Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska część 2

Influence of the features selection method on the results of objects classification using environmental data on Polish voivodeships part 2

Danuta Tarka

Politechnika Białostocka, Wydział Zarządzania, Katedra Informatyki Gospodarczej i Logistyki

Abstract

The main aim of the paper is comparison of the diagnostics features selection methods influence on regional objects linear classification using as an example official enviromental data on polish voivodeships. Three methods and their variations were used. Those used methods were: Hellwig method, median Hellwig method, inverse matrix. Spearmn;s coefficient of correlation was used to compare results of rankings.

Keywords: empirical features selection methods, objects classification and ranking.

Wstęp

Badanie typu taksonomicznego składa się z kilku podstawowych etapów. Pierwszy z nich to określenie celu i zakresu badania, i jest to etap specyficzny dla każdego badania. Następnym etapem jest dobór cech do badania. Jest to jedna z kluczowych decyzji w tego typu badaniu. Walesiak (2006) stwierdza np., że jest to jedno z „(…) najważniejszych, a zarazem najtrudniejszych zagadnień. Od jakości zestawu zmiennych zależy bowiem wiarygodność ostatecznych wyników klasyfikacji i trafność podejmowanych na ich podstawie decyzji." Nie jest to jednak problem zbyt często podejmowany w badaniach praktycznych i wielu badaczy podejmując

analizę nie omawia szerzej problemu wyboru cech diagnostycznych użytych w analizie.

Dobór cech (zmiennych) do badania dzieli się zasadniczo, na dwa etapy: merytoryczny i formalny. W większości badań empirycznych autorzy podają zestaw zmiennych dobranych w oparciu o kryteria merytoryczne lub formalne (najczęściej za etap formalny przyjmują użycie współczynnika zmienności) bez szerszej dyskusji problemu. Jednak już na etapie analizy merytorycznej czyli własności jakie powinny mieć „dobre" cechy diagnostyczne nie ma pełnej zgodności wśród autorów[1]

Etap doboru formalnego jest jeszcze słabiej reprezentowany w literaturze. W polskiej literaturze szerzej zajmuje się tym problemem Walesiak (2005), podobnie jednak jak inni autorzy[2], analizuje problem od strony teoretycznej, dla zmiennych stochastycznych, tzn. zakłada typ rozkładu i za pomocą symulacji analizuje wyniki klasyfikacji. Podstawowym jednak problemem w badaniach realnych procesów społeczno- gospodarczych jest to, iż nie znamy rzeczywistych rozkładów zmiennych oraz to, że mając do czynienia z cechami empirycznymi opisującymi[3] całą zbiorowość nie możemy przyjąć założenia o typie rozkładu zmiennych, zwłaszcza, że najpopularniejsze założenie o normalności rozkładu jest w przypadku cech społeczno-gospodarczych mocno wątpliwe do przyjęcia.

W badaniach praktycznych wszyscy, w zasadzie, autorzy zgadzają się jednak, że cechy diagnostyczne użyte do klasyfikacji i/lub porządkowania zbioru obiektów powinny od strony formalnej:

1. dobrze dyskryminować obiekty
2. być słabo skorelowane między sobą
3. być silnie skorelowane z cechami odrzuconymi (czyli być ich dobrymi reprezentantkami).

Do ceny stopnia dyskryminowania obiektów najczęściej używa się, na ogół, klasycznego współczynnika zmienności. Można jedna spotkać pogląd, iż bardziej właściwym punktem odniesienia jest nie średnia rozkładu a jego mediana[4]. Autorzy ci proponują użycie medianowego współczynnika względnego jako miary dyspersji cech.

Wielu autorów uważa, że wystarczy tylko uwzględnienie postulatu dyskryminacji cech co, jak już było wspomniane, sprowadzają do użycia wybranego współczynnika zmienności do doboru cech. Na ile jest to jednak wystarczające narzędzie

[1] Przegląd tych dyskusji autorka przedstawiła w pracy Tarka (2010).
[2] Np. Montanari A., Lizzani L.(2001) czy Steinly D., Brusco M.J. (2008).
[3] Przyjmujemy tu za Zeliasiem (2000), że cechy i zmienne w sensie matematycznym są tożsamymi nazwami.
[4] Patrz np. Lira J. i in. (2002), Młodak (2006).

jest pytaniem do dyskusji. Jednym z argumentów za więcej niż tylko merytoryczną analizą cech, związaną z dziedziną i zakresem badania, jest postulat by do klasyfikacji obiektów (zwłaszcza liniowej) dobierać cechy zgodnie z ogólną zasadą: maksymalny zasób informacji przy minimalnej liczbie cech. Zbyt liczny zbiór cech utrudnia lub uniemożliwia poprawną klasyfikację (Zeliaś 2002), to samo jednak można powiedzie gdy zbiór jest zbyt mały w sensie reprezentatywności.

Jeżeli obszar badania jest wąski lub dostępna jest mała liczba cech, do kilkunastu, wówczas nie ma, w zasadzie, problemu z doborem cech do badania. Badacz bierze, na ogół, wszystkie dostępne cechy, często nawet obniżając wymagania co do własności dyskryminacyjnej[5] gdy dostępnych cech jest niewiele.

Dobór cech zaczyna być istotny gdy cel (kryterium) badania jest określony szeroko (np. poziom rozwoju gospodarczego, społecznego, dobrobytu, życia itp.) i podstawy teoretyczne badanego zjawiska wymagają, do poprawnej analizy, dużego zbioru cech oraz badacz ma do dyspozycji duży zbiór wyjściowy cech. Analiza merytoryczna może wówczas okazać się niewystarczająca. Duża część cech wyjściowych jest w poziomach, a jak zauważali np. Fajferek (1966), Zeliaś (2004) do badań porównawczych należy brać cechy w postaci wskaźników natężenia eliminujących działania czynników ubocznych, z punktu widzenia badania, a wpływających na poziom cech. W badaniach regionalnych najczęściej eliminuje się wpływ wielkości obiektów (powierzchnie, liczba ludności itp.). Konstruując cechy w postaci wskaźników nierzadko powiększamy liczbę potencjalnych cech. Teoria nie zawsze „daje" jednoznaczne podstawy do konstrukcji wskaźników i do jednego merytorycznie zjawiska możemy skonstruować wiele alternatywnych wskaźników. Jako przykład weźmy określenie stopnia zanieczyszczenia wody ściekami możemy, w oparciu o dostępne dane, skonstruować m.in. następujące wskaźniki[6]: 1) ścieki przemysłowe i komunalne odprowadzone do wód lub ziemi w przeliczeniu na liczbę mieszkańców, 2) udział ścieków oczyszczonych w ściekach ogółem 3) udział ścieków wymagających oczyszczenia w ściekach odprowadzonych 4) udział ścieków oczyszczonych w ściekach wymagających oczyszczenia 5) podobne wskaźniki można skonstruować dla tylko ścieków przemysłowych i/lub komunalnych, niezależnie. Wszystkie te wskaźniki w sumie reprezentują jeden atrybut i zachodzi pytanie, przyjąć w badaniu wszystkie czy tylko jeden ze wskaźników by poprawnie scharakteryzować stopień zanieczyszczenia środowiska ściekami. Bez-

[5] W literaturze przyjmuje się, na ogół, minimalny poziom współczynnika zmienności cechy dyskryminującej pomiędzy 10%—20%. Jeśli jednak jest mała liczba cech w wielu badaniach za wystarczający uznaje się V=5% określając także w ten sposób stopień szczegółowości z jaką badacz ocenia wagę różnic pomiędzy obiektami.

[6] Nazwy cech są, ze względu na ilość miejsca, skrócone w stosunku do nazw z rocznika statystycznego.

pośrednio trudno ocenić, który miernik będzie bardziej reprezentatywny. W takiej sytuacji należy użyć jakiejś metody doboru cech[7] do badania, która wzięłaby pod uwagę relacje wskaźników alternatywnych między sobą oraz z pozostałymi cechami opisującymi inne atrybuty badanego zjawiska.

Od takiej metody należałoby oczekiwać, że:

1. pozostawi cechy niosące informacje komplementarne czyli jak najmniej ze sobą skorelowane;
2. wyeliminuje cechy niosące informacje substytucyjne, cechy pozostawione będą dobrze reprezentowały te wyeliminowane z badania, co wyraża się w założeniu, że cechy eliminowane będą silnie skorelowane z pozostawionymi.

Zakładając, że dokonano merytorycznego doboru cech, następnym etapem jest ocena siły dyskryminacji zbiorowości przez daną cechę. W tym celu używa się miar zmienności. Najpopularniejsza to klasyczny współczynnik zmienności (V_s), można także użyć współczynnika zmienności opartego o odchylenie przeciętne (V_d) lub któregoś ze współczynników pozycyjnych (V_{Me})[8]. Z grupy tych ostatnich coraz bardziej propagowane jest używanie medianowego odchylenia względnego[9].

$$V_{MOB} = \frac{Me\left|x_{ij} - Me\ (x_j)\right|}{Me\ (x_j)}100\ \% = \frac{MOB}{Me\ (x_j)}\%$$

Kukuła (1986, 2000) zaproponował by dodatkowo (obok współczynnika zmienności), używać współczynnika względnej amplitudy wahań jako kryterium odrzucenia cechy proponując jako wartość progową $A(x)\leq 1{,}2$[10].

Generalnie rzecz biorąc istnieją dwa podejścia do wyboru cech ze zbioru cech potencjalnych, jak określa się zbiór wyjściowy oparty o analizę merytoryczną[11]:

1. metodę doboru cech stosuje się do całego zbioru cech potencjalnych otrzymując zestaw cech reprezentujących badane zjawisko jako całość;
2. podejście dualne:
   - najpierw stosuje się dowolną metodę klasyfikacji by pogrupować cechy w podzbiory cech podobnych (reprezentujących substytucyjną informację o jakimś zjawisku cząstkowym),
   - następnie z każdej grupy wybiera się cechę –reprezentantkę grupy.

[7] Choć logiczniej byłoby powiedzieć metody eliminacji cech zbędnych.

[8] $V_{Me}=Me/Q$ gdzie $Q=Q_3-Q_1$

[9] Patrz np. Wysocki (2010), Młodak (2006).

[10] Jest to iloraz wartości maksymalnej do minimalnej cechy. Patrz dokładniejsze omówienie miary np. w Kukuła (2000). Z doświadczeń autorki wynika jednak, iż jest to miara mało użyteczna albowiem na ogół cechy poniżej proponowanej wartości progowej są też eliminowane przez współczynniki zmienności (zwłaszcza klasyczny).

[11] W naszym przypadku tak nazwiemy zbiór pozostały po zastosowaniu współczynników zmienności i odrzuceniu cech o zmienności poniżej progu 10%.

W przypadku podejścia dualnego należy podjąć dwie kluczowe decyzje: określić jaką metodą klasyfikować cechy oraz w jaki sposób wybrać reprezentantki grup. Wadą tego podejścia jest to, że a) podział na grupy może niepoprawnie odzwierciedlać zjawisko (atrybut) w sensie merytorycznym, czyli cechy, przynajmniej w części, mogą w danej grupie nie być merytorycznie związane ze sobą; b) występuje silna redukcja zbioru cech co może w sumie dać zestaw cech słabo (mało dokładnie, niewystarczająco) reprezentujący badane zjawisko od strony merytorycznej i prowadzić do niewłaściwych wniosków.

Istnieje trzecie podejście łączące oba powyższe. Niektórzy autorzy proponują najpierw pogrupowanie cech według kryteriów merytorycznych[12] a następnie do każdej z grup niezależnie zastosowanie metody doboru. W efekcie otrzymuje się pogrupowanie wg kryteriów cząstkowych (np. odrębny dobór wskaźników gospodarczych, odrębny społecznych) a następnie cechy je reprezentujące ale bez, narzuconej z góry, konieczności wybrania jednej reprezentantki[13].

Pytanie podstawowe zadane w tej pracy brzmi: jak bardzo metody doboru cech wpływają na wyniki klasyfikacji obiektów[14]. Aby precyzyjnie określić kryterium porównania zawężono badanie do analizy wyników porządkowania liniowego opartego o dobrane cechy a co za tym do analizy podobieństwa uporządkowania obiektów a nie podobieństwa uzyskanych zbiorów cech[15]

By ujednolić sposób uzyskania wyników tak, by tylko typ użytej metody różnicował wynik końcowy[16], przyjęto następujące założenia:

1. podstawowym punktem odniesienia do porównań będzie wynik rankingu uzyskany na zbiorze cech potencjalnych (bez użycia jakiejkolwiek metody doboru),
2. tam, gdzie to było niezbędne przyjęto jako progową wartość współczynnika korelacji liniowej Pearsona na poziomie $r^{*}=0,7$[17],
3. wszystkie otrzymane zbiory cech diagnostycznych standaryzowano i ujednolicono zmienne do postaci stymulant poprzez odwrotność,
4. porządkowanie liniowe przeprowadzono metodą Hellwiga (1968),

[12] Podział ekspercki.

[13] Dalszy ciąg postępowania może być dwojaki; albo badamy obiekty wg kryteriów cząstkowych, albo uzyskany zbiór cech scalamy w jeden, na podstawie którego badamy zjawisko w całości.

[14] Pomijamy tu odrębny problem kiedy poszczególne podejścia (jednoetapowe, dualne, mieszane) należy stosować i czy są one równoważne. Jest to decyzja badacza podejmowana w zależności od typu problemu badawczego i stopnia szczegółowości analizy.

[15] Problem od tej strony przedstawiano np. w pracach Hadasik (1993), Nowak (1981), tu autorzy prezentowali mierniki podobieństwa zbiorów cech.

[16] Ranking obiektów.

[17] Badanie dotyczy całej zbiorowości a nie próby losowej wobec tego przyjęto wartość progową współczynnika korelacji Pearsona uznawaną w podręcznikach statystyki za silną.

5. do porównania wyników rankingów użyto współczynnika korelacji rang Spearmana,
6. progowym kryterium odrzucenia był współczynnik zmienności nie większy niż 10%.

Materiał statystyczny i wstępna eliminacja cech

Do analizy przyjęto zbiór danych dotyczących stanu i ochrony środowiska w układzie wojewódzkim dla roku 2005. Ten rok jest ostatnim, w którym opublikowano szczegółowe dane dotyczące zanieczyszczeń powietrza w ujęciu wojewódzkim[18].

Przyjętym kryterium uporządkowania jest ocena stanu i ochrony środowiska w ujęciu wojewódzkim. Jest to kryterium wystarczająco szerokie by mieć do dyspozycji duży materiał wyjściowy a jednocześnie dosyć spójny merytorycznie.

Powodem przyjęcia tego obszaru i kryterium analizy była także chęć wyeliminowania dyskusji merytorycznej związanej z wyborem zbioru pierwotnego cech opisujących proces, na podstawie którego będzie przeprowadzona analiza *metod doboru cech pod kątem ich równoważności*. Przyjęto założenie, że dane opublikowane w roczniku są to dane opisujące stan i ochronę środowiska po „dyskusji" merytorycznej dokonanej przez ekspertów. Ekspercki jest także podział cech na grupy merytoryczne. Jednym słowem zawartość rocznika uznano za pierwotny zbiór cech diagnostycznych po selekcji merytorycznej i w merytorycznym podziale na grupy opisujące cząstkowe składowe stanu i ochrony środowiska w poszczególnych województwach jako obiektach. Eksperci wyodrębnili 8 zasadniczych działów (aspektów) w ochronie środowiska i w badaniu uznano ten podział za obowiązujący aczkolwiek badaniu podlegało 7 aspektów ochrony; ósmy – promieniowanie i hałas jako zbyt specyficzne, a w związku z tym nie spełniające jednego z warunków stawianego cechom, pominięto w analizie.

Przyjęty podział na działy w skrócie określimy jako[19]: I-ziemia, II-woda, III.-powietrze, IV-ochrona przyrody, V-odpady, VI- inspekcja sanitarna, VII-ekonomiczne aspekty ochrony. W oparciu o ten szeroki zbiór dokonano niezbędnych przekształceń cech do postaci wskaźników uzyskując zbiór k=80 cech potencjalnych. Następnie użyto wspomnianych wcześniej pięciu miar zmienności V_s, V_d, V_{Me}, V_{MOB} oraz A(x) do określenia stopnia zróżnicowania cech. Jako, że V_d jest

[18] Z ponad 20 cech pozostało parę dotyczących tylko emisji zanieczyszczeń z zakładów szczególnie uciążliwych.

[19] Są to skrócone i uproszczone nazwy działów w jakie są pogrupowane dane statystyczne w roczniku *Ochrona środowiska 2006*, GUS, Warszawa 2006, Informacje i Opracowania Statystyczne.

zawsze mniejsze niż V_s miary tej użyto do odrzucenia jednej cechy o współczynniku na granicy progu dopuszczalności V_s,. Oba współczynniki kwartylowe dawały w większości przypadków wartości na podobnym poziomie ale nie wykazywały regularności typu jeden stale większy od drugiego. W efekcie pozostawiono V_{MOB} jako bardziej zalecany w literaturze. Współczynnik A(x) okazał się mało przydatny przy sugerowanej przez Kukułę wartości progowej. Bardzo mało cech miało współczynnik względnej amplitudy wahań poniżej tej wartości a przy tym wszystkie one zostały wyeliminowane przez pozostałe współczynniki zmienności. Jeśli chodzi o relacje pomiędzy V_s i V_{MOB} to $V_s > V_{MOB}$ we wszystkich przypadkach poza trzema. Ponieważ kwestia, który ze współczynników jest lepszy w takich badaniach jest jednak nierozstrzygnięta do dalszego badania przyjęto dwa zbiory cech potencjalnych oznaczone jako Zb1- powstał w oparciu o zastosowanie V_s i zawiera n=74 cech i Zb2-powstał w wyniku użycia V_{MOB} a zawiera n=71 cech[20]. Następnie na zbiorze Zb1 zastosowano podejście mieszane doboru cech.

Metody doboru

Prezentowane tu wyniki są etapem drugim analizy. Dotyczą porównania rankingów utworzony na zbiorach z zastosowaniem mieszanego podejścia do doboru zmiennych. Jako miar doboru cech użyto[21]:

1. parametrycznej metody Hellwiga[22],
2. medianowej modyfikacji metody Hellwiga[23],
3. metody macierzy odwrotnej.

Zbiór cech potencjalnych był podzielony na siedem działów merytorycznych. Do każdego działu zastosowano jedną z trzech metod doboru i uzyskano zmienne reprezentujące zjawisko opisane danym działem[24]. Następnie a) zmienne scalono

[20] W pierwszym kroku zastosowano do tych dwóch zbiorów, niezależnie, te same procedury doboru cech, porangowano obiekty i porównano podobieństwo uzyskanych rankingów. W ten sposób sprawdzono także, na ile użycie różnych współczynników zmienności zróżnicowało wyniki rankingów. Następnie zastosowano szerszy zbiór metod i porównano wyniki tylko w ramach jednego zbioru (Zb1). eliminując tym samym wpływ doboru współczynnika zmienności na wyjściowy zbiór cech potencjalnych.. Wyniki tego fragmentu pracy oraz analizy porównawczej wyników otrzymanych przy użyciu metod klasyfikacji cech, a następnie wyboru reprezentantek przedstawiono w pracy złożonej w czasopiśmie *Taksonomia*.

[21] Opis wszystkich metod patrz np. Młodak (2006).

[22] Hellwig (1981).

[23] Modyfikacja polega na użyciu mediany zamiast średniej arytmetycznej.

[24] Cechy w zbiorze wyjściowym zostały przypisane do poszczególnych działów (ziemia, woda itd.) tak ja są umieszczone w roczniku (a więc grupowanie cech było merytoryczne, z góry ustalone) a

jako jeden zbiór i skonstruowano miarę syntetyczną, b) w oparciu o cechy reprezentujące poszczególne działy skonstruowano cząstkowe miary syntetyczne i obliczono średnią ważoną liczbą cech w dziale, jako końcową miarę syntetyczną. Rangowanie było więc przeprowadzone w oparciu o dwa typy miar syntetycznych:
- miary skonstruowane w oparciu o zbiór cech jako całość,
- miary skonstruowane jako średnia ważona z miar cząstkowych.

Pierwszym punktem odniesienia do porównań jest wynik rankingu uzyskanego bez eliminacji cech, czyli na zbiorze cech potencjalnych[25] Drugim punktem odniesienia jest ranking oparty o miarę syntetyczną uzyskaną jako ważona średnia z miar cząstkowych dla działów eksperckich (bez doboru)..

Wyniki poszczególnych rankingów porównano stosując współczynnik korelacji Pearsona. Tabela 1 pokazuje współczynniki korelacji rang uzyskane przy porównaniu uporządkowań na zbiorze cech jako całości.

Tabela 1 Korelacja rang Spearmana rankingów dla zbiorów cech jako całości

Metoda doboru	zb1.całość n=74	M.H całość n=47	MH medianowa całość n=46	M.odwrotna całość n=48
zb1.c	1			
M.H. całość	0,4853	1		
MH medianowa całość	0,4176	0,9029	1	
M.odwrotna całość	0,5765	0,7794	0,6588	1

Źródło: opracowanie własne. Litera n oznacza liczbę cech w zbiorze.

Ponieważ nie mamy kryterium „lepszości" z formalnego punktu widzenia, porównania wyników poszczególnych rankingów będziemy dokonywali w stosunku do wyników uzyskanych na zbiorze cech potencjalnych a więc bez doboru cech. Dobór, poprzez eliminację cech „zbędnych" powoduje pewną utratę informacji związaną z usuniętymi cechami. Zakładając, że zbiór wyjściowy zawiera pełną, a nawet nadmiarową z punktu widzenia procesu poznawczego, informację sprawdzamy jaką część tej informacji tracimy stosując dobór cech. „Utrata" informacji związana jest ze zróżnicowaniem zbiorów cech diagnostycznych, w oparciu o które porządkuje się obiekty[26].

Jak widać z tabeli 1 związek pomiędzy rankingami z doborem cech a rankingiem na zbiorze wyjściowym jest niezbyt duży. Wynik najbliższy rankingowi wyj-

następnie do każdego działu zastosowano metodę doboru cech uzyskując cechy reprezentujące poszczególne działy.

[25] Przypomnijmy jest to zbiór otrzymany po zastosowaniu współczynnika zmienności.

[26] Poszczególne metody redukowały od 35% do38% cech zbioru wyjściowego.

ściowemu dała metoda macierzy odwrotnej ale r_s=57,65% oznacza dużą różnicę w uporządkowaniach obiektów. Stosując w ten sposób metody doboru cech należy się liczyć z dużą utratą informacji[27]. Analizując współczynniki korelacji pomiędzy wynikami uzyskanymi za pomocą różnych metod doboru, czyli pod kątem substytucyjności metod, można uznać, że metoda Hellwiga i jej medianowa modyfikacja dając wyniki zbliżone na poziomie r_s=90,29% są w miarę substytucyjne. Wyniki metody macierzy odwrotnej różnią się zdecydowanie od pozostałych, choć w przypadku metody Hellwiga mniej niż w stosunku do jej medianowej modyfikacji. Jednak procedura doboru cech działami a następnie ich scalenie w jeden zbiór daje zdecydowanie różniące się wyniki od zbioru bez doboru cech.

Gdy badane zagadnienie jest bardzo szerokie jak np. poziom rozwoju[28] wówczas często sugeruje się analizę poszczególnych składowych tego zagadnienia a następnie scalenie wyników w postaci uśrednionej miary syntetycznej. W tabeli 2 przedstawiono zbiorcze wyniki porównań rankingów opartych o kryteria cząstkowe.

Tabela 2 Korelacja rang Spearmana rankingów w oparciu o średnią z miar cząstkowych

Metoda doboru	zb1.działy bez doboru	M.H działy	MH medianowa działy	M. odwrotna działy
zb1.działy bez doboru	1			
M.H. działy	0,7	1		
MH medianowa. działy	0,9265	0,8324	1	
M.odwrotna działy	0,7971	0,8471	0,8618	1

Źródło: opracowanie własne. Liczebność cech w zbiorze jak podana w nagłówkach tabeli 1.

Najsilniejszy związek z rankingiem bez doboru cech wykazuje medianowa modyfikacja Hellwiga, współczynnik korelacji r_s=92,65% pokazuje silne podobieństwo pomiędzy rankingami. Dużo słabiej plasuje się metoda macierzy odwrotnej, zaś metoda Hellwiga, co jest pewnym zaskoczeniem, nie tylko daje najbardziej różniące się wyniki ale też silnie różniące się od wyników dla zbioru wyjściowego, w przypadku jej stosowania do konstrukcji miary syntetycznej opartej o analizy cząstkowe .

Porównując też wyniki tych metod między sobą zauważamy, że użycie macierzy odwrotnej daje wyniki bliższe uzyskanym za pomocą medianowej metody

[27] Na ile jest ona istotna z merytorycznego punktu widzenia jest problem do odrębnej analizy. W niniejszej pracy porównujemy metody od strony formalnej.
[28] Patrz np. Zeliaś A. (2004).

Hellwiga niż oryginalnej. Oryginalna metoda Hellwiga w zastosowaniu do działów merytorycznych daje najbardziej odbiegające wyniki od rankingu bez doboru cech. W przypadku konstrukcji miary syntetycznej jako średniej z miar cząstkowych bardzo podobne wyniki daje użycie metody Hellwiga z medianową modyfikacją. Najgorzej wypada oryginalna metoda Hellwiga[29]. Należy oczywiście pamiętać, że na wyniki miały wpływ konkretne dane empiryczne, na podstawie których przeprowadzono badanie, wobec tego nie można przeprowadzać zbyt zdecydowanych uogólnień. Widać też wyraźnie, że decyzja czy badanie prowadzimy na całym zbiorze cech czy też ma miejsce dobór działami jest decyzją istotnie wpływającą na wyniki badania a więc i wnioski.

Podsumowanie

Jeżeli założymy, że zbiór cech potencjalnych daje uporządkowanie liniowe obiektów w oparciu o najpełniejszą informację to jego redukcja, jeżeli uznamy, że jest potrzebna lub niezbędna, powinna dawać taki zbiór cech, porządkowanie obiektów na którym powinno być zbliżone do uporządkowania na zbiorze wyjściowym. Na ile ono jest "poprawne' czy "dobre" nie jest kwestią metod ilościowych a wiedzy merytorycznej i intuicji badacza.

Prezentowane tu wyniki analizy pokazują, że wyniki porządkowania oparte o dobór cech dla poszczególnych działów odrębnie a następnie konstrukcję jednej miary syntetycznej są bardzo oddalone od miary obliczonej na całym zbiorze, bez doboru cech, wobec czego podejmując decyzję według jakiej procedury przeprowadzamy badanie musimy mieć świadomość, że będzie to miało wpływ na wyniki porządkowania.

Piśmiennictwo

1. Fajferek A., 1966. *Region ekonomiczny i metody analizy regionalnej*. PWE, Warszawa.
2. Gan G., Ma Ch., Wu J., 2007. *Data Clustering Theory, Algorithms, And Applications*. ASA-SIAM Series on Statistics and Applied Probability.

[29] Podobnie słabiej wypadła metoda Hellwiga (co nie znaczy, że źle) gdy ją stosowano do konstrukcji standardowej miary syntetycznej na całości zbioru cech w porównaniu do metody z medianą. Najlepiej wypadła jednak prosta metoda grafowa (pomimo silniejszej niż metoda Hellwiga redukcji cech). Dawała ona najbardziej zbieżne wyniki z całym zbiorem cech potencjalnych. Te wyniki przedstawiono w pracy *Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska* złożonej do czasopisma *Taksonomia*.

3. Grzega U., 2008. *Poziom życia ludności w Polsce i krajach ościennych*. Prace naukowe Akademii Ekonomicznej w Katowicach, Katowice.
4. Hadasik D., 1993. *Kilka uwag na temat porównywalności wyników różnych badań taksonomicznych*. Przegląd Statystyczny 2, s. 233-236.
5. Hellwig Z., 1968. *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr*. Przegląd Statystyczny 4.
6. Hellwig Z., 1981. *Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych*. (w:) W. Welfe (red.). *Metody i modele matematyczno-ekonomiczne w doskonaleniu zarządzania gospodarką socjalistyczną*. PWE, Warszawa.
7. Hellwig, Z., Siedlecka U., Siedlecki J., 1995. *Taksonometryczne modele zmian struktury gospodarczej Polski*. Instytut Rozwoju i Studiów Strategicznych, Warszawa.
8. Lira J., Wagner W., Wysocki F., 2002. *Mediana w zagadnieniach porządkowania obiektów wielocechowych*. (w:) J. Paradysz (red.). *Statystyka regionalna w służbie samorządu lokalnego i biznesu*. Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Akademia Ekonomiczna w Poznaniu, Poznań.
9. Malina A., 2004. *Wielowymiarowa analiza przestrzennego zróżnicowania struktury gospodarki Polski według województw*. Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
10. Młodak A., 2006. *Analiza taksonomiczna w statystyce regionalnej*. Difin, Warszawa.
11. Montanari A., Lizzani L., 2001. *A Projection Pursuit Approach to Variable Selection*. Computational Statistics And Data Analysis. 35, s. 463-473.
12. Nowak E., 1981. *Badanie zgodności metod wyboru cech diagnostycznych*. Przegląd Staystyczny 3-4, s. 301-309.
13. *Ochrona środowiska* 2006. GUS, Warszawa, Informacje i Opracowania Statystyczne
14. Romesburg H. Ch., 2004. *Cluster Analysis For Researches*. Lulu Press North Carolina.
15. Steinly D., Brusco M.J., 2008. *Selection of variables In Cluster Analysis: An Empirical Comparison of Eight Procedures*. Psychometrika 73 (1), s.125-144.
16. Tarka D., 2010. *Własności cech diagnostycznych w badaniach typu taksonomicznego*. Ekonomia i Zarządzanie, Politechnika Białostocka, Białystok 2 (4), s. 194-205.
17. Walesiak M., 2006. *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*. XXVII seminarium nt. „Przestrzenno-czasowe modelowanie zjawisk gospodarczych", s. 185-203.
18. Walesiak M., 2005. *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji*. Taksonomia 12, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.
19. Zeliaś A., 2002. *Uwagi na temat wyboru metody normowania zmiennych diagnostycznych*. (w:) Kufel T. i M. Piłatowska (red.). *Analiza szeregów czasowych na początku XXI wieku*. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.
20. Zeliaś A. (red.), 2004. *Poziom życia w Polsce i krajach Unii Europejskiej*. PWE, Warszawa.