymi funkcjami np. liniowymi co ułatwia ich identyfikację.

8 Akceptacja wyników przez społeczność naukową

W toku prac nad metodą mini-modeli powstało 11 publikacji w czasopismach i materiałach konferencyjnych. Wyniki badań zostały zaprezentowane na następujących konferencjach:

- „Wpływ młodych naukowców na osiągnięcia polskiej nauki” w Częstochowie 19.11.2011
- „Advanced Computer Systems” w Międzyzdrojach 30.05 – 1.06.2012.
- „Wpływ młodych naukowców na osiągnięcia polskiej nauki – IV Edycja” w Gdańsku 14.04.2013
- „Krajowa Konferencja Studentów i Młodych Pracowników Nauki” w Szczecinie 22.05 – 23.05.2014
- „Advanced Computer Systems” w Międzyzdrojach 22.10 – 24.10.2014
- 14th International Conference on Artificial Intelligence and Soft Computing ICAISC 2015 w Zakopanym 14.06 – 18.06.2015
- „Sejmik Młodych Informatyków X edycja” w Międzyzdrojach 17.09 – 19.09.2015
- 15th International Conference on Artificial Intelligence and Soft Computing ICAISC 2016 w Zakopanym 12.06 – 16.06.2016
- „Advanced Computer Systems” w Międzyzdrojach 19.10 – 21.10.2016
- International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, Marseille, Francja, 6.09 - 8.09.2017

9 Osiągnięcia zgłaszane w ramach obrony pracy

Oryginalny wkład własny autora pracy stanowi:

- Opracowanie wersji mini-modeli działających w przestrzeni wielowymiarowej ($> 3D$).

- Zastosowanie sferycznego układu współrzędnych w procesie określania bryły geometrycznej w przestrzeni wielowymiarowej używanej do definiowania obszaru domeny mini-modelu.
- Heurystyczny algorytm uczenia mini-modeli oparty na bryłach wielowymiarowych.
- Zastosowanie lokalnej redukcji wymiarowości w celu odrzucenia zmiennych lokalnie nieistotnych.
- Zastosowanie algorytmów klasteryzacji danych w celu podziału przestrzeni wejść na domeny mini-modeli.

10 Skrócona prezentacja struktury i układu pracy

Praca została podzielona na cztery główne rozdziały. W rozdziale pierwszym opisano wybrane metody i zagadnienia podstawowe wykorzystywane w pracy. Opisując metody, skupiono się w głównej mierze na algorytmach bazujących na próbkach. Opisano również kilka algorytmów klasteryzacji, używanych w jednej z wersji mini-modeli. W rozdziale drugim opisano wybrane zagadnienia związane z bryłami wielowymiarowymi, ze względu na charakter domeny mini-modeli. Rozdział trzeci zawiera opis metody mini-modeli w różnych wariantach. W rozdziale czwartym znalazły się wyniki eksperymentów numerycznych przeprowadzonych na metodzie mini-modeli oraz na metodach wchodzących w skład grupy porównawczej. Rozdział piąty zawiera podsumowanie. Do pracy został również dołączony Dodatek A zwierający przykłady mini-modeli bazujących na sferycznym układzie współrzędnych.

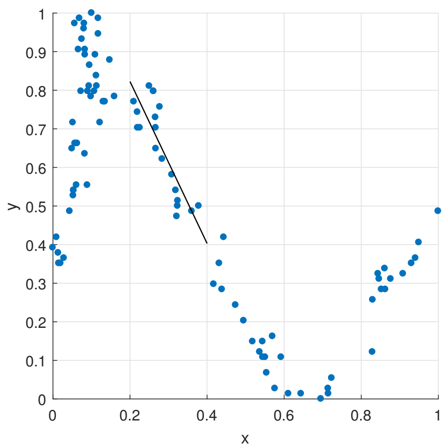
11 Zawartość pracy

W odróżnieniu od znanych i powszechnie stosowanych metod modelowania globalnego takich jak: sieci neuronowe, sieci neurorozmyte, aproksymacja wielomianowa, które dokonują aproksymacji w całej domenie systemu, metoda mini-modeli (metoda MM) nie tworzy modelu globalnego, jeżeli nie jest on wymagany. Główna zasada działania prezentowanej metody opiera się na założeniu, że w procesie modelowania danych bardzo często jesteśmy zainteresowani odpowiedzią na jedno tylko konkretne zapytanie. Wiedza na temat wartości zmiennej zależnej w całej modelowanej dziedzinie nie zawsze jest potrzebna. Dlatego metoda MM stara się zidentyfikować funkcję matematyczną opisującą zależność pomiędzy zmiennymi wejściowymi i wyjściowymi operując jedynie na danych znajdujących się w lokalnym otoczeniu punktu zapytania. Dla przykładowego zapytania: "Jaka wysoka będzie stopa bezrobocia w sytuacji, gdy podaż pieniądza będzie wynosić 950 mld zł, stopa inflacji 10%, liczba ludności 37.5 mln osób, a minimalne wynagrodzenie będzie na poziomie 1500 zł?", odpowiedź wymaga jedynie danych znajdujących się w otoczeniu punktu zapytania tj. danych, dla których podaż pieniądza wynosi około 950 mld zł, stopa inflacja jest na poziomie około 10%, liczba ludności wynosi około 37.5 mln osób, a minimalne wynagrodzenie wynosi około 1500 zł.

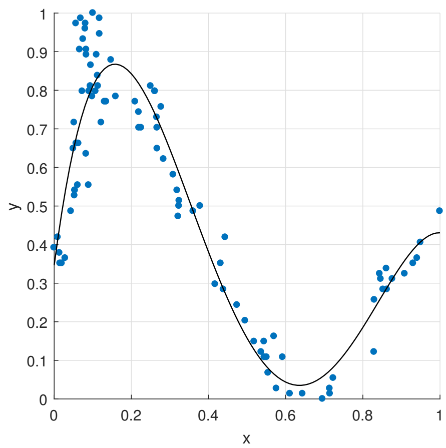
Metoda MM jest metodą modelowania lokalnego i operuje tylko i wyłącznie na danych znajdujących się w lokalnym otoczeniu punktu zapytania, które jest ściśle określone i nie obejmuje całej modelowanej dziedziny tak jak w przypadku modeli globalnych. Lokalne otoczenie puntu nazywamy *obszarem mini-modelu* lub *domeną mini-modelu*. Można je zdefiniować, jako obszar bryły geometrycznej określonej w przestrzeni wejść. Dla danych dwuwymiarowych (jedna zmienna niezależna, jedna zmienna zależna) będzie to odcinek, dla trójwymiarowych (dwie zmienne niezależne, jedna zmienna zależna) będzie to figura płaska (np. trójkąt lub czworokąt) dla danych czterowymiarowych będzie to bryła wypukła (np. czworościan, sześcian, ośmiościan). Obszar mini-modelu uogólniony do przestrzeni n-wymiarowej przyjmuje postać $(n-1)$-wymiarowej wypukłej i nieregularnej hiperbryły [26, 27, 29]. Domena mini-modelu może zostać zdefiniowana również jako hipersfera lub hiperelipsoida [31, 32].

11.1 Różnice pomiędzy mini-modelem a modelem globalnym

Na wstępie należy podkreślić zasadniczą różnicę pomiędzy mini-modelem, a modelem globalnym. Mini-model operuje tylko i wyłącznie na punktach znajdujących się w granicach jego zasięgu, które są ściśle określone i nie obejmują całej modelowanej dziedziny tak jak w przypadku modeli globalnych. Model globalny operuje i uczy się na wszystkich dostępnych próbkach. Omawianą różnicę w przestrzeni 2D obrazuje Rysunek 2 przedstawiający różnice pomiędzy mini-modelem, a modelem globalnym. Model globalny (regresja wielomianowa na Rysunku 2a) korzysta ze wszystkich dostępnych próbek i obejmuje całą dziedzinę problemu. Mini-model liniowy (Rysunek 2b) tworzony jest na podstawie próbek należących do ściśle określonego ciągłego obszaru przestrzeni wejść zwanego dalej otoczeniem punktu zapytania, ob-



(a) Mini-model



(b) Model globalny

Rysunek 2: Porównanie mini-modelu i modelu globalnego. *Źródło: opracowanie własne.*

szarem mini-modelu lub domeną mini-modelu.

Co prawda, w ekstremalnych przypadkach mini-model liniowy może rozszerzyć otoczenie punktu zapytania na całą dostępną dziedzinę, ale jest to sytuacja mogąca mieć miejsce jedynie podczas modelowania danych, w których zachodzi ścisła zależność liniowa, co jest raczej rzadkością.

Metoda MM jest metodą modelowania lokalnego i w opisanej wersji nie tworzy modelu globalnego takiego jak np. sieci neuronowe. Jeżeli istnieje potrzeba można dla mini-modeli również utworzyć model globalny poprzez "sklejenie" ze sobą mini-modeli dla kolejnych sąsiadujących odpowiednio rozmieszonych punktów zapytania. Należy jednak pamiętać że model globalny utworzony w ten sposób nie będzie gładkim modelem ciągłym. Powierzchnia tego typu modelu globalnego nie będzie gładka na granicach obszarów utworzonych przez niezależne i sąsiadujące ze sobą lokalne mini-modele. Sytuacja ta jest podobna do sytuacji, w której chcielibyśmy utworzyć model globalny przy pomocy metody k-NN. Utworzenie gładkiego modelu globalnego pozwala nam zobrazować jak kształtuje się modelowana przez nas wartość w całej domenie w przestrzeni dwu- lub trójwymiarowej.

Metoda mini-modeli należy do grupy metod bazujących na próbkach *(ang. memory-based learning)*. Algorytmy należące do tej grupy nie dokonują generalizacji na podstawie danych uczących, ale porównują nowe próbki danych z już istniejącymi danymi pochodzącymi z treningu, a przetrzymywanymi w pamięci. Dlatego niekiedy w angielskojęzycznej literaturze określa się te metody jako *memory-based learning* czyli metody bazujące na pamięci [1, 3, 16].

Podejście takie ma oczywiście swoje wady i zalety. Złożoność danego zagadnienia rośnie wraz z liczbą próbek uczących. Istnieje również ryzyko przeuczenia modelu oraz jego podatności na dane odstające lub zaszumione. Z drugiej jednak strony, algorytmy oparte na próbkach mają zdolność do szybkiego przystosowania się do nowych danych, które wcześniej nie były obserwowane. Nowa próbka danych jest po prostu dodawana do pamięci. Daje to przewagę nad innymi algorytmami uczenia maszynowego w sytuacji, w której do systemu ciągle spływają nowe dane.

11.2 Mini-modele w przestrzeni dwuwymiarowej

Początkowe prace nad mini-modelami obejmowały modelowanie problemów dwu- i trójwymiarowych [19, 20, 21, 22, 23]. Lokalne otoczenie punktu zapytania czyli domena mini-modelu przybiera wtedy kształt odcinka lub figury płaskiej w zależności od wymiarowości problemu. Definiowanie zakresu domeny polega na operowaniu wierzchołkami bryły geometrycznej. Zmiana położenia figury geometrycznej lub zmiana jej kształtu sprowadza się do zmiany położenia jej wierzchołków.

Lokalne otoczenie nie jest stałe i może przybierać różne wartości w zależności od położenia punktu zapytania, modelowanego problemu i użytego wariantu mini-modeli. Głównym parametrem definiującym domenę mini-modelu jest parametr minimalnej liczby próbek (parametr m), która powinna brać udział w uczeniu mini-modelu. Przyjmuje się, że parametr jest zdefiniowany następująco:

$$m \geqslant d + 1, \tag{1}$$

gdzie d jest to liczba wymiarów przestrzeni. W rozważanej przez nas przestrzeni 2D, $m \geqslant 3$. Zbiór danych uczących P definiujemy jako zbiór punktów p_i posortowanych rosnąco względem zmiennej x:

$$\begin{gathered} p_i = (x_i, y_i), \\ P = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}, \text{ gdzie } \quad x_1 \leqslant x_2 \leqslant \cdots \leqslant x_I. \end{gathered} \tag{2}$$

Lokalne otoczenie punktu zapytania zdefiniowanego jako x_0 definiujemy poprzez jego granice: dolną x_a i górną x_b:

$$\begin{gathered} x_a, x_b \in X, \\ x_0 \in [x_a; x_b]\,. \end{gathered} \tag{3}$$

Zbiór punktów, na którym będzie operował mini-model definiujemy jako M

$$\begin{aligned} M = \{p_i \in P : x_i \in [x_a; x_b]\}, \\ card(M) \geqslant m. \end{aligned} \tag{4}$$

W trakcie badań zostały opracowane dwa warianty metody: liniowa i nieliniowa.

11.2.1 Mini-modele liniowe 2D

Mini-model liniowy (MM) w procesie budowy modelu matematycznego używa regresji liniowej [20, 21, 22, 23]. Podczas uczenia MM powstaje pytanie jak zdefiniować optymalne lokalne otoczenie punktu zapytania. Otoczenie takie powinno być ciągłym podzbiorem danych uczących, na którym regresja liniowa będzie generowała jak najmniejszy błąd. Sprowadza się to w praktyce do odpowiedniego zdefiniowania puntów granicznych P_a, P_b. Jedną z odpowiedzi na to pytanie jest zachłanne przeszukanie wszystkich możliwych kombinacji spełniających podane wyżej warunki (1) - (4).

W odróżnieniu od mini-modeli bazujących na regresji liniowej, istnieje możliwość heurystycznego uczenia MM, w którym nie występuje problem zdefiniowania lokalnego otoczenia punktu zapytania. Jest ono tworzone na bieżąco wraz z uczniem MM. Zwalnia nas to z zachłannego przeszukiwania wszystkich możliwych kombinacji. Algorytm rozpoczyna się od wylosowania położenia punktów granicznych, tak aby spełniały warunki (1) - (4). Uczenie heurystyczne składa się z dwóch podstawowych operacji: przesuwania punktu granicznego wzdłuż osi OX oraz przesuwania wzdłuż osi OY.

11.2.2 Mini-modele nieliniowe 2D

Mini-modele nieliniowe, w odróżnieniu od metody opisanej powyżej, w procesie uczenia przybierają kształt wygiętego odcinka. Odcinek może mieć postać fragmentu krzywej wielomianowej, fragmentu łuku okręgu lub przyjmować inną matematyczną postać. Ze względu na swój nieregularny kształt mini-modele nieliniowe posiadają możliwość lepszego dopasowania się do modelowanych danych.

Sposób działania mini-modeli opartych o aproksymację wielomianową jest analogiczny do działania MM bazujących na regresji liniowej. W celu znalezienia optymalnego lokalnego otoczenia punktu zapytania, tutaj również niezbędne jest zachłanne przeszukanie wszystkich możliwych kombinacji punktów uczących spełniających warunki (1) - (4). Różnica polega na użyciu aproksymacji wielomianowej drugiego stopnia zamiast regresji liniowej.

Nie ma potrzeby używania aproksymacji wyższego rzędu, gdyż doprowadziłoby to jedynie do zwiększenia złożoności obliczeniowej algorytmu. Metoda mini-modeli ze względu na swoją zasadę działania jest w stanie modelować skomplikowany kształt funkcji przy pomocy kilku stosunkowo prostych mini-modeli.

Oprócz mini-modeli bazujących na aproksymacji wielomianowej istnieją także MM uczone heurystycznie. Heurystyczne MM nieliniowe w początkowej fazie uczone są tak samo jako ich liniowe odpowiedniki. Uczenie odbywa się poprzez cykliczne naprzemienne przemieszczanie punktów granicznych wzdłuż osi OX i OY. Procedura powtarza się, aż zostanie spełniony warunek zatrzymania. W efekcie otrzymujemy najlepszy mini-model będący odcinkiem pomiędzy punktami P_a i P_b. Następnie w celu najlepszego dopasowania znalezionego rozwiązania do modelowanych danych odcinek jest wyginany jako łuk okręgu.

11.2.3 Parametr minimalnej liczby próbek

Dobranie parametru minimalnej liczby próbek (parametr m) jest sprawą kluczową dla poprawnego działania mini-modelu. Niestety nie istnieje żadna reguła mówiąca w jaki sposób dobrać minimalną liczbę próbek do danego modelowanego problemu. Sytuacja jest tu analogiczna do problemu doboru parametru liczby sąsiadów w algorytmie k-NN. Ustawienie parametru *a priori* na wybraną wartość z reguły nie da optymalnych wyników.

Parametr m może zostać dobrany globalnie dla całego zbioru danych uczących. Warto zaznaczyć, że optymalna wartość parametru m z reguły jest różna dla różnych typów mini-modeli. Wartość parametru m może zostać ustawiona lokalnie, dla wybranego przez nas podzakresu. Dziedzinę można podzielić na kilka podprzedziałów i dla każdego z nich ustawić lokalną wartości parametru m. Dzięki tego typu podejściu możliwe jest lepsze dopasowanie mini-modeli do modelowanych danych. Obszary danych cechujące się większą zmiennością będą posiadały inną wartość parametru m od obszarów bardziej jednorodnych. Wyniki eksperymentów numerycznych uwodniły, że mini-modele, w których zastosowano lokalną wartość parametru m osiągały lepsze wyniki od mini-modeli, w których wartość parametru była ustawiana globalnie dla całego przedziału.

11.3 Mini-modele w przestrzeni wielowymiarowej

W początkowych badaniach obejmujących mini-modele w przestrzeni dwuwymiarowej, w procesie definiowania lokalnego otoczenia punktu zapytania operowano wierzchołkami bryły geometrycznej. Podejście to jest możliwe do zastosowania w przestrzeni trójwymiarowej, jednakże zastosowanie go w przestrzeni o większej liczbie wymiarów byłoby nieefektywne. Operowanie wierzchołkami bryły geometrycznej jest obarczone "przekleństwem wymiarowości", gdyż liczba wierzchołków w bryle opartej na hipersześcianie wzrasta wykładniczo. Kolejnym problemem jest problem współpłaszczyznowości wierzchołków należących do jednej ściany. Gdy ściana ma postać wieloboku, zmiana położenia jednego wierzchołka wymusza zmianę położenia innych wierzchołków należących do tej samej ściany w celu zachowania ich współpłaszczyznowości. Ponadto występuje szereg innych problemów takich jak: problem manipulowania ścianą jako zbiorem wierzchołków, problem rozstrzygania o zawieraniu się danego punktu we wnętrzu domeny mini-modelu opisanej jedynie przez jej wierzchołki. Złożoność tych zagadnień wzrasta wraz z wymiarowością przestrzeni. Wyżej wymienione problemy doprowadziły do powstania koncepcji manipulowania ścianą jako całością oraz przeniesienia większości obliczeń z kartezjańskiego do sferycznego układu współrzędnych [26, 27, 29]. Centrum układu stanowi punkt zapytania. Ściana zdefiniowana jest poprzez pojedynczy punkt zwany generatorem ściany. Istnieje założenie, że ściana jest ortogonalna do wektora utworzonego przez ww. punkt oraz początek układu współrzędnych. Koncepcja ta posiada następujące zalety:

- brak problemu współpłaszczyznowości wierzchołków należących do jednej ściany,
- stosunkowo proste obliczenia rozstrzygające o zawieraniu się punktu we wnętrzu domeny mini-modelu,
- zmniejszenie liczby parametrów potrzebnych do sterowania mini-modelem,
- stosunkowo proste manipulowanie położeniem ściany,
- stosunkowo prosta rozszerzalność mini-modelu do przestrzeni o wyższej wymiarowości.

Zastosowanie tego podejścia pozwala uniknąć przekleństwa wymiarowości (mini-model bazujący na obszarze hipersześcianu w przestrzeni $n-$wymiarowej będzie posiadał 2^{n-1} wierzchołków, ale tylko $2(n-1)$ ścian). Ma to duże znaczenie, ponieważ pozwala uprościć trudny heurystyczny proces uczenia się mini-modelu.

11.3.1 Domena mini-modelu jako bryła w przestrzeni wielowymiarowej

Pierwsza część algorytmu realizuje konwersję punktów danych z układu współrzędnych kartezjańskich do układu współrzędnych hipersferycznych [26, 27, 29]. Transformacja następuje jedynie w przestrzeni wejść, zmienna wyjściowa pozostaje bez zmian. Oznacza to, że trójwymiarowy punkt danych zostanie przetransformowany do układu współrzędnych biegunowych, a punkt czterowymiarowy do układu współrzędnych sferycznych. W ogólnym przypadku n-wymiarowy punkt zostanie przetransformowany do układu współrzędnych opartych na $(n-1)$-wymiarowej sferze. Punkt zapytania $Q = \{x_{Q1}, x_{Q2}, \ldots, x_{Qn}, y_Q\}$, staje się centrum układu współrzędnych. Wartości zmiennych wejściowych $x_{Q1}, x_{Q2}, \ldots, x_{Qn}$ są znane, wartość zmiennej wyjściowej y_Q jest niewiadomą, którą zostanie obliczona w wyniku działania algorytmu. Wszystkie punkty danych p_i są również konwertowane do hipersferycznego układu współrzędnych. Punkt jest określony przez promień $r \in [0, \infty)$, (odległość od centrum) i kąty $\varphi_{i1}, \varphi_{i2}, \ldots, \varphi_{i(n-2)} \in [0; \pi), \varphi_{i(n-1)} \in [0; 2\pi)$. Zbiór P składający się z I punktów oznaczamy jako:

$$\begin{aligned} &P = \{p_1, p_2, \ldots, p_i, \ldots, p_I\}, \\ &p_i = (x_{i1}, \ldots, x_{in}, y_i) = (r_i, \varphi_{i1}, \ldots, \varphi_{i(n-1)}, y_i). \end{aligned} \tag{5}$$

W ogólnym przypadku domena mini-modelu jest wypukłą nieregularną hiperbryłą składającą się z J ścian. Dla simpleksu liczba ścian wynosi $J = n + 1$, dla bryły bazującej na hipersześcianie $J = 2n$, gdzie n jest to wymiarowość przestrzeni. Każda ściana j hiperbryły jest częścią płaszczyzny F_j. W przyszłych rozważaniach całą płaszczyznę F_j będziemy nazywać ścianą. Płaszczyznę w sferycznym układzie współrzędnych możemy zdefiniować przy pomocy pojedynczego punktu G_j zwanego *generatorem ściany*. Istnieje założenie, że płaszczyzna F_j jest ortogonalna do wektora utworzonego przez generator ściany $\overrightarrow{QG_j}$. Każdą płaszczyznę F_j możemy zdefiniować jako:

$$F_j = \left\{ G_j, p_i : \varphi_{ij} < \frac{\pi}{2} \wedge r_i = \frac{r_j}{\cos \varphi_{ij}} \right\}, \tag{6}$$

gdzie φ_{ij} jest wartością kąta pomiędzy wektorami $\overrightarrow{QG_j}$ i $\overrightarrow{Qp_i}$. Płaszczyzna F_j dzieli całą przestrzeń wejść na dwie półprzestrzenie. Pierwsza półprzestrzeń zawiera punkty danych, które mogą zostać włączone w domenę mini-modelu. Możemy ją zdefiniować następująco:

$$I_j = \left\{ p_i : \varphi_{ij} \geqslant \frac{\pi}{2} \cup \left(\varphi_{ij} < \frac{\pi}{2} \wedge r_i \leqslant \frac{r_j}{\cos \varphi_{ij}} \right) \right\}. \tag{7}$$

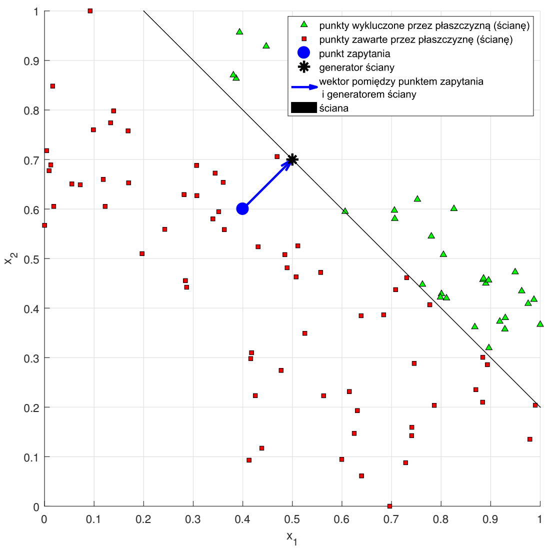
Druga półprzestrzeń zawiera punkty, które nie wejdą w skład domeny mini-modelu i jest zdefiniowana jako:

$$E_j = \left\{ p_i : \varphi_{ij} < \frac{\pi}{2} \wedge r_i > \frac{r_j}{\cos \varphi_{ij}} \right\}. \tag{8}$$

Każda płaszczyzna zawierająca w sobie ścianę bryły definiującej domenę mini-modelu dzieli przestrzeń wejść w ten sposób bez względu na wymiarowość. Cześć wspólna wszystkich półprzestrzeni I_j definiuje domenę mini-modelu i zawiera w sobie punkty uczące:

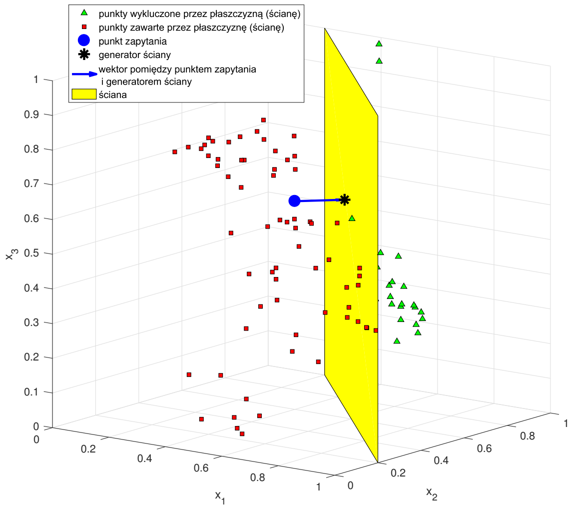
$$M = I_1 \cap I_2 \cap \ldots I_J. \tag{9}$$

Sposób, w jaki płaszczyzna dzieli przestrzeń wejść został przedstawiony na Rysunkach 3 i 4. Punkty oznaczone, trójkątami na pewno zostaną wyłą-

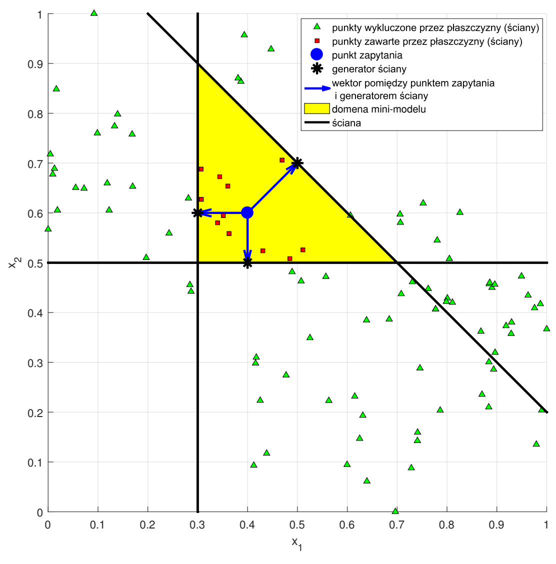


Rysunek 3: Przykład podziału przestrzeni 2D przez prostą (ścianę). *Źródło: opracowanie własne.*

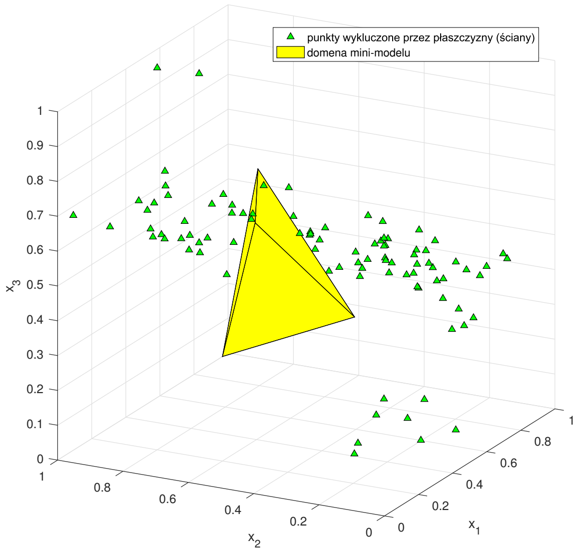
czone poza domenę mini-modelu natomiast punkty oznaczone kwadratami mogą zostać w niej zawarte. O tym czy punkt rzeczywiście wejdzie w skład domeny mini-modelu decyduje jego położenie względem innych płaszczyzn. Tylko punkty, które będą zawarte przez wszystkie płaszczyzny wejdą w skład domeny MM. Przykład jak ściany tworzą domenę mini-modelu dla problemu trójwymiarowego został pokazany na Rysunku 5. Na Rysunku 6 została



Rysunek 4: Przykład podziału przestrzeni 3D przez płaszczyznę (ścianę). *Źródło: opracowanie własne.*



Rysunek 5: Przykład domeny mini-modelu dla problemu 3D. *Źródło: opracowanie własne.*



Rysunek 6: Przykład domeny mini-modelu dla problemu 4D. *Źródło: opracowanie własne.*

przedstawiona przykładowa domena dla problemu czterowymiarowego. Punkty danych wyselekcjonowane w powyższy sposób stają się danymi uczącymi dla metody modelowania matematycznego używanej przez mini-model.

Bardzo ważną operacją używaną podczas uczenia mini-modeli jest rotacja całej domeny mini-modelu czyli rotacja całej bryły geometrycznej w przestrzeni wielowymiarowej. Rotacji można dokonać poprzez pomnożenie współrzędnych *generatorów ścian* przez odpowiednią macierz rotacji.

11.3.2 Algorytm uczenia mini-modeli

Uczenie mini-modelu składa się z dwóch części i polega na określeniu optymalnego lokalnego otoczenia punktu zapytania oraz nastrojeniu metody modelowania matematycznego używanego przez MM. Dzięki temu MM posiada zdolność do lokalnej adaptacji w wybranym obszarze dziedziny problemu. Jako algorytm modelowania matematycznego może zostać użyta dowolna metoda modelowania: regresja liniowa, regresja wielomianowa, wartość średnia z próbek, wnioskowanie rozmyte, sieć neuronowa, itd.

W praktyce jako algorytmu modelowania matematycznego bardzo często używa się regresji liniowej. Wybrany algorytm jest stosunkowo prosty i jest zdolny jedynie do modelowania zależności liniowej. Nie nadaje się on do modelowania złożonych zależności występujących w całej domenie rzeczywistego zbioru danych. Pomimo swej prostoty, regresja liniowa jest wystarczająca aby modelować lokalne zależności, które występują w małej ograniczonej części domeny.

Określenie optymalnego otoczenia punktu zapytania, czyli domeny MM jest procesem heurystycznym i polega na zmianie położenia ścian bryły otaczającej domenę. W praktyce sprowadza się jedynie do zmian położenia punktów zwanych *generatorami ścian* G_j. Istnieje wiele potencjalnych lokalnych otoczeń punktu zapytania, które mogą zostać osiągnięte w procesie uczenia. Po zdefiniowaniu domeny mini-modelu, metoda używa punktów wchodzących w jego skład, jako danych uczących dla algorytmu modelowania matematycznego. Następnie mini-model oblicza średni błąd generowany przez model matematyczny na próbkach uczących i obliczana jest wartość dla zmiennej zależnej punktu zapytania. W następnym kroku mini-model stara się odnaleźć kolejną domenę, następnie oblicza błąd generowany na próbkach oraz odpowiedź mini-modelu na punkt zapytania. Domena MM, z najmniejszym błędem zostaje uznana za optymalną. Należy jednak zaznaczyć, że nie każda domena mini-modelu jest poprawna. Punkt zapytania musi znajdować się w domenie, a jej powierzchnia musi być bryłą (hiperbryłą) wypukłą. Poza tymi podstawowymi założeniami domena powinna spełniać następujące warunki:

1. liczba punktów uczących wchodząca w skład mini-modelu powinna być większa lub równa niż zadana wartość minimalna,

2. liczba punktów uczących wchodząca w skład mini-modelu powinna być mniejsza lub równa niż zadana wartość maksymalna,

3. stosunek długości wektorów $\overrightarrow{QG_j}$ o minimalnej i maksymalnej długości, powinien być mniejszy niż zadana wartość,

4. punkt zapytania nie powinien być ekstrapolowany przez próbki uczące (nie zawsze jest to możliwe, a czasem nie jest wymagane).

11.4 Mini-modele z lokalną redukcją wymiarowości

Metoda mini-modeli może zostać rozszerzona o lokalną redukcję wymiarowości. Lokalna redukcja wymiarowości opiera się na założeniu, że wszystkie zmienne wejściowe nie muszą być tak samo istotne w poszczególnych częściach domeny. Dla przykładu pewna zmienna, statystycznie istotna w danej części domeny może być kompletnie nieistotna w innym jej obszarze, a tym samym może być źródłem szumu zakłócającego dokładność modelowania. W przestrzeni wielowymiarowej partycjonowanie przestrzeni w regularne sektory jest zadaniem trudnym, ze względu na zdolność pokrycia każdego sektora wystarczającą liczbą danych. Problem ten jest określany "klątwą wymiarowości". Z tego względu istotność zmiennych jest sprawdzana osobno dla każdego punktu zapytania [18].

Jak już wcześniej zostało napisane, uczenie mini-modelu składa się z dwóch etapów i polega na określeniu optymalnego lokalnego otoczenia punktu zapytania oraz nastrojeniu metody modelowania matematycznego używanego przez MM. Zaproponowana modyfikacja metody używa regresji liniowej jako algorytmu modelowania matematycznego. Użyty algorytm regresji liniowej jest stosunkowo prosty i potrafi modelować jedynie zależności liniowe. Pomimo to jest on wystarczający aby modelować lokalne zależności, które występują w małej ograniczonej części domeny. Ponadto istnieje dobrze znany zestaw narzędzi statystycznych potrafiących przetestować dokładność modelowania oraz poszczególne parametry metody. W celu dokonania lokalnej redukcji wymiarowości wykonywana jest regresja krokowa. Algorytm sprawdza statystyczną istotność zmiennych przy pomocy testu t-studenta. Zmienne statystycznie nieistotne zostają usunięte z modelu. Odrzucanie zmiennych dla wybranego punktu zapytania jest w praktyce lokalną redukcją wymiarowości. Istnieją oczywiście inne metody statystyczne pozwalające sprawdzić dopasowanie regresji, jak np. test F. Jednakże test t pozwala na sprawdzenie poszczególnych parametrów modelu [9, 11, 13, 33].

Procedura uczenia mini-modelu jest bardzo podobna do algorytmu uczenia mini-modeli opartych o bryły wielowymiarowe. Różnica polega jedynie na rozszerzeniu procedury uczenia modelu matematycznego. Wyliczenie współczynników równania regresji liniowej zostaje rozszerzone do regresji krokowej, która w oparciu o testy statystyczne badające istotność zmiennych, odrzuca kolejne zmienne nieistotne. Bez zmian zostaje natomiast część algorytmu odpowiedzialna za znajdowanie domeny mini-modelu. Polega ona tak jak wcześniej na losowej zmianie położenia punktów G_j, w wyniku czego zostaje znalezionych wiele potencjalnych domen MM.

Także w tej wersji mini-modeli należy każdy model traktować osobno. Model dla każdego kolejnego punktu zapytania może mieć różny zestaw zmiennych istotnych.

11.5 Mini-modele oparte o klasteryzację

Proces określania domeny mini-modelu będący częścią procesu jego uczenia może zastać zastąpiony przez algorytmy klasteryzacji danych [28, 30]. Użycie metod klasteryzacji modyfikuje jedynie pierwszy etap uczeni mini-modelu polegający na odnalezieniu obszaru optymalnej domeny dla wybranego punktu zapytania. Domena MM zawiera w sobie zastaw punktów, które zostaną później użyte w procesie uczenia. W przedstawionej modyfikacji metody obszar przestrzeni wejść problemu zostanie podzielony na domeny mini-modeli w wyniku działania algorytmu klasteryzacji. Każdy klaster będzie stanowił osobną domenę mini-modelu. W pracy skupiono się na metodach opartych o centroidy takich jak: algorytm k-średnich, k-medoidów, c-średnich.

Każdy klaster może być interpretowany jako oddzielna domena MM. Każdy MM używa jedynie takiego zestawu punktów uczących, który zawiera się w odpowiadającej mu domenie (klastrze). Należy pamiętać, że mini-modele nie są ze sobą połączone i każdy MM należy traktować osobno. Metoda nie tworzy spójnej gładkiej powierzchni funkcyjnej w całej modelowanej dziedzinie. Jest to zachowanie analogiczne do sposobu działania innych wersji mini-modeli opisanych powyżej.

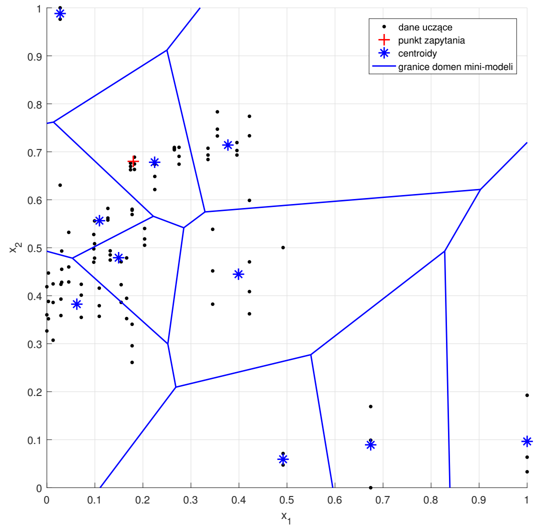
Ze względu na to, że metoda c-średnich potrafi przydzielić punkt danych do kilku klastrów jednocześnie, w dalszej części algorytmu możliwe jest zastosowanie kilku wariantów. Domenę mini-modelu dla wybranego punktu zapytania można zdefiniować w dwóch wersjach:

- Wyostrzenie granic - granice poszczególnych domen będę oddzielne i ściśle wyznaczone. Punkt danych będzie należał do domeny (klastra) z największą wartością funkcji przynależności.

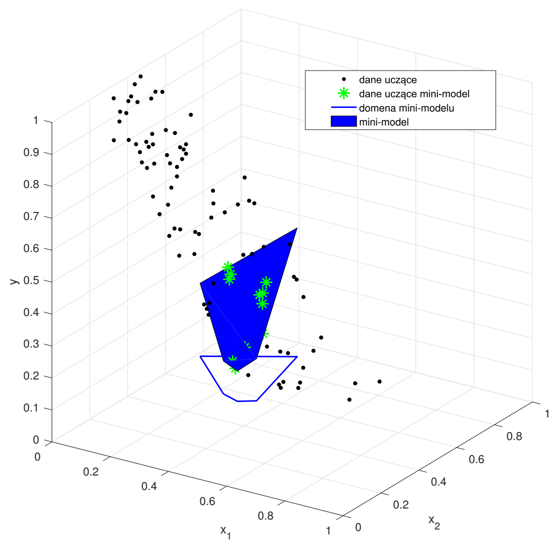
- Zastosowanie progu przynależności - granice poszczególnych domen będą rozmyte i będą na siebie nachodziły. Punkt danych będzie należał do tych domen, w których wartość funkcji przynależności będzie powyżej zdefiniowanego progu.

Zastosowanie algorytmu klasteryzacji danych możliwe jest w dwóch wariantach. W pierwszej wersji algorytm klasteryzacji używa zmiennych niezależnych i zmiennej zależnej tj.: $x_1, x_2, \ldots, x_n, y$. Jednakże w części przypisywania punktu do konkretnego klastra algorytm używa jedynie zmiennych niezależnych: $x_1, x_2, \ldots, x_n$. Jest to spowodowane faktem, że punkt zapytanie oczywiście nie posiada informacji na temat zmiennej zależnej y. W drugiej wersji algorytm używa cały czas tylko zmiennych niezależnych tj.: $x_1, x_2, \ldots, x_n$.

Rysunek 7 przedstawia przykładowy podział przestrzeni wejść przy użyciu algorytmu k-średnich. Każdy wielokąt jest domeną oddzielnego mini-modelu. Rysunek 8 prezentuje mini-model w przestrzeni trójwymiarowej.



Rysunek 7: Przykład podziału przestrzeni 2D przez algorytm k-średnich. *Źródło: opracowanie własne.*



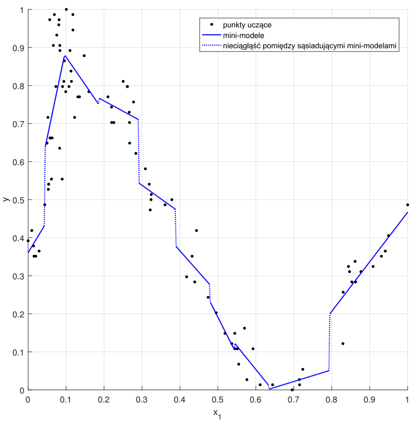
Rysunek 8: Przykład działania metody mini-modeli bazującej na algorytmie k-średnich w przestrzeni 3D. *Źródło: opracowanie własne.*

11.6 Mini-model jako klasyfikator

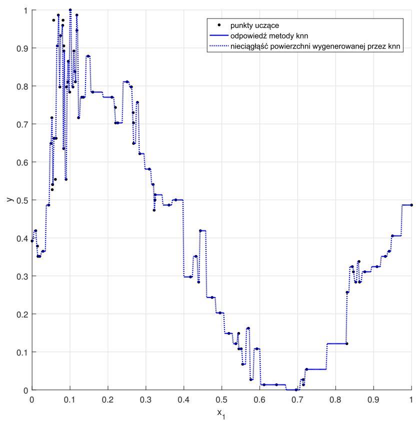
Pomimo że celem niniejszej rozprawy jest przedstawienie mini-modeli jako metody uczenia algorytmów regresyjnych, warto wspomnieć o możliwości zastosowania opisywanej metody w zadaniu klasyfikacji. Jak wcześniej zostało napisane uczenie mini-modelu składa się z dwóch części i polega na określeniu optymalnego lokalnego otoczenia punktu zapytania oraz nastrojeniu metody modelowania matematycznego używanego przez MM. Druga część algorytmu może zostać zmodyfikowana i zamiast uczyć metodę modelowania matematycznego można nastroić metodę klasyfikacji. W tego typu podejściu możemy użyć dowolnej metody klasyfikującej dane. Idea działania jest identyczna jak w przypadku użyciu MM w zadaniu aproksymacji. Metoda działa lokalnie i osobno dla każdego kolejnego punktu zapytania. Dokonywana klasyfikacja odbywa się jedynie w oparciu o punkty znajdujące się w ściśle określonym przedziale przestrzeni wejść zamkniętym w domenie mini-modelu. W zadaniu klasyfikacji jako punkt zapytania rozumiemy punkt danych w przestrzeni, którego wartość atrybutów wejściowych $x_1, x_2, \ldots, x_n$ jest nam znana, a któremu chcemy przypisać etykietę danej klasy. Punkt zapytania staje się centrum hipersferycznego układu współrzędnych. Jako algorytm klasyfikacji można przyjąć bardzo proste podejście, w którym mini-model został tak zaprogramowany aby starał się w swojej domenie uchwycić jedynie próbki należące do jednej klasy. Oczywiście nie zawsze taka sytuacja jest możliwa, wtedy należy zwrócić klasę, która jest najczęściej reprezentowana przez próbki uczące. Jako optymalną domenę mini-modelu rozumiemy taką, która jest wysoce jednorodna w rozumieniu próbek należących do tej samej klasy. Przyjmuje się, że klasyfikowana próbka powinna znajdować się w centrum domeny. Jednakże, nie zawsze jest to możliwe. Mini-model klasyfikujący powinien spełniać takie same warunki jak mini-model regresyjny.

11.7 Porównanie metod: *k*-najbliższych sąsiadów i mini-modeli

Metoda mini-modeli jest pod pewnymi względami podobna do metody k-najbliższych sąsiadów [26]. Oba algorytmy używają danych uczących jedynie z najbliższego otoczenia punktu zapytania, który jest interesujący dla naukowców. Obie metody należą do metod bazujących na próbkach i obliczają odpowiedź na bieżąco. Obie metody charakteryzuje także powierzchnia funkcyjna utworzona przez następujące po sobie modele w całej domenie modelowanego systemu, która nie jest gładka. Można zaobserwować wyraźne "skokowe" zachowanie pochodnych powierzchni funkcyjnej. Metoda MM nie jest w swej istocie przeznaczona do globalnego modelowania, ale może być użyta w tym celu. Rysunki 9a i 9b porównują powierzchnie funkcyjne wygenerowane przez obie metody w przestrzeni 2D. Rysunki 10a i 10b prezentują porównanie obu metod dla danych 3D.



(a) Metoda mini-modeli

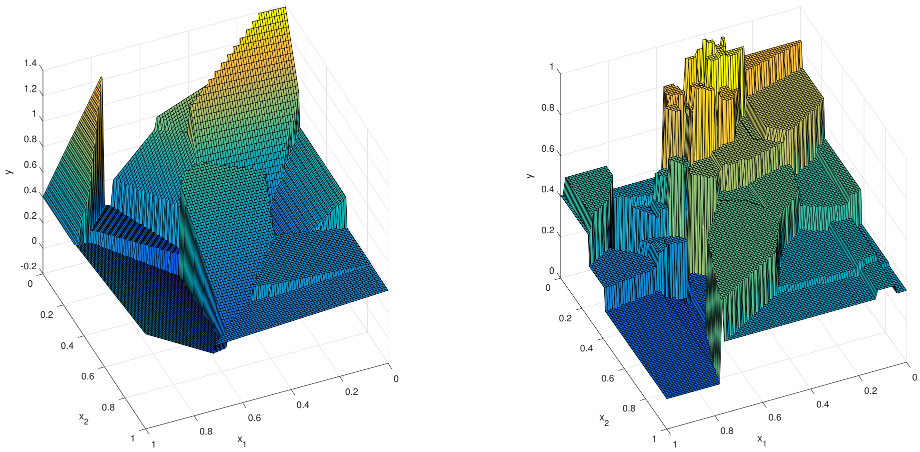


(b) Metoda k-NN (dla $k = 1$)

Rysunek 9: Porównanie nieciągłości powierzchni funkcyjnej metody mini-modeli oraz metody k-NN w przestrzeni 2D. *Źródło: opracowanie własne.*

Metoda k-najbliższych sąsiadów w czasie obliczania odpowiedzi na punkt zapytania bierze pod uwagę jedynie k najbliższych próbek. Podstawowa wersja metody używa metryki euklidesowej w celu obliczenia odległości pomiędzy próbkami, a punktem zapytania. W klasycznej wersji metoda oblicza odpowiedź modelu jako średnią wartość zmiennej zależnej z k najbliższych próbek, ewentualnie średnią ważoną przez odległość od punktu zapytania. W interpretacji geometrycznej, wszystkie k najbliższych punktów mieści się w kole dla problemu 3-wymiarowego, w sferze dla problemu 4-wymiarowego, itd. W ogólności będzie to $(n-1)$-wymiarowa hipersfera dla problemu n-wymiarowego. Dla porównania metoda mini-modeli bazująca na układzie hipersferycznym, zawiera punkty uczące w obszarze wypukłego wieloboku dla problemu 3-wymiarowego, wypukłego wielościanu dla problemu 4-wymiarowego. W ogólnym przypadku n-wymiarowym będzie to $(n-1)$-wymiarowa bryła wypukła. Liczba próbek wewnątrz mini-modelu nie jest stała, ale znajduje się w pewnym z góry ustalonym przedziale. Ma to pewne implikacje, ponieważ dla wybranego punktu zapytania istnieje wiele możliwych domen mini-modelu.

Metodę najbliższych sąsiadów można rozpatrywać jako pewien rodzaj metody mini-modeli. Biorąc pod uwagę takie podejście domeną mini-modelu w



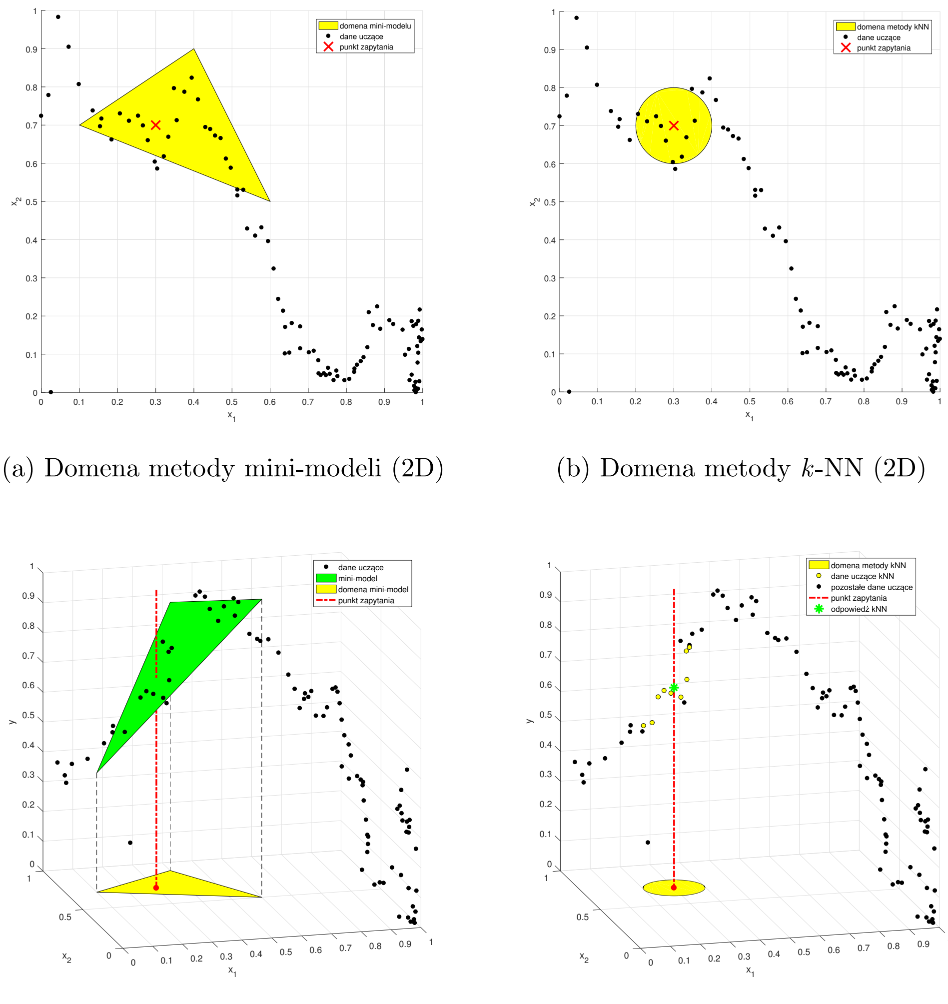
(a) Metoda mini-modeli (b) Metoda k-NN (dla $k = 1$)

Rysunek 10: Porównanie nieciągłości powierzchni funkcyjnej metody mini-modeli oraz metody k-NN w przestrzeni 3D. *Źródło: opracowanie własne.*

algorytmie najbliższych sąsiadów będzie $(n-1)$-wymiarowa hipersfera dla problemu n-wymiarowego. Algorytmem modelowania matematycznego będzie tutaj wartość średnia zmiennej zależnej punktów uczących albo średnia ważona przez odległość od punktu zapytania. W algorytmie najbliższych sąsiadów jest możliwa tylko jedna domena dla wybranego punktu zapytania i wartości k. Rysunki 11a i 11b przedstawiają różnicę pomiędzy metodami w przestrzeni wejść. Rysunki 11c i 11d porównują obie metody w pełnej przestrzeni problemu. Kolejną różnicą pomiędzy obiema metodami jest fakt, że algorytm najbliższych sąsiadów bierze pod uwagę jedynie średnią wartość próbek. Metoda mini-modeli w procesie obliczania odpowiedzi używa nie tylko wartości próbek uczących, ale bierze pod uwagę także tendencję występującą w okolicy punktu zapytania. Informacja na ten temat pozwala osiągać lepszą dokładność modelowania w obszarach, w których występują luki informacyjne.

11.8 Właściwości ekstrapolacyjne

Metoda mini-modeli posiada dobre właściwości ekstrapolacyjne. Mini-modele biorą pod uwagę tendencję występującą w otoczeniu punktu zapytania, w czym przejawia się ich przewaga nad metodą najbliższych sąsiadów. W skraj-



(a) Domena metody mini-modeli (2D)

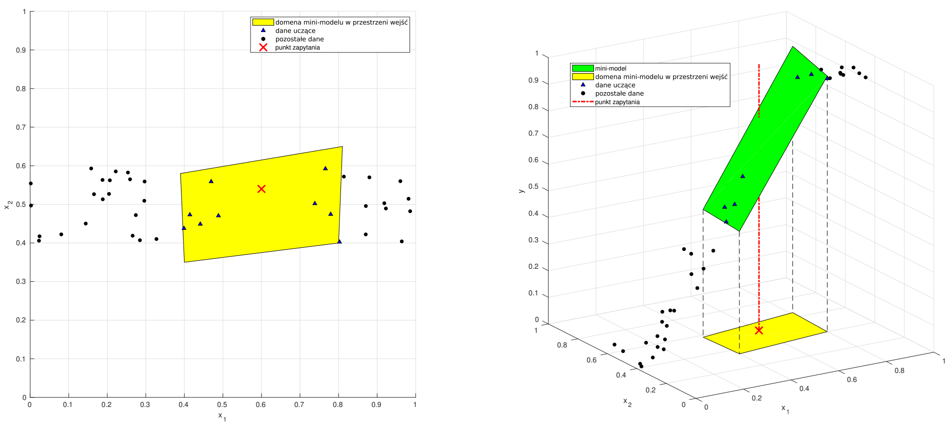
(b) Domena metody k-NN (2D)

(c) Metoda mini-modeli (3D)

(d) Metoda k-NN (3D)

Rysunek 11: Porównanie metody mini-modeli oraz metody k-NN w domenie problemu (2D) oraz w pełnej przestrzeni wejść (3D). *Źródło: opracowanie własne.*

nych częściach wykresu na Rysunkach 9 oraz 10 można porównać jak obie metody zachowują się w przypadku ekstrapolacji danych. W przypadku metody k-NN powierzchnia funkcyjna w obszarach, w których będzie wykonywana ekstrapolacja zawsze będzie płaska. W odmienny sposób zachowa się metoda mini-modeli, która w obszarach ekstrapolacyjnych będzie zachowywała linie trendu wygenerowaną na podstawie danych leżących najbliżej punktu zapytania. Dzięki wykorzystywaniu informacji na temat trendu danych metoda mini-modeli jest w stanie "przeskoczyć" lukę informacyjną i zawrzeć w swojej domenie dane znajdujące się po jej "drugiej stronie". Rysunek 12 przedstawia poglądową sytuację, w której metoda mini-modeli pracuje na danych znajdujących się po różnych stronach luki informacyjnej.



(a) Domena metody mini-modeli (2D)

(b) Metoda mini-modeli (3D)

Rysunek 12: Przykład metody mini-modeli, która używa danych znajdujących się po różnych stronach luki informacyjnej. *Źródło: opracowanie własne.*

11.9 Podsumowanie

Opisana w pracy metoda mini-modeli jest metodą modelowania ogólnego i może zostać zastosowana w procesie modelowania danych różnorakiego pochodzenia. Zbiory danych numerycznych mogą pochodzić z różnego rodzaju systemów, na przykład: fizycznych, ekonomicznych czy też biologicznych. Oczywiście metoda, nie we wszystkich przypadkach daje najlepsze wyniki i niekiedy jej dokładność nieznacznie ustępuje niektórym innym metodom modelowania. Jednak w przypadku MM typ pochodzenia danych nie ma wpływu na dokładności uzyskiwanych wyników.

Wszystkie wersje metody mini-modeli zostały przetestowane przy użyciu autorskiego oprogramowania. W celu przetestowania metod bazujących na próbkach, względem których porównywano skuteczność metody mini-modeli, użyto funkcji zaimplementowanych w środowisku MATLAB®.

Dokładność modelowania algorytmu przebadano w czterech głównych wariantach różniących się sposobem uczenia. Przeanalizowano: mini-modele 2D, mini-modele bazujące na bryłach wielowymiarowych, mini-modele bazujące na bryłach wielowymiarowych z redukcją wymiarowości, mini-modele bazujące na algorytmach klasteryzacji.

Podział ten może ulec rozszerzeniu jeżeli pod uwagę weźmiemy szereg parametrów definiujących sposób uczenia mini-modeli. Pod uwagę wzięto między innymi różne bryły definiujące domenę mini-modelu. W przypadku mini-modeli bazujących na metodach klasteryzacji użyto nie tylko różnych metod grupowania danych, ale także różne metryki oraz informacje czy zmienna y ma zostać użyta w procesie podziału przestrzeni na klastry. Rozdrabniając ten podział maksymalnie można przyjąć, że metodę mini-modeli łącznie przebadano w blisko czterdziestu wariantach.

Wyniki przeprowadzonych eksperymentów numerycznych potwierdzają skuteczność działania metody mini-modeli. Duża liczba prób testowych z mini-modelami 2D pozwoliła zobrazować jak metoda zachowuje się w przestrzeni, którą da się łatwo wizualizować. Jednakże szczególnie interesujące są wyniki dla pozostałych wersji metody mini-modeli operujących w przestrzeniach wielowymiarowych. Metoda i w tym przypadku udowodniła swoją skuteczność. Dla niektórych zbiorów danych metoda osiągała najlepsze rezultaty spośród porównywanych algorytmów. W przypadku gdy metoda mini-modeli nie była najdokładniejsza spośród porównywanych algorytmów, jej dokładność nie odbiegała od innych testowanych metod opartych na próbkach. Liczba zbiorów danych wielowymiarowych, na których przetestowano metodę może budzić zastrzeżenia należy jednak pamiętać, że metoda została przetestowana w dużej liczbie wariantów. Ponadto eksperymenty z mini-modelami 2D zostały wykonane na dużo większej liczbie danych, a poziom skuteczności pozostał porównywalny. Dodatkowo wyniki eksperymentów numerycznych oraz ich analiza są wystarczającą podstawą do udowodnienia hipotezy badawczej, że udoskonalona metoda uczenia mini-modeli umożliwia wykazanie co najmniej części przewidywanych zalet tej metody względem geometrycznych metod modelowania bazujących na próbkach.

11.9.1 Dalsze kierunki badań

Jako dalsze kierunki badań autor proponuje:

- Opracowanie algorytmów dzielących dane uczące na podgrupy na zasadzie działania algorytmu *kD*-drzew w celu zmniejszenia złożoności obliczeniowej mini-modeli bazującej na bryłach wielowymiarowych.

- Zastosowanie algorytmów optymalizacji w procesie obliczania optymalnej domeny mini-modelu. Zaproponowane rozwiązanie powinno być analogiczne do uczenia regresyjnej maszyny wektorów nośnych, w której parametry metody obliczane są przy pomocy metod optymalizacji.

- Zastosowanie innych algorytmów klasteryzacji danych. W pracy przetestowano jedynie metody oparte o centroidy. Metody oparte o gęstość próbek, takie jak np. algorytm DBSCAN, potrafią zidentyfikować punkty graniczne lub dane zaszumione, dla których obliczenie poprawnej odpowiedzi numerycznej może być niemożliwe.

- Opracowanie mini-modeli uwzględniających niepewność pomiarową próbek w oparciu o logikę rozmytą lub matematykę interwałową.

12 Lista publikacji

- Pietrzykowski, M. Comparison of effectiveness of linear mini-models with some methods of modelling. In *Młodzi Naukowcy dla Polskiej Nauki* (2011), Creativetime, pp. 113–123

- Pietrzykowski, M. The use of linear and nonlinear mini-models in process of data modeling in a 2d-space. In *Nowe trendy w Naukach Inżynieryjnych* (2011), Creativetime, pp. 100–108

- Pietrzykowski, M. Effectiveness of mini-models method when data modelling within a 2d-space in an information deficiency situation. *Journal of Theoretical and Applied Computer Science 6* (2012), 21–27

- Pietrzykowski, M. Use of the mini-model method in classification task on example of iris flower dataset. *Journal of Theoretical and Applied Computer Science 8* (2014), 27–36

- Pietrzykowski, M. Application of the mini-models based on n-dimensional simplex for modeling of buildings energy performance. *International Journal Of Computer Technology and Applications 6* (2015), 697–700

- Pietrzykowski, M. Comparison between mini-models based on multidimensional polytopes and k-nearest neighbor method: Case study of

4d and 5d problems. In *Soft Computing in Computer and Information Science* (2015), vol. 342 of *Advances in Intelligent Systems and Computing*, Springer Verlag, pp. 107–118

- Pietrzykowski, M., and Piegat, A. Geometric approach in local modeling: Learning of mini-models based on n-dimensional simplex. In *Artificial Intelligence and Soft Computing* (Cham, 2015), L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 460–470

- Pietrzykowski, M. Zastosowanie metody mini-modeli opartej na hiperszešcianie w procesie modelowania danych wielowymiarowych. *Zeszyty Naukowe. Studia Informatica / Uniwersytet Szczeciński Nr 38* (1 2015), 91–103

- Piegat, A., and Pietrzykowski, M. Local modeling with local dimensionality reduction: Learning method of mini-models. In *Artificial Intelligence and Soft Computing* (Cham, 2016), L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 375–383

- Pietrzykowski, M. Local regression algorithms based on centroid clustering methods. *Procedia Computer Science 112* (2017), 2363 – 2371. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France

- Pietrzykowski, M., and Pluciński, M. Mini-model method based on k-means clustering. *Przegląd Elektrotechniczny R. 93, nr 1* (1 2017), 73–76

Spis rysunków

Literatura

[1] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Broomhead, D. S., and Lowe, D. Multivariable functional interpolation and adaptive networks. *Complex Systems 2* (1988).

[3] Cichosz, P. *Systemy uczące się.* Wydawnictwa Naukowo-Techniczne, 2000.

[4] Cleveland, W. S., and Devlin, S. J. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association 83*, 403 (1988), 596–610.

[5] Donoho, D. L., and Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences 100*, 10 (2003), 5591–5596.

[6] Dowdy, S., and Wearden, S. *Statistics for research.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1991.

[7] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V. Support vector regression machines. In *NIPS* (1996).

[8] Fix, E., and Hodges, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique 57*, 3 (1989), 238–247.

[9] Greene, W. *Econometric Analysis.* Pearson Education, 2012.

[10] Hassani, S. *Mathematical Physics: A Modern Introduction to Its Foundations.* Springer International Publishing, 2013.

[11] Hellwig, Z. *Elementy rachunku prawdopodobieństwa i statystyki matematycznej.* Państwowe Wydaw. Naukowe, 1998.

[12] Hong, Y. P., and Pan, C.-T. Rank-Revealing QR Factorizations and the Singular Value Decomposition. *Mathematics of Computation 58*, 197 (1992), 213–232.

[13] Larose, D. T. *Metody i modele eksploracji danych.* Wydawnictwo Naukowe PWN, 2008.

[14] Law, M. H. C., and Jain, A. K. Incremental nonlinear dimensionality reduction by manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 3 (March 2006), 377–391.

[15] Martinez, A. M., and Kak, A. C. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 2 (Feb 2001), 228–233.

[16] Murphy, K. P. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[17] Paeth, A. *Graphics Gems V (Macintosh Version).* Graphics Gems - Macintosh. Elsevier Science, 2014.

[18] Piegat, A., and Pietrzykowski, M. Local modeling with local dimensionality reduction: Learning method of mini-models. In *Artificial Intelligence and Soft Computing* (Cham, 2016), L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 375–383.

[19] Piegat, A., Wąsikowska, B., and Korzeń, M. Differences between the method of minimodels and the k-nearest neighbors on example of modeling of unemployment rate in poland. In *Proceedings of 9th Conference on Information Systems in Management* (01 2011), pp. 34–43.

[20] Piegat, A., Wąsikowska, B., and Korzeń, M. Zastosowanie samouczącego się, trzypunktowego minimodelu do modelowania stopy bezrobocia w polsce. *Studia Informatica 27* (01 2011), 45–58.

[21] Pietrzykowski, M. Comparison of effectiveness of linear mini-models with some methods of modelling. In *Młodzi Naukowcy dla Polskiej Nauki* (2011), Creativetime, pp. 113–123.

[22] Pietrzykowski, M. The use of linear and nonlinear mini-models in process of data modeling in a 2d-space. In *Nowe trendy w Naukach Inżynieryjnych* (2011), Creativetime, pp. 100–108.

[23] Pietrzykowski, M. Effectiveness of mini-models method when data modelling within a 2d-space in an information deficiency situation. *Journal of Theoretical and Applied Computer Science 6* (2012), 21–27.

[24] Pietrzykowski, M. Use of the mini-model method in classification task on example of iris flower dataset. *Journal of Theoretical and Applied Computer Science 8* (2014), 27–36.

[25] Pietrzykowski, M. Application of the mini-models based on n-dimensional simplex for modeling of buildings energy performance. *International Journal Of Computer Technology and Applications 6* (2015), 697–700.

[26] Pietrzykowski, M. Comparison between mini-models based on multidimensional polytopes and k-nearest neighbor method: Case study of 4d and 5d problems. In *Soft Computing in Computer and Information Science* (2015), vol. 342 of *Advances in Intelligent Systems and Computing*, Springer Verlag, pp. 107–118.

[27] Pietrzykowski, M. Zastosowanie metody mini-modeli opartej na hipersześcianie w procesie modelowania danych wielowymiarowych. *Zeszyty Naukowe. Studia Informatica / Uniwersytet Szczeciński Nr 38* (1 2015), 91–103.

[28] Pietrzykowski, M. Local regression algorithms based on centroid clustering methods. *Procedia Computer Science 112* (2017), 2363 – 2371. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

[29] Pietrzykowski, M., and Piegat, A. Geometric approach in local modeling: Learning of mini-models based on n-dimensional simplex. In *Artificial Intelligence and Soft Computing* (Cham, 2015), L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 460–470.

[30] Pietrzykowski, M., and Pluciński, M. Mini-model method based on k-means clustering. *Przegląd Elektrotechniczny R. 93, nr 1* (1 2017), 73–76.

[31] Pluciński, M. Mini-models – local regression models for the function approximation learning. In *Artificial Intelligence and Soft Computing* (Berlin, Heidelberg, 2012), L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., Springer Berlin Heidelberg, pp. 160–167.

[32] Pluciński, M. Nonlinear ellipsoidal mini-models – application for the function approximation task. *Przegląd Elektrotechniczny R. 88, nr 10b* (10 2012), 247–251.

[33] Rawlings, J., Pantula, S., and Dickey, D. *Applied Regression Analysis: A Research Tool.* Springer Texts in Statistics. Springer New York, 2001.

[34] Rejer, I. *Metoda modelowania wielkowymiarowego system z użyciem metod sztucznej inteligencji na przykładzie bezrobocia w Polsce.* PhD thesis, Wydział Nauk Ekonomicznych i Zarządzania, Uniwersytet Szczeciński, Szczecin, 2003.

[35] Rodriguez, M. A., and Winternitz, P. Quantum superintegrability and exact solvability in n dimensions. *Journal of Mathematical Physics 43*, 3 (2002), 1309–1322.

[36] Specht, D. F. A general regression neural network. *IEEE transactions on neural networks 2 6* (1991), 568–76.

[37] Vapnik, V., and Chervonenkis, A. A note on one class of perceptrons. *Automation and Remote Control 25* (1964).