STUDIA INFORMATICA 2014
Volume 35 Number 2 (116)

Łukasz PAŚKO, Galina SETLAK
Politechnika Rzeszowska, Zakład Informatyki

# OCENA SEGMENTACJI RYNKU ZA POMOCĄ MIAR JAKOŚCI GRUPOWANIA DANYCH

**Streszczenie**. Celem niniejszego artykułu jest przedstawienie miar służących do badania jakości grupowania danych i zastosowanie tych miar do oceny segmentacji rynku. W wykonanych badaniach analizowano dane dotyczące rynków zbytu przedsiębiorstwa produkującego wyroby gospodarstwa domowego. Segmentację rynku przeprowadzono z wykorzystaniem sieci neuronowych Kohonena. W pracy przedstawiono wyniki grupowania danych oraz ich ocenę. Wnioski na temat jakości utworzonych klastrów są próbą ogólnej oceny przeprowadzonej segmentacji rynku.

**Słowa kluczowe**: eksploracja danych, grupowanie danych, ocena jakości grupowania, sieci neuronowe Kohonena

# EVALUATION OF MARKET SEGMANTATION USING MEASURES OF DATA CLUSTERING QUALITY

**Abstract**. The purpose of this paper is to present the measures used to evaluate the quality of data clustering and apply them to assess market segmentation. In the analysis the data of manufacturing companies that producing household products was used. The market segmentation was carried out using Kohonen neural network. This paper describes results of the clustering and evaluation of the clusters. The conclusions on the quality of clusters are attempt to overall assessment of the market segmentation.

**Keywords**: data mining, data clustering, evaluation of data clustering, Kohonen neural networks

## 1. Wstęp

Jednym z istotniejszych działań podejmowanych przez przedsiębiorstwo jest obserwacja rynków zbytu, na których ono funkcjonuje. Polega to m.in. na odkrywaniu charakterystyki

i potencjalnych możliwości rynków, a także na ustalaniu najkorzystniejszej dla nich strategii stymulacji zbytu. Inną ważną kwestią jest poszukiwanie nowych rynków zbytu, mogących przynieść przedsiębiorstwu ewentualne korzyści [15]. Wszystkie te działania wymagają nie tylko intuicji menedżerów, lecz muszą być także poparte odpowiednimi badaniami rynkowymi. Przykładem takiego badania jest segmentacja rynku.

Podstawowy cel segmentacji to poznanie potrzeb klientów tworzących rynek zbytu [14, 15]. Literatura wyróżnia segmentacje rynku opisową i predykcyjną. W przeprowadzonych badaniach, które opisano szczegółowo w pracy [18], dokonano najpierw grupowania danych za pomocą sieci Kohonena, co stanowiło segmentację opisową. Następnie, z wykorzystaniem modeli drzew decyzyjnych, przeprowadzono klasyfikację danych odpowiadającą segmentacji predykcyjnej.

Celem niniejszego artykułu jest uzupełnienie wcześniejszych analiz o ocenę grupowania danych. Przedstawiono tutaj miary służące do badania jakości grupowania. Sklasyfikowano powyższe miary oraz zastosowano każdą z nich do oceny rezultatów grupowania wykonanego za pomocą sieci Kohonena. Przeprowadzone grupowanie danych miało na celu zrealizowanie opisowej segmentacji rynku, dlatego podjęto również próbę wykorzystania miar jakości grupowania do ogólnej oceny wykonanej segmentacji.

## 2. Opis segmentacji rynku

Analizy opisane w artykule [18], które w tej pracy poddano ocenie, miały na celu wspomaganie segmentacji rynku zbytu sprzętu gospodarstwa domowego. Analizowany zbiór danych opracowano na podstawie badań marketingowych i rynkowych w latach 2003-2005. W zbiorze zebrano dane dotyczące cech charakterystycznych 194 dostępnych na rynku odkurzaczy, które są obiektami analizy.

Każdy produkt został opisany za pomocą dwunastu atrybutów, odgrywających rolę zmiennych niezależnych. Oprócz tych atrybutów w zbiorze danych została zapisana jedna zmienna zależna o nazwie CLASS. Zawiera ona informację o tym, do jakiego segmentu rynku należy każdy z produktów. Segmenty te ustalono na podstawie wspomnianych badań rynkowych i oznaczono je etykietami $\{m_1, m_2, m_3, m_4\}$.

Jak już wspomniano, przeprowadzone analizy polegały na dokonaniu opisowej i predykcyjnej segmentacji rynku. W niniejszym artykule przedstawiono ocenę jakości grupowania danych, wykorzystanego w segmentacji opisowej, dlatego omówienie segmentacji predykcyjnej zostanie w tej pracy pominięte.

Grupowanie danych za pomocą sieci Kohonena zrealizowano z wykorzystaniem oprogramowania *STATISTICA Neural Networks*. Sieć Kohonena, nazywana samoorganizującym

się odwzorowaniem cech (ang. *Self-Organizing Feature Maps – SOFM*), składa się z dwóch warstw neuronów. Pierwsza warstwa to neurony wejściowe, których zadaniem jest tylko przekazanie danych wejściowych do wszystkich neuronów warstwy drugiej (wyjściowej). Druga warstwa jest najważniejszym elementem sieci Kohonena, ponieważ jednocześnie pełni funkcję obliczeniową i prezentuje wyniki grupowania. Neurony są tutaj rozmieszczone najczęściej na kształt prostokątnej siatki, nazywanej mapą topologiczną (ang. *topological map*). W wyniku uczenia sieci każdemu obiektowi, który poddano grupowaniu, jest przyporządkowany neuron zwycięski. Obiekty, mające zwycięzców położonych blisko siebie na mapie topologicznej, są do siebie podobne i tworzą poszukiwane grupy [9, 10, 11].

Podczas segmentacji opisowej atrybut CLASS był niewidoczny dla sieci Kohonena. Potraktowano go jako atrybut porównawczy, z którym zostaną skonfrontowane grupy zidentyfikowane przez sieć. Grupowanie danych rozpoczęto od utworzenia i nauczenia kilkunastu sieci Kohonena, różniących się wielkością mapy topologicznej. Szczegółowe informacje na temat wykorzystanych sieci oraz sposobu ich uczenia przedstawiono w pracy [18].

Do dalszej analizy wybrano jedną sieć, której mapa topologiczna pozwalała najłatwiej zauważyć skupiska podobnych do siebie obiektów. Zidentyfikowane skupiska oznaczono etykietami $\{c_1, c_2, c_3, c_4\}$. Ostatni krok analizy polegał na przydzieleniu nowego segmentu rynku każdemu produktowi. Wykonano to, biorąc pod uwagę położenie produktu na mapie topologicznej. Informację o przynależności produktu do nowego segmentu umieszczono w drugiej zmiennej zależnej, której nadano nazwę CLUSTER.

Po zrealizowaniu segmentacji opisowej zbiór danych wzbogacił się o drugą zmienną zależną. Dzięki temu każdy produkt miał przydzielone dwa segmenty rynku:

- CLASS = $\{m_1, m_2, m_3, m_4\}$ – segment wynikający z badania marketingowego,
- CLUSTER = $\{c_1, c_2, c_3, c_4\}$ – segment ustalony przez sieć Kohonena.

## 3. Ocena jakości grupowania

Grupowanie, inaczej klasteryzacja (ang. *clustering*), ma na celu zidentyfikowanie naturalnych grup, nazywanych skupiskami lub klastrami, występujących w zbiorze danych. W wyniku grupowania obiekty o podobnych cechach powinny zostać umieszczone w tej samej grupie, a obiekty różne od siebie – w innych grupach [8]. Grupowanie dzieli zbiór obiektów na podzbiory (grupy) przy uwzględnieniu cechy charakterystycznych wykrytych podczas dokonywania podziału [2, 6, 20, 21].

Proces oceny uzyskanych wyników grupowania jest w literaturze nazywany badaniem jakości grupowania (ang. *cluster validity*). Wszystkie metody służące do walidacji struktury uzyskanych klastrów autorzy dzielą na (m.in. [1, 4, 5, 7, 13]):

- wzorcowe, wśród których wykorzystywane są metody oparte na:
  - wskaźnikach zewnętrznych (ang. *external validation*),
  - wskaźnikach względnych (ang. *relative validation*);
- bezwzorcowe, w przypadku których stosowane są wskaźniki wewnętrzne (ang. *internal validation*), dzielące się na:
  - miary spójności klastrów (ang. *measure of cluster cohesion*),
  - miary separacji klastrów (ang. *measure of cluster separation*).

Do zastosowania metod wzorcowych niezbędne jest posiadanie wzorca idealnej struktury klastrów. Uzyskany wynik grupowania jest porównywany z wzorcową strukturą grup i na tej podstawie ocenia się przeprowadzone grupowanie. Gdy wzorcem są klastry zaproponowane przez eksperta w danej dziedzinie lub pochodzące z innych, zewnętrznych źródeł wiedzy, wówczas mamy do czynienia ze wskaźnikami opartymi na kryterium zewnętrznym. Natomiast gdy wyniki grupowania są porównywane z wzorcową strukturą klastrów, którą uzyskano za pomocą tej samej techniki grupowania (ale przykładowo z użyciem innych parametrów algorytmu grupującego), wtedy mówi się o wskaźnikach biorących pod uwagę kryterium względne.

Metody bezwzorcowe generują ocenę rezultatów grupowania, korzystając tylko z informacji zawartych w zbiorze danych, który poddano grupowaniu. Wykorzystywane w tych metodach wskaźniki nazywane są wewnętrznymi, ponieważ wiedza na temat struktury klastrów jest wewnętrzna w stosunku do zbioru danych. Wśród metod wewnętrznych wyróżnia się dwie główne miary: spójność i separowalność. Badając, jak bardzo podobne do siebie są obiekty w danym klastrze, bazuje się na miarach spójności skupisk, natomiast sprawdzając, jak oddalone są od siebie poszczególne klastry, mówi się o miarach separacji skupisk.

Oprócz wymienionych wskaźników odrębną grupę metod stanowią miary pozwalające ustalić optymalną liczbę grup występujących w zbiorze danych. Wskaźniki te odgrywają bardzo istotną rolę, ponieważ jeśli zakładana w czasie grupowania liczba grup będzie się różnić od ich rzeczywistej liczby, to jakość zidentyfikowanych grup nigdy nie będzie odpowiednia. Miary wyznaczające liczbę skupisk występują w literaturze pod ogólną nazwą wskaźników jakości grupowania (ang. *cluster validity indices*) [3, 13, 16].

Zgodnie z zaprezentowaną klasyfikacją metod oceny rezultatów grupowania, w niniejszej pracy zastosowano zarówno metody wzorcowe, jak i bezwzorcowe. Sekcja 4 przedstawia wskaźniki wewnętrzne, mierzące spójność i separowalność skupisk. Wyniki tych miar poprzedzono dodatkowo opisem dwóch metod, pozwalających na sprawdzenie istnienia skupisk w zbiorze danych oraz na ustalenie liczby tych skupisk. W sekcji 5 opisano walidację zidentyfikowanych grup, wykorzystując zewnętrzne informacje o klastrach wzorcowych, ustalonych w wyniku badań rynkowych.

## 4. Wyniki badania klastrów z wykorzystaniem metod bezwzorcowych

Na początku niniejszej sekcji przedstawiono wykorzystanie statystyki Hopkinsa do ustalenia, czy zbiór danych poddany grupowaniu zawiera naturalne skupiska przypadków. Następnie, za pomocą błędu kwantyzacji wektorowej, podjęto próbę wyznaczenia optymalnej liczby klastrów. Główna część tej sekcji to miary jakości klastrów otrzymanych w wyniku grupowania siecią Kohonena. W poniższych analizach zbiór tych klastrów oznaczono jako $\boldsymbol{C} = \{c_1, c_2, c_3, c_4\}$.

Zgodnie z założeniem metod bezwzorcowych do oceny skupisk zidentyfikowanych przez sieć Kohonena nie wykorzystywano zewnętrznych informacji na temat grupowanych obiektów. Jednak dla porównania wykonano analogiczne badania na oryginalnych klasach znajdujących się w zbiorze danych i ustalonych za pomocą badań rynkowych. Klasy te potraktowano jak skupiska, a ich zbiór przyjęto oznaczać jako $\boldsymbol{M} = \{m_1, m_2, m_3, m_4\}$. Zatem dokonano tutaj podwójnej oceny, badając przynależność obiektów do klastrów zapisanych zarówno w zmiennej CLUSTER ($\boldsymbol{C}$), jak i w zmiennej CLASS ($\boldsymbol{M}$). Taki sposób analizy ma ułatwić końcowe porównanie segmentów rynku znalezionych podczas grupowania siecią Kohonena z segmentami pochodzącymi z badań rynkowych.

Miary jakości grupowania traktują każdy obiekt ze zbioru danych jako wektor $\boldsymbol{x}$, stąd mówimy, że analizowany zbiór danych $\boldsymbol{X}$ jest złożony z $n = 194$ wektorów. Dwie przedstawione wyżej struktury klastrów, $\boldsymbol{C}$ i $\boldsymbol{M}$, w pełni pokrywają zbiór $\boldsymbol{X}$, co oznacza, że każdy wektor $\boldsymbol{x}_n$ należy dokładnie do jednego z klastrów $\boldsymbol{C}$ oraz do jednego z klastrów $\boldsymbol{M}$.

Po znormalizowaniu i przekodowaniu zmiennych niezależnych każdy wektor opisano za pomocą $N = 14$ parametrów. W zrealizowanych badaniach podstawą wielu zastosowanych wskaźników jest odległość między wektorami $\boldsymbol{x}_i$ oraz $\boldsymbol{x}_j$, którą oznaczono jako $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Do jej obliczenia przyjęto miarę nazywaną odległością euklidesową, wyrażoną wzorem:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{k=1}^{N} \left(x_{ik} - x_{jk}\right)^2} \,. \tag{1}$$

### 4.1. Sprawdzenie istnienia naturalnych skupisk w zbiorze danych (test Hopkinsa)

Do przeprowadzenia testu Hopkinsa wybrano ze zbioru danych $p$ = 40 przypadków, z których utworzono zbiór $T$. Następnie wygenerowano taką samą liczbę przypadków o rozkładzie losowym, tworząc zbiór $L$. W kolejnym kroku znaleziono dla wszystkich przypadków ze zbiorów $T$ i $L$ najbliższego sąsiada w zbiorze oryginalnym. Po zidentyfikowaniu najbliższego sąsiada ustala się odległość od niego. Obliczono tutaj dwie wartości: $u_i$, oznaczającą odległość $i$-tego wektora ($i = 1, 2, \ldots, p$) ze zbioru $L$ od najbliższego sąsiada ze zbioru orygi-

nalnego, oraz $w_i$, która stanowi odległość $i$-tego wektora ze zbioru $T$ od najbliższego sąsiada z analizowanego zbioru. Dla tak zdefiniowanych wartości statystyka Hopkinsa ma postać:

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}. \tag{2}$$

Tabela 1

Odległości wyznaczane dla testu Hopkinsa

| Zbiór | $p$ | Odległość od najbliższego sąsiada | | | $\sum_{i=1}^{p} w_i$ | $\sum_{i=1}^{p} u_i$ |
|---|---|---|---|---|---|---|
| | | minimalna | maksymalna | średnia | | |
| $T$ | 40 | $1{,}11 \cdot 10^{-6}$ | 2,21 | 0,08 | **3,22** | - |
| $L$ | 40 | 1,01 | 5,16 | 3,33 | - | **133,39** |

Wyniki pomiarów odległości dla testu Hopkinsa przedstawiono w tabeli 1. Rezultat statystyki Hopkinsa wynosi $H \approx$ **0,023**. Jeśli jej wartość byłaby bliska 0,5, oznaczałoby to, że zbiór oryginalnych wektorów $T$ nie różni się zasadniczo od zbioru losowego $L$. Wniosek w takiej sytuacji to brak naturalnych skupisk. Wynik zbliżony do 0 lub 1 mówi, że w zbiorze występują naturalne skupiska obiektów, co stwierdzono w badanym zbiorze danych [16].

### 4.2. Ustalenie optymalnej liczby klastrów

Inny problem pojawiający się w zagadnieniu grupowania danych to ustalenie liczby klastrów występujących w zbiorze. Wykorzystano do tego miarę błędu kwantyzacji wektorowej, wyznaczoną dla różnej liczby klastrów $K$. Do jej obliczenia wymagane jest znalezienie centrum każdego z rozpatrywanych klastrów. Wektor centralny klastra $k$, stanowiący szukane centrum, to średnia wszystkich wektorów znajdujących się w klastrze $k$:

$$\boldsymbol{c}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{x}_i , \tag{3}$$

gdzie $n_k$ jest liczbą wektorów w klastrze $k$.

Błąd kwantyzacji wektorowej można przedstawić w postaci sumarycznej (wzór (4)) lub jednostkowej (wzór (5)):

$$E_q = \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in k_i} d^2(\boldsymbol{x}, \boldsymbol{c}_i), \tag{4}$$

$$e_q = \frac{1}{n} \sum_{i=1}^{K} \sum_{\boldsymbol{x} \in k_i} d^2(\boldsymbol{x}, \boldsymbol{c}_i). \tag{5}$$

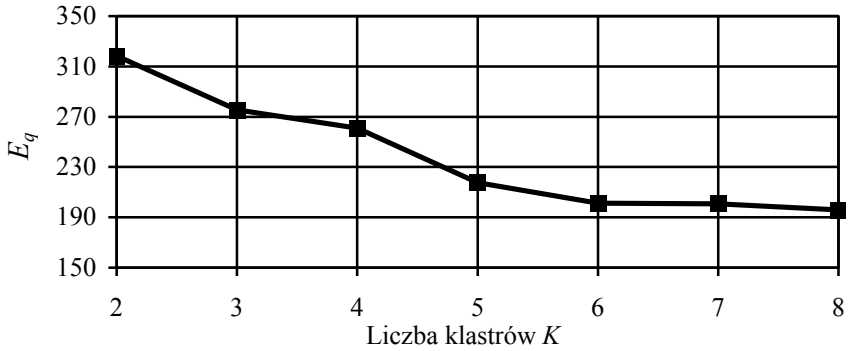
Im mniejsza jest wartość błędu, tym wektory położone są bliżej centrów skupisk, do których należą [16].

W badaniach przeanalizowano różne układy skupisk na mapie topologicznej sieci Kohonena. Wybrano kilka struktur skupisk, w których liczba klastrów była równa $K = 2, 3, \ldots, 8$. Wyniki przedstawiono w tabeli 2.

Tabela 2

Błędy kwantyzacji dla różnej liczby klastrów

| Liczba klastrów $K$ | 2 | 3 | **4** | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| $E_q$ | 318,45 | 275,49 | **260,87** | 217,57 | 201,23 | 200,69 | 195,92 |
| $e_q$ | 1,64 | 1,42 | **1,34** | 1,12 | 1,04 | 1,03 | 1,01 |




Rys. 1. Sumaryczny błąd kwantyzacji w funkcji liczby klastrów
Fig. 1. Total quantization error as a function of the number of clusters

Na podstawie sumarycznego błędu kwantyzacji wektorowej $E_q$ sporządzono wykres zależności tego błędu od przyjętej liczby klastrów $K$, pokazany na rysunku 1. W celu ustalenia optymalnej liczby klastrów można posłużyć się powyższym wykresem. Szukaną liczbę wskazuje miejsce, w którym błąd $E_q$ stabilizuje się. W tym przypadku wykres sugeruje przyjęcie 6 klastrów jako liczby optymalnej. Metoda ta jest tylko przybliżeniem najlepszej liczby skupisk, dlatego w dalszych badaniach przyjęto, że liczba klastrów wynosi 4. Ma to związek z liczbą segmentów rynku ustalonych w wyniku badań rynkowych. Dzięki temu łatwiejsze będzie porównanie obu sposobów segmentacji rynku.

### 4.3. Miary rozproszenia klastrów

W tej sekcji przedstawiono trzy miary rozproszenia klastrów. Obliczenia wykonano dla skupisk sieci Kohonena ($c_i$) i grup ustalonych za pomocą badań rynkowych ($m_j$).

Pierwszym wskaźnikiem jest **średnie rozproszenie klastra $k$ przy uwzględnieniu odległości między jego wektorami**, co wyraża wzór:

$$\sigma_1(k) = \frac{1}{m} \sum_{\substack{\boldsymbol{x} \in k \\ \boldsymbol{y} \in k}} d^2(\boldsymbol{x}, \boldsymbol{y}), \tag{6}$$

gdzie $m = \dfrac{n_k(n_k - 1)}{2}$, natomiast $n_k$ jest liczbą wektorów w klastrze $k$.

Tabela 3

Wyniki obliczeń miary rozproszenia klastrów $\sigma_1$

| Klaster $c_i$ | $n_i$ | Suma odległości | $\sigma_1(c_i)$ | Klaster $m_j$ | $n_j$ | Suma odległości | $\sigma_1(m_j)$ |
|---|---|---|---|---|---|---|---|
| $c_1$ | 36 | 1804,70 | **2,86** | $m_1$ | 34 | 1656,20 | **2,95** |
| $c_2$ | 43 | 1757,64 | **1,95** | $m_2$ | 42 | 3075,56 | **3,57** |
| $c_3$ | 81 | 10148,91 | **3,13** | $m_3$ | 72 | 7829,60 | **3,06** |
| $c_4$ | 34 | 1515,26 | **2,70** | $m_4$ | 46 | 3598,15 | **3,48** |

Drugi wskaźnik mierzy **rozproszenie klastra** $k$ **na podstawie odległości jego wektorów od centrum** $\boldsymbol{c}_k$. Miarę tę można zapisać następująco:

$$\sigma_2(k) = \frac{1}{n_k} \sum_{\boldsymbol{x} \in k} d^2(\boldsymbol{x}, \boldsymbol{c}_k). \tag{7}$$

Tabela 4

Wyniki obliczeń miary rozproszenia klastrów $\sigma_2$

| Klaster $c_i$ | $n_i$ | Suma odległości | $\sigma_2(c_i)$ | Klaster $m_j$ | $n_j$ | Suma odległości | $\sigma_2(m_j)$ |
|---|---|---|---|---|---|---|---|
| $c_1$ | 36 | 50,13 | **1,39** | $m_1$ | 34 | 48,71 | **1,43** |
| $c_2$ | 43 | 40,88 | **0,95** | $m_2$ | 42 | 73,23 | **1,74** |
| $c_3$ | 81 | 125,30 | **1,55** | $m_3$ | 72 | 108,74 | **1,51** |
| $c_4$ | 34 | 44,57 | **1,31** | $m_4$ | 46 | 78,22 | **1,70** |

Trzecia miara rozproszenia jest wyrażona jako **średnica klastra** $k$. Jest to maksymalna odległość pomiędzy wektorami tworzącymi klaster, co przedstawia wzór:

$$D_k = \max_{\substack{\boldsymbol{x} \in k \\ \boldsymbol{y} \in k}} \left\{ d^2(\boldsymbol{x}, \boldsymbol{y}) \right\}. \tag{8}$$

Tabela 5

Wyniki obliczeń średnicy klastrów $D_k$

| Klaster $c_i$ | $D_{ci}$ | Klaster $m_j$ | $D_{mj}$ |
|---|---|---|---|
| $c_1$ | **7,22** | $m_1$ | **7,31** |
| $c_2$ | **6,18** | $m_2$ | **8,07** |
| $c_3$ | **7,00** | $m_3$ | **7,15** |
| $c_4$ | **6,01** | $m_4$ | **7,25** |

Badając rozproszenie skupisk, otrzymano informację o tym, jak bardzo oddalone od siebie są wektory tworzące skupisko. Obiekty w klastrach powinny być położone jak najbliżej siebie, tworząc tym samym spójne struktury [4]. Przyjmuje się, że im większe jest rozproszenie klastra (mniejsza spójność), tym mniejsze jest podobieństwo obiektów do niego przypisanych. Podstawowa miara rozproszenia klastra może być wyrażona jako wariancja jego obiektów, która powinna dążyć do minimum [4]. Wariancji odpowiada miara $\sigma_2$, przedstawiająca zróżnicowanie obiektów względem centrum klastra. Wyniki obliczeń zebrane w tabeli 4 pokazują mniejsze rozproszenie klastrów $c_i$. Wyjątkiem jest skupisko $c_3$. Odpowiadający mu

klaster $m_3$ jest bardziej skoncentrowany wokół swojego centrum, jednak różnice te nie są znaczące.

Bardziej złożona obliczeniowo miara $\sigma_1$ bada odległości pomiędzy każdą parą wektorów należących do danego skupiska. Porównując klastry $c_i$ oraz $m_j$, można zauważyć, że z wyjątkiem klastra $c_3$ miary rozproszenia $\sigma_1$ (tabela 3) są większe dla klastrów $m_j$. Widać to szczególnie na przykładzie skupisk $c_2$ i $m_2$, mających zbliżoną do siebie liczność.

Na podstawie średnicy klastra $D_k$ również można wnioskować, jakie jest zróżnicowanie obiektów w klastrze. Mierzy ona odległość w przestrzeni pomiędzy najbardziej różniącymi się od siebie obiektami danego skupiska. Tabela 5 wskazuje, że wszystkie skupiska $c_i$ mają mniejsze średnice od odpowiadających im skupisk $m_j$. Również klaster $c_3$ ma mniejszą średnicę od $m_3$. Zatem wyższe wartości miar $\sigma_1$ i $\sigma_2$ klastra $c_3$ w porównaniu z $m_3$ mogą wynikać raczej z jego dużej liczności aniżeli z większego rozproszenia obiektów.

### 4.4. Miary separacji między klastrami

Druga grupa wskaźników wewnętrznych to miary separacji między klastrami. Niniejsza sekcja zawiera trzy takie miary. Ich wartości mówią, jak bardzo oddalone od siebie są skupiska. Podobnie jak w badaniach rozproszenia, wszystkie obliczenia przeprowadzono dla klastrów zidentyfikowanych z wykorzystaniem sieci Kohonena ($c_i$) i dla grup ustalonych za pomocą badań rynkowych ($m_j$).

Pierwsza miara wyznacza **separowalność między klastrami $k_i$ i $k_j$ na podstawie rozkładu wektorów tworzących te klastry**, co przedstawia wzór:

$$s_1(k_i, k_j) = \frac{1}{n_{k_i} n_{k_j}} \sum_{\substack{\boldsymbol{x} \in k_i \\ \boldsymbol{y} \in k_j}} d^2(\boldsymbol{x}, \boldsymbol{y}). \tag{9}$$

Tabela 6

Wyniki obliczeń miary separacji $s_1$

| Klaster $c_i$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | Klaster $m_j$ | $m_1$ | $m_2$ | $m_3$ | $m_4$ |
|---|---|---|---|---|---|---|---|---|---|
| $c_1$ | 0 | 3,56 | 4,59 | 7,46 | $m_1$ | 0 | 3,59 | 4,11 | 6,71 |
| $c_2$ | 3,56 | 0 | 3,61 | 6,51 | $m_2$ | 3,59 | 0 | 3,58 | 5,43 |
| $c_3$ | 4,59 | 3,61 | 0 | 4,94 | $m_3$ | 4,11 | 3,58 | 0 | 4,77 |
| $c_4$ | 7,46 | 6,51 | 4,94 | 0 | $m_4$ | 6,71 | 5,43 | 4,77 | 0 |

Drugi sposób obliczenia separacji bierze pod uwagę **odległości między centrami klastrów**. Dla klastrów $k_i$ i $k_j$ tak zdefiniowaną miarę wyraża wzór:

$$s_2(k_i, k_j) = d^2(\boldsymbol{c}_{k_i}, \boldsymbol{c}_{k_j}). \tag{10}$$

Tabela 7

Wyniki obliczeń miary separacji $s_2$

| Klaster $c_i$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | Klaster $m_j$ | $m_1$ | $m_2$ | $m_3$ | $m_4$ |
|---|---|---|---|---|---|---|---|---|---|
| $c_1$ | 0 | 1,21 | 1,65 | 4,76 | $m_1$ | 0 | 0,41 | 1,16 | 3,58 |
| $c_2$ | 1,21 | 0 | 1,11 | 4,25 | $m_2$ | 0,41 | 0 | 0,33 | 1,99 |
| $c_3$ | 1,65 | 1,11 | 0 | 2,08 | $m_3$ | 1,16 | 0,33 | 0 | 1,56 |
| $c_4$ | 4,76 | 4,25 | 2,08 | 0 | $m_4$ | 3,58 | 1,99 | 1,56 | 0 |

Trzeci wskaźnik jest wyrażony jako **najkrótsza odległość pomiędzy wektorami klastrów $k_i$ oraz $k_j$**. Obliczenia tej miary wykonano na podstawie wzoru:

$$d\left(k_i, k_j\right) = \min_{\substack{\boldsymbol{x}\in k_i \\ \boldsymbol{y}\in k_j}} \left\{d^2\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\}. \tag{11}$$

Tabela 8

Wyniki obliczeń miary separacji $d(k_i, k_j)$

| Klaster $c_i$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | Klaster $m_j$ | $m_1$ | $m_2$ | $m_3$ | $m_4$ |
|---|---|---|---|---|---|---|---|---|---|
| $c_1$ | 0 | 1,004 | 1,005 | 4,046 | $m_1$ | 0 | 0,006 | 0,016 | 2,072 |
| $c_2$ | 1,004 | 0 | 1,000 | 3,085 | $m_2$ | 0,006 | 0 | 0,004 | 0,131 |
| $c_3$ | 1,005 | 1,000 | 0 | 1,004 | $m_3$ | 0,016 | 0,004 | 0 | 0,006 |
| $c_4$ | 4,046 | 3,085 | 1,004 | 0 | $m_4$ | 2,072 | 0,131 | 0,006 | 0 |

Powyższe miary separacji w różny sposób ujmują odległości pomiędzy klastrami w przestrzeni danych. Jednak wszystkie te wskaźniki łączy jedna zasada: im większa jest separacja dwóch skupisk, tym mniejsze jest ich podobieństwo, a więc obiekty umieszczone w jednym skupisku charakteryzują się większą odmiennością od obiektów ze skupiska drugiego [3, 8]. Dlatego w praktyce wyraźnie odseparowane klastry uważa się za optymalne.

Dwie najpopularniejsze metody pomiaru separacji zostały oznaczone jako $s_1$ i $s_2$. Wskaźnik $s_1$ bada odległości pomiędzy każdą parą wektorów, które należą do dwóch rozpatrywanych klastrów, natomiast wskaźnik $s_2$ mierzy dystans, jaki w przestrzeni danych dzieli ich punkty centralne. Wyniki tych miar, zebrane w tabelach 6 i 7, wskazują, że klastry znalezione przez sieć Kohonena są bardziej odseparowane od siebie w porównaniu z klastrami $m_j$. Taką zależność widać pomiędzy każdą parą skupisk. Wyjątek stanowi miara $s_1$ dla grup $c_1$ i $c_2$. Odpowiadające im klastry $m_1$ i $m_2$ dzieli mniejszy dystans, a różnica w odległościach wnosi jedynie 0,03.

Miarę $d(k_i, k_j)$ zaliczono do wyznaczników separacji, ponieważ pokazuje ona, jak blisko „sąsiadują" ze sobą dwa skupiska. Miara ta określa odległość pomiędzy dwoma najbardziej podobnymi do siebie obiektami z klastrów $k_i$ i $k_j$, tak więc na jej podstawie można stwierdzić, która struktura klastrów zawiera wyraźniej oddalone skupiska. Patrząc na wyniki przedstawione w tabeli 8, widać, że struktura klastrów $c_i$ zawiera znacznie bardziej odseparowane od siebie grupy. Niewielkie odległości pomiędzy najbliższymi obiektami skupisk $m_j$ mogą sugerować nakładanie się na siebie tych klastrów. Oznaczałoby to, że w dwóch różnych grupach,

ustalonych w czasie badań rynkowych, znajdują się obiekty o prawie identycznych cechach. Nie jest to korzystne zjawisko, dlatego klastry $c_i$ można uznać za optymalne.

### 4.5. Miary jakości segmentacji całej przestrzeni danych

Sekcja ta przedstawia cztery następujące miary jakości, z których dwie bazują na rozproszeniu klastrów, a dwie kolejne odnoszą się do separacji między klastrami [16]:

- rozproszenie całkowite klastrów $k_i$ w całej przestrzeni danych dla miary $\sigma_1$:

$$r(\sigma_1) = \sum_{i=1}^{K} \sigma_1(k_i), \tag{12}$$

- rozproszenie całkowite klastrów $k_i$ w całej przestrzeni danych dla miary $\sigma_2$:

$$r(\sigma_2) = \sum_{i=1}^{K} \sigma_2(k_i), \tag{13}$$

- miara separacji międzyklastrowej w całej przestrzeni danych dla miar $s_1$ oraz $\sigma_1$:

$$s(s_1) = \sum_{\substack{i,j=1 \\ j \neq i}}^{K} \frac{s_1(k_i, k_j)}{\sigma_1(k_i)}, \tag{14}$$

- miara separacji międzyklastrowej w całej przestrzeni danych dla miary $s_2$:

$$s(s_2) = \sum_{\substack{i,j=1 \\ j \neq i}}^{K} s_2(k_i, k_j). \tag{15}$$

Tabela 9

Wyniki miar jakości segmentacji całej przestrzeni danych

| Klastry | $r(\sigma_1)$ | $r(\sigma_2)$ | $s(s_1)$ | $s(s_2)$ |
|---|---|---|---|---|
| $c_i$ | 10,64 | 5,20 | 23,67 | 30,14 |
| $m_j$ | 13,06 | 6,39 | 17,34 | 18,05 |

W porównaniu z miarami przedstawionymi w sekcjach 4.3 i 4.4 te wskaźniki mogą jeszcze łatwiej ocenić jakość grupowania, ponieważ odnoszą się do całej przestrzeni danych. Oznacza to, że nie rozpatrują one rozproszenia i separacji pojedynczych skupisk, ale badają całościowo obie struktury klastrów: $c_i$ oraz $m_j$.

Przyjmuje się, że bardziej optymalna jest ta struktura skupisk, w której odległości pomiędzy klastrami są większe, a rozproszenie klastrów jest mniejsze. Taka sytuacja jest pożądana, gdyż wtedy w poszczególnych skupiskach znajdują się obiekty bardziej do siebie podobne, zatem jakość klasteryzacji jest większa [3, 8].

Po skonfrontowaniu wyników powyższych miar, zebranych dla obu badanych struktur klastrów w tabeli 9, potwierdzają się spostrzeżenia wynikające z poprzednich sekcji. Podsu-

mowując, można stwierdzić, że na podstawie rezultatów miar separacji i rozproszenia, klastry oznaczone jako $c_i$ można uznać za korzystniejsze w porównaniu z klastrmi $m_j$.

## 5. Wyniki badania klastrów na podstawie wskaźników zewnętrznych

W tej sekcji przedstawiono miary, których celem jest zbadanie relacji zachodzących między klastrami $c_i$, zidentyfikowanymi przez sieć Kohonena, a oryginalnymi klasami $m_j$, które zostały przypisane wszystkim obiektom w wyniku badań rynkowych. Ponieważ niniejsza sekcja prezentuje wskaźniki wzorcowe, przyjęto tutaj, że klastry $m_j$ są wzorcową strukturą skupisk.

Najczęściej spotykane miary wzorcowe to:

- **precyzja klasowa** (ang. *precision*), obliczana jako:

$$pr(c,m)=\frac{n_{cm}}{n_c}, \tag{16}$$

gdzie $n_{cm}$ to liczba elementów klastra wzorcowego $m$ należących do klastra badanego $c$, a $n_c$ to liczba elementów klastra badanego $c$,

- **miara odtworzeniowa** (ang. *recall*), wyznaczana na podstawie wzoru:

$$r(c,m)=\frac{n_{cm}}{n_m}, \tag{17}$$

gdzie $n_m$ to liczba elementów klastra wzorcowego $m$.

Tabela 10

Wyniki obliczeń precyzji klasowej i miary odtworzeniowej

| $pr(c, m)$ | | | | | $r(c, m)$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Klastry | $c_1$ | $c_2$ | $c_3$ | $c_4$ | Klastry | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
| $m_1$ | 0,75 | 0,16 | 0 | 0 | $m_1$ | 0,79 | 0,21 | 0 | 0 |
| $m_2$ | 0,25 | 0,49 | 0,15 | 0 | $m_2$ | 0,21 | 0,50 | 0,29 | 0 |
| $m_3$ | 0 | 0,35 | 0,65 | 0,12 | $m_3$ | 0 | 0,21 | 0,73 | 0,06 |
| $m_4$ | 0 | 0 | 0,20 | 0,88 | $m_4$ | 0 | 0 | 0,35 | 0,65 |

Powyższe miary przedstawiają, w jaki sposób rozkładają się klastry wzorcowe w klastrach badanych [17]. Precyzja klasowa pokazuje, jaki odsetek wektorów skupiska $c_i$ pochodzi z poszczególnych klastrów $m_j$. Podczas losowania pewego obiektu ze skupiska $c_i$ precyzja $pr(c_i, m_j)$ wskaże prawdopodobieństwo tego, że wybrany zostanie element należący do $m_j$ [12], natomiast miara odtworzeniowa mówi, jaka część wektorów klastra $m_j$ została przypisana do każdego ze skupisk $c_i$ [17]. W sytuacji gdy występuje pełna zgodność pomiędzy obiema strukturami klastrów, miary $pr(c_i, m_j)$ i $r(c_i, m_j)$ dla $i = j$ mają wartość 1, a dla $i \neq j$ wynoszą 0.

Wyniki precyzji klastrowej zawarte w tabeli 10 wskazują, że większość obiektów należących do skupisk $c_1$, $c_3$, $c_4$ pochodzi z odpowiadających im klastrów $m_j$. Skupisko $c_2$ zawiera najbardziej różnorodne obiekty, z których niecała połowa jest przypisana do klastra $m_2$. Z drugiej część tabeli wynika, że najlepszą miarą odtworzeniową charakteryzuje się klaster $m_1$, którego 79% przypadków zalicza się do właściwego mu skupiska $c_1$. Wielkości obliczone dla klastra $m_2$ potwierdzają dużą niezgodność jego elementów ze skupiskiem $c_2$.

Obie miary pokazują, że istotna część obiektów nie przynależy do odpowiadających sobie klastrów. Świadczą o tym znaczne wartości pozadiagonalne w lewej i prawej części tabeli 10. Potwierdza to fakt, że struktura skupisk $c_i$ jest odmienna od struktury $m_j$.

Kolejne miary wzorcowe to:

- **entropia klastra**, opisana za pomocą zależności:

$$e_c = -\sum_{m=1}^{K} \frac{n_{cm}}{n_c} \log_2 \left( \frac{n_{cm}}{n_c} \right), \tag{18}$$

- **jednorodność klastra**, mierzona jako:

$$p_c = \max_m \left\{ \frac{n_{cm}}{n_c} \right\}. \tag{19}$$

Tabela 11
Wyniki obliczeń entropii i jednorodności klastrów

| Klaster | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| $e_c$ | 0,81 | 1,46 | 1,27 | 0,52 |
| $p_c$ | 0,75 | 0,49 | 0,65 | 0,88 |

Entropia przyjmuje zawsze wartość nieujemną, wyrażoną w bitach. Na wielkość entropii danego skupiska ma wpływ przynależność jego obiektów do klastrów wzorcowych. Jeśli w skupisku $c_i$ znajdowałyby się obiekty pochodzące tylko z jednego klastra $m_j$, wówczas entropia $c_i$ byłaby równa 0 [12]. W przypadku badanych skupisk maksymalna wartość $e_c$ może wynieść 2. Będzie tak, gdy w skupisku $c_i$ znajdzie się taka sama liczba elementów przypisanych do każdego z czterech klastrów $m_j$. Z kolei jednorodność skupiska jest definiowana jako maksymalna wartość precyzji klasowej, obliczonej dla poszczególnych klastrów wzorcowych. Z rezultatów tych miar (tabela 11) wynika, że najbardziej jednorodny jest klaster $c_4$. Potwierdzają to najmniejsza wartość entropii i największa miara $p_c$ tego klastra.

Powyższe dwie miary można obliczyć również dla całego zbioru danych. Entropia całkowita $e$ jest wyrażona jako suma entropii wszystkich skupisk, obliczona z wagą biorącą pod uwagę liczność każdego skupiska. W analogiczny sposób sumowane są wartości $p_c$ wszystkich klastrów, czego wynikiem jest jednorodność całkowita $p$. Rezultaty wynoszące $e = 1{,}09$ oraz $p = 0{,}66$ świadczą o znaczącym zróżnicowaniu skupisk $c_i$.

Ostatni wskaźnik jest nazywany **miarą *F* dla pojedynczego klastra**, którą wyznacza się według wzoru:

$$F(c,m)=\frac{2pr(c,m)\cdot r(c,m)}{pr(c,m)+r(c,m)}. \tag{20}$$

Tabela 12

Wyniki obliczeń miary *F*

| Klastry | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| $m_1$ | 0,77 | 0,18 | - | - |
| $m_2$ | 0,23 | 0,49 | 0,20 | - |
| $m_3$ | - | 0,26 | 0,69 | 0,08 |
| $m_4$ | - | - | 0,25 | 0,75 |

Wskaźnik *F*, definiowany jako średnia harmoniczna precyzji i odtworzenia, przyjmuje zawsze wartość z zakresu [0, 1]. Gdy zachodzi idealna zgodność klastra badanego z klastrem wzorcowym, wtedy miara *F* wynosi 1 [16]. Wyniki zebrane w tabeli 12 wskazują, że żaden klaster $c_i$ nie jest w pełni reprezentatywny dla elementów z odpowiadającego mu klastra $m_j$. Jest to potwierdzeniem wcześniejszych spostrzeżeń, świadczących o odmienności rezultatów grupowania siecią Kohonena od segmentów wyłonionych podczas analiz rynku.

## 6. Podsumowanie

Poszukiwanie naturalnych skupisk podobnych do siebie obiektów w zbiorze danych jest często pierwszym etapem odkrywania informacji zawartych w danych. Nie jest to łatwe zadanie, ponieważ grupowanie jest badaniem nieukierunkowanym, w czasie którego trudno znaleźć wskazówki mówiące, jak powinna wyglądać prawidłowa struktura skupisk i jaka ma być ich liczba. Klasteryzacja zwykle wymaga wiedzy *a priori* na temat zbioru danych, której na wstępnym etapie badań często brakuje. Grupując dane, należy zdecydować, ile skupisk występuje w zbiorze oraz czy otrzymana za pomocą algorytmu grupującego struktura klastrów odpowiada rzeczywistej. Dlatego istotną kwestią jest poznanie metod oceny zidentyfikowanych skupisk zarówno w kontekście wewnętrznym, jak i przy porównaniu ich z odrębną strukturą skupisk, stanowiącą wzorzec lub będącą wynikiem innej techniki grupowania. Metody oceny są szczególnie istotne, gdy rozpatrywana przestrzeń danych jest wielowymiarowa, przez co sprawdzenie poprawności klasteryzacji za pomocą wizualizacji danych jest niemożliwe.

Celem niniejszego artykułu była ocena rezultatów klasteryzacji, opisanej w artykule [18] i wykonanej przy użyciu sieci Kohonena. Zadaniem klasteryzacji było odnalezienie grup podobnych do siebie obiektów w zbiorze danych obejmującym wyroby gospodarstwa domowe-

go, tworzących ten sam segment rynku. Z racji tego, że każdy element zbioru miał już wstępnie przypisany segment rynku ustalony w wyniku badań rynkowych, postanowiono porównać obie struktury segmentów. Wnioski z porównania miały być próbą oceny segmentacji rynku wykonanej z zastosowaniem metod sztucznej inteligencji.

Zrealizowane zadanie grupowania było wielowymiarowym problemem eksploracji danych. Jego ocena opierała się na dwojakiego rodzaju miarach: wewnętrznych, oceniających rozproszenie i separację utworzonych skupisk, oraz zewnętrznych, porównujących zidentyfikowane skupiska z wzorcową strukturą klastrów. Jako wzorzec wykorzystano segmenty rynku będące wynikiem analiz marketingowych.

Porównując podobieństwo obiektów znajdujących się w obu badanych strukturach klastrów, można stwierdzić, że grupy znalezione przez sieć Kohonena składają się z bardziej jednorodnych elementów, natomiast wzorcowe segmenty rynku charakteryzowały się mniejszą spójnością, co przemawiało za gorszą jakością tej segmentacji. Odmienność otrzymanej struktury klastrów od struktury wzorcowej pokazały także miary zewnętrzne.

Oceniając jakość segmentacji rynku wykonanej z zastosowaniem sztucznej sieci neuronowej Kohonena, można przyjąć, że jest ona korzystniejsza niż segmentacja przeprowadzona za pomocą badań rynkowych.

Otrzymane klastry wykorzystano do klasyfikacji analizowanych wyrobów. Klasyfikatorami były modele drzew decyzyjnych, utworzone za pomocą algorytmów CART i CHAID. Wyniki zadania klasyfikacji danych opisano w artykule [18]. Kontynuując w przyszłości przeprowadzone badania, można ocenić zastosowane metody klasyfikacji. Literatura proponuje wiele miar jakości, które mogą być użyteczne w przypadku drzew decyzyjnych. Rezultaty tego badania przedstawiłyby w innym świetle analizowane tutaj struktury klastrów, pokazując, jak wpływają one na wynik klasyfikacji danych.

## BIBLIOGRAFIA

1. Brun M., Sima C., Hua J., Lowey J., Carroll B., Suh E., Dougherty E. R.: Model-based evaluation of clustering validation measures. Pattern Recognition, Vol. 40, No. 3, Elsevier, 2007, s. 807÷824.
2. Cios K., Pedrycz W., Świniarski R.: Data mining methods for knowledge discovery. Kluwer, Norwell, MA 1998.
3. Everitt B. S., Landau S., Leese M.: Cluster analysis. Wiley Publishing, New York 2009.
4. Halkidi M., Batistakis Y., Vazirgiannis M.: On clustering validation techniques. Journal of Intelligent Information Systems, Vol. 17, No. 2÷3, Springer, 2001, s. 107÷145.

5. Halkidi M., Vazirgiannis M.: Clustering validity assessment: finding the optimal partitioning of a data set. Proceedings IEEE International Conference on Data Mining, ICDM, 2001, s. 187÷194.
6. Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa 2005.
7. Jain A. K., Dubes R. C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Jersey 1988.
8. Jain A. K., Murty M. N., Flynn P. J.: Data clustering: a review. ACM Computing Surveys, Vol. 31, No. 3, 1999, s. 264÷323.
9. Kohonen T.: Self-organization and associative memory. Springer Verlag, Berlin 1989.
10. Kohonen T.: The self-organizing maps. Proceedings of the IEEE, 1990, Vol. 78, No. 9, s. 1464÷1480.
11. Larose D. T.: Odkrywanie wiedzy z danych. Wyd. Nauk. PWN, Warszawa 2006.
12. Meila M.: Comparing clusterings – an information based distance. Journal of Multivariate Analysis, Vol. 98, No. 5, 2007, s. 873÷895.
13. Migdał-Najman K.: Ocena jakości wyników grupowania – przegląd bibliografii. Przegląd Statystyczny, Vol. 58, no. 3÷4, 2011, s. 281÷299.
14. Migut G.: Zastosowanie technik analizy skupień i drzew decyzyjnych do segmentacji rynku. Materiały Seminarium StatSoft „Zastosowanie nowoczesnej analizy danych w marketingu i badaniach rynku", Kraków 2010.
15. Mynarski S.: Metody ilościowe i jakościowe badań rynkowych i marketingowych. StatSoft, Kraków 2010.
16. Osowski S.: Metody i narzędzia eksploracji danych. Wyd. BTC, Legionowo 2013.
17. Rokach L., Maimon O.: Clustering methods. Data Mining and Knowledge Discovery Handbook, Springer, 2005.
18. Setlak G., Paśko Ł.: Zastosowanie metod eksploracji danych do segmentacji rynków. Studia Informatica, Vol. 34, No. 2A (111), Gliwice 2013, s. 311÷323.
19. Stąpor K.: Automatyczna klasyfikacja obiektów. Wyd. Exit, Warszawa 2005.
20. Żurada J., Barski M., Jędruch W.: Sztuczne sieci neuronowe. PWN, Warszawa 1996.

Wpłynęło do Redakcji 30 stycznia 2014 r.

**Abstract**

The main goal of this paper is to present the measures of data clustering quality and to apply them to assess results of clustering. The assessment is a continuation of data analysis described in the paper [18], where market segmentation using data mining methods was

made. In the analysis the data set containing characteristics of household products was used. The first chapter introduces theoretical foundation on market segmentation. The market segmentation has been developed based on data clustering using Kohonen neural network. Therefore, second section describes its results. After clustering process, four clusters were identified. They have been treated as the market segments.

The third section concerns the assessing of clustering quality. In this place the classification of quality measures was introduced and each type of these measures was described. During the clustering validation, we focused on two of them: external and internal.

The fourth section firstly shows how can we check occurrence of natural groups in the data set. Next, we present a method used to determine optimal number of clusters. The main part of this section presents several indicators belonging to the internal validation. Before the clustering, each product has been assigned to one of the four market segments that were established during marketing research. All of presented indicators have been used to asses both clusters identified by Kohonen network and clusters from marketing research. Second part of the validation process is described in section five, where the external validation is shown.

The last section compares the quality of clusters found by Kohonen neural network with clusters identified during marketing research. The conclusions on the quality are attempt to overall assessment of the market segmentation.

**Adresy**

Galina SETLAK: Politechnika Rzeszowska, Zakład Informatyki, al. Powstańców Warszawy 8, 35-959 Rzeszów, Polska, gsetlak@prz.edu.pl.
Łukasz PAŚKO: Politechnika Rzeszowska, Zakład Informatyki, al. Powstańców Warszawy 8, 35-959 Rzeszów, Polska, lpasko@prz.edu.pl.