Jacek Czerniak

WIZUALIZACJA WIELOWYMIAROWOŚCI ZA POMOCĄ „*RURY PRÓBEK*" – TEORIA ZAGADNIENIA.

Streszczenie

*Artykuł porusza zagadnienie wizualizacji wielowymiarowej. Powstał on w związku z badaniami nad przygotowaniem i dyskretyzacją danych poprzedzającymi ich przetwarzanie metodą zbiorów przybliżonych (*ang. rough sets*).Treścią publikacji jest zarys teoretyczny metody „*Rury Próbek*" jako sposobu na wizualizację wielowymiarową rzeczywistych zestawów danych eksperymentalnych. Pokazuje ona sposób wizualizowania do 360 wymiarów, co w tradycyjnych ujęciach wydawało się mocno kłopotliwe, czy wręcz nierealne. Artykuł dotyka również implikacji* post factum *zobrazowania tejże wielowymiarowości. Wytycza drogi dalszych badań wkraczających w dziedzinę m.in. grafiki komputerowej.*

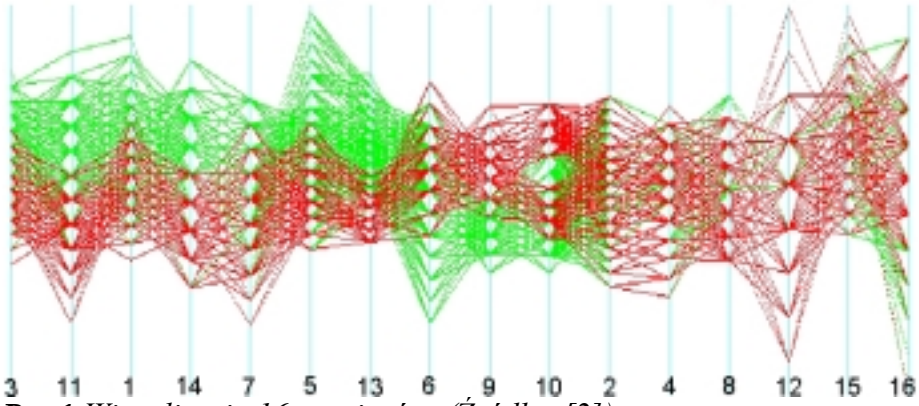
Słowa kluczowe:

*wizualizacja wielowymiarowa, wielowymiarowość, zbiory przybliżone (*ang. rough sets*), funkcje wielowymiarowe, dyskretyzacja.*

1. Wstęp

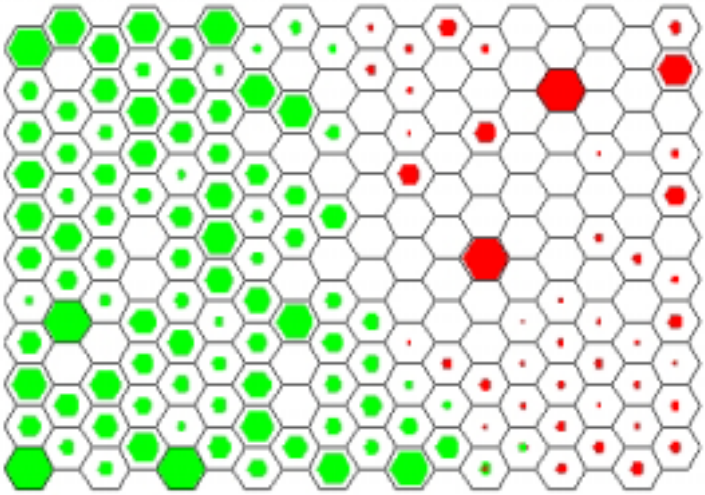
Zapewne każdy, kto zetknął się z problemem wielowymiarowości zastanawiał się jak wyobrazić sobie czwarty wymiar. Trywialne ujęcie kartezjańskie, którym posługuje się matematyka zdaje się zupełnie niewystarczające do zobrazowania rzeczywistości przekraczającej trzy wymiary. Populistyczne podejście do zagadnienia spycha wszelkie rozważania na ten temat do dysput filozoficznych. Może i nie bez przyczyny boć historia filozofii zna wielu badaczy, którzy dotykali problemu wielowymiarowości, jednak cytowanie tych poszukiwań nie jest celowym w tym opracowaniu. Fizycy po opublikowaniu Ogólnej Teorii Względności częściej postrzegają czas jako czwarty wymiar. Lecz co z wymiarem piątym, trzydziestym, lub setnym?

Jakież zatem stosuje się obecnie sposoby i wzorce dla wizualizacji wielowymiarowych zjawisk? Są to przeważnie próby translacji wielu wymiarów do dwuwymiarowej płaszczyzny. Dobrym przykładem takiego podejścia jest metoda prezentowana na rysunku Rys.1 gdzie zastosowano analizę węzłów i interwałów [3].



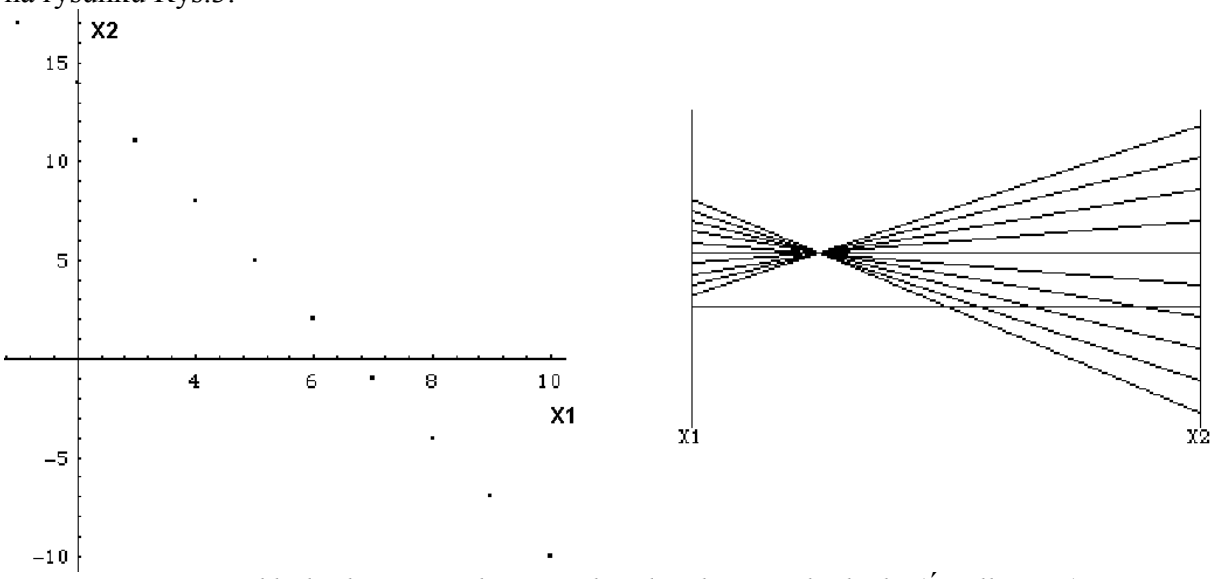
Rys.1 *Wizualizacja 16 wymiarów. (Źródło: [2])*

Warto nadmienić, że przedstawiony wykres jest przykładem wizualizacji szesnastu wymiarów. Zwiększenie wymiarowości rozważanego zjawiska pogorszy i tak już niełatwą interpretację wykresu. Kolejny rysunek Rys.2 pokazuje ciekawe podejście zwane mapą Kohonen'a. Jest to również przykład wizualizacji przestrzeni szesnastowymiarowej. Z pozoru uzyskany obraz nie wydaje się zbyt czytelnym lecz autorzy [3][6] wskazują go jako bardzo użyteczny przed poddaniem danych przetwarzaniu za pomocą sieci neuronowych.



Rys.2 *Mapa Kohonen'a. (Źródło: [2])*

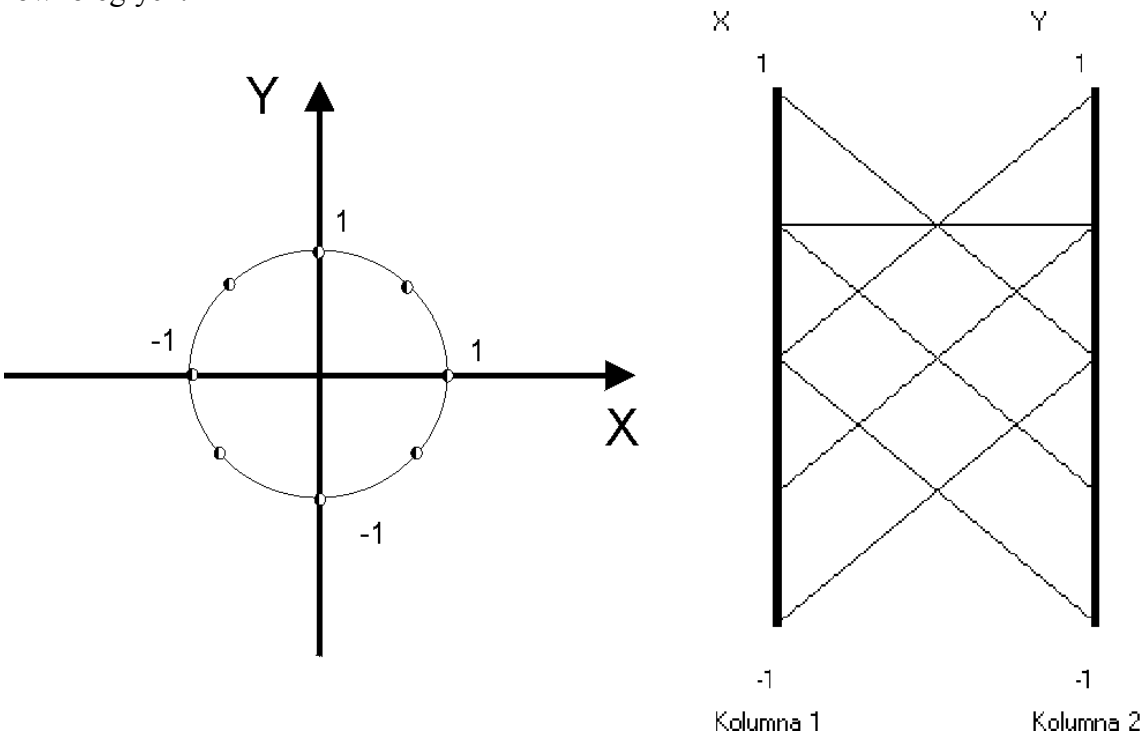
Wśród sposobów wizualizacji wykresów funkcji na uwagę zasługuje system współrzędnych równoległych [2][5][7][8]. Metoda ta powstała jako alternatywa dla kartezjańskiego układu współrzędnych i właśnie jej zestawienie z tym klasycznym podejściem możemy obserwować na rysunku Rys.3.



Rys.3 *Zestawione układy; kartezjański i współrzędnych równoległych. (Źródło: [2])*

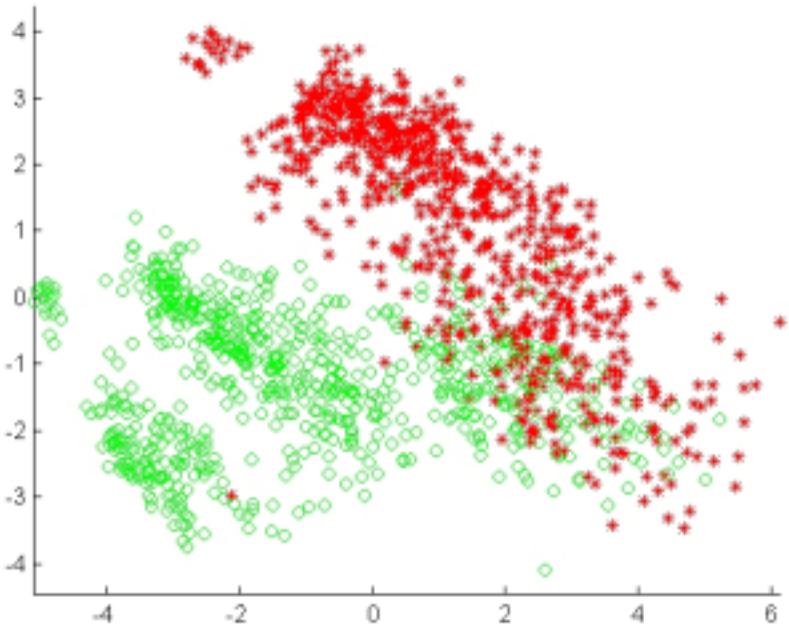
Jest to bardzo ciekawe podejście będące istotnym wydarzeniem wśród alternatywnych metod wizualizacji współrzędnych warto więc przeanalizować jeszcze jeden przykład użycia tej

metody. Rysunek Rys.4 pokazuje zestawienie okręgu z zaznaczonymi kilkoma punktami oraz tego co jest obrazem zależności współrzędnych tychże punktów w ujęciu współrzędnych równoległych.



Rys.4 *Okrąg tradycyjnie i w układzie współrzędnych równoległych. (Źródło: część B wygenerowana na podstawie [1])*

Kończąc rozważania o stosowanych metodach wizualizacji wielowymiarowości chciałbym pokazać przykład często stosowanego sposobu pokazania dwóch powiązanych ze sobą atrybutów. Rozpowszechnianym sposobem wizualnej analizy zależności jest rozważanie „chmury" punktów, które otrzymujemy podając na osie OX i OY dwa atrybuty. Na rysunku Rys.5 można zobaczyć przykład wzmiankowanej „chmury". Jest to metoda znana i szeroko stosowana lecz w dalszej części artykułu zaprezentowane zostanie podejście alternatywne, które może okazać się ciekawym uzupełnieniem powyższego.



Rys. 5 *Chmura próbek rzutowana do dwóch wymiarów. (Źródło: [3]).*

2. Wyjaśnienie zagadnienia „*Rury Próbek*”.

Ogólna idea metody wizualizacji wielowymiarowej zwanej metodą „Rury próbek” pojawiła się podczas pracy nad dyskretyzacją danych wejściowych przed obróbką metodami zbiorów przybliżonych (*ang. rough sets*). Postawiono m.in. pytanie czy nie można w sposób wizualny przewidywać wstępnie istnienia zależności pomiędzy atrybutami.

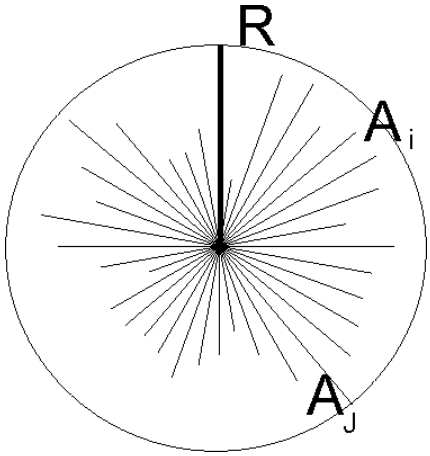
Jeżeli zatem rozważymy ogólny zestaw N atrybutów A, gdzie $a \in \{a_1,...,a_n\}$. W wyniku otrzymujemy zbiór D składający się z M atrybutów decyzyjnych, gdzie $d \in \{d_1,...,d_m\}$. Pozostaje tylko dodać, że poruszamy się w systemie rzeczywistym opisanym wynikami K eksperymentów. Co daje nam w efekcie rozszerzenie wektora A,D o wymiarowości n+m do macierzy opisanej wymiarowością (n+m)xk. Pokazuje to schematyczne ujęcie w poniższej tabeli.

Lp.	a_1	...	a_n	d_1	...	d_m
1		...			...	
...		...			...	
k		...			...	

Tab.1 Schematyczne pokazanie danych wejściowych.

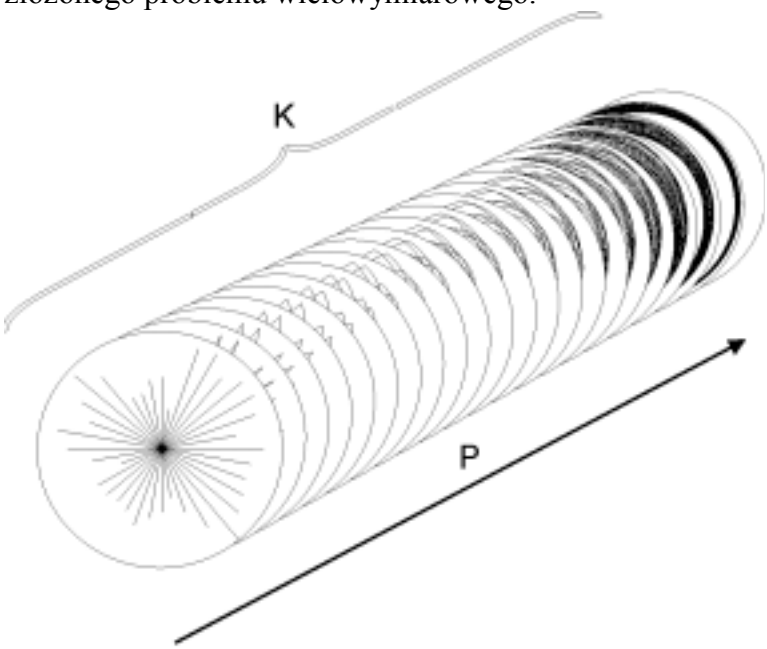
Spróbujmy na początek zwizualizować wektor będący zestawem danych z jednej próbki J. Jako przestrzeń wizualizacji przyjmijmy koło o promieniu R. Należy tu na wstępie dokonać przeskalowania zakresu atrybutów wektora a_i tak aby wartości poszczególnych jego elementów można było ukazać jako odcinki o wartości od 0 do R. Przy czym należy tak dobrać przekształcenie skalujące żeby uwzględnić także ewentualne wartości ujemne zbioru atrybutów a_i. Następnym krokiem jest umieszczenie każdego z atrybutów próbki na kole promieniście od środka. Mamy zatem teoretyczną możliwość przedstawienia próbki 360-cio wymiarowej co

wydaje się liczbą wystarczającą dla większości systemów rzeczywistych choć faktycznie ograniczoną. Przykładowy „plaster” jednej próbki z atrybutami o wartościach przeskalowanych w zależności od R przedstawia poniższy rysunek.



Rys.6 *Zestaw przeskalowanych wartości atrybutów.*

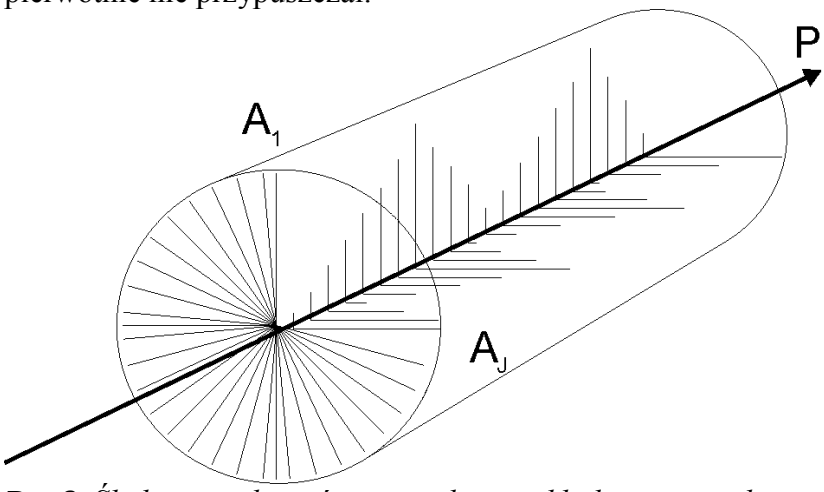
Jak zatem wyglądałby cały zestaw K próbek pomiarowych pokazany na jednym zbiorczym wykresie. Taką propozycję zawiera kolejny rysunek (Rys.7). Jest to walec mający na osi symetrii opisane próbki pomiarowe, natomiast na płaszczyźnie prostopadłej od osi symetrii pojawiają się kolejne okręgi („plastry”) z opisanymi na okręgach wartościami poszczególnych atrybutów. Należy przy tym założyć stałą pozycję atrybutu, tzn. stałe przypisanie mu stopnia z 360 możliwych. Jak zatem widać system wielowymiarowy (ok. 360D) został sprowadzony do systemu trójwymiarowego (3D), który na trwałe wszedł już do kanonu. Takie odwzorowanie zbioru w zbiór pozwala na łatwiejszą percepcję przez człowieka, złożonego problemu wielowymiarowego.



Rys 7 *Wiele próbek wielowymiarowych.*

Jakie dodatkowe korzyści możne osiągnąć badacz posługując się takim systemem odwzorowania. Otóż jak pokazuje rysunek Rys.8 można wizualnie starać się ocenić czy pomiędzy dwoma szeregami wartości atrybutów nie rysuje się jakaś zależność, którą w wyniku

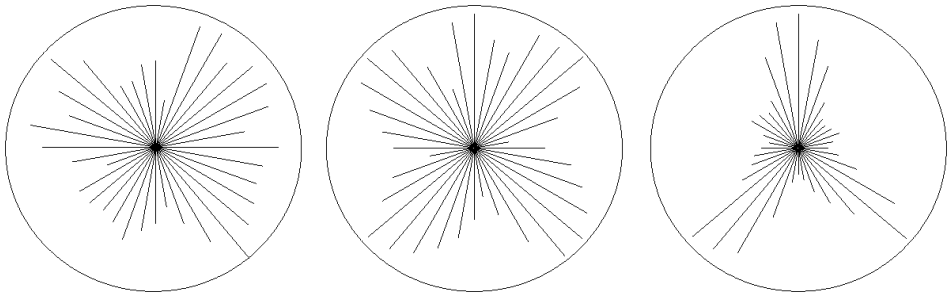
dalszych badań metodami zbiorów przybliżonych (*ang. rough sets*) czy też innymi metodami numerycznymi będzie można potwierdzić lub odrzucić jako zbyt słabą. Takie podejście powoduje zatem wpływ na swoisty determinizm algorytmów wykrywających zależności i pozwala zwrócić uwagę badacza na istnienie potencjalnych zależności, których istnienia pierwotnie nie przypuszczał.



Rys.8 *Śledzenie zależności pomiędzy przykładowymi atrybutami; A_1 oraz A_j*

3. Implikacje i wytyczenie kierunków dalszych badań.

Ciekawą zatem poznawczo wydaje się możliwość stworzenia oprogramowania, które bazując na przytoczonej wyżej idei pozwoli dowolnie zmieniać m.in. obserwowane pary atrybutów, kolejność przypisania do stopni itp. Warto przy tym zwrócić uwagę na możliwość analizowania kształtów poszczególnych rozgwiazd wpisanych w okręgi próbek. Jak widać na poniższych rysunkach mogą one przybierać różne kształty.

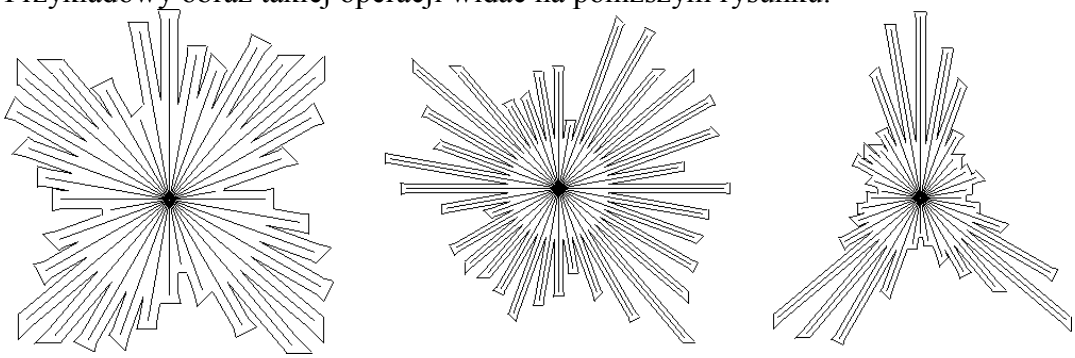


Rys. 9 *Przykładowe rozgwiazdy dla losowo dobranych próbek.*

Do dalszego wyjaśnienia pozostaje zatem czy uda się zaprzęgając w zagadnienia metody rozpoznawania obrazów [4], znajdować taki rdzeń rozgwiazdy, który będziemy potrafili powiązać jednoznacznie z przestrzenią decyzyjną rozpatrywanego eksperymentu. Dało by to możliwość wypracowania nowej, opartej przede wszystkim na przetwarzaniu obrazów metody wykrywania zależności układu atrybutów decyzyjnych (wyjściowych) od przestrzeni atrybutów wejściowych.

Warto przy tym rozważyć czy zastosowanie kolorów przy rysowaniu plastrów nie wniesie nowej jakości do powyższych rozważań dając nowe kierunki predykcji związane ze zjawiskiem mieszania się barw. Także ciekawym i wartym rozpatrzenia wydaje się podejście

polegające na badaniu samych obwiedni, jakie pozostaną po wymaskowaniu rozgwiazd. Przykładowy obraz takiej operacji widać na poniższym rysunku.



Rys. 10 *Przykładowe obwiednie na rozgwiazdach.*

Być może stosując wspomniane już metody rozpoznawania uda się sformułować jednoznaczne zakresy pomiędzy którymi powinna znaleźć się próbka w formie obejmy aproksymowanej z rozgwiazdy aby można ją było jednoznacznie przypisać do danej grupy decyzyjnej.

Literatura

1. A.Szustalewicz, P.Zuchowski, Wizualizacja wielowymiarowych danych metoda 'Parallel Coordinate Plots', program komputerowy w Delphi, *Raport Instytutu Informatyki Uniwersytetu Wroclawskiego* No. 1, styczen 2001, 7 str.
2. C.V.Jones, Parallel coordinates, Chesapeake Supply Chain Division, New Providence, NJ 07974, 1998. http://catt.bus.okstate.edu/jones98/parallel.html
3. A.Szustalewicz,Visualisation of multivariate data using parallel coordinate plots and Kohonen's SOM's. Which is better?, *Advanced Computer Systems, part I, Eighth International Conference, ACS'2001, Mielno, Poland, October 17--19, 2001*, Technical University of Szczecin, pp 89--98.
4. G.Kuchariew, Przetwarzanie i analiza obrazów cyfrowych, Wydawnictwo Uczelniane Politechniki Szczecińskiej, Szczecin 1998.
5. A.Alexakis, M. Deftereos, Y. Samiotakis, Visualization techniques in statistics. The parallel coordinates visualizer (PARCOVI), In: Ph. Nanopoulos, *NTTS'98 International Seminar on New Techniques and Technologies for Statistics, Sorrento, Italy 4-6 November 199*8, Contributed Papers, pp 19–24.
6. A.Inselberg, N-dimensional graphics, part I – lines and hyperplanes. Technical Report G320-2711, IBM Scientific Center, 9045 Lincoln Boulevard, Los Angeles (CA), 900435, 1981.
7. *T.Avidan and S.Avidan,* ParallAX – A data mining tool based on parallel coordinates, Computational Statistics 14 (1999) 1, 79-89
8. A.Inselberg, Don't panic...just do it in parallel! Computational Statistics 14 (1999) 1,53-77