

Impact of metrics on the effectiveness of Kohonen network clustering

Wpływ metryki na skuteczność klasteryzacji sieci Kohonena

Krystian Wypart*, Edyta Łukasik

Department of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract

The research paper focuses on investigating the impact of different metrics on the clustering process in Kohonen networks, also known as self-organized maps (SOM). The theoretical foundations of Kohonen networks, including their structure and algorithm of operation, are presented. Various metrics, such as Euclidean distance, Manhattan and cosine distance and their potential impact on clustering process are then discussed. Experiments were conducted using the Iris dataset constrained to two dimensions using PCA.

Keywords: clustering; Kohonen network; metrics

Streszczenie

Artykuł koncentruje się na badaniu wpływu różnych metryk na proces klasteryzacji w sieciach Kohonena, znanych również jako samoorganizujące się mapy (SOM). Przedstawiono teoretyczne podstawy sieci Kohonena, w tym ich strukturę i algorytm działania. Następnie omówiono różne metryki, takie jak odległość euklidesowa, Manhattan i kosinus oraz ich potencjalny wpływ na proces klasteryzacji. Przeprowadzone zostały eksperymenty z wykorzystaniem zbioru danych Iris, ograniczonym do dwóch wymiarów za pomocą PCA.

Słowa kluczowe: klasteryzacja; sieć Kohonena; metryki

*Corresponding author

Email address: krvstian.wypart@pollub.edu.pl (K. Wypart)

Published under Creative Common License (CC BY 4.0 Int.)

1. Wstęp

W erze cyfrowej, gdzie ogromne ilości danych są generowane każdego dnia, kluczowym wyzwaniem jest ich efektywna analiza i interpretacja. Klasteryzacja czyli, proces grupowania na podstawie podobieństwa, jest jednym z podstawowych narzędzi w analizie danych. Sieci Kohonena, nazywane samoorganizującymi się mapami (SOM), są zaawansowaną techniką stosowaną w nienadzorowanym uczeniu maszynowym [1-2], skuteczność tej techniki może być zależna od doboru odpowiedniej metryki odległości wykorzystywanej w algorytmie sieci. Artykuł skupi się na zbadaniu oraz analizie wspomnianej skuteczności.

Sieci Kohonena, opracowane przez Teuvo Kohonena w latach 80-tych, to rodzaj sztucznych sieci neuronowych, używanych do nienadzorowanego uczenia się. Pomimo szerokiego zastosowania sieci Kohonena w różnych dziedzinach nauki i technologii [2-6], często brak jest szerszego zrozumienia wpływu różnych metryk na proces klasteryzacji. Istniejące badania i literatura naukowa koncentrują się głównie na praktycznych zastosowaniach tych sieci. Niniejszy artykuł powstał w celu wypełnienia luki badawczej i dostarczenia możliwie cennych informacji na ten temat.

Celem artykułu jest zbadanie, jak wybór metryki odległości wpływa na proces klasteryzacji realizowany przez sieć neuronową, jaką jest sieć Kohonena. Analizie zostaną poddane trzy metryki: euklidesowa, Mahalanobisa oraz kosinusowa. Zbadana zostanie hipoteza mówiąca o tym, że metryka Euklidesa użyta w sieci Kohonena daje najlepsze rezultaty grupowania danych.

2. Przegląd literatury

Praca [7] koncentruje się na udoskonaleniu map Kohonena w celu klastrowania danych pochodzących z czujników rejestrujących szeregi czasowe. Autorzy wprowadzili dwie modyfikacje do standardowego algorytmu Kohonena: pierwsza polega na zastosowaniu miary odległości Dynamic Time Warping (DTW) zamiast tradycyjnej metryki euklidesowej, a druga na wprowadzeniu schematu próbkowania, który wybiera najbardziej zróżnicowane elementy, jako początkowych reprezentantów klastrow. W artykule omówiono, jak te mapy mogą być efektywnie używane do klastrowania danych o wysokiej wymiarowości, takich jak dane z metabolomiki i chemometrii. Eksperymenty porównują zmodyfikowane mapy Kohonena z metodami hierarchicznymi klastrowania bazującymi na DTW jako miarze odległości. Wprowadzone modyfikacje wykazały lepszą jakość klastrowania i większą szybkość w porównaniu do klastrowania hierarchicznego, sugerując wyższą efektywność nowego podejścia w analizie szeregów czasowych. Do oceny jakości klastrowania autorzy wykorzystali wariację wewnątrzgrupową, indeks Calińskiego-Harabasa oraz indeks Randa. W końcowej części pracy autorzy podsumowują swoje badania, podkreślając przewagę zmodyfikowanego algorytmu map Kohonena nad tradycyjnymi technikami. Zwracają uwagę na większą skalowalność i efektywność ich metody, co czyni ją bardzo atrakcyjną do analizy danych szeregów czasowych. Wskazują także na możliwość dalszych badań nad analizą innych funkcji aktualizacji wektorów oraz optymalizacją funkcji celu.

Głównym celem pracy [8] jest opracowanie metody grupowania obrazów na podstawie podobieństwa, z uwzględnieniem cech takich jak kolory, kompozycja oraz wrażenia. Autorzy analizują dwa główne aspekty obrazów: wrażenia (opisane słowami) i kolory. Słowa opisujące wrażenia i kolory są używane do tworzenia wektorów reprezentujących obrazy. Na przykład, obraz może być opisany jako: świeży, przyjemny, rytmiczny, gęsty, zakłopotanie, pożądanie, cierpienie, chwila, bunt, złość, kłamstwo. Metoda określania podobieństwa między obrazami opiera się na wykorzystaniu wektorów i odległości ε . Przeprowadzono eksperymenty, które klasyfikowały obrazy na podstawie podobieństwa, tworząc grupy obrazów o podobnych cechach. Obrazy są przekształcane na wektory, co pozwala na zastosowanie metod matematycznych i statystycznych do analizy ich podobieństwa. Na podstawie obliczonych odległości, obrazy są klasyfikowane do grup od 1 do 4. Omówiono również zastosowanie samoorganizujących się map (SOM) do grupowania obrazów, co umożliwia uporządkowaną reprezentację danych wysokowymiarowych. Autorzy stwierdzili, że zaproponowana metoda grupowania obrazów jest skuteczna i może być pomocna w organizowaniu kolekcji obrazów. Metoda SOM okazała się użyteczna w redukcji wymiarowości danych i wizualizacji grup podobieństwa.

W artykule [9] zbadano wydajność samoorganizujących się map (SOM) przy użyciu różnych topologii i metryk odległości. Do badań wykorzystano trzy zestawy danych: Iris, Breast Cancer Wisconsin i Car Evaluation, reprezentujące małe, średnie i duże zbiory danych. Omówiono architekturę oraz zasady działania SOM, w tym procesy uczenia i testowania. Przedstawiono etapy inicjalizacji wag, obliczania najlepszej jednostki (BMU), aktualizacji węzłów sąsiednich oraz osiągania zbieżności. W pracy opisano trzy główne typy topologii kratowych dla sieci SOM: trójkątną, prostokątną i heksagonalną, oraz ich wpływ na działanie sieci. W topologii trójkątnej neurony są ułożone w trójkątną siatkę, każdy z trzema bezpośrednimi sąsiadami. W topologiach prostokątnych i heksagonalnych neurony są rozmieszczone odpowiednio w sieci prostokątnej i sześciokątnej, każdy z czterema bezpośrednimi sąsiadami. W badaniu zastosowano odległości euklidesową, Manhattan, Czebyszewa i Canberra. Na zestawie danych Iris najwyższą dokładność klasyfikacji osiągnięto przy użyciu metryki euklidesowej z topologią sześciokątną, uzyskując wynik 100%. Pozostałe wyniki dla tej metryki przy użyciu topologii trójkąta i prostokąta to odpowiednio 95,56% i 97,78%. Drugie miejsce zajęła odległość Czebyszewa z wynikiem 97,78% dla wszystkich topologii. Metryki Manhattan i Canberra zanotowały niższe wyniki na poziomie 77,78% dla wszystkich topologii, z wyjątkiem metryki Manhattan dla sześciokąta, która osiągnęła 84,44%. Na zbiorze danych Breast Cancer Wisconsin wyniki były bardzo interesujące, ponieważ trzy metryki: euklidesowa, Manhattan i Czebyszewa uzyskały podobne wyniki na wszystkich topologiach, wynoszące 88,76%. Odległość Canberra zanotowała najgorszy wynik 71% i 79,89% dla topologii sześciokąta. Dla zbioru

danych Car Evaluation wyniki dla metryki euklidesowej i topologii trójkąta, prostokąta i sześciokąta wyniosły odpowiednio 90,91%, 91,86% i 96,60%. Dla Manhattan odpowiednio 90,91%, 94,70% i 90,91%, dla Czebyszewa 90,91%, 91,86% i 85,23%, a dla Canberra 85,23%, 71,97% i 75,76%. Można zauważyć, że topologia sześciokąta w połączeniu z metryką euklidesową zapewniała najwyższe wyniki we wszystkich testowanych zestawach danych. Metryka euklidesowa wydaje się być najbardziej skuteczna przy współpracy z różnymi topologiami krat w SOM, ponieważ wymaga mniej obliczeń niż pozostałe metryki, co zmniejsza złożoność obliczeniową systemu. Mniejsza złożoność obliczeń przekłada się na lepszą wydajność i dokładność klasyfikacji.

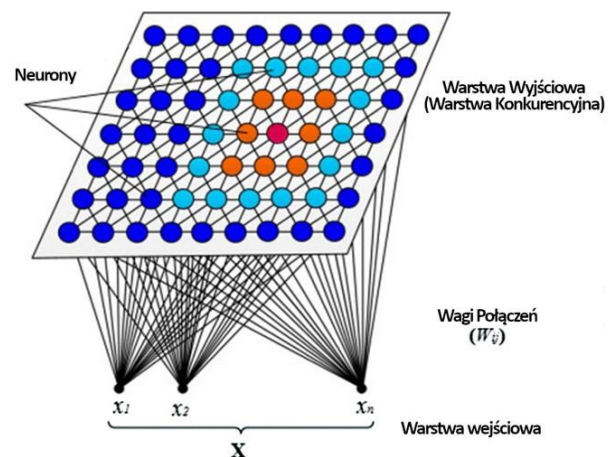
3. Wstęp do sieci Kohonena

Ten rozdział został poświęcony przedstawieniu zasad działania sieci Kohonena oraz wykorzystanych w późniejszych badaniach metryk odległości.

3.1. Podstawy działania sieci

Sieć Kohonena, powszechnie znana jako samoorganizująca się mapa (ang. Self-Organizing Map, SOM), jest istotnym narzędziem w dziedzinie sztucznej inteligencji i uczenia maszynowego. Została opracowana przez fińskiego naukowca Teuvo Kohonena w latach 80. XX wieku, jako środek do wizualizacji i analizy skomplikowanych, wielowymiarowych zbiorów danych. SOM, jako typ sieci neuronowych, jest głównie stosowana w zadaniach uczenia nienadzorowanego, gdzie dane nie są oznaczone etykietami ani klasami [10].

Struktura sieci Kohonena składa się z dwóch podstawowych warstw (Rysunek 1).



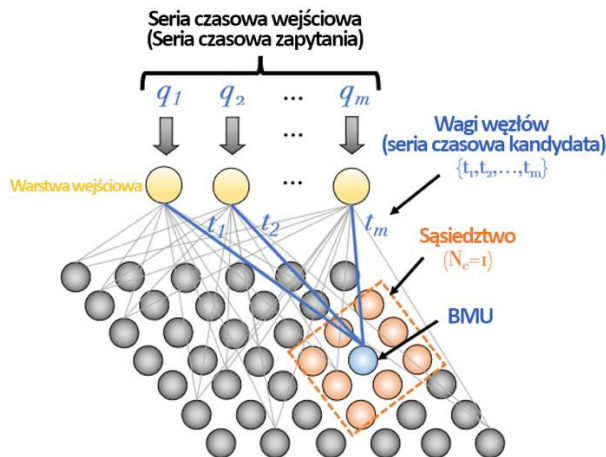
Rysunek 1: Struktura sieci Kohonena [11].

Pierwsza warstwa to warstwa wejściowa, której zadaniem jest odbieranie danych wejściowych. Danymi mogą być proste wartości po złożone wzorce i tekstury. Drugą ważną warstwą jest warstwa konkurencyjna, na której znajdują się neurony, konkurujące ze sobą o stanie się neuronem zwyciężskim (ang. Best Matching Unit, BMU). Każdy neuron w tej warstwie jest powiązany ze wszystkimi neuronami w warstwie wejściowej i posiada

unikalny wektor wag, który jest modyfikowany podczas procesu uczenia.

3.2. Uczenie sieci (trening)

Trening sieci (Rysunek 2) jest istotnym elementem, który umożliwia efektywne działanie w analizie oraz ewentualnym późniejszym grupowaniu danych. Proces ten można podzielić na kilka bardziej szczegółowych kroków, które pozwalają zrozumieć, jak sieć się uczy i od losowych wag na początku staje się zorganizowaną mapą cech, która reprezentuje struktury danych wejściowych.



Rysunek 2: Struktura sieci neuronów [12].

Uczenie obejmuje następujące etapy:

1. Inicjalizacja wag: Na początku następuje inicjalizacja wag wszystkich neuronów w sieci. Wartości te są losowe i pochodzą z niskiego zakresu od 0 do 1, aby umożliwić mapie jak najlepsze dostosowanie się do danych w trakcie treningu.
2. Konkurencja między neuronami: Podczas każdej iteracji procesu uczenia, dostarczany jest do sieci pojedynczy wektor danych wejściowych. Każdy neuron na mapie mierzy swoją odległość od tego wektora, używając wcześniej określonej metryki odległości. Neuron, którego wagi są najbliższe do wektora wejściowego, staje się neuronem zwycięskim, czyli BMU (Best Matching Unit). Ten proces jest kluczowy, ponieważ determinuje, które informacje z danych wejściowych są najważniejsze dla każdego neuronu.
3. Aktualizacja wag BMU i sąsiadujących neuronów: Po wyłonieniu BMU, wagi tego neuronu oraz sąsiadujących z nim neuronów są aktualizowane, aby zbliżyć je do wektora wejściowego. Proces ten jest regulowany przez funkcję sąsiedztwa, która decyduje, które neurony i w jakim stopniu będą aktualizowane. Zazwyczaj funkcja ta ma kształt krzywej Gaussa lub prostokąta, gdzie największa aktualizacja ma miejsce w pobliżu BMU, a mniejsze aktualizacje dotyczą bardziej oddalonych neuronów.
4. Proces adaptacji: Aktualizacja wag jest procesem adaptacyjnym, podczas którego wagi neuronów są sukcesywnie przesuwane w kierunku wektora wejściowego. Wielkość tego przesunięcia zależy od

współczynnika uczenia (learning rate), który określa stopień modyfikacji wag neuronów w odpowiedzi na wektor wejściowy. Zwykle współczynnik ten zmniejsza się w miarę postępu procesu uczenia, co pozwala na dokładniejszą regulację wag na dalszych etapach.

5. Uczenie konkurencyjne: Proces konkurencji, aktualizacji i adaptacji jest powtarzany przez wiele iteracji z wykorzystaniem całego zestawu danych wejściowych. W miarę trwania uczenia, mapa staje się coraz bardziej zorganizowana. Neurony na mapie zaczynają organizować się w różne wzorce lub klasy danych wejściowych, tworząc strukturę, odzwierciedlającą topologię i różnorodność tych danych.
6. Zakończenie: Proces uczenia trwa do momentu, gdy wagi neuronów się stabilizują, co oznacza, że mapa nie wykazuje już znaczących zmian, pomimo dalszych iteracji. W idealnej sytuacji każdy neuron na mapie reprezentuje różne wzorce lub klasy danych wejściowych, a podobne wzorce są mapowane na sąsiednie neurony. Powstaje w ten sposób topologicznie uporządkowana mapa cech, która skutecznie odzwierciedla strukturę przestrzeni wejściowej.

3.3. Metryki odległości

Metryki w sieci Kohonena określają, który neuron najlepiej odpowiada danemu wektorowi danych, wpływając tym samym na kształtowanie się mapy. W sieciach SOM metryki wykorzystywane są w dwóch przypadkach: neuron-dane czyli dopasowanie i wybór BMU oraz neuron-neuron czyli aktualizacja wag.

3.3.1. Odległość euklidesowa

Odległość euklidesowa między dwoma punktami w przestrzeni n -wymiarowej jest równa pierwiastkowi kwadratowemu z sumy kwadratów różnic ich odpowiednich współrzędnych [13]. Jest to najbardziej znana metryka, często wykorzystywana w geometrii. Matematycznie dla dwóch punktów: $P = (p_1, p_2, \dots, p_n)$ i $Q = (q_1, q_2, \dots, q_n)$ odległość euklidesowa obliczana jest jako:

$$d(P, Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (1)$$

3.3.2. Odległość Mahalanobisa

Odległość Mahalanobisa różni się znacząco od pozostałych metryk, ponieważ bierze pod uwagę korelację między zmiennymi w danych i jest skalowana odwrotnie proporcjonalnie do ich wariancji. Jest często używana w analizie wielowymiarowej, gdzie różne cechy mogą mieć różne jednostki i zakresy zmienności. Odległość Mahalanobisa staje się równoważna odległości euklidesowej, gdy zmienne są niezależne i mają takie same odchylenia standardowe. Odległość ta jest często wykorzystywana do oceny podobieństwa między sklasyfikowanymi próbkami statystycznymi, a próbką nieznaną [14]. Dla wektora P , średniej μ oraz macierzy kowariancji S , odległość Mahalanobisa jest obliczana jako:

$$d(P, \mu) \sqrt{(P - \mu)^T S^{-1} (P - \mu)} \quad (2)$$

3.3.3. Odległość kosinusowa

Odległość kosinusowa, znana także jako podobieństwo kosinusowe, jest wykorzystywana w analizie danych do oceny podobieństwa między dwoma wektorami o niezerowych wartościach, zdefiniowanymi w przestrzeni z iloczynem skalarnym. To podobieństwo opiera się na kosinusie kąta pomiędzy dwoma wektorami, co odpowiada stosunkowi ich iloczynu skalarnego do iloczynu ich długości. Podobieństwo kosinusowe przenosi ideę kosinusa kąta pomiędzy dwoma wektorami z geometrii analitycznej do przestrzeni wielowymiarowej [15]. Odległość ta dla wektorów P i Q jest obliczana jako:

$$d(P, Q) = 1 - \frac{\sum_{i=1}^n p_i q_i}{\sqrt{\sum_{i=1}^n p_i^2} * \sqrt{\sum_{i=1}^n q_i^2}} \quad (3)$$

3.4. Redukcja wymiarowości zbioru danych

PCA (ang. Principal Component Analysis), jest techniką stosowaną w analizie danych. Pozwala na zmniejszenie liczby zmiennych w zbiorze danych, zachowując przy tym jak najwięcej informacji o nim. PCA tworzy nowe zmienne (składowe główne), które są kombinacjami liniowymi oryginalnych zmiennych, ale są ze sobą nieskorelowane (ortogonalne) [16].

4. Wyniki badań i analiza

Ten rozdział koncentruje się na analizie wpływu różnych metryk odległości na proces klasteryzacji w sieci Kohonena. Celem przeprowadzonych badań jest wyjaśnienie, jak wybór metryki wpływa na zdolność sieci do efektywnego grupowania danych i odkrywania ich struktury. Wyniki opierają się na ocenie map wag neuronów, macierzy-U oraz wskaźnika sylwetki.

4.1. Narzędzia do analizy wyników

Macierz-U (ang. U-matrix), prezentuje odległości między neuronami na SOM, oferując wgląd w podobieństwo i różnice między poszczególnymi jednostkami i ich sąsiadami. Macierz-U tworzy się poprzez obliczenie odległości między sąsiadującymi neuronami na mapie. Dla każdego neuronu określone są odległości do jego najbliższych sąsiadów, następnie są one reprezentowane jako wysokości macierzy [17]. Wartości te mogą być interpretowane jako miary niepodobieństwa – wyższa wartość, równa się większej różnicy między sąsiadującymi neuronami. Niskie wartości (doliny) w macierzy, wskazują na obszary gdzie neurony są do siebie podobne i mogą formować klastry. Wysokie wartości (wzniesienia) oznaczają granice między klastrami.

Wskaźnik sylwetki (ang. Silhouette index) jest metodą oceny spójności wewnętrznej klastrów w analizie skupień. Jest wskaźnikiem, który mierzy jak dobrze każdy obiekt pasuje do swojego klastra, w porównaniu

do innych klastrów [18]. Wartość wskaźnika sylwetki mieści się w przedziale od -1 do 1. Wartość bliska 1 oznacza, że punkt jest dobrze dopasowany do swojego klastra i słabo dopasowany do sąsiednich klastrów. Wartość bliska 0 wskazuje, że punkt leży na granicy decyzyjnej między dwoma klastrami, z kolei wartość bliska -1 świadczy o tym, że punkty zostały przypisane do niewłaściwych klastrów.

4.2. Metodologia przeprowadzenia badań

Badania zostały przeprowadzone z wykorzystaniem takich samych wartości współczynnika uczenia oraz współczynnika sąsiedztwa dla każdej ze zbędnych metryk wynoszącymi odpowiednio 0.5 oraz 1. Celem ustalenia takich wartości było zachowanie identycznych warunków testowych. Warto zauważyć, że wartości tych współczynników zostały dobrane na podstawie licznych testów próbnych i wspomniane wcześniej wartości prezentowały najlepsze wyniki w kontekście efektywności grupowania dla każdej metryki. Kod wykorzystanej sieci Kohonena obejmował 100 iteracji treningowych, a finalne badanie zostało powtórzone 5 razy z każdą metryką, w celu zweryfikowania istnienia możliwych odchyleń w ostatecznym wyniku.

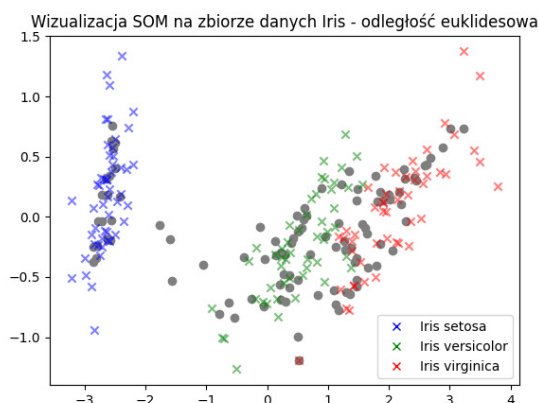
W przeprowadzonych badaniach wykorzystano gotowy zbiór danych Iris zawierający informacje o 3 gatunkach irysów. Jest to zbiór czterowymiarowy składający się ze 150 rekordów. Na potrzeby badań został on ograniczony do dwóch wymiarów przy użyciu narzędzia PCA.

4.3. Wyniki badań dla metryki euklidesowej

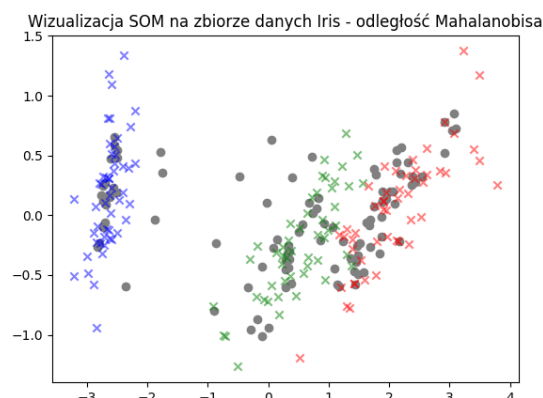
Na Rysunku 3 z mapą neuronów, można od razu zaobserwować, że sieć dobrze poradziła sobie z oddzieleniem gatunków irysów. Najlepiej oddzielonym jest gatunek *Iris setosa*. Dwa pozostałe gatunki reprezentują przypuszczalnie podobne cechy z tego względu, że widoczne jest zjawisko nakładania się na siebie tych zbiorów w połowie oraz górnej ich części. W pustej przestrzeni pomiędzy zbiorem *Iris setosa* oraz *Iris versicolor*, widocznych jest kilka neuronów, które nie wydają się być przypisane w widoczny sposób do któregoś z gatunków. Ponadto w każdym z gatunków nie zostaje dobrze pokryta górna część zbioru.

Na macierzy z Rysunku 4 widać wyraźne skupiska kwadratów koloru czarnego, które symbolizują doliny otoczone jaśniejszymi kwadratami, czyli wzniesieniami, co sugeruje istnienie granic klastrów.

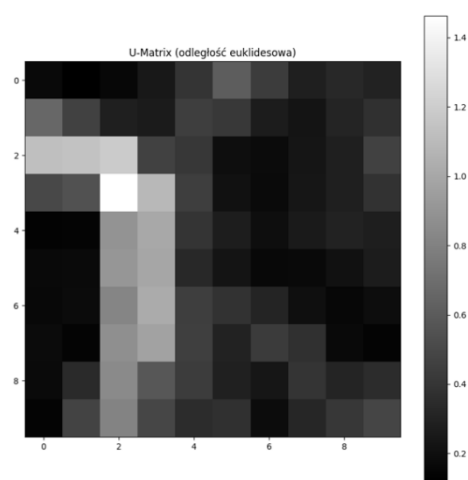
Widać tu trzy takie skupiska: dwa z lewej strony macierzy i jedno z prawej. Po lewej stronie macierzy można zauważyć wyraźne wzniesienie, które oznacza dobrze zdefiniowaną granicę klastra. Jest to najlepiej odseparowany gatunek *Iris setosa*. Wskaźnik sylwetki dla sieci Kohonena z metryką euklidesową wynosi 0.1276, co sugeruje, że neurony znajdują się na granicy decyzyjnej pomiędzy dwoma klastrami, co potwierdza obserwacja Rysunków 3 i 4.



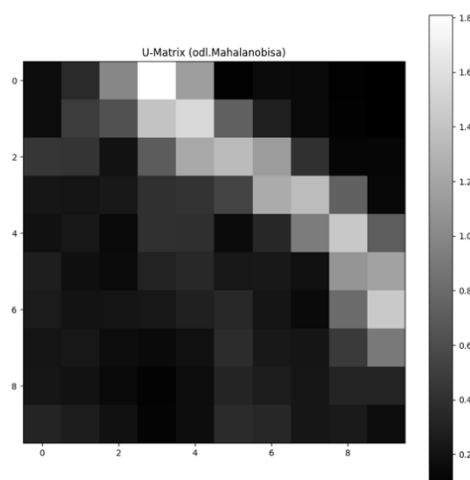
Rysunek 3: Mapa neuronów dla metryki euklidesowej.



Rysunek 5: Mapa neuronów dla metryki Mahalanobisa.



Rysunek 4: Macierz-U dla metryki euklidesowej.



Rysunek 6: Macierz-U dla metryki Mahalanobisa.

4.4. Wyniki badań dla metryki Mahalanobisa

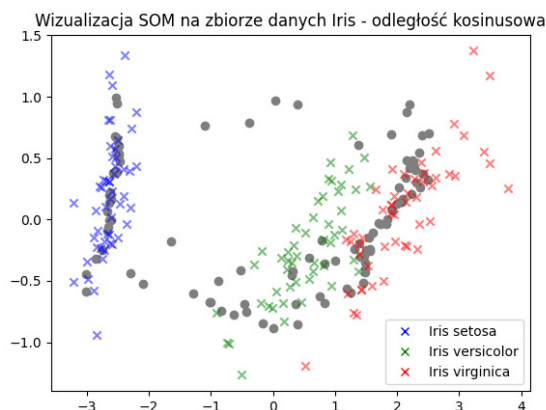
Na Rysunku 5 można zaobserwować intensywne zagęszczenie neuronów w środku każdego z gatunków. W *Iris setosa* jest widoczne wyraźnie gorsze pokrycie góry jak i dołu, a także występuje więcej neuronów w przestrzeni wolnej między gatunkami. Zdecydowanie gorzej niż w przypadku metryki euklidesowej prezentuje się gatunek *Iris versicolor*, gdzie w górnej jego części znajduje się jeszcze mniej neuronów. Natomiast gatunek *Iris virginica* wydaje się być praktycznie identyczny, jak w badaniu z metryką euklidesową.

Na macierzy z Rysunku 6, można zauważyć bardzo dobrze *Iris setosa* z wyraźną granicą. Widoczny jest również najbardziej zaciemniony obszar lewej dolnej części macierzy, co może wiązać się z istnieniem kolejnego klastra, na co wskazują jaśniejsze kwadraty dookoła niego, prawdopodobnie będzie to gatunek *Iris virginica*. Większe natężenie podobnych do siebie neuronów występuje również w lewej górnej części macierzy, a także nieco poniżej co może oznaczać, że te neurony są bardziej podobne do siebie i mogą reprezentować ten sam klaster.

Wskaźnik sylwetki dla metryki Mahalanobisa wyniósł -0.0008, wartość ta jest bliska zeru co oznacza, że punkty znajdują się na granicy decyzyjnej pomiędzy dwoma klastrami. Obrazuje to również fakt, że na macierzy wiele małych ciemnych punktów jest losowo „rozrzucana” po macierzy tym samym nie tworząc większych skupisk jasno sygnalizujących istnienie klastra.

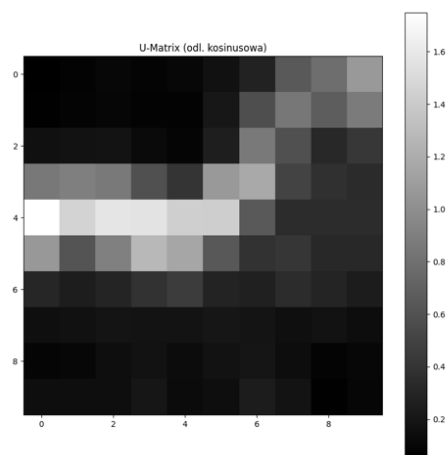
4.5. Wyniki badań dla metryki kosinusowej

Na Rysunku 7 widoczny jest dobrze odseparowany zbiór *Iris setosa*, tak jak w przypadku dwóch wcześniej zbadanych metryk. Natomiast można zaobserwować zdecydowanie jak do tej pory największą liczbę neuronów w przestrzeni wolnej między gatunkami. *Iris versicolor* został w tym badaniu najgorzej odseparowanym gatunkiem, gdyż neurony występuje praktycznie tylko w dolnej jego części. Gatunek *Iris virginica* ma stosunkowo dobrze wydzielony środek zbioru, jednak jego górna część nie została wystarczająco dobrze pokryta, tak jak w poprzednich badaniach. Już na tym etapie analizy można dojść do wniosku, że ta metryka najgorzej wpłynęła na skuteczność grupowania danych.



Rysunek 7: Mapa neuronów dla metryki kosinusowej.

Macierz z Rysunku 8 prezentuje wyraźnie oddzielony klastrowy w górnej jego części, co najprawdopodobniej wskazuje na gatunek *Iris setosa*. Natomiast w dolnej części macierzy jest widoczne nagromadzenie czarnych kwadratów jednakże bez widocznych granic. Może to świadczyć o istnieniu tylko dwóch klastrow, co potwierdza sytuacja z Rysunku 7, gdzie gatunek *Iris versicolor* został praktycznie bez pokrycia, nie licząc kilku neuronów w jego dolnej części. Wskaźnik sylwetki dla tej metryki wynosi -0.2522 co jest najgorszym wynikiem spośród wszystkich zbadanych metryk. Jest to wartość bliższa zeru, a więc mówi o tym, że neurony znajdują się na granicy decyzyjnej pomiędzy dwoma klastrowami lub zostały przypisane do nieodpowiednich klastrow.



Rysunek 8: Macierz-U dla metryki kosinusowej.

5. Podsumowanie

Po analizie wyników przeprowadzonych badań pierwszy z nasuwających się wniosków jest taki, że metryka kosinusowa wypadła zdecydowanie najgorzej spośród wszystkich zbadanych metryk. Praktycznie całkowicie pominięty został gatunek *Iris versicolor*. Może to mieć związek z charakterystyką tej metryki, która w przeciwieństwie do innych zbadanych nie bazuje na

bezpośredniej odległości między punktami, a operuje na kącie między wektorami. Jej wykorzystanie w procesie klastrowacji może okazać się nieefektywne, co sugerują otrzymane wyniki.

Metryka Mahalanobisa pomimo swojego zaawansowania, gdyż jak zostało opisane wcześniej bierze pod uwagę korelację między zmiennymi, nie zaprezentowała tak dobrze jak metryka euklidesowa, która jest zdecydowanie najmniej skomplikowaną z metryk. Pomimo niewysokiej rozbieżności w wartościach wskaźnika sylwetki dla tych dwóch metryk z wizualnej oceny można śmiało wnioskować, że metryka euklidesowa wypadła najlepiej, jeśli chodzi o jakość klastrow, co potwierdza tym samym postawioną wcześniej hipotezę.

Należy również zaznaczyć specyfikę zbioru *Iris* wykorzystanego w tym badaniu, ponieważ został on ograniczony do dwóch wymiarów za pomocą narzędzia PCA, które zachowuje jego charakterystykę, jednak mimo tego taki zabieg może mieć to wpływ na końcowy wynik procesu klastrowacji.

Ponadto w literaturze związanej z klastrowacją oraz sieciami Kohonena, odległość euklidesowa jest najczęściej pojawiająca się metryką, co może być związane z tym, że jak pokazują powyższe badania jest ona najbardziej skuteczną metryką, będąc przy tym najmniej złożoną obliczeniowo oraz najbardziej wydajną, co udo- wodniono w pracy [9].

Literatura

- [1] R. Ponmalai, C. Kamath, Self-Organizing Maps and Their Applications to Data Analysis, Lawrence Livermore National Laboratory (LLNL) (2019), <https://doi.org/10.2172/1566795> [15.04.2024]
- [2] M. Saraee, S. Vahid Moosavi, S. Rezapour, Application of Self Organizing Map (SOM) to model a machining process, Journal of Manufacturing Technology Management 26 (2004) 818-830, <https://doi.org/10.1108/17410381111149666> [10.04.2024]
- [3] S. C. Sheridan, C. C. Lee, The Self-Organizing Map in Synoptic Climatological Research, Progress in Physical Geography Earth and Environment 35 (2011) 109-119, <https://doi.org/10.1177/0309133310397582>
- [4] E. N. Cassano, J. M. Glisan, J. J. Cassano, W. J. Gutowski Jr., M.W. Seefeldt, Self-organizing map analysis of widespread temperature extremes in Alaska and Canada, Climate Research 62 (2015) 199-218, <https://doi.org/10.3354/cr01274>
- [5] W. F. de Carvalho Rocha, Ch. B. do Prado, N. Blonder, Comparison of Chemometric Problems in Food Analysis using Non-Linear Methods, Molecules 25 (2020) 3025, <https://doi.org/10.3390/molecules25133025>
- [6] I. Zankinski, K. Kolev, T. Balabanov, Alternatives for Neighborhood Function in Kohonen Maps, In Large-Scale Scientific Computing (LSSC) (2019) 540-544, https://doi.org/10.1007/978-3-030-41032-2_62
- [7] K. J. Krishnan, K. Mitra, Clustering Time Series Sensor Data Using Modified Kohonen Maps, In Seventh Indian Control Conference (ICC) (2021) 129-134, [doi: 10.1109/ICC54714.2021.9703173](https://doi.org/10.1109/ICC54714.2021.9703173).

- [8] N. Ishii, Y. Tokuda, I. Torii, T. Kanda, Similarity Grouping of Paintings by Distance Measure and Self Organizing Map, In Knowledge-Based and Intelligent Information and Engineering Systems (KES) (2009) 713-720, https://doi.org/10.1007/978-3-642-04592-9_88
- [9] S. Ramiah, Effect of Lattice Topologies and Distance Measurements in Self-Organizing Map for Better Classification, In Emerging Research in Computing Information, Communication and Applications, (2019) 183-191, https://doi.org/10.1007/978-981-13-5953-8_16
- [10] K. L. Priddy, P. E. Keller, Artificial Neural Networks S: An Introduction, SPIE, 2005.
- [11] R. Shirani Faradonbeh, S. Shaffiee Haghsheenas, A. Taheri, R. Mikaeil, Application of self-organizing map and fuzzy c-mean techniques for rockburst clustering in deep underground projects, Neural Computing and Applications 32 (2020) 8545-8559, <https://doi.org/10.1007/s00521-019-04353-z>
- [12] A. Javed, D. M. Rizzo, B. Suk. Lee, R. Gramling, Sometimes: self organizing maps for time series clustering and its application to serious illness conversations, Data Mining Knowledge Discovery 38 (2024) 813-839, <https://doi.org/10.1007/s10618-023-00979-9>
- [13] K. J. Smith, Precalculus: A Functional Approach to Graphing and Problem Solving ed 6 Jones & Bartlett Learning 6th edition, 2011.
- [14] H. Rabal, N. Cap, C. Criado, N. Alamo, Holodiagrams using Mahalanobis Distance Optik 123 (2012) 1725-1731, <https://doi.org/10.1016/j.ijleo.2011.11.077>
- [15] M. E. Masoud, M. M. A. Mahfouz, Protection scheme for transmission lines based on alienation coefficients for current signals, IET Generation Transmission and Distribution 4 (2010) 1236-1244 <https://doi.org/10.1049/iet-gtd.2009.0648>
- [16] H. Abdi, L. J. Williams, Principal component analysis, WIREs Computational Statistics 2 (2010) 433-459, <https://doi.org/10.1002/wics.101>
- [17] J. Lötsch, A. Ultsch, Exploiting the Structures of the U-matrix, In Advances in Self-Organizing Maps and Learning Vector Quantization. Advances in Intelligent Systems and Computing 295 (2014) 249-257, https://doi.org/10.1007/978-3-319-07695-9_24
- [18] P. J. Rousseeuw, Silhouettes, A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53-65, [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)