Autoreferat
Redukcja wymiarowości, klasteryzacja i wykrywanie danych odstających w uczeniu maszynowym

Dominik Olszewski

22 czerwca 2023

Część I

1 Posiadane dyplomy, stopnie naukowe lub artystyczne – z podaniem podmiotu nadającego stopień, roku ich uzyskania oraz tytułu rozprawy doktorskiej

- Stopień doktora nauk technicznych (z wyróżnieniem) w dyscyplinie naukowej Automatyka i Robotyka nadany przez Politechnikę Warszawską w 2012 roku. Tytuł rozprawy doktorskiej: „Modelowanie statystyczne i pseudoodległości w wybranych zagadnieniach analizy danych". Promotor: dr hab. inż. Bartłomiej Beliczyński, prof. uczelni. Recenzenci: prof. dr hab. Jacek Koronacki oraz dr hab. inż. Marcin Iwanowski, prof. uczelni.

- Tytuł zawodowy magistra inżyniera po ukończeniu jednolitych studiów magisterskich na kierunku Elektrotechnika w specjalności Automatyka i Inżynieria Komputerowa nadany przez Politechnikę Warszawską w 2007 roku. Tytuł pracy magisterskiej: „Wykorzystanie analizy falek w przetwarzaniu obrazów". Promotor: dr inż. Sławomir Skoneczny. Recenzent: dr hab. inż. Bartłomiej Beliczyński, prof. uczelni.

2 Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych lub artystycznych

- Adiunkt badawczo-dydaktyczny w Zakładzie Sterowania w Instytucie Sterowania i Elektroniki Przemysłowej na Wydziale Elektrycznym Politechniki Warszawskiej w okresie od października 2012 do chwili obecnej.
- Adiunkt badawczo-dydaktyczny w Pracowni Systemów Inteligentnych w Instytucie Badań Systemowych Polskiej Akademii Nauk w okresie od października 2012 do czerwca 2013.
- Asystent badawczo-dydaktyczny w Zakładzie Sterowania w Instytucie Sterowania i Elektroniki Przemysłowej na Wydziale Elektrycznym Politechniki Warszawskiej w okresie od stycznia 2012 do października 2012.

Część II

1 Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 z późn. zm.)

W niniejszych rozdziałach autoreferatu zaprezentowane zostaną główne osiągnięcia naukowe autora, które wybrane zostały jako jego istotny wkład w rozwój dziedziny badawczej sztucznej inteligencji i uczenia maszynowego.

Ta część autoreferatu podzielona została na trzy rozdziały odpowiadające trzem obszarom uczenia maszynowego, w których to autor zaproponował własne rozwiązania, będące przedmiotem jego najważniejszych publikacji naukowych.

W odpowiednich podrozdziałach przybliżone są innowacyjne elementy, które autor wprowadza w kolejno rozważanych metodach i algorytmach analizy danych. Całość dorobku oryginalnego może być natomiast ujęta określeniem – „wkład w rozwój obszaru badawczego redukcji wymiarowości, klasteryzacji i wykrywania danych odstających w uczeniu maszynowym".

Poszczególne osiągnięcia autora, prezentowane w niniejszym autoreferacie mają charakter udoskonaleń i rozszerzeń istniejących metod i algorytmów

Tabela 1: Zestawienie danych dotyczących publikacji wnioskodawcy prezentowanych w ramach cyklu powiązanych tematycznie artykułów naukowych (ostatnia aktualizacja cytowań: 22 czerwca 2023)

Artykuł	Wkład wnioskodawcy	Impact Factor	CiteScore	Punkty MEiN	Cytowania (WoS)	Cytowania (Scopus)	Rok publikacji
[1]	100%	9.657	13.1	200	3	4	2021
[2]	100%	7.802	11	140	5	9	2021
[3]	70%	Nie dotyczy	Nie dotyczy	20	3	2	2016
[4]	100%	8.665	12.2	140	0	0	2023
[5]	50%	8.518	15.5	140	16	17	2014
[6]	100%	2.565	4.2	70	3	5	2016
[7]	70%	0.402	1.8	20	5	6	2013
[8]	100%	8.139	12	200	70	94	2014
[9]	100%	8.139	12	200	45	57	2012

uczenia maszynowego. Każdej takiej modyfikacji z kolei towarzyszy teoretyczne uzasadnienie, którego celem jest przekonujące wytłumaczenie podejmowanych kroków i decyzji, zmierzających ku uzyskaniu nowych, poprawionych wersji rozważanych technik analizy danych, tak, aby finalnie otrzymać wyższą skuteczność ich działania, zweryfikowaną i potwierdzoną stosownie przeprowadzonym studium eksperymentalnym.

W ten sposób, autor ma zamiar wyjaśnić przyczyny i cel zmian wprowadzanych w analizowanych podejściach z dziedziny uczenia maszynowego od strony czysto teoretycznej. Ocena empiryczna skuteczności działania nowych, proponowanych wersji rozważanych metod będzie zatem jedynie sprawdzeniem i weryfikacją słuszności przyjętych i przedstawionych wcześniej postulatów w sferze teoretycznej.

Tabela 1 przedstawia dane bibliometryczne dotyczące publikacji wnioskodawcy prezentowanych w ramach cyklu powiązanych tematycznie artykułów naukowych.

Zagadnienia uczenia maszynowego będące treścią prac [1, 2, 3, 4, 5, 6, 8, 9] podejmowane są w sposób gruntowny, wnikliwy i wyczerpujący, zaś sam wnioskodawca starał się dołożyć wszelkich starań, aby prace te miały charakter

kompletnego i analitycznego studium naukowego o istotnym i zauważalnym wkładzie oryginalnym, zarówno w sferze rozważań teoretycznych, jak i praktycznych badań i eksperymentów.

Główne osiągnięcia wnioskodawcy można ująć w skróconej formie następująco:

- sformułowanie adaptacyjnej postaci metody NeRV opartej o wstępną klasteryzację wejściowych danych (artykuł [1]),
- zaproponowanie udoskonalenia metody SOM, w którym zachowywane jest rozproszenie wejściowych danych (artykuł [2]),
- zaproponowanie udoskonalenia metody SOM bazującego na charakterystyce częstościowej wejściowych danych (artykuł [3]),
- zaprojektowanie asymetrycznej wersji metody NeRV, która odzwierciedla asymetryczne powiązania pomiędzy wejściowymi danymi oraz dąży do zachowania ich topologii poprzez wykorzystanie geometrycznej reprezentacji wejściowych danych (artykuł [4]),
- opracowanie asymetrycznej odmiany metody klasteryzacji danych k-centroidów (artykuł [5]),
- wprowadzenie połączenia asymetrycznej wersji algorytmu SOM oraz asymetrycznej postaci techniki klasteryzacji danych k-średnich (artykuł [6]),
- przystosowanie asymetrycznego podejścia SOM do wizualizacji szeregów czasowych (artykuł [7]),
- zaproponowanie metody wykrywania danych odstających za pomocą mapy SOM oraz algorytmu klasyfikacji progowej (artykuł [8]),
- sformułowanie techniki wykrywania danych odstających z wykorzystaniem rozkładu probabilistycznego LDA (artykuł [9]).

Wszystkie te proponowane rozwiązania i udoskonalenia istniejących metod uczenia maszynowego omówione są w kolejnych rozdziałach niniejszego autoreferatu.

2 Redukcja wymiarowości danych

Pierwszym omawianym w niniejszym autoreferacie obszarem badawczym jest redukcja wymiarowości danych lub też inaczej mówiąc redukcja liczby wymiarów danych. Jest to poddziedzina uczenia maszynowego.

Celem redukcji wymiarowości danych jest sformułowanie przekształcenia o charakterze rzutowania prowadzącego z wejściowego, oryginalnego, wielowymiarowego zbioru danych do wejściowego, wynikowego, niskowymiarowego zbioru danych, w taki sposób, aby w wyjściowej przestrzeni danych zachowane zostały podobieństwa pomiędzy punktami danych, które istnieją w wejściowej przestrzeni danych. Do dyspozycji są zarówno techniki liniowe, jak i nieliniowe oraz podejścia nadzorowane, jak i nienadzorowane.

O redukcji wymiarowości danych można mówić zawsze, gdy liczba wymiarów w przestrzeni wyjściowej jest mniejsza od liczby wymiarów w przestrzeni wejściowej. Jednakże, w szczególnym przypadku, gdy liczba wymiarów w przestrzeni wyjściowej jest równa jeden, dwa lub trzy, wówczas mówić można o wizualizacji danych, ponieważ zbiór danych posiadających nie więcej niż trzy wymiary może zostać przedstawiony w formie graficznej. W przypadku tak rozumianej i tak zdefiniowanej wizualizacji danych, najczęściej wykorzystywana liczba wymiarów w przestrzeni wyjściowej to dwa, ze względu na łatwość zilustrowania graficznego oraz wystarczającą wygodę późniejszej analizy i interpretacji.

Wizualizacja danych wielowymiarowych poprzez rzutowanie do nowoskonstruowanej przestrzeni danych o mniejszej liczbie wymiarów niż oryginalna przestrzeń wejściowa zyskała już dużą popularność ze względu na szeroką feerię potencjalnych możliwości praktycznego wykorzystania w rozmaitych obszarach i dziedzinach analizy danych. Jedną z podstawowych zalet tego rozwiązania w sztucznej inteligencji i uczeniu maszynowym jest fakt, iż transformacja danych w formie redukcji ich wymiarowości pozwala na uzyskanie ilustracji danych w postaci obrazu, a taka postać danych staje się przedmiotem możliwej analizy, oceny i interpretacji przez osoby niebędące ekspertami, ani w obszarze sztucznej inteligencji i uczenia maszynowego, ani w specyficznej dziedzinie, z której pochodzą same dane. Ten efekt możliwy jest dzięki wizualnej analizie danych.

Obszar redukcji wymiarowości danych oraz ścisłe z nim związanej wizualizacji danych został wybrany przez autora niniejszego autoreferatu jako dyscyplina naukowa stwarzająca możliwości wprowadzenia nowych rozwiązań oraz udoskonalania tych już istniejących, w której wykorzystywany aparat matematyczny w znacznym stopniu opiera się na statystyce matematycznej i rachunku prawdopodobieństwa oraz na wybranych narzędziach geometrii różniczkowej.

W zakresie redukcji wymiarowości i wizualizacji danych, omawianych w tym rozdziale autoreferatu, autor chciałby przedstawić swoje trzy osiągnięcia stanowiące, jego zdaniem, istotny wkład w rozwój ogólnej dziedziny uczenia maszynowego.

2.1 Adaptacyjna metoda NeRV

Metoda określana anglojęzycznym terminem Neighborhood Retrieval Visualizer (NeRV) należy do grupy probabilistycznych technik redukcji wymiarowości i wizualizacji danych, którym początek dała metoda Stochastic Neighbor Embedding (SNE) wprowadzona w pracy [10]. W podejściu SNE dąży się do zachowania lokalnego otoczenia każdego punktu z oryginalnej, wejściowej, wielowymiarowej przestrzeni danych w wynikowej, wyjściowej, niskowymiarowej przestrzeni danych. Cel ten jest osiągany poprzez minimalizowanie niepodobieństwa pomiędzy rozkładami prawdopodobieństwa opisującymi lokalne otoczenia punktów w przestrzeni wejściowej oraz wyjściowej. Wspomniane rozkładu dotyczą prawdopodobieństwa, że dany punkt będzie znajdował się w kolejnym otoczeniu aktualnie rozpatrywanego punktu – inaczej mówiąc, że będzie jego sąsiadem. Tak sformułowana procedura minimalizacji pozwoli na wyznaczenie lokalizacji odpowiednich (poszukiwanych) punktów w wyjściowej przestrzeni danych. Algorytm SNE zdobył dużą popularność i pojawił się szereg jego rozszerzeń i udoskonaleń, wśród których, jako jedno ze szczególnie istotnych, należy wskazać podejście NeRV. To rozwiązanie z kolei zostało wybrane jako przedmiot rozważań mających na celu kontynuację procesu udoskonalania przez autora niniejszego autoreferatu.

Metoda NeRV w standardowej postaci została sformułowana w pracy [11] w oparciu o definicję funkcji kosztu (1). Zatem cała procedura postępowania jest odmienna niż w przypadku metody SNE i zyskuje charakter sformalizowanego, rygorystycznego i profesjonalnego wywodu matematycznego, którego finalnym wynikiem jest probabilistyczna forma funkcji kosztu (2), będącą przedmiotem minimalizacji przeprowadzanej przez technikę NeRV. Choć więc ostatecznie techniczne czynności wykonywane w ramach działania metod SNE i NeRV mogą być uznane za podobne (gdy mowa już o samej końcowej optymalizacji), to teoretyczne uzasadnienie obu tych metod leżące u ich podstaw oraz ogólny sposób rozumowania tłumaczący ich działanie są już w istocie zgoła odmienne.

$$E_i \;=\; N_{\mathrm{FP},i}\, C_{\mathrm{FP}} \;+\; N_{\mathrm{FN},i}\, C_{\mathrm{FN}}\,, \tag{1}$$

gdzie i jest indeksem danego punktu, $N_{\mathrm{FP},i}$ jest liczbą wszystkich punktów uznanych za fałszywie pozytywne, C_{FP} jest przyjętym kosztem błędu typu

fałszywie pozytywnego, $N_{\mathrm{FN},i}$ jest liczbą wszystkich punktów uznanych za fałszywie negatywne, natomiast C_{FN} jest przyjętym kosztem błędu typu fałszywie negatywnego.

$$E_{\mathrm{NeRV}} \;\approx\; \lambda \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \;+\; (1 \;-\; \lambda) \sum_i \sum_{j \neq i} q_{j|i} \log \frac{q_{j|i}}{p_{j|i}}\,, \qquad (2)$$

gdzie $D\,(\cdot,\,\cdot)$ jest dywergencją Kullbacka—Leiblera, $q_{j|i}$ jest prawdopodobieństwem wyboru $j$tego punktu w sąsiedztwie $i$tego punktu w przestrzeni wyjściowej, $p_{j|i}$ jest prawdopodobieństwem wyboru $j$tego punktu w sąsiedztwie $i$tego punktu w przestrzeni wejściowej, natomiast λ jest pomocniczym parametrem pozwalającym na zróżnicowanie, który z błędów jest dla na bardziej istotny i szkodliwy – błąd precyzji, czy błąd czułości (innymi słowy, który z tych błędów jest dla nas ważniejszy).

Adaptacyjna wersja algorytmu NeRV została zaproponowana przez autora niniejszego autoreferatu w artykule [1]. Zgodnie z koncepcją wprowadzoną w tej pracy, szerokości sąsiedztwa punktów, zarówno w przestrzeni wejściowej, jak i wyjściowej określane sposób w sposób adaptacyjny – na podstawie odpowiednich własności danych wejściowych. Sam etap właściwej redukcji wymiarowości za pomocą tradycyjnej metody NeRV, poprzedzony jest wstępną klasteryzacją danych wejściowych, dzięki której określane zostają własności rozproszenia danych wejściowych. Dla każdego klastra danych w przestrzeni wejściowej obliczona zostaje wariancja wewnątrzklastrowa i przypisywana jest ona każdemu punktowi należącemu do tego klastra. Owa wartość wariancji wewnątrzklastrowej jest następnie wykorzystana do określenia szerokości sąsiedztwa punktów w przestrzeniach – wejściowych i wyjściowej podczas klasycznego działania metody NeRV już w kolejnym etapie analizy danych.

Teoretyczne uzasadnienie takiego postępowania w celu udoskonalenia techniki NeRV jest następujące. Początkowa klasteryzacja danych pozwala określić i wyodrębnić skupienia w obrębie wejściowego zbioru danych, a zatem jednocześnie pozwala zaprezentować informację na temat rozproszenia danych wejściowych. Precyzyjny, liczbowy sposób wyrażenia tego rozproszenia uzyskać można dzięki obliczeniu wartości wariancji wewnątrzklastrowych dla każdego klastra. Te wartości z kolei będą pomocne w określeniu szerokości sąsiedztwa punktów w wejściowej i wyjściowej przestrzeni danych podczas standardowego działania algorytmu NeRV, ponieważ zapewnią one odpowiednią korespondencję pomiędzy rozmiarem lokalnego otoczenia w metodzie NeRV oraz charakterystyką rozproszenia danych w przestrzeni wejściowej. Tak zmodyfikowana technika NeRV zyskuje cechy adaptacyjności, ponieważ ma zdolność adaptacyjnego dopasowania sposobu swojego działania do własności danych wejściowych. Konkretnym parametrem dobieranym w sposób

Tabela 2: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji słów

	q	U_d	s	p-value
SOM & k-means	8,174,844/10,868,000 = 0.7522	2,459,849/10,868,000 = 0.2263	2.2406	$< 10^{-4}$
t-SNE & k-means	8,005,903/10,868,000 = 0.7366	2,513,558/10,868,000 = 0.2313	2.4352	$< 10^{-4}$
Traditional NeRV & k-means	8,910,547/10,868,000 = 0.8199	1,979,504/10,868,000 = 0.1821	2.1213	$< 10^{-4}$
Proposed NeRV & k-means	9,230,362/10,868,000 = 0.8493	1,497,091/10,868,000 = 0.1378	2.2812	——-

Tabela 3: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji sygnałów EKG

	q	U_d	s	p-value
SOM & k-means	45/63 = 0.7143	18/63 = 0.2857	0.6824	$< 10^{-4}$
t-SNE & k-means	53/63 = 0.8413	10/63 = 0.1587	0.7420	$< 10^{-4}$
Traditional NeRV & k-means	57/63 = 0.9048	7/63 = 0.1111	0.6701	$< 10^{-4}$
Proposed NeRV & k-means	60/63 = 0.9365	5/63 = 0.0794	0.6688	——-

adaptacyjny staje się już omawiana szerokość sąsiedztwa w metodzie NeRV, a uzyskanym korzystnym i pożądanym efektem jest stosowanie małych szerokości sąsiedztwa dla danych o dużym skupieniu w przestrzeni wejściowej, zaś dużych szerokości sąsiedztwa w przypadku danych o znacznym rozproszeniu w przestrzeni wejściowej.

Badania eksperymentalne zostały przeprowadzone na trzech zbiorach danych. Zbiory te reprezentowały dane tekstowe, medyczne (sygnały rytmu ludzkiego serca EKG) oraz dane dźwiękowe (muzyka poważna).

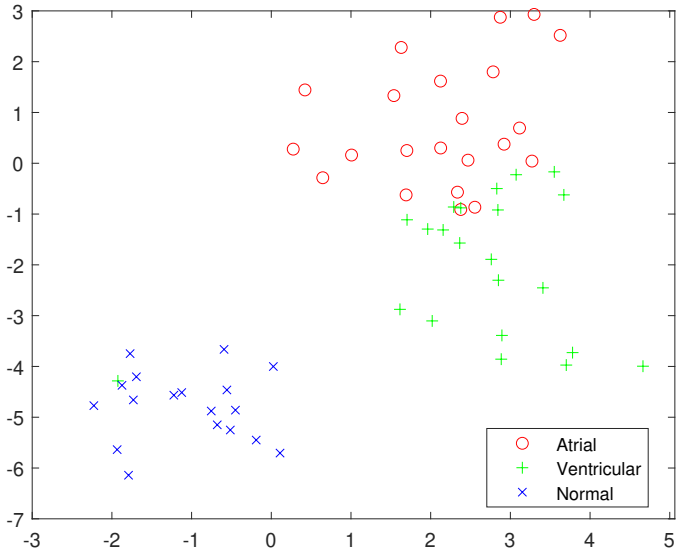
Proponowana metoda oceniona była na podstawie porównania z innymi wybranymi metodami odniesienia.

Sam proces redukcji wymiarowości danych i wynikającej z niej wizualizacji danych, kontynuowany był w postaci klasteryzacji *a posteriori* w celu umożliwienia zmierzenia jakości redukcji wymiarowości poprzez pomiar dokładności klasteryzacji danych w przestrzeni dwuwymiarowej, ponieważ trafność klasteryzacji można ocenić w nietrudny sposób za pomocą konkretnych kryteriów liczbowych, mianowicie, za pomocą miary trafności (ang. accuracy rate) oraz stopnia niepewności (ang. uncertainty degree).

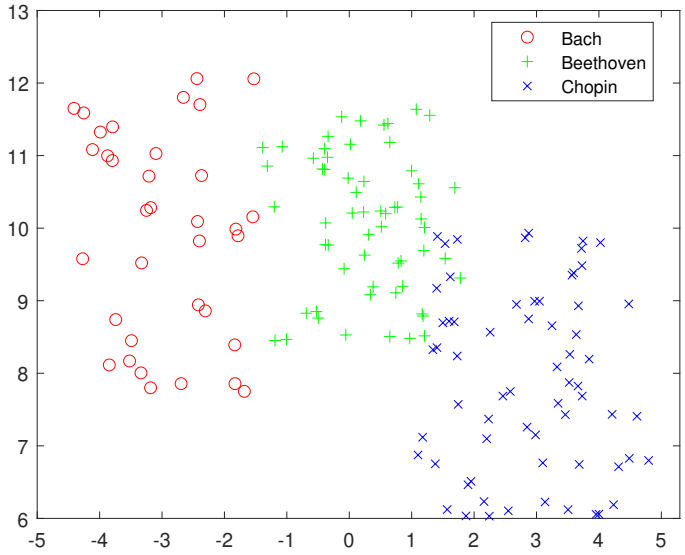
Część wyników badań eksperymentalnych przedstawiona jest w Tabelach 2, 3 i 4 oraz na Rysunkach 1 i 2, które prezentują jedynie fragment przeprowadzonych prac badawczych, zaś całość dostępna jest w pracy [1].

Tabela 4: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji utworów muzyki poważnej

	q	U_d	s	p-value
SOM & k-means	$135/160 = 0.8438$	$13/160 = 0.0813$	1.3056	$< 10^{-4}$
t-SNE & k-means	$128/160 = 0.8000$	$16/160 = 0.1000$	1.1678	$< 10^{-4}$
Traditional NeRV & k-means	$144/160 = 0.9000$	$10/160 = 0.0625$	1.1830	$< 10^{-4}$
Proposed NeRV & k-means	$155/160 = 0.9688$	$9/160 = 0.0563$	1.2447	——-



Rysunek 1: Wyniki wizualizacji sygnałów EKG za pomocą proponowanej adaptacyjnej metody NeRV



Rysunek 2: Wyniki wizualizacji utworów muzyki poważnej za pomocą proponowanej adaptacyjnej metody NeRV

2.2 Adaptacyjna metoda SOM

Mapa samoorganizująca się (ang. Self-Organizing Map (SOM)) to architektura sztucznej sieci neuronowej wprowadzona przez fińskiego uczonego Teuvo Kohonena w jego pracy [12], gdzie rozważane są zagadnienia samoorganizacji i kwantyzacji wektorów. Sama nazwa tej metody (ang. Self-Organizing Map) oraz odpowiadający jej skrót SOM pojawiają się natomiast później, w pracy [13] również autorstwa Kohonena. Kohonen podczas formułowania projektu mapy SOM opierał się w pewnym stopniu na badaniach Christopha von der Malsburga [14], które co prawda dotyczyły dziedziny neurologii, lecz jak się okazało mogły również stanowić ważną i skuteczną inspirację dla studiów w obszarze sztucznej inteligencji, uczenia maszynowego i inteligentnej analizy danych.

Metoda SOM stanowiącą swoistą topologię sztucznej sieci neuronowej, może być również interpretowana jako nielinowe przekształcenie prowadzące do redukcji liczby wymiarów oryginalnego zbioru danych. A zatem taka perspektywa analizy i oceny techniki SOM pozwala niewątpliwie zakwalifikować ją jako podejście z dziedziny redukcji wymiarowości danych.

2.2.1 Adaptacja metody SOM w oparciu o wstępną klasteryzację danych wejściowych

Autor niniejszego autoreferatu zaproponował w pracy [2] udoskonalenie sieci SOM, w którym parametr szerokości sąsiedztwa neuronów w regularnej, dwuwymiarowej strukturze mapy wizualizacji SOM dobierana jest w sposób adaptacyjny. Sposób ten bazuje na informacji na temat stopnia rozproszenia danych w przestrzeni wejściowej. Stopień ten określany jest w efekcie przeprowadzenia nienadzorowanego procesu klasteryzacji danych wejściowych. Dla każdego klastra obliczany jest parametr wariancji wewnątrzklastrowej i ta właśnie wartość przypisana jest każdemu punktowi w danym klastrze, a następnie wykorzystana podczas określania rozmiaru otoczenia neuronów w trakcie uczenia sztucznej sieci neuronowej SOM.

Teoretyczne uzasadnienie wprowadzonej modyfikacji w metodzie SOM odnosi się do postulatu zachowania ścisłej relacji pomiędzy rozproszeniem danych wejściowych – wyrażonym jako wariancja wewnątrzklastrowa uzyskana po wstępnej klasteryzacji w przestrzeni wejściowej – a parametrem szerokości sąsiedztwa neuronów w procesie uczenia struktury mapy SOM. Obszary w wejściowej przestrzeni danych o niewielkim skupieniu będą poprawnie i skutecznie wizualizowane na mapie SOM, gdy także i neurony będące wynikiem ich rzutowania będą rozproszone. Taki efekt będzie miał miejsce, gdy dla tych neuronów będą przeznaczone odpowiednio duże obszary na ekranie SOM. A to z kolei jest możliwe dzięki zastosowaniu stosownie dużych szerokości sąsiedztwa neuronów podczas procesu uczenia SOM. Analogicznie w przypadku obszarów o wysokim stopniu skupienia powinny korespondować z dużymi obszarami na mapie SOM, czyli z wysokimi wartościami parametru szerokości sąsiedztwa neuronów w metodzie SOM.

Eksperymenty mające na celu weryfikację i potwierdzenie proponowanego rozwiązania przeprowadzone zostały na trzech skrajnie różniących się zbiorach danych: zbiór danych tekstowych, zbiór danych dźwiękowych (zadanie rozpoznawania mówcy) oraz zbiór danych reprezentujących aktywności robota mobilnego, wykonującego zadane czynności.

Proponowana metoda oceniona była na podstawie porównania z innymi wybranymi metodami odniesienia.

Sam proces redukcji wymiarowości danych i wynikającej z niej wizualizacji danych, kontynuowany był w postaci klasteryzacji *a posteriori* w celu umożliwienia zmierzenia jakości redukcji wymiarowości poprzez pomiar dokładności klasteryzacji danych w przestrzeni dwuwymiarowej, ponieważ trafność klasteryzacji można ocenić w nietrudny sposób za pomocą konkretnych kryteriów liczbowych, mianowicie, za pomocą miary trafności (ang. accuracy rate) oraz stopnia niepewności (ang. uncertainty degree).

Tabela 5: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji słów

	q	U_d	time[sec]	s	Bonferroni p-value
SOM & k-means	8,175,273/10,868,000 = 0.7522	2,459,195/10,868,000 = 0.2263	2985	2.6496	3.3726e-38
TASOM & k-means	8,389,009/10,868,000 = 0.7719	2,304,016/10,868,000 = 0.2120	2921	3.0194	6.9446e-39
ASOM & k-means	8,948,773/10,868,000 = 0.8234	1,957,675/10,868,000 = 0.1801	2763	2.1154	4.5263e-41
FBSOM & k-means	9,183,460/10,868,000 = 0.8450	1,523,471/10,868,000 = 0.1402	2940	2.8730	2.2312e-40
LAMA & k-means	9,195,801/10,868,000 = 0.8461	1,454,947/10,868,000 = 0.1339	2320	2.3501	3.5082e-38
LARFSOM & k-means	8,992,470/10,868,000 = 0.8274	1,649,030/10,868,000 = 0.1517	2865	3.0041	7.6093e-38
RBSOM & k-means	8,972,611/10,868,000 = 0.8256	1,894,781/10,868,000 = 0.1743	2260	2.0802	3.7914e-41
Proposed SOM & k-means	**9,262,579/10,868,000 = 0.8523**	**1,248,742/10,868,000 = 0.1149**	**2889**	**2.3841**	——

Tabela 6: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji mówców

	q	U_d	time[sec]	s	Bonferroni p-value
SOM & k-means	134/160 = 0.8375	15/160 = 0.0938	0.0370	1.3518	1.8918e-39
TASOM & k-means	137/160 = 0.8563	16/160 = 0.1000	0.0400	1.5904	3.1430e-42
ASOM & k-means	139/160 = 0.8688	18/160 = 0.1125	0.0390	1.3482	3.9373e-37
FBSOM & k-means	145/160 = 0.9063	10/160 = 0.0625	0.0380	1.2838	6.8351e-38
LAMA & k-means	144/160 = 0.9000	15/160 = 0.0938	0.0390	1.5703	1.1820e-40
LARFSOM & k-means	138/160 = 0.8625	12/160 = 0.0750	0.0380	1.7162	1.1080e-36
RBSOM & k-means	137/160 = 0.8563	13/160 = 0.0813	0.0360	1.3720	1.4935e-38
Proposed SOM & k-means	**154/160 = 0.9625**	**5/160 = 0.0313**	**0.0480**	**1.4011**	——

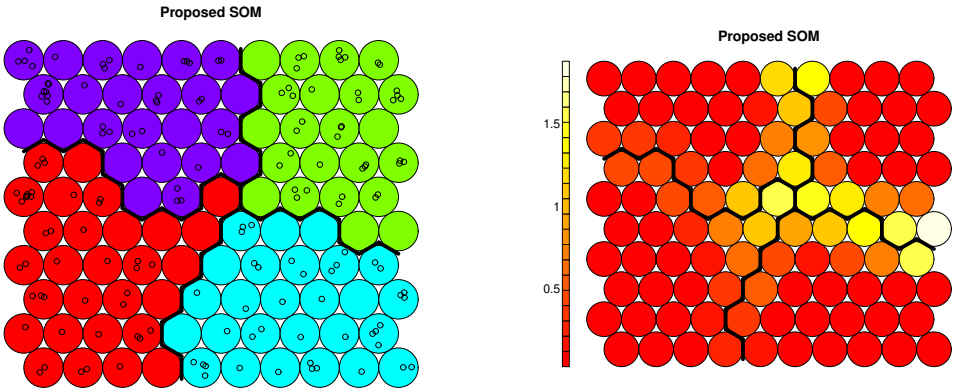
Część wyników badań eksperymentalnych przedstawiona jest w Tabelach 5, 6 i 7 oraz na Rysunkach 3 i 4, które prezentują jedynie fragment przeprowadzonych prac badawczych, zaś całość dostępna jest w pracy [2].

2.2.2 Adaptacja metody SOM w oparciu o charakterystykę częstościową danych wejściowych

Adaptacyjna wersja algorytmu SOM może być również uzyskana dzięki zastosowaniu podejścia wykorzystującego informację na temat częstości występowania poszczególnych punktów w wejściowym zbiorze danych [3]. Innymi słowy, można rzec, że budowana jest charakterystyka częstościowa danych wejściowych. Charakterystyka ta dopuszcza pewne niewielkie rozbieżności pomiędzy punktami określanymi jako „w przybliżeniu takie same", co matematycznie wyrażone jest poprzez zastosowanie pewnego progu niepodobieństwa określającego stopień tolerancji podczas porównywania punktów w

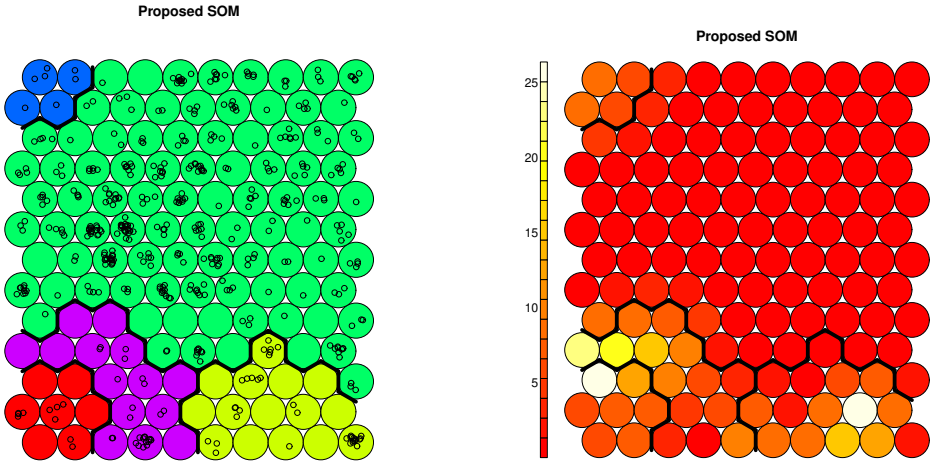
Tabela 7: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji aktywności robota

	q	U_d	time[sec]	s	Bonferroni p-value
SOM & k-means	371/463 = 0.8013	43/463 = 0.0929	0.1470	1.7903	4.3058e-41
TASOM & k-means	375/463 = 0.8099	25/463 = 0.0540	0.1850	1.7411	2.9004e-41
ASOM & k-means	374/463 = 0.8078	16/463 = 0.0346	0.1430	1.6960	5.3556e-41
FBSOM & k-means	438/463 = 0.9460	13/463 = 0.0281	0.1460	1.7712	1.4626e-39
LAMA & k-means	429/463 = 0.9266	29/463 = 0.0626	0.1440	1.7055	1.7227e-38
LARFSOM & k-means	436/463 = 0.9417	12/463 = 0.0750	0.1490	1.6482	4.1262e-40
RBSOM & k-means	417/463 = 0.9006	25/463 = 0.0540	0.1440	1.6831	2.0047e-38
Proposed SOM & k-means	**441/463 = 0.9525**	**9/463 = 0.0194**	**0.1450**	**1.7395**	——



(a) Struktura mapy (b) Macierz U

Rysunek 3: Wyniki wizualizacji i klasteryzacji mówców za pomocą proponowanej metody SOM



(a) Struktura mapy (b) Macierz U

Rysunek 4: Wyniki wizualizacji i klasteryzacji aktywności robota za pomocą proponowanej metody SOM

celu stwierdzenia, czy są „w przybliżeniu takie same", czy różne. Koncepcja wykorzystania takiego progu niepodobieństwa wprowadzona została w pracy [7], omówionej w Rozdziale 3.3, gdzie prezentowana jest, między innymi, asymetryczna postać sieci neuronowej SOM.

Informacja na temat częstości występowania poszczególnych punktów w wejściowej przestrzeni danych wykorzystana jest podczas określania szerokości sąsiedztwa neuronów na mapie SOM.

Teoretyczne uzasadnienie tak sformułowanej propozycji udoskonalenia techniki SOM jest następujące. Często występujące punkty w wejściowej przestrzeni danych są zlokalizowane blisko siebie w przestrzeni wejściowej, a zatem mapowanie SOM powinno dążyć do zachowania tego powiązania pomiędzy wejściowymi punktami danych i poprawnego odzwierciedlenia i przedstawienia go w otrzymanej wizualizacji danych. A zatem obszar na ekranie wizualizacji przypisany tym punktom powinien być nieduży. A tak może się stać dzięki użyciu małej szerokości sąsiedztwa neuronów reprezentujących te punkty na mapie SOM. Naturalnie, odwrotny proces i odwrotne postępowanie powinny mieć miejsce w przypadku punktów występujących rzadko w wejściowej przestrzeni danych.

Badania eksperymentalne w tej części analiz autora niniejszego autoreferatu przeprowadzone zostały z wykorzystaniem zbioru danych tekstowych, zbioru danych reprezentujących sygnały rytmu ludzkiego serca EKG oraz sygnały dźwiękowe reprezentujące utwory różnych kompozytorów muzyki kla-

Tabela 8: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji słów

Investigated method	q_{total}	s	p-value	U_d
TASOM	0.7719	2592.2	$< 10^{-4}$	0.2120
Auto-SOM	0.7479	3005.4	$< 10^{-4}$	0.2298
SOAN	0.7360	2767.2	$< 10^{-4}$	0.3301
Batch SOM	0.7416	2946.0	$< 10^{-4}$	0.1845
ADSOM	0.7285	3094.9	$< 10^{-4}$	0.2038
BTASOM	0.7798	3048.9	$< 10^{-4}$	0.1812
A-GHSOM	0.7381	2629.3	$< 10^{-4}$	0.2073
t-SNE & LDA	0.6955	2826.7	$< 10^{-4}$	0.2830
t-SNE & NMF	0.6929	2906.9	$< 10^{-4}$	0.2759
NeRV & LDA	0.6903	2446.9	$< 10^{-4}$	0.2578
NeRV & NMF	0.6894	2548.4	$< 10^{-4}$	0.3047
Proposed adaptive SOM	0.8450	2738.9	———	0.1402

sycznej.

Praca [3] posiada trzech autorów i jest wynikiem współpracy autora niniejszego autoreferatu z prof. dr. hab. inż. Januszem Kacprzykiem oraz z prof. dr. hab. Sławomirem Zadrożnym z Instytutu Badań Systemowych Polskiej Akademii Nauk.

Wkład autora autoreferatu jest następujący. Opracowanie teoretycznej koncepcji proponowanego udoskonalenia metody SOM; sformułowanie konkretnej metodologii postępowania w celu budowy docelowej architektury nowej sieci SOM; zaprojektowanie badań eksperymentalnych; implementacja proponowanej wersji algorytmu SOM; przeprowadzenie i nadzorowanie badań eksperymentalnych; ocena i interpretacja wyników przeprowadzonych badań eksperymentalnych; sformułowanie wniosków wynikających, zarówno z rozważań teoretycznych, jak i z wyników eksperymentów; napisanie manuskryptu artykułu. Szacunkowy wkład Dominika Olszewskiego wyrażony w procentach: 70%.

Stosowne oświadczenia są w załączeniu.

Proponowana metoda oceniona była na podstawie porównania z innymi wybranymi metodami odniesienia.

Sam proces redukcji wymiarowości danych i wynikającej z niej wizualizacji danych, kontynuowany był w postaci klasteryzacji *a posteriori* w celu umożliwienia zmierzenia jakości redukcji wymiarowości poprzez pomiar dokładności klasteryzacji danych w przestrzeni dwuwymiarowej, ponieważ trafność klasteryzacji można ocenić w nietrudny sposób za pomocą konkretnych kryteriów liczbowych, mianowicie, za pomocą miary trafności (ang. accuracy rate) oraz stopnia niepewności (ang. uncertainty degree).

Część wyników badań eksperymentalnych przedstawiona jest w Tabelach 8, 9 i 10, które prezentują jedynie fragment przeprowadzonych prac badawczych, zaś całość dostępna jest w pracy [3].

Tabela 9: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji utworów muzyki poważnej

Investigated method	q_{total}	s	p-value	U_d
TASOM	0.8563	0.6987	$< 10^{-4}$	0.0688
Auto-SOM	0.8750	0.7730	$< 10^{-4}$	0.0938
SOAN	0.7625	0.7140	$< 10^{-4}$	0.0625
Batch SOM	0.8375	0.6652	$< 10^{-4}$	0.0750
ADSOM	0.7688	0.7273	$< 10^{-4}$	0.0563
BTASOM	0.8813	0.6047	$< 10^{-4}$	0.1063
A-GHSOM	0.8375	0.7354	$< 10^{-4}$	0.0563
t-SNE & LDA	0.7563	0.7384	$< 10^{-4}$	0.2375
t-SNE & NMF	0.7438	0.7117	$< 10^{-4}$	0.2500
NeRV & LDA	0.7000	0.6704	$< 10^{-4}$	0.3375
NeRV & NMF	0.6438	0.6713	$< 10^{-4}$	0.3563
Proposed adaptive SOM	0.9125	0.7656	———	0.0313

Tabela 10: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji sygnałów EKG

Investigated method	q_{total}	s	p-value	U_d
TASOM	0.8175	1.1933	$< 10^{-4}$	0.1687
Auto-SOM	0.6905	1.1952	$< 10^{-4}$	0.1270
SOAN	0.6595	1.1865	$< 10^{-4}$	0.2063
Batch SOM	0.7778	1.2164	$< 10^{-4}$	0.1905
ADSOM	0.7857	1.1911	$< 10^{-4}$	0.2063
BTASOM	0.7937	1.2014	$< 10^{-4}$	0.1429
A-GHSOM	0.7302	1.1851	$< 10^{-4}$	0.1190
t-SNE & LDA	0.8095	1.3248	$< 10^{-4}$	0.0079
t-SNE & NMF	0.8071	1.3529	$< 10^{-4}$	0.0079
NeRV & LDA	0.7952	1.1493	$< 10^{-4}$	0.1349
NeRV & NMF	0.7778	1.2349	$< 10^{-4}$	0.1587
Proposed adaptive SOM	0.9286	1.3334	———	0.0476

3 Asymetria w analizie danych

Zjawisko asymetrii występuje w analizowanym zbiorze danych, gdy mamy do czynienia z punktami o różnym znaczeniu i istotności. Można wówczas rzecz, że charakteryzuje je różna prominencja. Efektem takiego zróżnicowania danych wejściowych jest powstanie niesymetrycznych powiązań pomiędzy punktami danych w przestrzeni wejściowej. Przykładem danych cechujących się tym rodzajem powiązań mogą być dane posiadające hierarchiczną strukturę. W takim przypadku, punkty w wejściowym zbiorze danych będą charakteryzowały się różnym stopniem ogólności i będziemy mieli do czynienia z punktami reprezentującymi wzorce danych na niskim oraz na wysokim poziomie ogólności. Taki rodzaj danych ma z definicji asymetryczną naturę i był przedmiotem rozważań i analiz w pracach autora niniejszego autoreferatu.

Aby wytłumaczyć i zilustrować zjawisko asymetrii w analizie danych, można posłużyć się prostym przykładem z dziedziny analizy tekstu. Mianowicie, tekst zazwyczaj posiada zróżnicowanie pod względem stopnia ogólności występujących w nim słów. I tak, w przypadku tekstu z dziedziny matematyki, można wyobrazić sobie występowanie słów „Bayes" oraz „statystyka". Naturalnie, słowo „Bayes" ma niższy poziom ogólności niż słowo „statystyka" i w konsekwencji relacja między tymi słowami będzie hierarchiczna, a zatem również asymetryczna. Potwierdzeniem tej obserwacji może być próba określenia miary niepodobieństwa pomiędzy tymi dwoma słowami. Gdy rozważamy niepodobieństwo w kierunku od słowa „Bayes" do słowa „statystyka", wówczas otrzymana wartość niepodobieństwa powinna być mniejsza niż w przeciwnym kierunku. Taka rozbieżność wartości zwracanych przez miarę niepodobieństwa w zależności od jej kierunku (kolejności umieszczenia jej argumentów wywołania) świadczy niepodważalnie o jej asymetrycznym charakterze.

3.1 Asymetryczna, zachowująca topologię metoda NeRV

Jako element dorobku w zakresie asymetrycznej analizy danych, autor niniejszego autoreferatu zaproponował w pracy [4] asymetryczną, zachowującą topologię wejściowych danych wersję metody NeRV.

U podstaw owej nowej postaci algorytmu NeRV leży postulat dążenia do adaptacyjnego dopasowania sposobu budowy rozkładów prawdopodobieństwa w metodzie NeRV do charakteru wejściowych danych. W pierwszym kroku konstruowany jest graf K-najbliższych-sąsiadów reprezentujący wejściowy zbiór danych. Graf ten jest dyskretną aproksymacją niskowymiarowej rozmaitości różniczkowej zanurzonej w wielowymiarowej, liniowej wejściowej

przestrzeni Euklidesa. Pozwala on zatem na wydobycie wewnętrznej geometrii wejściowych danych, a w konsekwencji ich topologii.

Graf typu K-najbliższych-sąsiadów ilustrujący wejściowe dane jest bazą dla wyznaczenia współczynników asymetrii (3) wykorzystanych jako istotne parametry Alfa-Beta dywergencji (4), która w ten sposób zyskuje asymetryczną postać (5) i może być wykorzystana do budowy funkcji gęstości prawdopodobieństwa (6) oraz (7). W ten sposób powstaje asymetryczna forma metody NeRV.

$$a_i \;=\; \frac{1}{K_i} \sum_{j=1}^{K_i} d_{\mathrm{Euc}} \left(x_{x_i, x_j} \right) , \tag{3}$$

gdzie x_i jest punktem wejściowym, x_j są punktami w sąsiedztwie punktu x_i, K_i jest liczbą najbliższych sąsiadów punktu x_i, natomiast d_{Euc} jest odległością Euklidesa.

Współczynnik asymetrii o indeksie i jest przypisany do $i$tego punktu wejściowego.

Wzór opisujący Alfa-Beta dywergencję ma następującą postać:

$$D_{\mathrm{AB}}^{(\alpha,\beta)} \left(\mathbf{P} \,||\, \mathbf{Q} \right) \;=\; -\frac{1}{\alpha\beta} \sum_{i=1}^{d} \left(p_i^{\alpha}\, q_i^{\beta} \;-\; \frac{\alpha}{\alpha+\beta}\, p_i^{\alpha+\beta} \;-\; \frac{\beta}{\alpha+\beta}\, q_i^{\alpha+\beta} \right) , \tag{4}$$

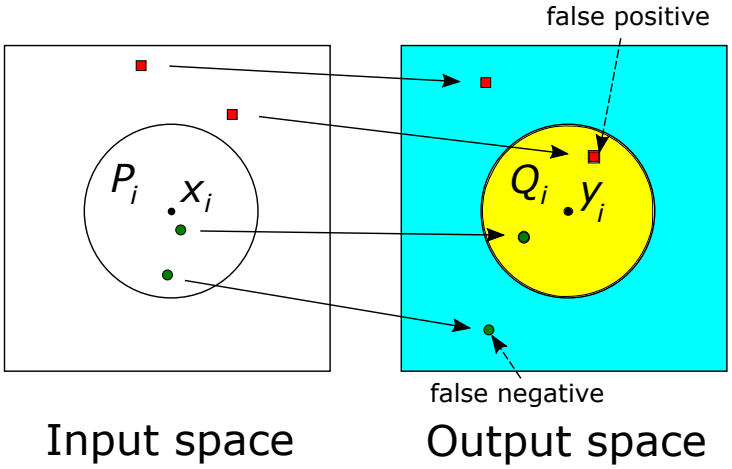
$$\alpha + \beta \neq 0 \, ,$$

gdzie $\mathbf{P} = (p_1, p_2, \ldots, p_d)$ i $\mathbf{Q} = (q_1, q_2, \ldots, q_d)$ są dodatnimi miarami (co jest terminem bardziej ogólnym niż rozkład prawdopodobieństwa), p_i i q_i $(i = 1, \ldots, d)$ reprezentują $i$te współrzędne miar $\mathbf{P}$ i $\mathbf{Q}$, podczas gdy d jest liczbą wymiarów w wejściowej przestrzeni danych.

Wprowadzona nowa asymetryczna miara niepodobieństwa pomiędzy punktami x_i oraz x_j z przypisanymi współczynnikami asymetrii – odpowiednio – a_i oraz a_j jest zdefiniowana w następujący sposób:

$$\begin{aligned} & D_{\mathrm{Asymmetric}}^{(a_i,a_j)} \left(x_i \,||\, x_j \right) \\ & \qquad = \; -\frac{a_j}{a_i} \sum_{l=1}^{d} \left(x_{i,l}\, x_{j,l}^{\frac{a_i}{a_j}} \;-\; \frac{a_j}{a_i + a_j}\, x_{i,l}^{1+\frac{a_i}{a_j}} \;-\; \frac{a_i}{a_i + a_j}\, x_{j,l}^{1+\frac{a_i}{a_j}} \right) , \end{aligned} \tag{5}$$

$$\frac{a_i}{a_j} \neq -1 \, .$$



Rysunek 5: Ilustracja graficzna mapowania danych przeprowadzanego przez metodę NeRV w celu redukcji wymiarowości danych wejściowych

$$p_{j|i} = \frac{\exp\left(-\frac{D^{(a_i,a_j)}_{\text{Asymmetric}}(x_i \,||\, x_j)^2}{\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{D^{(a_i,a_j)}_{\text{Asymmetric}}(x_i \,||\, x_k)^2}{\sigma_i^2}\right)}, \tag{6}$$

gdzie σ_i jest szerokością sąsiedztwa wykorzystywaną w projekcji dokonywanej przez metodę NeRV, zarówna w wejściowej, jak i w wyjściowej przestrzeni danych (nie ma ona nic wspólnego z szerokością sąsiedztwa punktów wejściowych używaną w grafie K-najbliższych-sąsiadów reprezentującym wejściowy zbiór danych). Szerokość sąsiedztwa σ_i jest zilustrowana jako promień okręgów przedstawionych na Rysunku 5.

$$q_{j|i} = \frac{\exp\left(-\frac{d_{\text{Euc}}(y_i, y_j)^2}{\sigma_i^2}\right)}{\sum_{k\neq i}\exp\left(-\frac{d_{\text{Euc}}(y_i, y_k)^2}{\sigma_i^2}\right)}, \tag{7}$$

gdzie y_i, y_j i y_k są punktami w niskowymiarowej, wyjściowej przestrzeni danych, to znaczy, są wynikiem mapowania punktów z wejściowej, wielowymiarowej przestrzeni danych.

Główny koncept metody NeRV przedstawiony jest w formie graficznej na Rysunku 5.

Teoretyczne uzasadnienie tak zaprojektowanej asymetrycznej techniki NeRV odnosi się bezpośrednio do sposobu odzwierciedlania zjawiska asymetrii w danych wejściowych. Asymetryczne współczynniki (3) odgrywają

podwójną rolę w proponowanej modyfikacji metody NeRV. Z jednej strony, ekstrahują informację na temat asymetrycznych powiązań danych wejściowych, a z drugiej strony, tworzą reprezentację geometrii (nieformalnie mówiąc, kształtu) i topologii wejściowych danych. Zatem ostatecznie, w ten sposób sformułowana metoda NeRV jest asymetryczna i zachowująca topologię oryginalnego, wejściowego zbioru danych.

Studium eksperymentalne w tym zakresie prac badawczych przeprowadzone zostało z wykorzystaniem trzech rzeczywistych zbiorów danych: zbioru danych tekstowych, zbioru danych dźwiękowych reprezentujących dane głosowe w zadaniu rozpoznawania mówcy oraz dane reprezentujące aktywność robota mobilnego wykonującego zadane czynności należące do pięciu różnych kategorii.

Proponowana metoda oceniona była na podstawie porównania z innymi wybranymi metodami odniesienia.

Sam proces redukcji wymiarowości danych i wynikającej z niej wizualizacji danych, kontynuowany był w postaci klasteryzacji *a posteriori* w celu umożliwienia zmierzenia jakości redukcji wymiarowości poprzez pomiar dokładności klasteryzacji danych w przestrzeni dwuwymiarowej, ponieważ trafność klasteryzacji można ocenić w nietrudny sposób za pomocą konkretnych kryteriów liczbowych, mianowicie, za pomocą miary trafności (ang. accuracy rate) oraz stopnia niepewności (ang. uncertainty degree).

Część wyników badań eksperymentalnych przedstawiona jest w Tabelach 11, 12 i 13 oraz na Rysunkach 6, 7 8 i 9, które prezentują jedynie fragment przeprowadzonych prac badawczych, zaś całość dostępna jest w pracy [4].

3.2 Asymetryczna klasteryzacja danych

Autor niniejszego autoreferatu zaproponował w pracy [5] asymetryczną wersję algorytmu klasteryzacji danych k-centroidów.

Standardowa postać tego algorytmu jest ograniczona, między innymi, przez problem formowania klastrów danych o zbliżonym rozmiarze, wyrażanym, zarówno poprzez liczbę punktów w klastrach, jak i obszar w przestrzeni danych zajmowany przez dany klaster.

Proponowane asymetryczne podejście ma na celu przezwyciężenie tego niepożądanego zjawiska. Asymetryczna wersja algorytmu k-centroidów sformułowana jest z wykorzystaniem Alfa-Beta dywergencji jako miary niepodobieństwa w metodzie k-centroidów. Alfa-Beta dywergencja jest wygodnym przedmiotem asymetryzacji, ponieważ zawiera parametry pozwalające na swobodny wpływ na jej własności i charakter w szerokim zakresie, w szczególności pod względem jej cechę symetrii/asymetrii. Wartości parametrów Alfa-Beta dywergencji dobierane są na podstawie aktualnego rozmiaru

Tabela 11: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji słów

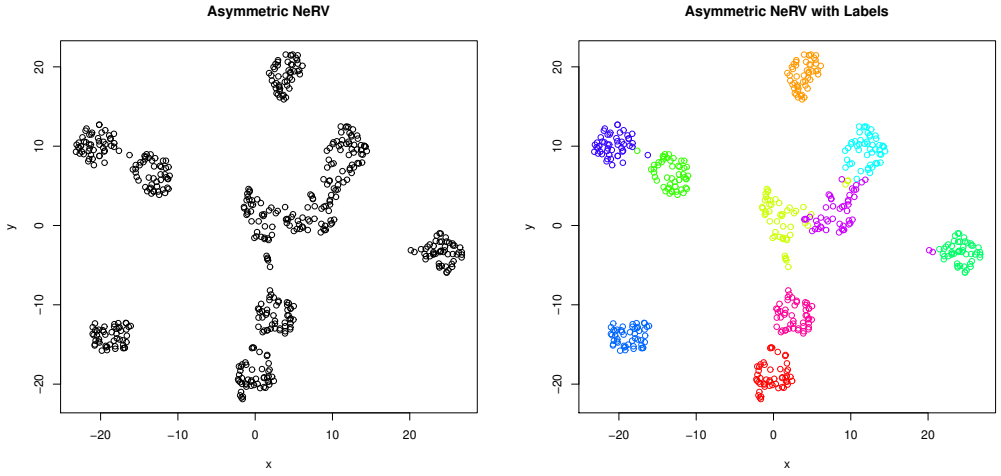
	q	U_d	s	Bonferroni p-value
LLE & k-means	9,192,725/10,868,000 = 0.8459	1,481,701/10,868,000 = 0.1363	2.5350	6.8504e-38
LLE & DBSCAN	8,982,036/10,868,000 = 0.8267	1,125,827/10,868,000 = 0.1036	3.0921	2.5083e-39
ISOMAP & k-means	8,592,473/10,868,000 = 0.7906	1,679,135/10,868,000 = 0.1545	3.2638	6.9145e-40
ISOMAP & DBSCAN	8,287,310/10,868,000 = 0.7625	1,480,638/10,868,000 = 0.1362	3.3911	5.2601e-42
t-SNE & k-means	8,005,971/10,868,000 = 0.7367	2,513,598/10,868,000 = 0.2313	3.0825	9.3859e-43
t-SNE & DBSCAN	8,001,178/10,868,000 = 0.7362	2,205,708/10,868,000 = 0.2030	3.1681	3.8427e-38
Traditional NeRV & k-means	8,910,543/10,868,000 = 0.8199	1,979,532/10,868,000 = 0.1821	2.5502	1.1533e-38
Traditional NeRV & DBSCAN	8,903,285/10,868,000 = 0.8192	1,955,731/10,868,000 = 0.1800	3.1052	3.1513e-39
UMAP & k-means	8,209,418/10,868,000 = 0.7554	2,560,348/10,868,000 = 0.2356	3.6381	3.0081e-41
UMAP & DBSCAN	8,093,142/10,868,000 = 0.7447	2,472,507/10,868,000 = 0.2275	2.4742	1.6427e-41
CBNeRV & k-means	9,230,487/10,868,000 = 0.8493	1,497,093/10,868,000 = 0.1378	2.0521	5.2988e-38
CBNeRV & DBSCAN	9,011,684/10,868,000 = 0.8292	1,484,370/10,868,000 = 0.1366	2.8554	1.0926e-39
Proposed NeRV & k-means	**9,708,124/10,868,000 = 0.8933**	**885,537/10,868,000 = 0.0815**	**3.4025**	——
Proposed NeRV & DBSCAN	9,392,411/10,868,000 = 0.8642	1,028,925/10,868,000 = 0.0947	2.7941	2.6324e-37

Tabela 12: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji mówców

	q	U_d	s	Bonferroni p-value
LLE & k-means	423/500 = 0.8460	41/500 = 0.0820	1.7403	8.5221e-42
LLE & DBSCAN	198/500 = 0.3960	19/500 = 0.0380	1.4835	3.6846e-37
ISOMAP & k-means	458/500 = 0.9160	34/500 = 0.0680	1.2481	1.4944e-39
ISOMAP & DBSCAN	254/500 = 0.5080	7/500 = 0.0140	1.4841	1.1644e-39
t-SNE & k-means	372/500 = 0.7440	78/500 = 0.1560	1.0845	2.9275e-41
t-SNE & DBSCAN	260/500 = 0.5200	48/500 = 0.0960	1.0142	1.2847e-40
Traditional NeRV & k-means	342/500 = 0.6840	71/500 = 0.1420	1.9674	4.6076e-37
Traditional NeRV & DBSCAN	329/500 = 0.6580	93/500 = 0.1860	1.1730	4.4016e-39
UMAP & k-means	432/500 = 0.8640	38/500 = 0.0760	1.7043	2.7799e-42
UMAP & DBSCAN	425/500 = 0.8500	20/500 = 0.0400	1.3110	2.1543e-39
CBNeRV & k-means	456/500 = 0.9120	33/500 = 0.0660	1.0355	1.7034e-42
CBNeRV & DBSCAN	448/500 = 0.8960	26/500 = 0.0520	1.3819	3.9348e-36
Proposed NeRV & k-means	**493/500 = 0.9860**	**20/500 = 0.0400**	**1.3159**	——
Proposed NeRV & DBSCAN	464/500 = 0.9280	24/500 = 0.0480	1.2037	5.1112e-41

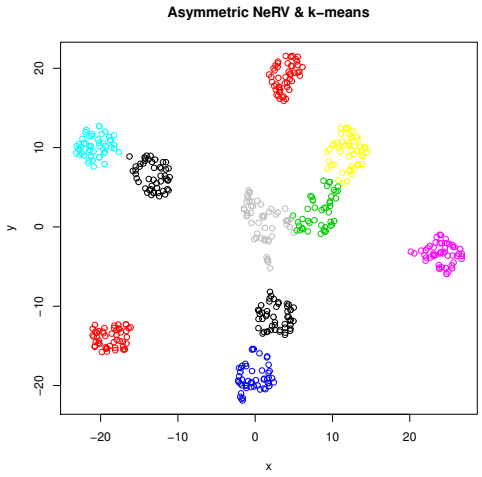
Tabela 13: Miary trafności, stopnie niepewności, odchylenia standardowe i p-wartości dla wizualizacji i klasteryzacji aktywności robota

	q	U_d	s	Bonferroni p-value
LLE & k-means	418/463 = 0.9028	32/463 = 0.0691	1.4025	1.4433e-37
LLE & DBSCAN	384/463 = 0.8294	12/463 = 0.0259	1.2724	6.4552e-36
ISOMAP & k-means	398/463 = 0.8596	31/463 = 0.0670	1.1703	6.3705e-37
ISOMAP & DBSCAN	391/463 = 0.8445	34/463 = 0.0734	1.5572	1.3762e-40
t-SNE & k-means	184/463 = 0.3974	79/463 = 0.1706	1.4804	3.8134e-38
t-SNE & DBSCAN	422/463 = 0.9114	0/463 = 0.0000	1.6003	2.3424e-40
Traditional NeRV & k-means	407/463 = 0.8790	41/463 = 0.0886	1.8610	2.1389e-42
Traditional NeRV & DBSCAN	391/463 = 0.8445	38/463 = 0.0821	0.9720	4.5358e-43
UMAP & k-means	414/463 = 0.8942	22/463 = 0.0475	0.9307	3.0277e-37
UMAP & DBSCAN	410/463 = 0.8855	21/463 = 0.0454	1.2888	3.9361e-39
CBNeRV & k-means	425/463 = 0.9179	28/463 = 0.0605	1.0618	6.6972e-41
CBNeRV & DBSCAN	417/463 = 0.9006	22/463 = 0.0475	1.5761	1.0762e-35
Proposed NeRV & k-means	**448/463 = 0.9676**	**19/463 = 0.0410**	**1.4207**	——
Proposed NeRV & DBSCAN	434/463 = 0.9374	8/463 = 0.0173	1.5993	8.7465e-40

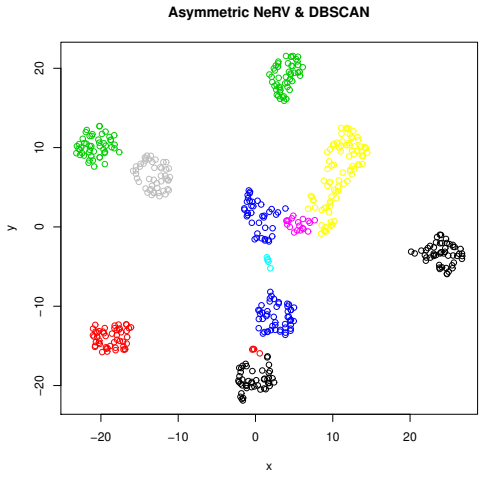


(a) Proponowany NeRV (b) Proponowany NeRV z etykietami

Rysunek 6: Wyniki wizualizacji mówców za pomocą proponowanej metody NeRV

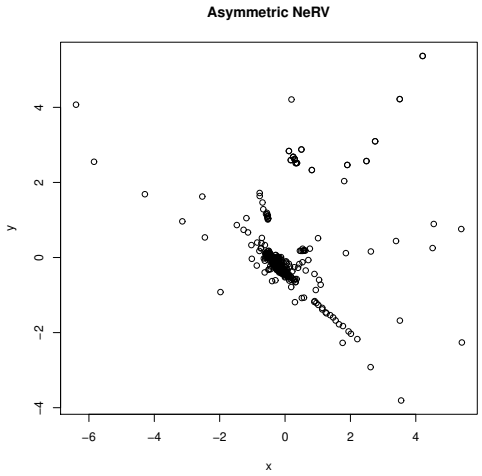


(a) Proponowany NeRV & k-means

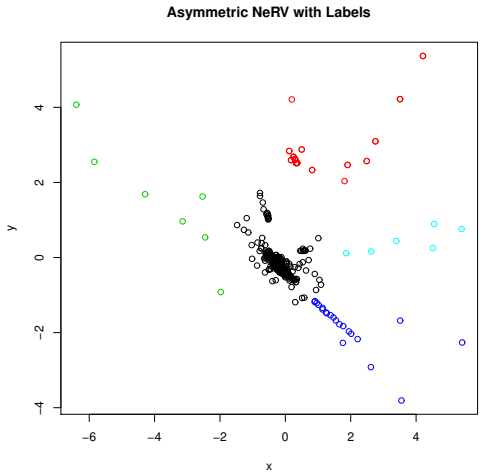


(b) Proponowany NeRV & DBSCAN

Rysunek 7: Wyniki wizualizacji mówców za pomocą proponowanej metody NeRV

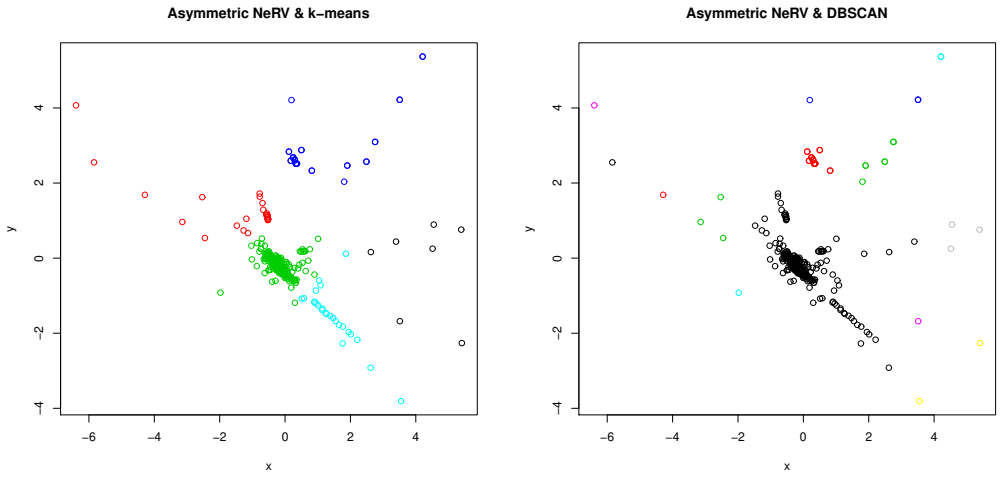


(a) Proponowany NeRV



(b) Proponowany NeRV z etykietami

Rysunek 8: Wyniki wizualizacji aktywności robota za pomocą proponowanej metody NeRV



(a) Proponowany NeRV & k-means (b) Proponowany NeRV & DBSCAN

Rysunek 9: Wyniki wizualizacji aktywności robota za pomocą proponowanej metody NeRV

klastrów w kolejnych cyklach techniki k-centroidów. Rozmiar klastrów definiowany jest jako obszar zajmowany przez klastry w przestrzeni danych, a nie jako liczba punktów w klastrach.

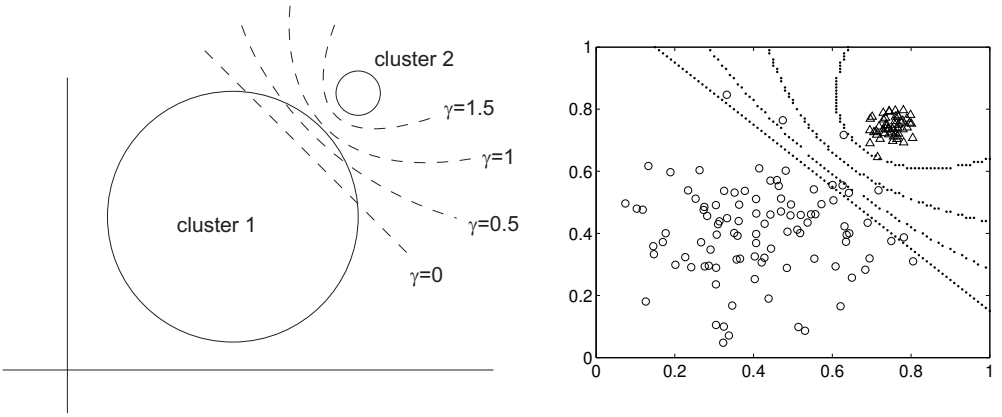
Uzasadnienie tak sformułowanego udoskonalania metody k-centroidów bazuje na koncepcji dopasowywania miary niepodobieństwa algorytmu k-centroidów do rozmiaru klastrów i w ten sposób uodpornienie podejścia k-centroidów na problem sztucznego i niepożądanego równoważenia rozmiaru generowanych klastrów. Asymetryzacja tej metody za pomocą parametrów uzależnionych od aktualnego rozmiaru klastrów pozwoli osiągnąć ten cel i stanowi zarazem przekonujące wytłumaczenie rozwiązania wprowadzonego przez autora niniejszego autoreferatu.

Proponowane rozwiązanie zilustrowane jest graficznie na Rysunku 10.

Eksperymenty przeprowadzone zostały na trzech zbiorach danych: zbiór danych tekstowych, zbiór danych medycznych przedstawiających sygnały rytmu ludzkiego serca EKG oraz zbiór danych dźwiękowych reprezentujących utwory muzyczne trzech różnych kompozytorów muzyki poważnej.

Praca [5], w której opublikowano omawiane nowe rozwiązania, posiada dwóch autorów i jest wynikiem współpracy wnioskodawcy z profesorem Branko Šterem z Uniwersytetu w Lublanie w Słowenii.

Wkład autora niniejszego autoreferatu jest następujący. Opracowanie teoretycznej koncepcji wykorzystania asymetrycznego podejścia w dziedzinie klasteryzacji danych; zaprojektowanie badań eksperymentalnych; implemen-



(a) Schematycznie

(b) Z danymi

Rysunek 10: Ilustracja wpływu parametru γ

Tabela 14: Wyniki klasteryzacji symulowanych dwuwymiarowych danych

	q_{total}	s	p-wartość
k-medoids	81.6%	7.0%	$< 10^{-4}$
Asymmetric k-means	91.7%	5.9%	6.7×10^{-3}
Asymmetric hierarchical	85.2%	5.3%	$< 10^{-4}$
DBSCAN	92.6%	3.1%	5.6×10^{-3}
Proponowana metoda	95.7%	4.9%	———

tacja proponowanej asymetrycznej wersji algorytmu k-centroidów; przeprowadzenie i nadzorowanie prac empirycznych; analiza, ocena i interpretacja wyników przeprowadzonych eksperymentów; sformułowanie wniosków wynikających, zarówno z rozważań teoretycznych, jak i z wyników przeprowadzonych eksperymentów; napisanie manuskryptu artykułu. Szacunkowy wkład Dominika Olszewskiego wyrażony w procentach: 50%.

Stosowne oświadczenia autorów są w załączeniu.

Część wyników badań eksperymentalnych przedstawiona jest w Tabelach 14, 15, 16 i 17 oraz na Rysunkach 11, 12 i 13, które prezentują jedynie fragment przeprowadzonych prac badawczych, zaś całość dostępna jest w pracy [5].

3.3 Asymetryczna klasteryzacja asymetrycznej wersji mapy SOM

Asymetryczna wersja metody SOM została wprowadzona w pracy [15] i uzyskana jest dzięki zastosowaniu współczynników asymetrii obliczanych w wy-

Tabela 15: Wyniki klasteryzacji utworów muzyki poważnej

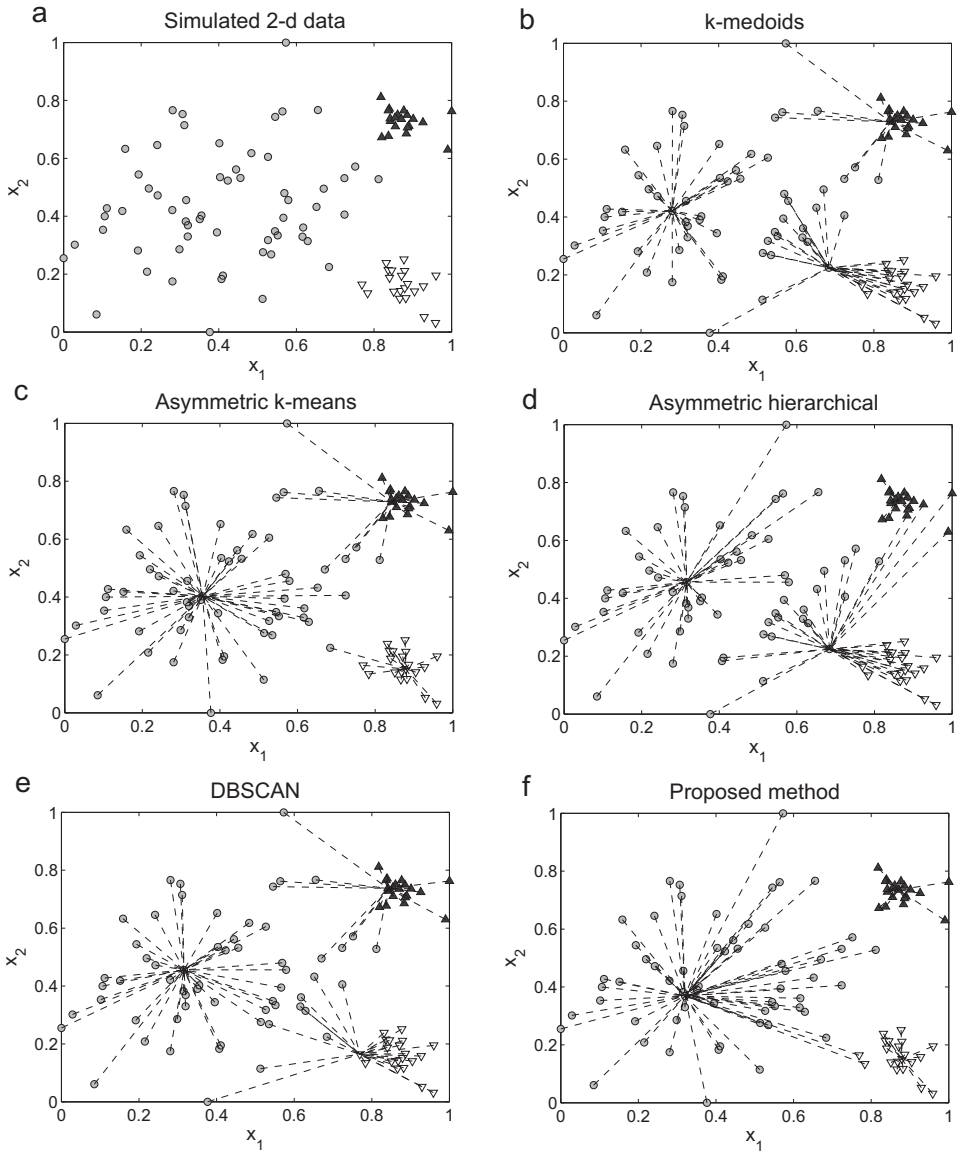
	q_{total}	s	p-wartość
k-medoids	81.5% (856/1050)	3.2	$< 10^{-4}$
Asymmetric k-means	87.5% (919/1050)	2.3	$< 10^{-4}$
Asymmetric hierarchical	83.6% (878/1050)	3.0	$< 10^{-4}$
DBSCAN	88.7% (931/1050)	3.0	4.6×10^{-3}
Proponowana metoda	90.2% (948/1050)	2.6	——–

Tabela 16: Wyniki klasteryzacji sygnałów EKG

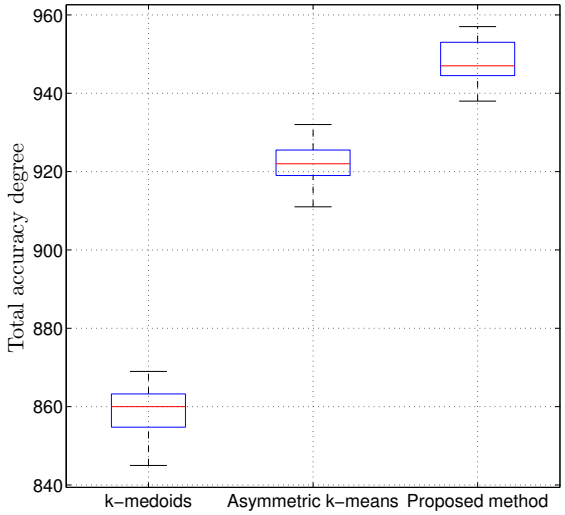
	q_{total}	s	p-wartość
k-medoids	73.0% (46/63)	1.1	$< 10^{-4}$
Asymmetric k-means	82.5% (52/63)	1.1	$< 10^{-4}$
Asymmetric hierarchical	71.4% (45/63)	1.3	$< 10^{-4}$
DBSCAN	98.4% (62/63)	1.1	$< 10^{-4}$
Proponowana metoda	93.7% (59/63)	0.8	——–

Tabela 17: Wyniki klasteryzacji słów

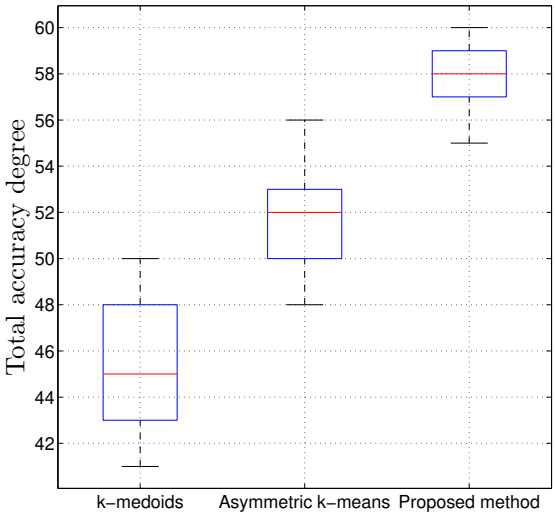
	q_{total}	s	p-wartość
k-medoids	75.38% (8,192,319/10,868,000)	61.1	$< 10^{-4}$
Asymmetric k-means	78.41% (8,521,727/10,868,000)	55.4	$< 10^{-4}$
Asymmetric hierarchical	77.82% (8,457,118/10,868,000)	53.8	$< 10^{-4}$
DBSCAN	84.13% (9,143,682/10,868,000)	60.6	$< 10^{-4}$
Proponowana metoda	91.37% (9,930,502/10,868,000)	61.8	——–



Rysunek 11: (a) Trzy klastry symulowanych dwuwymiarowych danych. Wyniki klasteryzacji symulowanych dwuwymiarowych danych za pomocą (b) algorytmu k-medoidów, (c) asymetrycznego algorytmu k-średnich, (d) asymetrycznego algorytmu hierarchicznego, (e) algorytmu DBSCAN, i (f) proponowanej metody.



Rysunek 12: Wykresy pudełkowe liczb poprawnie przypisanych utworów muzyki poważnej w klasteryzacji utworów muzyki poważnej odpowiadające pięciu badanym metodom.



Rysunek 13: Wykresy pudełkowe liczb poprawnie przypisanych sygnałów EKG w klasteryzacji sygnałów EKG odpowiadające pięciu badanym metodom.

niku pomiaru częstości występowania punktów w wejściowym zbiorze danych. Częstości te mają odzwierciedlać i wyrażać stopień ogólności punktów w wejściowym zbiorze danych, a zatem również ich znaczenie i prominencję. Są więc nośnikiem informacji dotyczącej asymetrycznych powiązań pomiędzy wejściowymi danymi. Tak zdefiniowane asymetryczne współczynniki stają się częścią wzoru wyznaczania miary niepodobieństwa zbudowanej na bazie tradycyjnej odległości Euklidesa.

Asymetryczna forma metody k-średnich to już oryginalna propozycja autora autoreferatu, przedstawiona w artykule [6]. W metodzie tej wykorzystane są wcześniej wspomniane w tym rozdziale współczynniki asymetrii, ale tym razem stanowiące rozszerzenie i udoskonalanie algorytmu klasteryzacji danych k-średnich.

W kolejnym kroku rozważań, autor niniejszego autoreferatu postanowił obrać jako przedmiot badań połączenie asymetrycznej wersji metody SOM oraz asymetrycznego algorytmu k-średnich (opublikowane również w artykule [6]). Kombinacja ta polega na tym, że oryginalne dane w pierwszym etapie analizy i przetwarzania podlegają redukcji wymiarowości za pomocą asymetrycznej metody SOM, a następnie są przedmiotem klasteryzacji z wykorzystaniem asymetrycznej techniki k-średnich przeprowadzającej proces klasteryzacji danych w dwuwymiarowej przestrzeni wizualizacji mapy SOM.

Uzasadnienie teoretyczne proponowanego sprzężenia asymetrycznych postaci algorytmów SOM oraz k-średnich sprowadza się do stwierdzenia, iż redukcja wymiarowości wykonana za pomocą metody SOM asymetryzowanej w określony sposób z wykorzystaniem stosownie zdefiniowanych współczynników asymetrii powinna być kontynuowana za pomocą metod analizy danych o podobnej naturze i podobnym charakterze. Zatem w przypadku kontynuacji analizy w postaci klasteryzacji danych na ekranie wizualizacji, uzasadnione jest wykorzystanie asymetrycznej metody klasteryzacji o tej samej technice asymetryzacji, co w przypadku asymetrycznej wersji sieci neuronowej SOM. W ten sposób, zapewniona zostanie koncepcyjna i metodologiczna spójność i jednolitość pomiędzy dwiema technikami analizy danych, kluczowymi w tym zadaniu, to znaczy, techniką redukcji wymiarowości (wizualizacji danych) oraz techniką klasteryzacji danych w dwuwymiarowej przestrzeni wizualizacji mapy SOM.

Badania eksperymentalne na wybranych zbiorach danych posłużyły weryfikacji i potwierdzeniu stwierdzeń i propozycji sformułowanych i postulowanych w sferze rozważań teoretycznych. Wykorzystane zbiory danych to: zbiór danych tekstowych, zbiór danych reprezentujących zużycie energii elektrycznej w przykładowym gospodarstwie domowym, zbiór danych dźwiękowych (utwory muzyki klasycznej), zbiór danych medycznych (sygnały rytmu ludzkiego serca EKG).

Kolejną propozycją w zakresie badań nad asymetryczną wersją mapy SOM była praca [7], w której wprowadzona została postać asymetrycznej mapy SOM przystosowanej do wizualizacji szeregów czasowych. W pracy tej zauważono, że w przypadku niektórych danych, na przykład, w przypadku szeregów czasowych, trudno oczekiwać równości dwóch obiektów będących szeregami czasowymi, ze względu na dużą zazwyczaj liczbę próbek danego szeregu czasowego. W proponowanym rozwiązaniu wprowadzono zatem próg tolerancji wspomagający porównywanie szeregów czasowych. Wynik porównywania dwóch szeregów czasowych jest dokonywany względem tego progu tolerancji, tak aby dopuścić pewną niewielką rozbieżność pomiędzy dwoma szeregami czasowymi i określić je wówczas jako „w przybliżeniu takie same", jak to już wspominano w Rozdziale 2.2.2 niniejszego autoreferatu. Koncepcja ta wiąże się w pewnym stopniu z założeniami, technikami i mechanizmami w dziedzinie logiki rozmytej.

Eksperymenty w pracy [7] zostały przeprowadzone przy wykorzystaniu zbioru danych zbudowanego z szeregów czasowych reprezentujących wartości akcji firm notowanych na amerykańskiej giełdzie papierów wartościowych.

Praca [7] posiada trzech autorów i jest wynikiem współpracy autora niniejszego autoreferatu z prof. dr. hab. inż. Januszem Kacprzykiem oraz z prof. dr. hab. Sławomirem Zadrożnym z Instytutu Badań Systemowych Polskiej Akademii Nauk.

Wkład autora autoreferatu jest następujący. Opracowanie teoretycznej koncepcji proponowanego udoskonalenia metody SOM; sformułowanie konkretnej metodologii postępowania w celu budowy docelowej architektury nowej sieci SOM; zaprojektowanie badań eksperymentalnych; implementacja proponowanej wersji algorytmu SOM; przeprowadzenie i nadzorowanie badań eksperymentalnych; ocena i interpretacja wyników przeprowadzonych badań eksperymentalnych; sformułowanie wniosków wynikających, zarówno z rozważań teoretycznych, jak i z wyników eksperymentów; napisanie manuskryptu artykułu. Szacunkowy wkład Dominika Olszewskiego wyrażony w procentach: 70%.

Stosowne oświadczenia autorów są w załączeniu.

4 Wykrywanie danych odstających

Dane odstające (ang. outliers) to punkty w zbiorze danych różniące się od głównego skupienia danych w tak znacznym stopniu, iż uzasadnione jest podejrzenie, że powstały w wyniku działania innego mechanizmu niż główne skupienie danych.

Wykrywanie danych odstających to szczególna postać zadania klasyfika-

cji danych, w którym mamy do czynienia jedynie z dwiema klasami danych: dane odstające oraz dane pozostałe. Z uwagę na fakt, iż dane określone jako „pozostałe" nie są zdefiniowane precyzyjnie, a jedyne stanowią uzupełnienie zbioru danych odstających w całym zbiorze danych, czasem w przypadku tego zagadnienia klasyfikacji i wykrywania danych odstających używa się określenia „klasyfikacji do jednej klasy (ang. one-class classification)", aby zwrócić uwagę na fakt, że w istocie w tak sformułowanym zadaniu, wyodrębniona została tylko jedna klasa – klasa danych odstających, a pozostałe dane nie posiadają etykiety i nie należą zatem do konkretnej klasy. Co więcej, w obrębie danych niebędących danymi odstającymi, można uformować szereg kolejnych klas i w konsekwencji kontynuować proces klasyfikacji danych.

Zagadnienie wykrywania danych odstających zyskuje szczególnie dużą wartość i znaczenie, a jego wyniki mogą być uznane za cenne i przydatne, z uwagi na możliwość wykorzystania w dziedzinie wykrywania i sygnalizowania potencjalnych zagrożeń i niebezpieczeństw związanych z nieuprawnionych, nielegalnym, bądź też jedynie budzącym podejrzenia zachowaniem użytkowników korzystających z określonych usług. Jako przykład podać można nieuprawnione korzystanie z usług telekomunikacyjnych, bankowych, czy ataki w sieciach komputerowych.

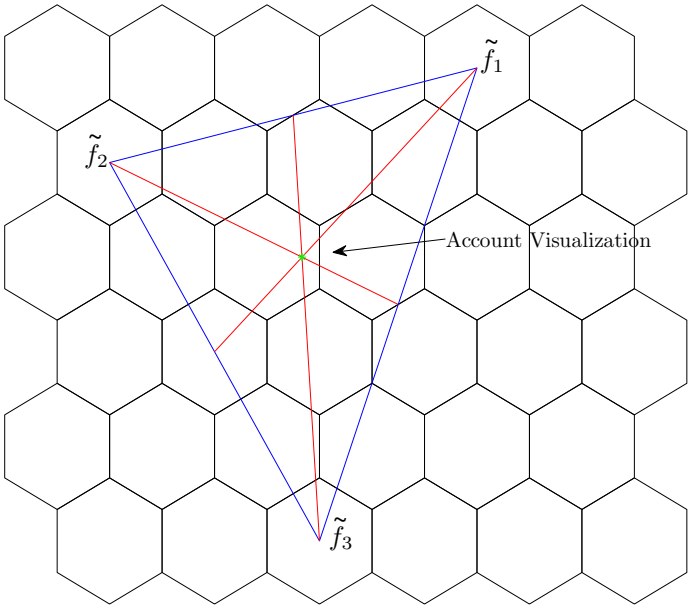
4.1 Wykrywanie danych odstających za pomocą metody SOM wizualizującej profile użytkowników

Autor niniejszego autoreferatu zaproponował w pracy [8] wykorzystanie metody SOM do generowania wizualizacji profilów użytkowników, a w kolejnej fazie analizy danych, do wykrywania aktywności użytkowników, która budzi podejrzenie i może mieć potencjalnie nielegalny charakter.

Wkład oryginalny autora to przystosowanie techniki SOM do projekcji wzorców danych reprezentowanych przez macierze, a zatem struktury dwuwymiarowe, a nie jedynie przez wektory cech (struktury jednowymiarowe), jak to ma miejsce w przypadku klasycznej mapy SOM. Inaczej mówiąc, cechy w analizowanym zbiorze danych miały swoją wewnętrzną wymiarowość. W ten sposób nowa, wprowadzana wersja mapy SOM była w stanie wizualizować całe konta użytkowników, zbudowane w oparciu o sekwencje zapisów poszczególnych aktywności użytkowników. Klasyczna mapa SOM nie posiadała takiej zdolności.

Wizualizacja kont użytkowników za pomocą proponowanej przez autora metody zilustrowana jest graficznie na Rysunku 14.

Drugim oryginalnym rozwiązaniem autora zaproponowanym w artykule [8] jest technika automatycznego wyznaczania wartości progu w algoryt-



Rysunek 14: Wizualizacja konta przykładowego użytkownika

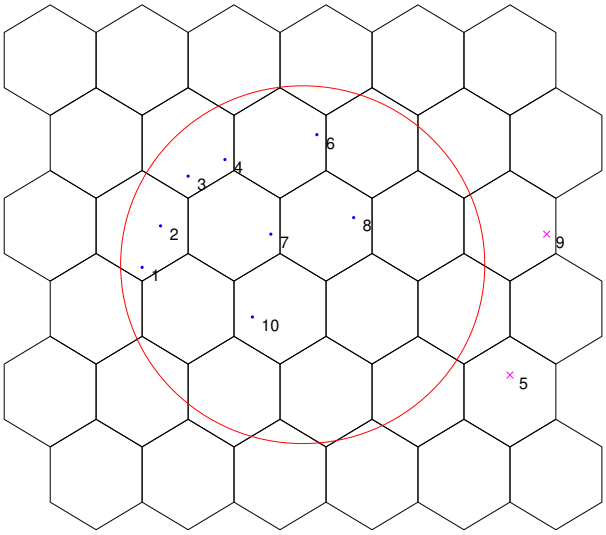
mie klasyfikacji typu progowego, wykorzystanym do finalnego wykrywania danych odstających, a w konsekwencji profilów użytkowników, związanych z podejrzeniem przekroczenia uprawnień.

Finalna klasyfikacja mająca na celu wykrycie danych odstających zaprezentowana jest na Rysunku 15.

Uzasadnienie teoretyczne wprowadzonego przez autora podejścia jest następujące. Metoda SOM w standardowej postaci rzutuje jeden (ten większy) z wymiarów macierzy reprezentujących konta użytkowników, a powstały zbiór punktów na mapie SOM sprowadzany jest do reprezentacji za pomocą jednego punktu, który jest centroidem tego zbioru. Automatyczny wybór progu w klasyfikacji progowej odbywa się w oparciu o wartości komórek macierzy U, będącej graficzną formą wyświetlania mapy SOM. W macierzy U identyfikowany jest ciąg wysokich wartości (tzw. grzbiet (ang. ridge)) i to on właśnie stanowi linię podziału separującą dane odstające od pozostałych.

Empiryczna weryfikacji i ocena proponowanych rozwiązań, rozszerzeń i udoskonaleń w części teoretycznej przeprowadzona została na kilku różnorodnych rzeczywistych zbiorach danych.

Rysunek 16 prezentuje wizualizację kont użytkowników telekomunikacyjnych z zaznaczeniem kont legalnych (litery N od ang. Non-fraudulent) oraz podejrzanych o nielegalne aktywności (litery F od ang. Fraudulent).



Rysunek 15: Ilustracja graficzna progowego binarnego algorytmu klasyfikacji

4.2 Wykrywanie danych odstających za pomocą rozkładu probabilistycznego LDA

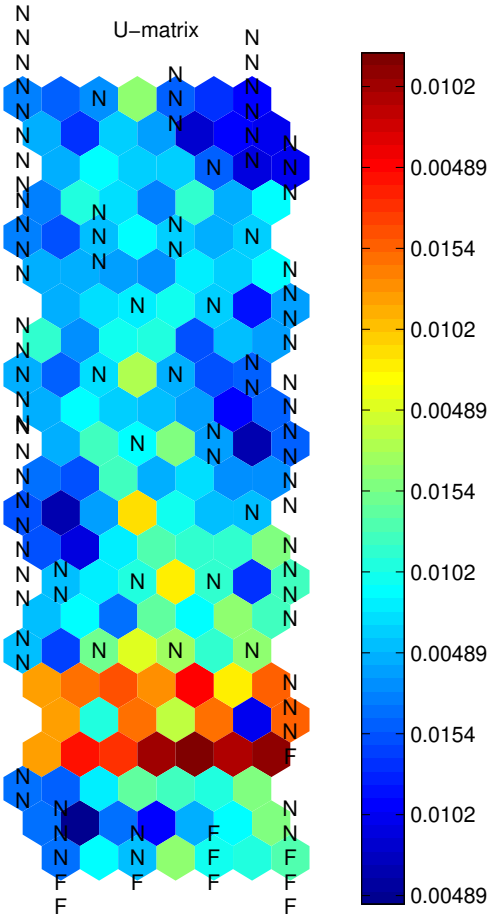
W tej części prac badawczych, autor niniejszego autoreferatu wprowadził w pracy [9] metodę wykrywania danych odstających sformułowaną w oparciu o modelowanie za pomocą rozkładu probabilistycznego Latent Dirchlet Allocation (LDA) oraz algorytm klasyfikacji.

Konta użytkowników telekomunikacyjnych modelowane są z wykorzystaniem rozkładu probabilistycznego LDA, a przypadki danych odstających identyfikowane są w oparciu o obliczanie miary niepodobieństwa do modelu LDA konta odniesienia, reprezentującego sposób zachowań typowego użytkownika usług telekomunikacyjnych.

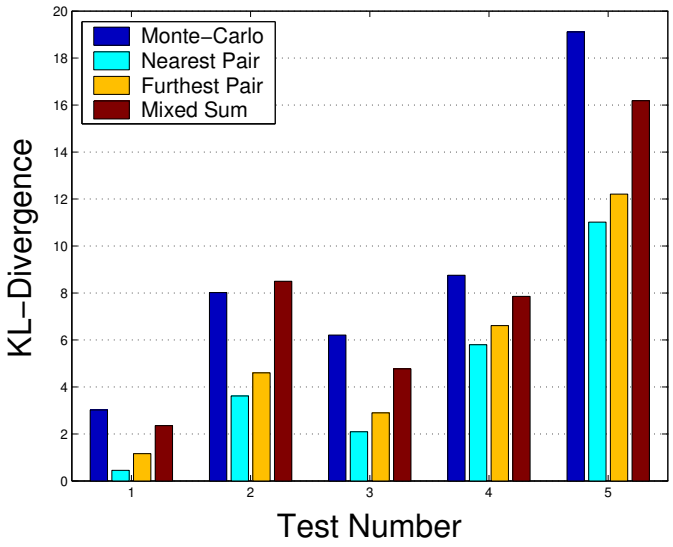
Wykorzystana miara niepodobieństwa to dywergencja Kullbacka—Leiblera [16] – klasyczna wielkość z pogranicza dziedziny rachunku prawdopodobieństwa oraz teorii informacji.

Wkładem oryginalnym autora jest opracowanie samej koncepcji klasyfikacji kont telekomunikacyjnych z zamiarem wykrywania danych odstających reprezentujących potencjalne nadużycia oraz zaproponowanie aproksymacji dywergencji Kullbacka—Leiblera pomiędzy dwoma rozkładami probabilistycznymi LDA, ponieważ tradycyjna postać dywergencji Kullbacka—Leiblera zdefiniowana jest jedynie dla dwóch funkcji gęstości prawdopodobieństwa.

Wyniki analizy porównawczej wybranych czterech metod aproksymacji



Rysunek 16: Macierz U reprezentująca mapę SOM wizualizującą badany zbiór danych telekomunikacyjnych z zaznaczeniem kont legalnych oraz podejrzanych o nielegalne aktywności

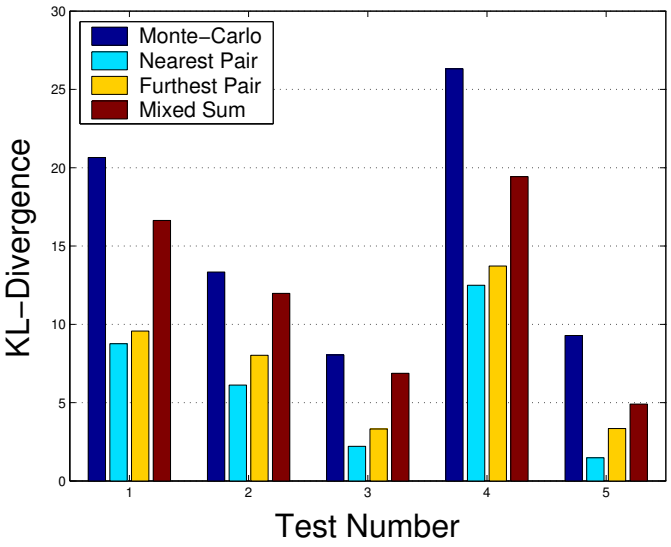


Rysunek 17: Wyniki aproksymacji dywergencji Kullbacka—Leiblera pomiędzy dwoma rozkładami probabilistycznymi LDA dla trzech klas ukrytych

dywergencji Kullbacka—Leiblera przedstawione są na Rysunkach 17 i 18.

Uzasadnienie teoretyczne tak zaprojektowanej metody wykrywania danych odstających odnosi się do faktu, że dane telekomunikacyjne cechują się pewnym specyficznym charakterem. Otóż w przypadku problemu wykrywania nieuprawnionego korzystania z usług telekomunikacyjnych zazwyczaj dysponujemy niewielką ilością danych uczących w porównaniu z rozmiarem zbioru danych testowych. A to w istotny sposób ogranicza możliwości wykorzystania metod opierających się na sztucznych sieciach neuronowych. Uzasadnionym i racjonalnym wyborem będą zatem matematyczne modele probabilistyczne użyte jako podstawa do formułowania metod i algorytmów wykrywania danych odstających w obrębie zbioru danych telekomunikacyjnych. Szczególnym zainteresowaniem i dużą uwagą mogą cieszyć się generyczne modele bayesowskie o formie skończonych rozkładów mieszaninowych, jak już wspominany model LDA wprowadzony w pracy [17]. Natomiast pod względem wyboru dywergencji Kullbacka—Leiblera jako miary niepodobieństwa, decyzję tę można uzasadnić jej użytecznością w określeniu rozbieżności pomiędzy rozkładami prawdopodobieństwa potwierdzaną nieustająco na przestrzeni wielu lat, która nierozerwalnie wiąże tej wielkość z obszarami badawczymi statystyki i rachunku prawdopodobieństwa, a w dalszej perspektywie również z dziedzinami teorii informacji, sztucznej inteligencji i uczenia maszynowego.

Badania eksperymentalne tym razem dotyczyły zbioru danych telekomu-



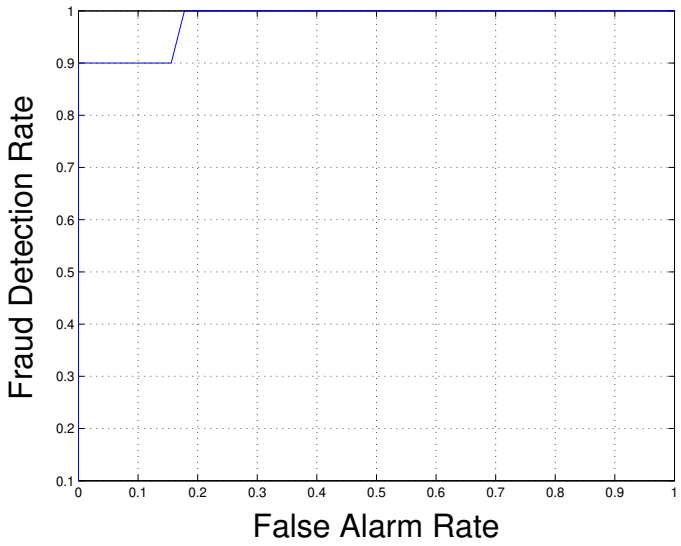
Rysunek 18: Wyniki aproksymacji dywergencji Kullbacka—Leiblera pomiędzy dwoma rozkładami probabilistycznymi LDA dla pięciu klas ukrytych

nikacyjnych reprezentujących konta użytkowników jako sekwencje aktywności (wykonywanych połączeń telefonicznych). Celem eksperymentów była naturalnie ocena skuteczności nowej metody wykrywania danych odstających w dziedzinie danych telekomunikacyjnych oraz weryfikacja i potwierdzenie słuszności i poprawności tez, postulatów i stwierdzeń w sferze teorii rozważanej i prezentowanej w pracy [9] przez autora niniejszego autoreferatu.

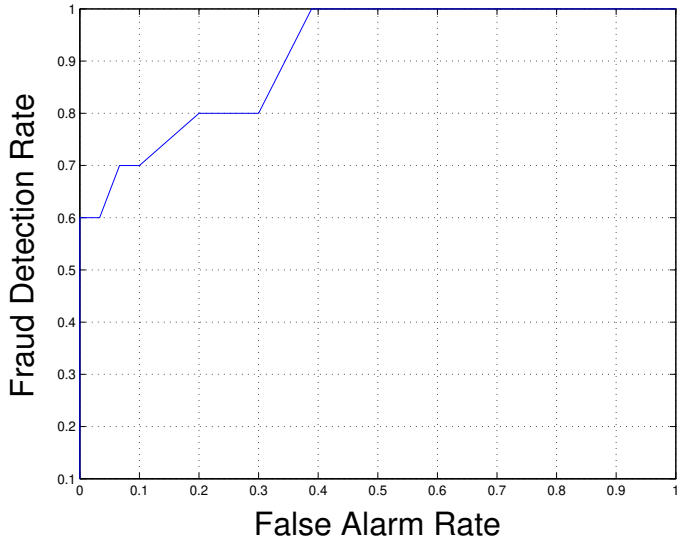
Fragmentaryczne wyniki badań eksperymentalnych w zakresie zagadnień omawianych w tym rozdziale przedstawione są na Rysunkach 19 i 20, na których ukazane są charakterystyki typu Receiver Operating Characteristics (ROCs) odpowiadające rozważnym metodom wykrywania danych odstających, sformułowanym w oparciu o modele probabilistyczne LDA oraz GMM. W tym konkretnym zadaniu analizy danych, dane odstające odpowiadały kontom użytkowników telekomunikacyjnych, budzącym podejrzenia nieuprawnionego korzystania z usług telekomunikacyjnych.

5 Podsumowanie osiągnięć naukowych

Podsumowując, można stwierdzić, iż cykl powiązanych tematycznie artykułów naukowych prezentowanych i omawianych w niniejszym autoreferacie dotyczy prac z dziedziny redukcji wymiarowości danach, klasteryzacji danych i wykrywania danych odstających. Przedstawione zostało dziewięć oryginalnych propozycji autora, które, zdaniem autora niniejszego autoreferatu,



Rysunek 19: Charakterystyka ROC dla proponowanej metody (AUROC=0.9833)



Rysunek 20: Charakterystyka ROC dla metody wykrywania danych odstających opartej na modelu GMM (AUROC=0.9111)

mogą być uznane za istotny wkład w rozwój dziedzin sztucznej inteligencji i uczenia maszynowego. Owe propozycje dotyczą metod redukcji wymiarowości, klasteryzacji danych i wykrywania danych odstających jako odrębnych dziedzin naukowych lub też powiązania tych trzech metodologii poprzez kontynuowanie procesu redukcji wymiarowości w postaci klasteryzacji lub wykrywania danych odstających następujących *a posteriori*. Niekiedy operacja klasteryzacji poprzedzała również samą redukcję wymiarowości, aby dostarczyć cenne i przydatne informacje wstępne służące następnie udoskonalaniu działania danej techniki redukcji wymiarowości danych, tak, jak to było w pracach [2, 1].

Redukcja wymiarowości danych rozważna była jako narzędzie do generowania graficznej ilustracji danych, innymi słowy, obrazu danych. Mówić zatem można wówczas o wizualizacji danych, która przekształca dane do postaci zbioru w przestrzeni jedno- dwu- lub trójwymiarowej (zazwyczaj dwu-), a zatem możliwej do przedstawienia wizualnego na ekranie i ułatwiającej analizę, ocenę, obserwację i interpretację nawet przez osoby niebędące ekspertami w dziedzinie uczenia maszynowego i redukcji wymiarowości danych.

Wszystkie proponowane przez autora metody i rozwiązania cechuje wyraźnie zaznaczony i zaakcentowany wkład oryginalny, jak również czytelne, przejrzyste, logiczne i przekonujące uzasadnienie teoretyczne leżące u podstaw wprowadzanych innowacji. Tak sformułowane teoretyczne uzasadnienie określa motywację autora podczas opracowywania nowych wersji rozważanych metod i algorytmów analizy danych, a także wytycza precyzyjny kierunek postępowania badawczego, tłumaczy konkretne wybory, tezy i decyzje i stwarza podstawy dla potwierdzenia słuszności powziętych kroków badawczych w sferze teoretycznej.

Prace autora skupione były, zarówno na wprowadzaniu modyfikacji, rozszerzeń i udoskonalań rozważanych i omawianych metod i algorytmów redukcji wymiarowości danych oraz klasteryzacji danych, jak i na badaniach o charakterze praktycznym z nowatorskimi elementami, jak w przypadku prac [8, 7, 9].

Badania eksperymentalne przygotowywane i przeprowadzane były w taki sposób, aby możliwie jak najszerzej i jak najbardziej gruntownie zweryfikować i ocenić skuteczność i przydatność proponowanych rozwiązań. A zatem wykorzystywane były zbiory danych różniące się w znacznym stopniu rozmiarem (liczbą instancji danych) oraz wymiarowością (liczbą wymiarów danych), a także naturą, typem i charakterem danych (dane tekstowe, medyczne (reprezentujące rytmy ludzkiego serca EKG), dźwiękowe (muzyczne i głosowe), dane reprezentujące aktywności robota mobilnego wykonującego zadane czynności, dane telekomunikacyjne, giełdowe, bankowe (dotyczące kart kredytowych), internetowe, dane dotyczące zużycia energii elektrycz-

nej w przykładowym gospodarstwie domowym). Badania eksperymentalne zaprojektowane w ten sposób stwarzały możliwość weryfikacji skalowalności proponowanych podejść, to znaczy zdolności do poprawnego i skutecznego funkcjonowania i działania w różnorodnych środowiskach i scenariuszach testowych i eksperymentalnych. W celu potwierdzenia cechy skalowalności wprowadzanych technik wykorzystywane były właśnie tak różnorodne pod wieloma względami zbiory danych, jak to było opisane wcześniej w tym rozdziale.

Literatura

[1] D. Olszewski, A clustering-based adaptive Neighborhood Retrieval Visualizer, Neural Networks 140 (2021) 247–260.

[2] D. Olszewski, A data-scattering-preserving adaptive self-organizing map, Engineering Applications of Artificial Intelligence 105 (2021) 104420.

[3] D. Olszewski, J. Kacprzyk, S. Zadrożny, An Improved Adaptive Self-Organizing Map, in: G. de Tré, P. Grzegorzewski, J. Kacprzyk, J. W. Owsiński, W. Penczek, S. Zadrożny (Eds.), Challenging Problems and Solutions in Intelligent Systems, Vol. 634 of Studies in Computational Intelligence, Springer, Cham, 2016, pp. 75–102.

[4] D. Olszewski, An asymmetric topology-preserving Neighborhood Retrieval Visualizer, Expert Systems with Applications 225 (2023) 120175.

[5] D. Olszewski, B. Šter, Asymmetric Clustering Using the Alpha–Beta Divergence, Pattern Recognition 47 (5) (2014) 2031–2041.

[6] D. Olszewski, Asymmetric k-Means Clustering of the Asymmetric Self-Organizing Map, Neural Processing Letters 43 (2016) 231–253.

[7] D. Olszewski, J. Kacprzyk, S. Zadrożny, Time Series Visualization Using Asymmetric Self-Organizing Map, in: M. Tomassini, A. Antonioni, F. Daolio, P. Buesser (Eds.), Adaptive and Natural Computing Algorithms, Vol. 7824 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 40–49.

[8] D. Olszewski, Fraud Detection Using Self-Organizing Map Visualizing the User Profiles, Knowledge-Based Systems 70 (2014) 324–334.

[9] D. Olszewski, A Probabilistic Approach to Fraud Detection in Telecommunications, Knowledge-Based Systems 26 (2012) 246–258.

[10] G. Hinton, S. T. Roweis, Stochastic Neighbor Embedding, Advances in Neural Information Processing Systems 14 (2002) 833–840.

[11] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization, Journal of Machine Learning Research 11 (2010) 451–490.

[12] T. Kohonen, Self-Organized Formation of Topologically Correct Feature Maps, Biological Cybernetics 43 (1) (1982) 59–69.

[13] T. Kohonen, The Self-Organizing Map, in: Proceedings of the IEEE, Vol. 28, 1990, pp. 1464–1480.

[14] C. von der Malsburg, Self-Organization of Orientation Sensitive Cells in the Striate Cortex, Kybernetik 14 (1973) 85–100.

[15] M. Martín-Merino, A. Muñoz, Visualizing Asymmetric Proximities with SOM and MDS Models, Neurocomputing 63 (2005) 171–192.

[16] S. Kullback, R. A. Leibler, On Information and Sufficiency, The Annals of Mathematical Statistics 22 (1) (1951) 79–86.

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

Część III

1 Informacja o wykazywaniu się istotną aktywnością naukową albo artystyczną realizowaną w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej

- Współpraca z Uniwersytetem w Oviedo w Hiszpanii w okresie od stycznia 2021 do chwili obecnej w ramach projektu "Time Series Classification Using Deep Learning" koordynowanego przez zespół badawczy w Uniwersytecie w Oviedo. Kierownik zespołu badawczego w Uniwersytecie w Oviedo: profesor José Ramón Villar Flecha.

- Współpraca z Uniwersytetem w Lublanie na Słowenii w okresie od czerwca 2012 do stycznia 2014 w ramach wymiany bilateralnej. W trakcie okresu tej współpracy powstała publikacja [5] – wspólna praca z profesorem Branko Šterem z Uniwersytetu w Lublanie.

- Zatrudnienie w Instytucie Badań Systemowych Polskiej Akademii Nauk (Pracownia Systemów Inteligentnych) w okresie od października 2014 do maja 2015 w ramach projektu "International PhD Projects in Intelligent Computing" finansowanego przez Fundację na Rzecz Nauki Polskiej. Projekt był też finansowany przez Unię Europejską w ramach programu "Innovative Economy Operational Programme 2007-2013 and European Regional Development Fund".

- Współpraca z Instytutem Podstaw Informatyki Polskiej Akademii Nauk w okresie od sierpnia 2012 do kwietnia 2013 w ramach projektu "Information Technologies: Research and Their Interdisciplinary Applications" of the Human Capital Operational Programme współfinansowany z Europejskiego Funduszu Socjalnego.

Część IV

1 Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę lub sztukę

W zakresie pracy dydaktycznej, wnioskodawca poświęcił szczególną uwagę obszarom sztucznej inteligencji, uczenia maszynowego oraz programowania wielowątkowego.

W dziedzinie programowania wielowątkowego, wnioskodawca przygotował oryginalny, autorski program przedmiotów dydaktycznych „Programowanie współbieżne i obiektowe" oraz „Programowanie współbieżne w języku Java", który uznaje i prezentuje w niniejszym autoreferacie jako swoje szczególne osiągnięcie dydaktyczne. Językiem programowania, wybranym przez wnioskodawcę jako przedmiot analizy implementacji rozwiązań wielowątkowych w programowaniu, był język Java. Konkretne frameworki i konstrukcje programistyczne rozważane w ramach tychże zajęć dydaktycznych to: framework high-level concurrency (wielowątkowość wysokopoziomowa), framework Fork-Join, framework operacje agregacyjne (ang. aggregate operations) oraz zagadnienie optymalizacji wykorzystania rzeczywistej mocy obliczenio-

wej komputera (w rozumieniu liczby dostępnych jednostek obliczeniowych CPU i GPU) przez program w języku Java.

Przedmioty dydaktyczne prowadzone w całym okresie pracy w Politechnice Warszawskiej, z zaznaczeniem tych, których program był autorskim wkładem prowadzącego:

- Artificial Intelligence (przedmiot w języku angielskim, autorski program opracowany samodzielnie przez prowadzącego), 3. semestr studiów magisterskich, kierunki: Informatyka, Automatyka i Robotyka.
- Wizualizacja danych (autorski program opracowany samodzielnie przez prowadzącego), 5. semestr studiów inżynierskich, kierunek: Informatyka.
- Zaawansowane metody wizualizacji danych (autorski program opracowany samodzielnie przez prowadzącego), 3. semestr studiów magisterskich, kierunek: Informatyka.
- Programowanie współbieżne i obiektowe (autorski program opracowany samodzielnie przez prowadzącego), 1. semestr studiów magisterskich, kierunek: Informatyka.
- Programowanie współbieżne w języku Java (autorski program opracowany samodzielnie przez prowadzącego), 1. semestr studiów magisterskich, kierunek: Automatyka i Robotyka.
- Testowanie oprogramowania, 5. semestr studiów inżynierskich, kierunek: Informatyka.
- Bezpieczeństwo oprogramowania (autorski program opracowany samodzielnie przez prowadzącego), 5. semestr studiów inżynierskich, kierunek: Informatyka.
- Podstawy inżynierii oprogramowania, 2. semestr studiów inżynierskich, kierunek: Informatyka.
- Języki i metodyka programowania (autorski program opracowany samodzielnie przez prowadzącego), 2. semestr studiów inżynierskich, kierunek: Automatyka i Robotyka.
- Computer Science (przedmiot w języku angielskim, autorski program opracowany samodzielnie przez prowadzącego), 1. i 2. semestr studiów inżynierskich, kierunek: Electrical Engineering (w języku angielskim).

W dziedzinie prowadzonych projektów studenckich, prac inżynierskich i magisterskich, wnioskodawca zajmował się współpracą ze studentami w zakresie następujących obszarów dydaktycznych: sztuczna inteligencja i uczenie maszynowe (redukcja wymiarowości danych, klasteryzacja, klasyfikacja, wykorzystania sztucznych sieci neuronowych), programowanie wielowątkowe oraz testowanie oprogramowania.

Wnioskodawca był promotorem następujących prac dyplomowych inżynierskich oraz magisterskich:

2 Prace dyplomowe magisterskie

Dyplomant(ka)	Tytuł pracy	Rok obrony
Piotr Woś	Analiza porównawcza koncepcji obsługi błędów logiki biznesowej w obszarze programowania objętym paradygmatem funkcyjnym	2023
Mateusz Sykut	Analiza i porównanie dwóch wybranych narzędzi automatycznego testowania aplikacji webowych: Selenium oraz TestCafe	2023
Karolina Wójcik	Opracowanie algorytmu optymalnego wyboru hulajnogi na podstawie lokalizacji użytkownika	2022
Kamil Kucharski	Analiza wybranych algorytmów optymalizacji w kontekście aplikacji do zarządzania obowiązkami domowymi	2022
Olga Sałagacka	Analiza porównawcza wybranych metod rozpoznawania tekstu na obrazach cyfrowych	2022
Paweł Talaga	Badanie wykorzystania sieci Generative Adversarial Network (GAN) w celu zwiększania rozdzielczości obrazów cyfrowych	2022
Ewa Kulesza	Analiza i porównanie dwóch wybranych narzędzi testowania oprogramowania opierających się na strategii testowania Behavior-Driven Development	2020

3 Prace dyplomowe inżynierskie

Dyplomant(ci)	Tytuł pracy	Rok obrony
Jakub Bączek	Rozproszony system przeprowadzający testy obciążeniowe aplikacji webowej typu REST	2020
Przemysław Sobolewski Krzysztof Gustalik	Aplikacja mobilna w języku C++ analizująca i przetwarzająca sygnały dźwiękowe w celu wspomagania osób z niepełnosprawnością słuchu	2020
Mateusz Sykut	Aplikacja webowa w języku Python wspomagająca proces optymalnego planowania podróży	2019
Do Nguyen Tien	Przewidywanie cen akcji z wykorzystaniem rekurencyjnej sieci neuronowej typu Long Short-Term Memory	2019
Dariusz Porowski	Implementacja algorytmu ewolucyjnego optymalizującego ruch robota w przestrzeni dwuwymiarowej	2018
Paweł Podgórski Mateusz Staszków	Zastosowanie wybranej metody rozpoznawania obrazów w aplikacji mobilnej do oceny składu biochemicznego produktów gastronomicznych	2018
Michał Dąbrowski	Wieloplatformowa aplikacja mobilna stworzona z wykorzystaniem technologii React-Native, której zadaniem jest wspomaganie procesu pakowania przed wyjazdem	2018

4 Aktywność o charakterze organizacyjnym:

- Współorganizowanie konferencji naukowej International Conference on Hybrid Artificial Intelligence Systems (HAIS) 2023.
- Współorganizowanie konferencji naukowej International Conference on Hybrid Artificial Intelligence Systems (HAIS) 2021.
- Współorganizowanie konferencji naukowej International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO) 2021.
- Pełnienie funkcji sekretarza seminarium Zakładu Sterowania w Instytucie Sterowania i Elektroniki Przemysłowej na Wydziale Elektrycznym Politechniki Warszawskiej.

5 Dodatkowe informacje

Przyznane nagrody i wyróżnienia

- Nagroda Best Paper za najlepszą publikację naukową w roku 2021 przyznana w ramach projektu „Inicjatywa Doskonałości – Uczelnia Badawcza (IDUB)" za artykuł pod tytułem "Clustering-based adaptive Neighborhood Retrieval Visualizer" opublikowany w czasopiśmie naukowym Neural Networks (Impact Factor JCR 2022: 9.657, 200 punktów MEiN).
- Certyfikat programistyczny "Sun Certified Programmer for the Java Platform 6.0 SE" wydany przez firmę Sun Microsystems.
- Nagroda trzeciego stopnia przyznana przez Politechnikę Warszawską za osiągnięcia naukowe w latach 2011-2012. Nagroda przyznana w październiku 2013.

.....
Podpis wnioskodawcy