Dokumentacja końcowa - MOW

Artur M. Brodzki

Adam Małkowski

1. Założenia wstępne

1.1 Wstęp – interpretacja tematu projektu

W ramach projektu dokonamy porównania jakości grupowania uzyskanego przy pomocy samo-organizujących się map (sieci SOM) oraz bardziej klasycznych metod grupowania: metody k-średnich, metody DBSCAN oraz metody najdalszego sąsiedztwa (ang. *complete-linkage clustering*). Przy wykorzystaniu dostępnych powszechnie zbiorów danych zamierzamy wyznaczyć klastry metodami klasycznymi oraz za pomocą sieci SOM, a następnie porównać jakość tych klastrów z użyciem kilku powszechnie wykorzystywanych w tym celu metryk.

1.2 Wykorzystywane algorytmy

Samo-organizujące się mapy zrealizujemy w środowisku R za pomocą pakietów som, kohonen. W problemie grupowania ważną kwestią jest przyjęta definicja klastra. Nie ma jednej definicji klastra uznanej za standardową, a różne algorytmy przyjmują na swój użytek różne definicje:

1. Sieci SOM mogą grupować na dwa sposoby. Ponieważ każdy neuron sieci SOM odpowiada pewnemu zapamiętanemu wzorcowi danych, można założyć, że każdy taki neuron - wzorzec definiuje pewien wytworzony przez sieć SOM klaster danych. Ponieważ liczba neuronów - klastrów jest z góry określona, takie podejście charakteryzuje się znacznym podobieństwem do algorytmu k-średnich (w przypadku granicznym, gdy promień sąsiedztwa sieci SOM jest równy 0, obie metody są równoważne).
   Drugie podejście do grupowania na sieciach SOM wykorzystuje fakt, że neurony tej sieci są zanurzone w przestrzeni euklidesowej (zazwyczaj dwuwymiarowej). Dynamika sieci wymusza rozmieszczanie podobnych sobie wzorców blisko siebie. Dzięki temu, same neurony - wzorce mogą być obiektem grupowania. Na takich neuronach można uruchomić jeden z klasycznych algorytmów grupowania i porównać jakość uzyskanych klastrów. Taka procedura pozwala też zweryfikować skuteczność sieci SOM do odnajdowania powiązanych ze sobą wzorców w danych.
2. Podejście oparte o centroidy - klaster jest geometrycznym środkiem zbioru bliskich sobie punktów. W celu analizy podejścia opartego o centroidy wykorzystamy algorytm k-średnich.
3. Podejście hierarchiczne - klastry są wyznaczane na różnych poziomach jako złączone elementy z poziomu niższego. Do analizy podejścia hierarchicznego wykorzystamy algorytm najdalszego sąsiedztwa.
4. Podejście gęstościowe - klastry to zagęszczone skupiska położonych blisko siebie punktów. Granice między klastrami leżą w obszarach o mniejszej gęstości punktów. Analizę podejścia gęstościowego przeprowadzimy na algorytmie DBSCAN.

1.3 Przygotowanie danych

Do analizy metod grupowania wykorzystamy 3 spośród zbiorów danych dostępnych powszechnie zbiorów danych:

1. Car Evaluation Data Set - zbiór liczący 1728 elementów opisanych parametrami słownikowymi http://archive.ics.uci.edu/ml/datasets/Car+Evaluation.
   Zbiór posiada parametr opisujący klasę samochodu w 4 stopniowej skali – niska, średnia, wysoka oraz bardzo wysoka, który zostanie wykorzystany jako klasyfikacja wzorcowa.
2. Iris Data Set - zbiór liczący 150 elementów opisanych parametrami rzeczywistymi https://archive.ics.uci.edu/ml/datasets/iris.
   Zbiór posiada parametr określający gatunek, do którego należy dany okaz rośliny, który zostanie wykorzystany jako klasyfikacja wzorcowa.
3. Adult Data Set - zbiór liczący 48 842 elementy opisane parametrami słownikowymi oraz całkowitoliczbowymi http://archive.ics.uci.edu/ml/datasets/Adult
   Jako klasyfikacja wzorcowa wykorzystamy informacje dotyczące dochodu – roczny dochód wyższy bądź niższy od $50.000. Istotna jest dysproporcja elementów klas – występuje między nimi stosunek rzędu 1:5.

Zbiory zostały dobrane tak, aby rozwiązywać problem grupowania korzystając z różnej ilości danych oraz korzystać z różnych rodzajów parametrów (słownikowych, całkowitoliczbowych, rzeczywistych).

Wykorzystywane do grupowania algorytmy opierają się na pojęciu odległości pomiędzy punktami i jest to najczęściej odległość euklidesowa. Wymaga to, by składowe analizowanych punktów były typu numerycznego. W rzeczywistych zbiorach danych wiele parametrów jest typu słownikowego. Najprościej jest kodować typ binarny: wartość $true$ jako 1 i wartość $false$ jako 0. Takie kodowanie jest dobrze określone - dla dowolnych $x, y \in \{0,1\}$: $(x-y)^2 \in \{0,1\}$ i posiada sensowną interpretację - wartość składnika $(x-y)^2$jest równa 1 dla $x \neq y$ i równa 0 dla $x = y$. W przypadku parametrów słownikowych o liczbie możliwych wartości $n \geq 2$ dokonamy zamiany ich na n parametrów typu binarnego.

Oprócz tego, wartości w danych zostaną zestandaryzowane korzystając z funkcji *scale* pakietu R tak, aby były podobnego rzędu wielkości – dotyczy to w szczególności zbioru danych Adult.

1.4 Metody przeprowadzania eksperymentów

W celu porównania jakości grupowania sieci SOM i algorytmów klasycznych, przeprowadzimy następująca procedurę:

1. Na wybranym zbiorze danych wyznaczymy klastry metodami klasycznymi: k-średnich, DBSCAN oraz najdalszego sąsiedztwa.
2. Tego samego zbioru danych użyjemy do nauczenia sieci SOM.
3. Mając wyznaczone klastry danych, możemy wyznaczyć i porównać ich jakość. Wykorzystamy dwa sposoby ewaluacji jakości klastra: wewnętrzny i zewnętrzny.

1. Ewaluacja wewnętrzna nie korzysta w żaden sposób z zewnętrznej wzorcowej klasyfikacji i mierzy jedynie jakość grupowania jako podziału spełniającego następujący warunek: elementy należące do tego samego klastra powinny być do siebie nawzajem dużo bardziej podobne niż elementy należące do różnych klastrów. W celu pomiaru jakości klastrów wykorzystamy dwa powszechnie używane w tym celu wskaźniki:
    - indeksu Dunna – posiada wartości między 0 a nieskończoność i powinien być minimalizowany – jest mały, gdy grupy są skoncentrowane oraz dobrze od siebie odseparowane.
    - indeksu Daviesa – Bouldina – posiada wartości między 0 a nieskończonością i powinien być maksymalizowany – jest duży, gdy grupy są skoncentrowane oraz dobrze od siebie odseparowane
2. Metody zewnętrzne korzystają z zadanej - nomen omen - z zewnątrz klasyfikacji uznanej za wzorcową. Dla tych zbiorów, w celu wyznaczenia zgodności uzyskanej klasyfikacji ze wzorcową, wykorzystamy dwa wskaźniki powszechnie wykorzystywane do oceniania jakości testów statystycznych:
    - indeks Randa – posiada wartości pomiędzy zero a jeden, powinien być maksymalizowany, obliczany ze wzoru $\frac{TP+TN}{TP+TN+FP+FN}$
    - F – indeks – posiada wartości pomiędzy zero a jeden, powinien być maksymalizowany, obliczany ze wzoru $\frac{2TP}{2TP+FP+FN}$

   O ile klastry dla metod klasycznych dane są jednoznacznie, to dla sieci SOM będziemy grupować na dwa sposoby, opisane w sekcji 1.2: traktując każdy neuron jako osobny klaster, lub na nauczonej sieci SOM wykonać klasyczne grupowanie przy użyciu metody najdalszego sąsiedztwa.
4. Powyższe operacje powtórzymy dla wybranych przez nas zbiorów danych oraz dla różnych zestawów parametrów wykorzystywanych algorytmów:
    1. liczby neuronów w sieci SOM i promienia sąsiedztwa
    2. liczby klastrów w przypadku grupowania neuronów sieci SOM
    3. liczby klastrów *k* w algorytmie k-średnich
    4. poziomu w metodzie najdalszego sąsiedztwa
    5. maksymalnego promienia sąsiedztwa (*eps*) oraz minimalnej liczby punktów do utworzenia grupy (*minPts*) w algorytmie DBSCAN

Tak zaplanowany eksperyment pozwala na porównanie różniących się od siebie paradygmatów grupowania za pomocą spójnych metryk oceny jakości. Powtórzenie procedury dla wybranych zbiorów danych i różnych parametrów pozwoli porównać jakość grupowania w różnych sytuacjach badawczych. Uzyskane wnioski i korelacje opiszemy szczegółowo w dokumentacji końcowej projektu.

2. Wyniki eksperymentów:

2.1 Zbiór danych IRIS

2.1.1 Algorytm K-średnich

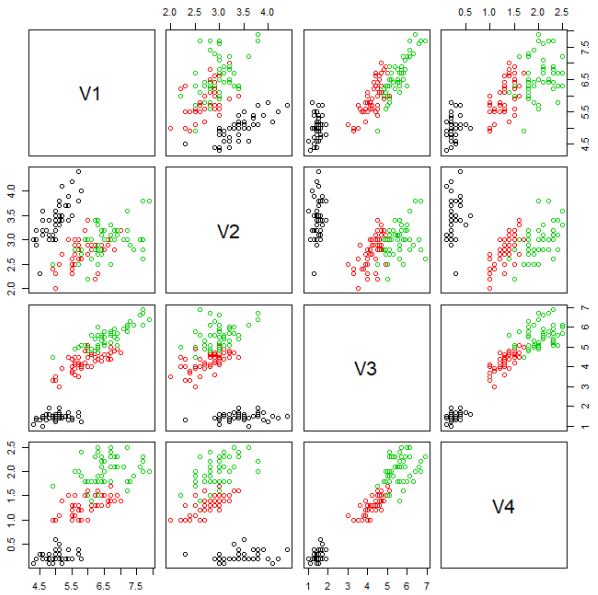
n – liczba grup	Davies - Bouldin	Dunn	Rand	F-indeks
1	0,00	∞	0,33	0,49
2	0,40	0,08	0,76	0,73
3	0,95	0,05	0,72	0,65
4	0,57	0,05	0,82	0,71
5	0,89	0,06	0,79	0,61
6	0,78	0,08	0,77	0,55
7	0,95	0,05	0,80	0,65
8	0,57	0,12	0,77	0,53
9	0,69	0,10	0,78	0,51
10	0,99	0,10	0,78	0,49

Tabela 1 Algorytm K-średnich, IRIS

Wskaźnik Daviesa Bouldina dla zbioru danych IRIS i parametru n z przedziału 1:10 mówi niewiele – oczekujemy jego możliwie najniższej wartości, która oznaczała by, że powstałe grupy są silnie skupione oraz dobrze odseparowane (dla przykładu z jedną grupą wynosi oczywiście 0) – dla kolejnych iteracji wartość ta jest niemonotoniczna i zmniejsza się lub zwiększa około dwukrotnie.

Indeks Dunna dla przytoczonych danych osiąga najmniejsze wartości dla 3 i 4 grup – jest to o tyle dziwne, że zbiór danych IRIS jest naturalnie podzielony na 3 grupy.

W celu dokonania ewaluacji zewnętrznej wykorzystujemy wzorcową klasyfikację zbioru Iris na trzy rodzaje roślin.



Rysunek 1 Referencyjny podział na grupy zbioru IRIS

Zgodnie z zamieszczonym obrazkiem ilustrującym referencyjny (dziedzinowy) podział przykładów na grupy można zaobserwować, że o ile grupa czarna jest dobrze oddzielona od pozostałych, o tyle grupy czerwona i zielona nachodzą na siebie – w związku z tym, metody grupowania oparte na odległości mają trudność z znalezieniem w zbiorze grupowania zgodne z referencyjnym. Jakość tej korelacji może być opisywana różnymi wskaźnikami – między innymi użytymi w tej pracy wskaźnikami Rand i F-indeks.

Wskaźnik Randa jest na poziomie 70-80% procent i wynika to z poprawnego zaklasyfikowania punktów klasy „czarnej" do jednej grupy niezależnie od liczby grup oraz uwzględnienia faktu, że wraz z wzrostem liczby grup klasa „zielono-czerwona" jest dzielona na coraz więcej grup – coraz mniejszych, ale niektóre zawierające się w większości w jednej z grup referencyjnych.

Wskaźnik F dla algorytmu k-średnich i zbioru danych IRIS przyjmuje najwyższe wartości w okolicy n = 3 co jest spowodowany tym, że wraz z wzrostem n spada liczba punktów trafnie zaklasyfikowanych (grupy stają się coraz mniejsze).

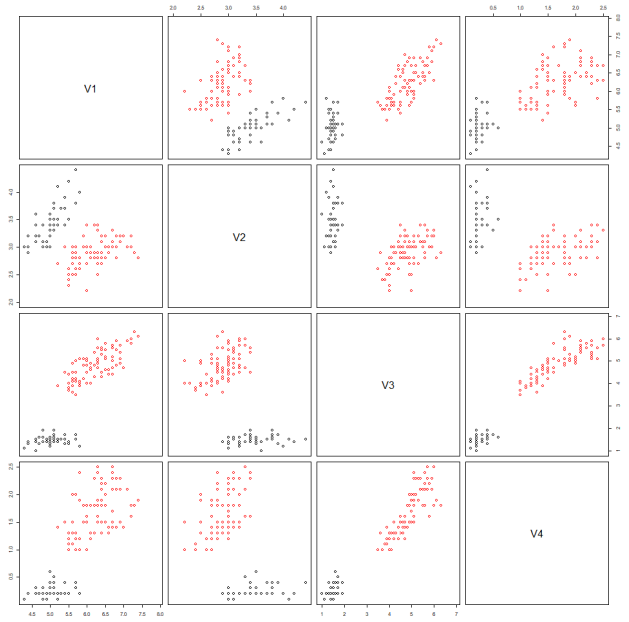
2.1.2 Algorytm DBSCAN

Lp.	Eps	Minimalna liczba punktów	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	0.25	1	0,49	0,19	0,73	0,32	81
2	0.5	1	0,37	0,15	0,77	0,70	12
3	0.75	1	0,36	0,17	0,78	0,74	3
4	1	1	0,38	0,34	0,78	0,75	2
5	1.25	1	0,38	0,34	0,78	0,75	2
6	1.5	1	0,38	0,34	0,78	0,75	2
7	1.75	1	0,00	∞	0,33	0,49	1
8	2	1	0,00	∞	0,33	0,49	1
9	0.25	5	2,90	0,03	0,56	0,52	3
10	0.5	5	10,53	0,08	0,77	0,70	2
11	0.75	5	0,48	0,17	0,78	0,74	2
12	1	5	0,38	0,34	0,78	0,75	2
13	1.25	5	0,38	0,34	0,78	0,75	2
14	1.5	5	0,38	0,34	0,78	0,75	2
15	1.75	5	0,00	∞	0,33	0,49	1
16	2	5	0,00	∞	0,33	0,49	1
17	0.25	10	0,67	0,02	0,46	0,49	2
18	0.5	10	5,38	0,04	0,79	0,69	2
19	0.75	10	1,42	0,08	0,78	0,73	2
20	1	10	0,38	0,34	0,78	0,75	2
21	1.25	10	0,38	0,34	0,78	0,75	2
22	1.5	10	0,38	0,34	0,78	0,75	2
23	1.75	10	0,00	∞	0,33	0,49	1
24	2	10	0,00	∞	0,33	0,49	1

Tabela 2 Algorytm DBSCAN, IRIS

Pomijając pierwszy przykład, w którym grupy są średnio dwuelementowe, podczas grupowania algorytmem DBSCAN często popada się w następujące skrajności – albo większość elementów trafia do szumu (istnienie szumu jest jedną z ciekawszych rzeczy wyróżniających algorytm)

w przypadku gdy minimalna liczba punktów nie jest niska, a *eps* jest niski albo tworzą się dwie duże grupy odpowiadające czarnej grupie referencyjnej oraz grupie zielono-czerwonej (w przypadkach w których grup jest 3, nadmiarowa grupa jest bardzo mało liczna). Sytuacje w których *eps* jest wysoki, a minimalna liczba punktów niska prowadzą do stworzenia jednej grupy zawierającej większość elementów. Indeks Daviesa Bouldina posiada dla większości eksperymentów niskie wartości. Wyjątkami w tej kwestii są podpunkty 9, 10, 18 i 19. Jest to spowodowane pominięciem elementów pomiędzy dużymi grupami referencyjnymi i oddaleniem od siebie powstałych grup.



Rysunek 2 Grupowanie zbioru danych IRIS nr. 19 algorytmem DBSCAN

Indeks Dunna osiąga maksymalną wartość dla przypadku z dwoma dużymi grupami (5, 6, 12, 13, 14, 20, 21, 22).

Metoda DBSCAN dla danych IRIS nie jest w stanie wykryć sensownie wszystkich trzech klas referencyjnych – najlepsze rozwiązania w kontekście wskaźników Randa i F to podział na dwie duże grupy – co ciekawe, daję on wyższa wartość wskaźnika F-indeks niż grupowanie z poprzedniej metody.

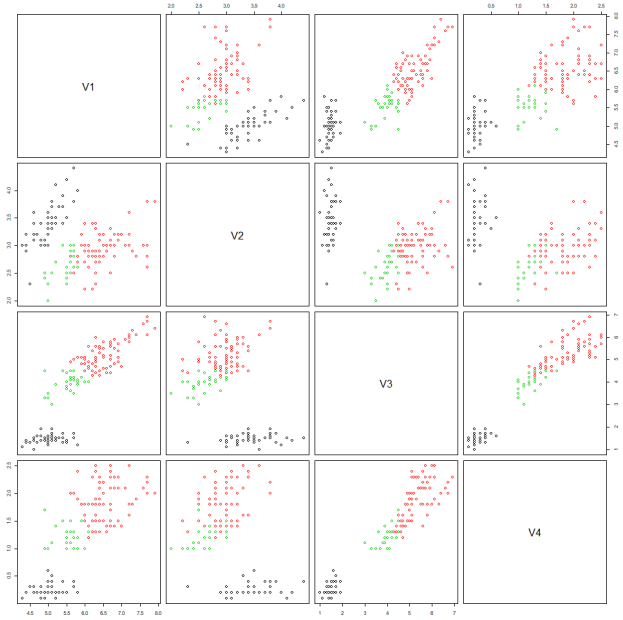
2.1.3 Algorytm najdalszego sąsiedztwa

n – liczba grup	Davies - Bouldin	Dunn	Rand	F-measure
1	0,00	∞	0,33	0,49
2	0,66	0,08	0,71	0,65
3	0,63	0,10	0,84	0,77
4	0,62	0,14	0,82	0,72
5	0,53	0,10	0,77	0,60
6	0,53	0,13	0,78	0,57
7	0,54	0,13	0,78	0,55
8	0,51	0,15	0,79	0,56
9	0,32	0,15	0,80	0,56
10	0,33	0,15	0,79	0,55

Tabela 3 Algorytm najdalszego sąsiedztwa, IRIS

Algorytm najdalszego sąsiedztwa podczas swojego działania łączy grupy, których maksymalna odległość pomiędzy dwoma punktami jest najmniejsza – w efekcie, minimalizuje ten parametr. Wskaźnik Daviesa - Bouldina systematycznie spada wraz z wzrostem n – jest to spowodowane pojawianiem się nowych grup będących bardzo blisko siebie, co za tym idzie, słabo separowanych. Wskaźnik Dunna maleje wraz ze spadkiem n.

Metoda najdalszego sąsiedztwa bardzo dobrze sprawdza się na tym zbiorze danych – osiąga najwyższe wartości współczynników Randa i F-indeks z dotychczas omawianych dla przypadku z trzema grupami, mimo dość niskich wartości Daviesa - Bouldina i Dunna.



Rysunek 3 Grupowanie zbioru danych IRIS nr. 3 algorytmem najdalszego sąsiedztwa

Wartości współczynnika oscylują w okolicach 80% - podobnie jak w przypadku k-średnich – jest to spowodowane koniecznością uwzględnienia w grupach wszystkich punktów co przy większych n powoduje wzrost poprawnie niezaklasyfikowanych przykładów. F-indeks zgodnie z oczekiwaniami osiąga największą wartość dla poprawnej liczby grup równej 3.

2.1.4 Algorytm SOM

Sieci neuronowe typu SOM można wykorzystywać do grupowania na dwa różne sposoby – pierwszy z nich interpretuje poszczególne neurony sieci jako reprezentantów poszczególnych grup. W przypadku wykorzystywanej implementacji rodzi to problem, gdyż sieć definiuje się jako prostokąt o określonej wysokości i szerokości – w efekcie tego, nie ma pełnej dowolności w wyborze liczby klastrów (oczywiście zawsze można stworzyć sieć wymiaru 1xN, jednakże w przypadku sieci SOM sąsiednie neurony wpływają na siebie i sieć 1xN posiada inne własności niż równoliczna sieć mająca kształt bliższy kwadratowi).

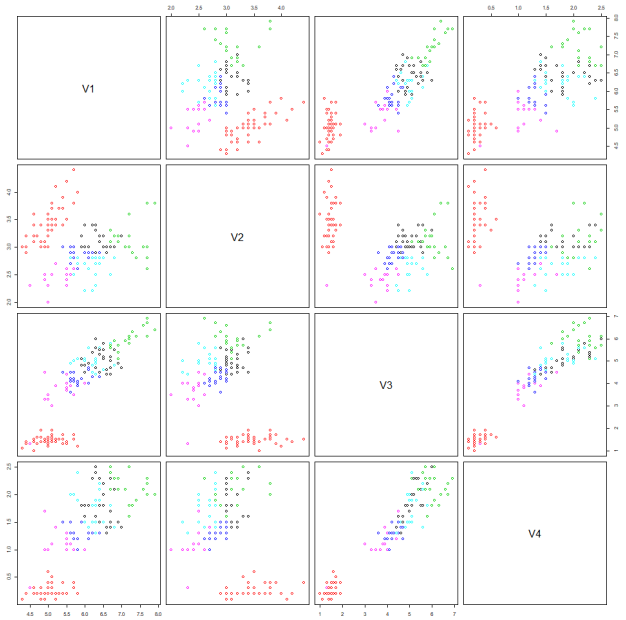
Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	1	2	0,384	0,339	0,776	0,746	2
2	1	3	0,335	0,045	0,828	0,738	3
3	1	4	0,629	0,045	0,777	0,630	4
4	2	2	0,999	0,042	0,814	0,690	4
5	2	3	0,860	0,049	0,815	0,648	6
6	2	4	0,481	0,063	0,768	0,509	8
7	3	3	0,733	0,062	0,793	0,557	9
8	3	4	0,600	0,040	0,760	0,443	12

Tabela 4 Algorytm SOM (neurony jako grupy), IRIS

Tabela z wynikami pokazuje, że najlepsze wartości współczynników występują dla sieci o wymiarach 1x3.

Indeks Daviesa Bouldina przyjmuje wyższe wartości dla bardziej kwadratowych sieci.

Indeks Dunna w podanych przykładach utrzymuje się na niskim poziomie – wynika to z nierównomiernego rozkładu analizowanych punktów. Większe sieci SOM rozdzielają naturalne grupy tak, że sąsiednie posiadają elementy bardzo blisko siebie co źle wpływa na wartość indeksu.



Rysunek 4 Grupowanie zbioru danych IRIS nr. 5 algorytmem SOM

Współczynniki ewaluacji zewnętrznej osiągają wartości podobne jak w przypadku metody najdalszego sąsiedztwa – współczynnik Randa utrzymuje się około 80%.

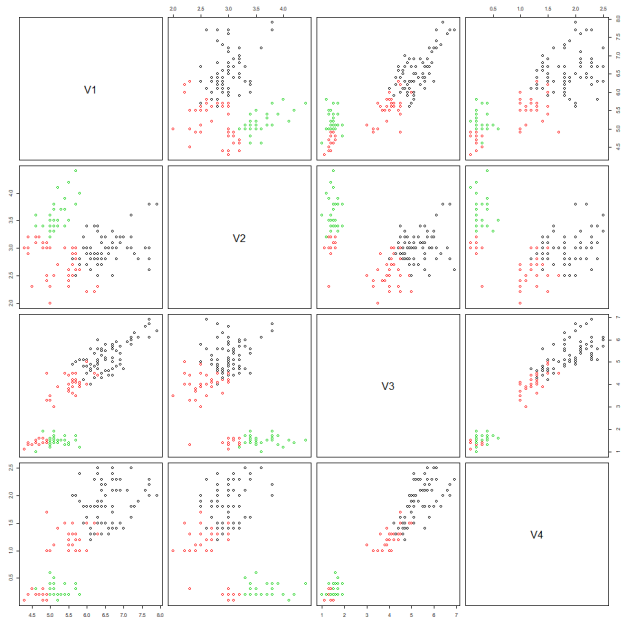
Drugim podejściem do grupowania korzystając z sieci SOM jest nauczenie sieci zawierającej znacznie więcej neuronów niż oczekiwanych grup, a następnie pogrupowanie samych neuronów jednym z klasycznych algorytmów.

2.1.5 Algorytm SOM + grupowanie

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	2	1	0,880	0,036	0,552	0,517	2
2	2	2	0,657	0,030	0,706	0,644	2
3	2	4	0,803	0,040	0,506	0,525	2
4	2	10	0,609	0,057	0,715	0,658	2
5	3	1	0,584	0,025	0,674	0,538	3
6	3	2	0,791	0,031	0,709	0,584	3
7	3	4	0,744	0,049	0,741	0,624	3
8	3	10	0,694	0,025	0,707	0,607	3
9	4	1	0,772	0,073	0,730	0,621	4
10	4	2	0,429	0,063	0,789	0,681	4
11	4	4	0,598	0,068	0,773	0,645	4
12	4	10	0,587	0,114	0,824	0,741	4
13	8	1	0,415	0,094	0,767	0,560	8
14	8	2	0,407	0,068	0,771	0,594	8
15	8	4	0,433	0,074	0,751	0,512	8
16	8	10	0,744	0,089	0,751	0,517	8

Tabela 5 Algorytm SOM + Algorytm najdalszego sąsiedztwa, IRIS

Przytoczone powyżej eksperymenty miały pokazać zależność wyniku od kształtu sieci neuronowej – kolejne czwórki przykładów opisują próbę osiągnięcia konkretnej liczby grup korzystając z różnych struktur sieci. Widoczne są różnice które potwierdzają wizualizacje grupowania – dodatkowo istotne są czynniki losowe – inicjalizacja sieci oraz kolejność podawania przykładów. Nie widać zaś jednoznacznej prawidłowości pomiędzy jakością grupowania wyrażoną w postaci wartości 4 analizowanych współczynników. Najbardziej interesująca grupa przykładów 5-8 (ponieważ liczba grup odpowiada grupowaniu referencyjnemu) nie posiada przedstawiciela cechowanego dobrymi (względem innych algorytmów) wartościami współczynników Rand i F-indeks.



Rysunek 5 Grupowanie zbioru danych IRIS nr. 7 algorytmami SOM + najdalszego sąsiedztwa

Niskie wartości są spowodowane małym dopasowaniem grupowania do grupowania referencyjnego – wynika to nie tylko z działania sieci SOM, a również z wykorzystanego algorytmu grupowania neuronów – metoda najdalszego sąsiedztwa, która w dotychczasowych testach sprawdziła się najlepiej grupuje elementy centralne otrzymane w wyniku działania SOM w sposób niedopasowany do zbioru danych.

W kontekście zbioru danych IRIS ciężko o wysunięcie wniosków mówiących o przewadze sieci neuronowych typu SOM względem klasycznych algorytmów grupowania. Zauważalnym jest przewaga metody wykorzystującej neurony sieci SOM bezpośrednio wobec metody, która następnie je grupuje.

2.2 Zbiór danych CAR Evaluation

2.2.1 Algorytm K-średnich

n – liczba grup	Davies - Bouldin	Dunn	Rand	F-indeks
1	0,00	∞	0,33	0,49
2	0,40	0,08	0,76	0,73
3	0,81	0,10	0,88	0,82
4	0,93	0,10	0,84	0,73
5	0,32	0,07	0,84	0,73
6	0,59	0,11	0,86	0,73
7	0,84	0,11	0,82	0,67
8	0,86	0,11	0,84	0,68
9	0,94	0,07	0,82	0,65
10	0,82	0,12	0,77	0,47

Tabela 6 Algorytm K-średnich, CAR Evaluation

Indeksy Daviesa Bouldina i Dunna zwracają nieintuicyjne wyniki – w obu nie widać monotoniczności ani najwyższe wartości nie trafiają na przykład 4 (zgodny z referencyjną liczbą klas).

Jakość grupowania w porównaniu z grupowaniem referencyjnym przyjmuje najwyższe wartości dla przykładów 3, 4 i 5. Maksymalna wartość wskaźnika RAND jest bliska 90% skuteczności dopasowania. Wraz ze wzrostem liczby grup maleją wskaźniki RAND i F-indeks co jest spowodowane zmniejszeniem się grup a co za tym idzie, gorszym dopasowaniem do grup referencyjnych.

2.2.2 Algorytm DBSCAN

Lp.	Eps	Minimalna liczba punktów	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	$\sqrt{2}$	10	0,00	∞	0,25	0,40	1
2	$\sqrt{2}$	30	2,72	0,41	0,47	0,35	3
3	$\sqrt{2}$	150	0,00	∞	0,25	0,40	0
4	$\sqrt{4}$	150	0,00	∞	0,25	0,40	1
5	$\sqrt{6}$	250	0,00	∞	0,25	0,40	1
6	$\sqrt{8}$	500	0,00	∞	0,25	0,40	1

Tabela 7 Algorytm DBSCAN, CAR Evaluation

Analizowany zbiór danych w wersji przed przekształceniem parametrów posiadał 7 atrybutów oraz klasę. Po przekształceniu posiada 21 atrybutów binarnych – liczba ta jest redundantna, przykład musi posiadać dokładnie jedną jedynkę wśród kolumn odpowiadających oryginalnym atrybutom. Efektem takiej postaci jest fakt, że wartości odległości między dwoma punktami to odpowiednio $\sqrt{n}\ dla\ n = 2, 4, 6, 8, 10, 12, 14$. Taki stan rzeczy uniemożliwia sensowne wykorzystywanie algorytmu DBSCAN – ciężko znaleźć wartości parametrów, dla których sytuacja jest inna niż

a) wszystkie elementy należą do szumu,
b) wszystkie elementy należą do grupy 1,
c) prawie każdy element jest w innej grupie.

W ramach eksperymentów jedynie dla parametrów $eps = \sqrt{2}$ i minimalnej liczby punktów równej 30 udało się znaleźć jakkolwiek sensowne grupowanie.

	Szum	1	2	3
Graniczne	341	166	25	78
Ziarna	0	1100	3	15
Łącznie	341	1266	28	93

Tabela 8 DBSCAN Liczba punktów = 1728 MinPts=30 eps=1.42

Wartości indeksów Daviesa - Bouldina i Dunna dla tego grupowania pokazują, że grupy są słabo izolowane od siebie. Również wyniki indeksów porównujących nie napawają optymizmem – wyniki poniżej 0.5 (między innymi dlatego, że aż 341 punktów nie trafiło do żadnej z grup a aż 1266 elementów jest w jednej grupie).

2.2.3 Algorytm najdalszego sąsiedztwa

n – liczba grup	Davies - Bouldin	Dunn	Rand	F-indeks
1	0,00	∞	0,33	0,49
2	0,66	0,08	0,71	0,65
3	0,63	0,10	0,84	0,77
4	0,62	0,14	0,82	0,72
5	0,53	0,10	0,77	0,60
6	0,53	0,13	0,78	0,57
7	0,54	0,13	0,78	0,55
8	0,51	0,15	0,79	0,56
9	0,32	0,15	0,80	0,56
10	0,33	0,15	0,79	0,55

Tabela 9 Algorytm najdalszego sąsiedztwa, CAR Evaluation

Wyniki otrzymane algorytmem najdalszego sąsiedztwa są podobne do tych otrzymanych przy wykorzystaniu k-średnich – pomijając F-indeks większość wartości jest minimalnie lepszych dla najdalszego sąsiedztwa. Indeks Rand na poziomie 80% (co jest głównie zasługą poprawne niesklasyfikowanych). Najlepsze grupowanie tą metodą jest gorsze niż najlepsze grupowanie metodą k-średnich – jest to związane z opisaną przy okazji DBSCAN dyskretyzacją w przestrzeni możliwych odległości – algorytm najdalszego sąsiedztwa łączy zbiory na podstawie skrajnych odległości a w tym zbiorze każda wartość wiąże się z wieloma parami elementów. Znacznie większe możliwości znalezienia sensownego grupowania daje algorytm k-średnich w którym odległości są liczone od centroidów, one zaś mogą posiadać wartości różne niż 0 i 1, gdyż powstają w wyniku uśrednienia elementów grupy.

2.2.4 Algorytm SOM

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	1	2	2,42	0,41	0,47	0,36	2
2	1	3	2,41	0,41	0,59	0,30	3
3	1	4	2,35	0,41	0,62	0,28	4
4	2	2	2,28	0,41	0,62	0,27	4
5	2	3	2,36	0,45	0,66	0,22	6
6	2	4	1,99	0,45	0,67	0,21	8
7	3	3	2,10	0,45	0,68	0,20	9
8	3	4	2,19	0,45	0,71	0,15	12

Tabela 10 Algorytm SOM (neurony jako grupy), podejście ze strukturą heksagonalną, CAR Evaluation

Grupowanie korzystające z sieci SOM (o strukturze heksagonalnej, tak samo dla jak zbioru IRIS) daje bardzo słabe rezultaty – mało zależne od wymiarów sieci (większość indeksów utrzymuje się na podobnych, niskich wartościach). Rozczarowujące są słabe wyniki dla sieci 4 neuronowych które, mogło by się wydawać, powinny najlepiej oddawać grupy referencyjne – zaś indeksy Randa rzędu 60% oraz F-indeks 30% nie napawają optymizmem.

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	1	2	2,42	0,41	0,47	0,36	2
2	1	3	2,15	0,41	0,48	0,35	3
3	1	4	2,12	0,41	0,63	0,26	4
4	2	2	2,52	0,41	0,60	0,29	4
5	2	3	2,10	0,41	0,66	0,23	6
6	2	4	1,99	0,45	0,67	0,21	8
7	3	3	2,20	0,45	0,68	0,20	9
8	3	4	1,63	0,45	0,71	0,14	12

Tabela 11 Algorytm SOM (neurony jako grupy), podejście ze strukturą prostkątną, CAR Evaluation

Przetestowana została również siec o strukturze kwadratowej, z identycznymi parametrami jak poprzednia. Zmiana ta nie wpłynęła istotne nie wyniki – potencjalne zmiany mógłby wynikać z innej odległości pomiędzy neuronami a co za tym idzie innym propagowaniem się zależności – jednakże w tym przypadku wyniki działania sieci są słabe i zmiana ta wprowadza jedynie „szum” we współczynnikach.

2.2.5 Algorytm SOM + grupowanie

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup
1	10	40	2,08	0,41	0,29	0,39	2
2	15	15	2,08	0,41	0,29	0,39	2
3	20	20	2,08	0,41	0,29	0,39	2
4	10	40	2,04	0,41	0,33	0,39	2
5	15	15	2,04	0,41	0,33	0,39	3
6	20	20	2,36	0,41	0,33	0,39	3
7	10	40	2,19	0,45	0,49	0,35	3
8	15	15	2,06	0,41	0,49	0,35	3
9	20	20	2,11	0,45	0,49	0,35	4
10	10	40	2,03	0,45	0,50	0,34	4
11	15	15	2,12	0,45	0,61	0,28	4
12	20	20	1,91	0,41	0,57	0,31	4
13	10	40	1,74	0,45	0,67	0,21	8
14	15	15	2,09	0,45	0,67	0,21	8
15	20	20	2,12	0,45	0,68	0,20	8

Tabela 12 Algorytm SOM + Algorytm najdalszego sąsiedztwa, CAR Evaluation

W powyższych testach sprawdzano, jak poradzi sobie połączenie SOM z algorytmem najdalszego sąsiedztwa redukującego liczbę grup do 4.

Wyniki, jeżeli chodzi o indeksy analizujące geometryczną jakość grupowania, się nie zmieniły – oba indeksy utrzymują się na podobnym poziomie – oznacza to, że od strony geometrycznej powstałe grupy wyglądają podobnie.

Wskaźniki Randa oraz F-indeks przyjmują jeszcze niższe wartości niż w poprzednim podpunkcie. Co ciekawe, indeks Randa rośnie wraz z wzrostem liczb grup co wynika z większej liczby elementów poprawnie niezaklasyfikowanych – elementu tego nie uwzględnia F-indeks czego efektem jest monotoniczny spadek wartości wraz z liczbą grup.

Wnioskami dla tego zbioru danych są:

- trudność grupowania danych o samych atrybutach kategorycznych, ze względu na dyskretyzację odległości euklidesowej
- sieci SOM gorzej radzą sobie z tego typu danymi niż algorytm k-średnich
- algorytm DBSCAN nie radzi sobie w tego typu sytuacjach, z uwagi na zbyt równomierne rozmieszczenie punktów: albo każdy obiekt tworzy osobną grupę, albo grupa rozciąga się na cały zbiór.

2.3 Zbiór danych Adult

Zbiór danych Adult posiada zarówno wartości liczbowe (zawierające potencjalnie duże liczby, rzędu dziesiątek lub setek tysięcy, jak również liczby bliskie zera) jak i kategoryczne. W ramach wstępnego przetwarzania usunięta została kolumna *fnlwgt*, ponieważ prawie każdy wartość w tej kolumnie jest inna od pozostałych, a zatem nie dawałaby się grupować. Z uwagi na wysoką złożoność czasową i pamięciową wykorzystywanych algorytmów zbiór danych został ograniczony do 12tys próbek – jest to mała liczba, ewidentnie niewystarczająca do pokrycia przestrzeni o wymiarowości około 100 atrybutów, w związku z tym wyniki trzeba oglądać przez pryzmat zjawiska przekleństwa wymiarowości.

2.3.1 Algorytm K-średnich

n – liczba grup	Davies - Bouldin	Dunn	Rand	F-indeks
1	0,00	∞	0,64	0,78
2	1,62	0,10	0,63	0,77
3	5,09	0,00	0,48	0,47
4	2,28	0,03	0,56	0,59
5	3,88	0,00	0,55	0,54
6	3,57	0,03	0,46	0,41
7	4,38	0,01	0,47	0,42
8	3,44	0,00	0,46	0,41
9	3,26	0,00	0,45	0,35
10	2,32	0,01	0,45	0,38
11	2,41	0,00	0,42	0,28
12	3,27	0,00	0,43	0,31
13	3,62	0,00	0,45	0,30
14	4,08	0,00	0,41	0,22
15	3,00	0,00	0,42	0,23

Tabela 13 Algorytm K-średnich, Adult

Indeks Daviesa - Bouldina osiąga najniższą wartość dla n = 2, co może oznaczać istnienie pewnej korelacji między klasą referencyjną a ułożeniem w przestrzeni przykładów – zwłaszcza zważywszy na dość wysokie wartości indeksów porównawczych Rand oraz F-indeks.

Pozostałe współczynniki wraz ze wzrostem n pogarszają się – jest to spowodowane powstawaniem większej liczby grup a co za tym idzie, zmniejszaniem się minimalnej odległości między parą grup.

Wskaźniki oceny zewnętrznej również maleją, tutaj ze znacznie większą monotonicznością. Zwłaszcza F-indeks jest na niskim poziomie – wartość współczynnika mniejsza niż 0.5 oznacza, że jest mniej poprawnych oszacowań klasy pozytywnej niż błędów obu rodzajów.

Patrząc na wiersze 1 i 2 można zauważyć, że wartości RAND i F-indeks się zmniejszyły dla przykładu 2, co jest wiadomością niepokojącą – oznacza to, że powstała druga grupa przecina się w znacznym stopniu z oboma grupami referencyjnymi.

Rozwiązania z jedną grupą są premiowane z uwagi na występująca dysproporcje w liczności grup – jedna ma znaczną przewagę i wskaźniki zachowują się lepiej w sytuacji, w której pomijamy mniejszą grupę niż w sytuacji, w której grupę większą rozbijemy (jednocześnie nie zapewniając odpowiedniego pokrycia grupy mniejszej).

2.3.2 Algorytm DBSCAN

Lp.	Eps	Min_points	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup	|Szum|	|Grupa 1|
1	15	20	1,869	0,114	0,599	0,735	15	318	11205
2	10	40	1,960	0,060	0,575	0,707	5	953	10770
3	5	80	2,892	0,004	0,470	0,467	10	5909	3978
4	5	20	2,199	0,003	0,427	0,328	35	3671	4054
5	4	40	1,830	0,001	0,483	0,499	28	7433	294
6	2	20	1,785	0,020	0,403	0,196	1180	3835	15

Tabela 14 Algorytm DBSCAN, Adult

Kalibracja parametrów DBSCAN również w tym przypadku przysporzyła najwięcej trudności – bardzo łatwo popadało się w skrajności jak w poprzednim przypadku –

a) wszystkie elementy w szumie,
b) wszystkie (albo znaczna większość) elementów w grupie pierwszej,
c) bardzo duża liczba grup.

Przestrzeń posiada bardzo dużo atrybutów przy dość małej liczności próbek – efektem tego jest duże rozstrzelenie punktów – w takiej sytuacji małe wartości *eps* powodują w połączniu z małymi wartościami *minPts* dużą liczbę grup, w połączeniu z dużymi wartościami *minPts* brak znalezionych grup. Duże wartości *eps* powodują zaś połączenie większości elementów w jedną grupę.

Jeżeli chodzi o wartości wskaźników – indeksy Daviesa - Bouldina i Dunna dla przykładów z porównywalna liczba grup są znacznie lepsze – wynika to z pominięcia znacznej liczby elementów, co umożliwiło zmniejszenie oraz oddalenie od siebie wielu grup.

Również korzystając z tej metody nie udało się oddać grupowania referencyjnego – najwyższe wartości wskaźników Randa i F występują dla przykładów, w których jedna grupa dominuje – z analogicznego powodu co w przypadku k-średnich.

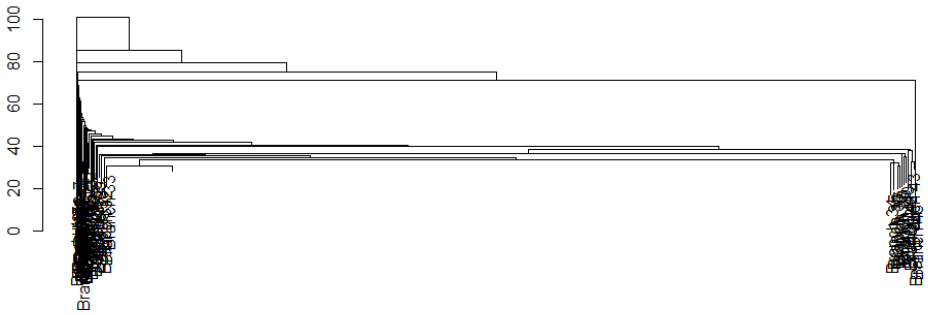
2.3.3 Algorytm najdalszego sąsiedztwa

n – liczba grup	Davies - Bouldin	Dunn	Rand	F-indeks
1	0,00	∞	0,64	0,78
2	0,16	0,91	0,64	0,78
3	0,21	0,80	0,64	0,78
4	0,21	0,73	0,64	0,78
5	0,21	0,77	0,64	0,78
6	0,32	0,69	0,64	0,78
7	0,29	0,71	0,64	0,78
8	0,29	0,75	0,64	0,78
9	0,29	0,71	0,64	0,78
10	0,29	0,62	0,64	0,78
11	0,29	0,63	0,64	0,78
12	0,29	0,66	0,64	0,78
13	0,29	0,70	0,64	0,78
14	0,29	0,72	0,63	0,78
15	0,29	0,70	0,63	0,77

Tabela 15 Algorytm najdalszego sąsiedztwa, Adult

Analizując wyniki metody najdalszego sąsiedztwa widać praktycznie identyczne wskaźniki dla większości wygenerowanych liczb grup – jest to spowodowane faktem, że grupy które różnią sąsiednie przykłady są bardzo mało liczne – w procesie grupowania hierarchicznego na poziomie 15 istnieje jedna duża grupa i reszta znacznie mniejszych i z każdym kolejnym poziomem albo małe grupy się łączą albo do są dołączane do grupy dużej – zważywszy na fakt, ze w klasyfikacji referencyjnej istnieją dwie grupy w tym jedna duża – zrozumiała jest mała różnica w wartościach wskaźników.

Wartości indeksów Daviesa - Bouldina oraz Dunna są na znacznie lepszym poziomie niż w przypadku pozostałych metod – jest to zauważalne w mniejszym lub większym stopniu dla wszystkich analizowanych zbiorów – gdyż w trakcie procesu grupowania metoda najdalszego sąsiedztwa łączy zbiory tak, aby powstawały zbiory możliwie najmniej szerokie. Dodatkowo, w przypadku zbioru Adult najwyższe poziomy hierarchii, czyli te przedstawione w tabeli znajdują się już w sytuacji dobrze odizolowanych zbiorów (trzonu i obserwacji odstających) co jest zauważone w wartościach indeksów.



Rysunek 6 Hierarchia grupowania zbioru danych Adult metoda najdalszego sąsiedztwa

2.3.4 Algorytm SOM

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup	Liczebność grup
1	1	2	0,16	0,91	0,64	0,78	2	[11998,2]
2	1	3	1,43	0,11	0,63	0,77	3	[115, 11828, 57]
3	1	4	2,30	0,08	0,60	0,74	4	[26, 96, 406, 11472]
4	2	2	1,73	0,11	0,62	0,76	4	11789, 25, 76, 110]
5	2	3	0,30	0,03	0,57	0,59	6	[106, 4732, 6216, 411, 532, 3]
6	2	4	3,11	0,03	0,54	0,57	8	[4558, 35, 643, 25, 34, 39, 6288, 378]
7	3	3	2,92	0,05	0,57	0,68	9	[128, 632, 375, 19, 26, 106, 42, 10170, 502]
8	3	4	3,28	0,01	0,53	0,51	12	[207, 627, 4143, 16, 107, 357, 63, 52, 346, 470, 59, 5553]

Tabela 16 Algorytm SOM (neurony jako grupy), Adult

Sieci SOM w podejściu – neuron jako grupa – również słabo poradziły sobie z odtworzeniem grupowania referencyjnego – jedynie eksperymenty 5, 6 i 8 znalazły podział posiadający dwie większe grupy (jednakże w złej proporcji oraz ze złym pokryciem grup referencyjnych – o czym świadczą niskie wartości indeksu Randa i F). Pozostałe przykłady opisują sytuacje z jedna klasą dominującą – co jest cechą wspólną dla wszystkich opisanych algorytmów na tym zbiorze danych. Wskaźniki posiadają wartości analogiczne do innych metod w przypadku podobnych rozkładów grup – przykłady z jedna klasą dominującą posiadają największe wskaźniki referencyjne.

2.3.5 Algorytm SOM + grupowanie

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup	Liczebność grup
1	15	60	0,48	0,35	0,64	0,78	2	[11990, 10]
2	25	25	1,24	0,49	0,64	0,78	2	[11991, 9]
3	30	30	0,26	0,54	0,64	0,78	2	[11996, 4]
4	15	60	1,82	0,33	0,64	0,78	3	[11960, 15, 25]
5	25	25	0,36	0,41	0,64	0,78	3	[11986, 7, 7]
6	30	30	0,44	0,36	0,64	0,78	3	[11983, 8, 9]
7	15	60	0,29	0,32	0,64	0,78	4	[11976, 12, 7, 5]
8	25	25	0,43	0,33	0,64	0,78	4	[11977, 5, 7, 11]
9	30	30	1,38	0,28	0,64	0,78	4	[11953, 7, 23, 17]
10	15	60	0,40	0,43	0,64	0,78	5	[11976, 7, 3, 7, 7]
11	25	25	0,58	0,28	0,64	0,78	5	[11943, 11, 23, 16, 7]
12	30	30	0,75	0,29	0,64	0,78	5	[11959, 7, 15, 5, 14]
13	15	60	0,37	0,32	0,64	0,78	8	[11941, 5, 8, 11, 6, 12, 9, 8]
14	25	25	0,77	0,27	0,64	0,78	8	[11926, 15, 19, 16, 5, 7, 4, 8]
15	30	30	0,49	0,27	0,64	0,78	8	[11921, 11, 16, 7, 14, 15, 11, 5]

Tabela 17 Algorytm SOM + Algorytm najdalszego sąsiedztwa, Adult

Testy korzystając z podejścia tworzenia grup jako grup neuronów sieci SOM przyniosły wyniki obciążone tym samym zjawiskiem które wystąpiło podczas testowania metody najdalszego sąsiedztwa, użytej tutaj do grupowania neuronów – wszystkie przykłady wyglądają prawie identycznie z uwagi na istnienie jednej grupy zawierającej większość przykładów oraz paru grup mniejszych, dołączanych do trzonu lub do siebie nawzajem.

Lp.	SOM X	SOM Y	Davies - Bouldin	Dunn	Rand	F-indeks	Liczba grup	Liczebność grup
1	15	60	3,05	0,13	0,63	0,77	2	[259, 11741]
2	25	25	3,68	0,06	0,62	0,76	2	[342, 11658]
3	30	30	3,49	0,07	0,61	0,75	2	[411, 11589]
4	15	60	2,68	0,06	0,57	0,71	3	[10896, 512, 592]
5	25	25	3,97	0,13	0,62	0,75	3	[146, 11496, 358]
6	30	30	4,45	0,05	0,56	0,69	3	[10657, 675, 668]
7	15	60	2,75	0,02	0,53	0,56	4	[5983, 4978, 440, 599]
8	25	25	4,24	0,00	0,53	0,55	4	[266, 5735, 4857, 1142]
9	30	30	4,05	0,01	0,55	0,59	4	[488, 4795, 6385, 332]
10	15	60	2,43	0,01	0,51	0,51	5	[5486, 4665, 276, 1032, 541]
11	25	25	3,85	0,01	0,55	0,56	5	[479, 6226, 3973, 148, 1174]
12	30	30	3,36	0,01	0,57	0,62	5	[10, 294, 4980, 82, 6634]
13	15	60	3,35	0,01	0,54	0,55	8	[5880, 92 ,71, 65, 83, 4591, 325, 893]
14	25	25	3,08	0,01	0,55	0,57	8	[114, 4626, 220, 6096, 373, 108, 276, 187]
15	30	30	2,74	0,01	0,44	0,37	8	[451, 4211, 2682, 335, 555, 3191, 237, 338]

Tabela 18 Algorytm SOM + Algorytm k-średnich, Adult

W celu dokładniejszego sprawdzenia tego podejścia, zaprezentowane zostały również wyniki grupowania neuronów korzystając z algorytmu k-średnich. W stosunku do poprzedniego podejścia widać znaczną różnicę, jeżeli chodzi o indeksy Daviesa Bouldina i Dunna świadczącą na korzyść algorytmu najdalszego sąsiedztwa – jest to jednak o tyle złudne, że algorytm ten optymalizuje w sposób pośredni te indeksy. Wykorzystanie k-średnich spowodowało większą równomierność, jeżeli chodzi o rozkład grup. Również to podejście nie na poprawne odtworzenie grup referencyjnych – indeksy porównujące grupowanie sukcesywnie spadają (w związku z stopniowym zmniejszeniem grupy dominującej) – jest to o tyle mylące, że maksymalne osiągnięte wartości są dla przypadku jednogrupowego – nie udało się poprawić tego rezultatu.

3. Wnioski

W wyniku przeprowadzonych eksperymentów ciężko o wyciąganie ogólnych prawidłowości na temat różnicy w działaniu sieci SOM a klasycznymi algorytmami grupowania. Okazuje się, że określone algorytmy preferują pewne sytuacje: np. nie należy używać zamiennie DBSCAN i algorytmu k-średnich. Sieci neuronowe SOM (neurony jako grupy) są w działaniu podobne do algorytmu k-średnich, jednakże osiągają nieco lepsze rezultaty. Podejście w których neurony

są grupowane zewnętrznym algorytmem jest silnie zależne od wyboru algorytmu i w ramach wykonanych eksperymentów ciężko wskazać na pozytywy tego rozwiązania.

W ramach pracy nad projektem zrozumieliśmy zasadę działania oraz sposób wykorzystania opisanych algorytmów grupowania – w szczególności poznaliśmy sieci neuronowe typu SOM. Nauczyliśmy się, że algorytm DBSCAN jest bardzo niestabilny i o ile dla niektórych zbiorów danych jest nie do zastąpienia, tak dla niektórych się kompletnie nie nadaje. Dodatkowo, projekt był kolejną motywacją do poprawiania warsztatu pracy w środowisku R i poznania nieznanych dotąd funkcjonalności.