PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Mirosława Lasek
Uniwersytet Warszawski

PROBLEMY DOBORU ZMIENNYCH W ZAGADNIENIACH GRUPOWANIA OBIEKTÓW METODA *FEATURESELECTOR* HEURYSTYCZNEGO DOBORU CECH*

1. Wstęp

W licznych badaniach statystycznych – w medycynie, biologii, ekonomii, zarządzaniu – zachodzi potrzeba grupowania obiektów. Do grupowania obiektów stosowane są różne metody, np. analizy skupień: hierarchiczne i niehierarchiczne, rozmytego skupiania, a obecnie coraz częściej: sieci neuronowe, algorytmy genetyczne, systemy generowania reguł podziału obiektów. Niezależnie od przyjętej metody grupowania (klasyfikacji bezwzorcowej) stajemy przed problemem wyboru zmiennych (kryteriów, cech)[1], które będziemy brać pod uwagę, grupując obiekty. Problem doboru zmiennych do grupowania obiektów nie jest trywialny, ponieważ wyniki algorytmów grupowania są bardzo wrażliwe na to, jakie zmienne zostaną wybrane do analizy [Kim, Street, Menczer 2002; Piramuthu]. Z problemem mamy do czynienia w grupowaniu różnorodnych obiektów. W prowadzonych przeze mnie pracach było to grupowanie klientów bankowych pod względem zdolności kredytowej i przynoszenia dochodów bankom [*Metody* ... 2002], powiatów woj. świętokrzyskiego pod względem ich atrakcyjności turystycznej [Lasek, Nowak, Pęczkowski 2003], przedsiębiorstw pod względem kondycji ekonomicznej (finansowej i majątkowej) [*Metody*

* Prace prowadzone na Wydziale Nauk Ekonomicznych Uniwersytetu Warszawskiego w ramach realizacji projektu badań statutowych BST-981/2004: „Statystyczne i ekonometryczne metody klasyfikacji danych. Zastosowania w badaniach porównawczych kondycji finansowej i majątkowej przedsiębiorstw" (kierownikiem projektu była Mirosława Lasek).

[1] Są tu stosowane różne terminy jako odpowiedniki angielskich pojęć: *features, variables, measurements, attributes.*

... 2002]. Przy dużej liczbie cech obiektów badania symulacyjne, polegające na doborze różnych zestawów cech grupujących, mogą być kłopotliwe, a nawet niemożliwe do realizacji. W polskiej literaturze można znaleźć wiele interesujących artykułów na temat wpływu doboru zmiennych na wyniki zarówno grupowania, jak i klasyfikacji w kolejnych wydaniach „Taksonomii" (por. [Rozmus, Wójcik 2002, s. 341--352]) oraz w innych pracach polskich autorów (por. np. [Gatnar 2001, rozdz. 3]).

Celem artykułu jest przedstawienie podstawowych przyczyn, skłaniających nas do ograniczania liczby zmiennych w przypadku grupowania obiektów oraz zaproponowanie stosowania heurystycznej metody wyboru zmiennych spośród większego ich zbioru, dających jak najlepsze pogrupowanie obiektów. Twórcami metody, algorytmu komputerowego i programu realizującego metodę jest zespół pracowników firmy MIT – Management Intelligenter Technologien GmbH z Aachen. W metodzie tej rozpoczyna się od niewielkiej liczby cech (np. dwóch lub nawet jednej), do których dokładane są cechy do momentu uzyskania takiej jakości grupowania, że dołożenie jakiejkolwiek nowej cechy z założonego zbioru nie daje poprawy tej jakości. W artykule zostanie przedstawiony algorytm metody, a także jej zalety i wady w porównaniu z innymi metodami doboru zmiennych w zagadnieniach grupowania obiektów. Rozważania będą zilustrowane przykładem doboru cech, które były stosowane na potrzeby tworzenia grup (skupień) przedsiębiorstw podobnych pod względem kondycji ekonomicznej.

2. Przyczyny selekcji zmiennych na potrzeby grupowania obiektów i stosowane metody

Praktyczne próby grupowania obiektów wydają się jednoznacznie wskazywać, że do podstawowych przyczyn zmuszających nas do ograniczenia liczby zmiennych i wyboru takiego ich zestawu, który możliwie najlepiej różnicuje obiekty, należą następujące: (i) metoda grupowania wymaga odpowiednio dużej liczby obiektów w stosunku do liczby rozpatrywanych cech, (ii) zbyt duża liczba zmiennych może powodować trudności interpretacji wyników grupowania (np. dużo skupień o małej liczebności), (iii) zmienne mogą być silnie skorelowane, (iv) łączne uwzględnienie zmiennych daje mniej informacji niż ich oddzielna analiza, (v) koszty pozyskania danych dla wszystkich zmiennych byłyby zbyt duże, (vi) czas realizacji algorytmów grupowania silnie wzrasta wraz z liczbą zmiennych. Do przeprowadzenia doboru cech na potrzeby grupowania obiektów są wykorzystywane takie metody, jak: analiza korelacji, analiza głównych składowych, badanie i porównywanie „pojemności informacyjnej" różnych zestawów kryteriów. Metodami dostosowanymi bardziej do przypadków doboru zmiennych w problemach klasyfikacji niż grupowania są testy na podstawie kryterium R^2, Chi^2, regresja krokowa, drzewa decyzyjne. Jak wskazują moje własne doświadczenia, dyskusje podczas konferencji oraz liczne opisy w literaturze, żadna z metod nie daje obiektyw-

nego, w pełni satysfakcjonującego rozwiązania. Istnieje potrzeba dalszych poszukiwań. Jednym z obiecujących kierunków, ze względu na słabo ustrukturalizowany, wymagający wiedzy, doświadczenia i intuicji charakter problemu, wydaje się skierowanie uwagi na możliwości, jakie mogą dać algorytmy heurystyczne.

3. Algorytmy heurystycznego wyszukiwania cech

Rozwój technik obliczeniowych (sprzętu i oprogramowania) daje możliwość zastosowania algorytmów heurystycznego wyszukiwania cech i przezwyciężenia w ten sposób problemów występujących w przypadku stosowania „tradycyjnych" metod dobierania cech różnicujących obiekty. Dzięki dysponowaniu wystarczającą pojemnością pamięci i szybko działającym komputerom i oprogramowaniu możemy, w możliwym do zaakceptowania czasie, przeszukać przestrzeń rozwiązań i przetestować znaczną liczbę różnych zestawów cech. Sposoby postępowania przy wyszukiwaniu cech mogą być następujące [Strackeljan, Behr 2002]: (a) rozpoczynamy od niewielkiej liczby cech (np. dwóch lub jednej), do których dokładamy kolejne cechy aż do momentu uzyskania wymaganej jakości grupowania lub (b) rozpoczynamy od zestawu wszystkich posiadanych cech i kolejno eliminujemy cechy o niewielkim wpływie (wkładzie) na jakość uzyskanego grupowania, a więc takie, z których rezygnacja w jak najmniejszym stopniu pogorszy wynik uzyskanych skupień (grup).

4. Kryteria wyboru zestawów cech na potrzeby grupowania

Ponieważ naszym celem jest znalezienie takich zestawów cech, które dadzą jak najlepsze pogrupowanie obiektów, kryteriami wyboru zestawów będą kryteria umożliwiające ocenę jakości grupowania [Strackeljan, Behr 2002; Strackeljan, Behr, Dedro 2003]: (1) kryterium określające liczbę obiektów ze zbioru treningowego, tj. zbioru obiektów o znanej przynależności do grup, zaklasyfikowanych poprawnie za pomocą danej kombinacji zmiennych – kryterium zwane wskaźnikiem *Reclassification Rate* (*ReClass*) oraz (2) kryterium przeciętnej odległości między obiektami, należącymi do różnych skupień (im odległości między obiektami z różnych skupień są większe, tym odległości między grupami obiektów możemy uznać za większe, a więc zestawy zmiennych za lepiej różnicujące obiekty)[2]; miernik ten jest interpretowany jako miara określająca naszą pewność (*certainty*) przydziału obiektu do grupy – kryterium zwane miernikiem *Average Distance of Membership*

[2] W przypadku stosowania zbiorów rozmytych wskaźnik może być obliczany jako różnica (*difference*) między wielkościami stopni przynależności: przynależnością obiektu do grupy, do której należy z największym stopniem przynależności, a przynależnością do grupy, do której należy z kolejnym co do wielkości stopniem przynależności.

(*Distance*). Próby zastosowania tylko jednego kryterium, pierwszego z wymienionych – *ReClass*, wskazały, że chociaż jest ono dobrym miernikiem jakości grupowania, okazuje się niewystarczające. W praktyce pojawiają się bowiem następujące problemy: (i) w przypadku małej liczby obiektów różne zbiory zmiennych (cech) często dają taki sam błąd grupowania, (ii) wyniki klasyfikacji mierzone wskaźnikiem *ReClass* poprawiają się często znacznie wraz ze wzrostem liczby rozważanych cech, równie często znaczny wzrost liczby cech może powodować tylko niewielkie zmiany błędu klasyfikacji, (iii) często mamy do czynienia z wieloma zestawami cech nie wpływających znacznie na zmiany w otrzymywanych grupowaniach, (iv) często występują różne zestawy cech, które dają taki sam podział na grupy.

5. Opis algorytmu

Ze względu na mniejszą zajętość pamięci i czas obliczeń oraz to, że w momencie zakończenia każdego kolejnego kroku poszukiwania zestawu cech zawsze jest dostępne poprawne rozwiązanie, na potrzeby własnych zastosowań wybrałam metodę „dokładania" kolejnych cech (pierwszy z wymienionych uprzednio sposobów postępowania). Załóżmy, że X jest zbiorem cech, spośród których dokonujemy wyboru: $X = \{ x_l \mid l = 1,2,...,N \}$, a M jest podzbiorem wybranych przez nas cech: $M = \{ m_i \mid i = 1,2,...,N \}$. Zakładamy także, że nie ma innego podzbioru X z taką samą lub większą liczbą cech niż M i takiego, który dałby klasyfikację lepszej jakości (przy założeniu określonego kryterium jakości klasyfikacji). Mówiąc prościej – dołożenie jakiejkolwiek cechy ze zbioru X nie dałoby poprawy jakości grupowania. W przyjętym algorytmie postępowania punktem wyjścia jest rozważanie pojedynczych cech. Następnie rozpatrywane są kombinacje każdej z tych cech z pozostałymi cechami. Poprzez dokładanie kolejnych cech na kolejnych etapach realizacji algorytmu brane są pod uwagę kombinacje o różnej liczbie cech. Te kombinacje – zestawy cech, które mają wskaźniki *ReClass* lub *Distance* lepsze od już utworzonych zestawów cech – są włączane do zbioru najlepszych rozwiązań o ustalonej liczebności, porządkowanego według wielkości *ReClass* i *Distance*. W zbiorze najlepszych rozwiązań mogą znaleźć się pojedyncze cechy, jeżeli tylko mają odpowiednio wysokie *ReClass* i/lub *Distance*.

6. Optymalizacja algorytmu w kierunku zmniejszenia liczby rozważanych kombinacji cech (zestawów cech)

Przeanalizowanie wszystkich zestawów cech, które mogłyby tworzyć zbiór M, jest często w praktyce niemożliwe ze względu na zbyt dużą liczbę kombinacji cech

do rozpatrzenia. Liczba kombinacji cech, którą powinniśmy przeanalizować, wynosiłaby bowiem [Strackeljan, Behr 2002; Strackeljan, Behr, Dedro 2003]:

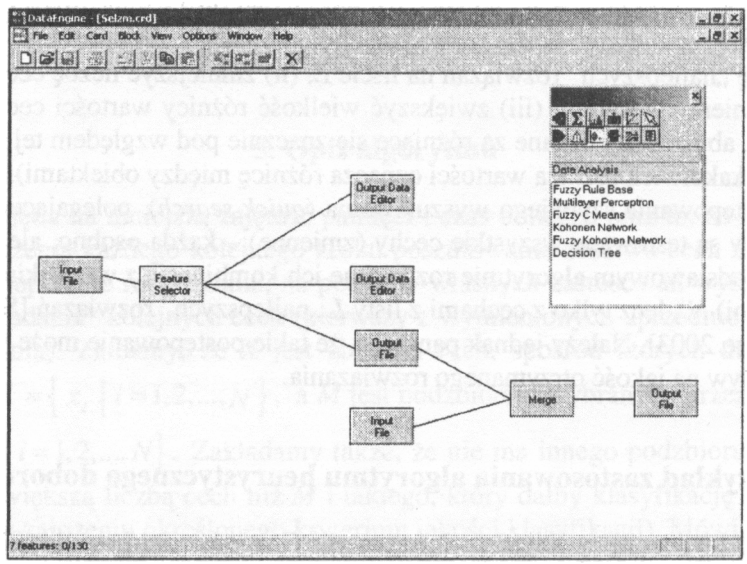
$$K=(N-1)\left(L\cdot N-\frac{L(L+1)}{2}-(N-2)+N\right),$$

gdzie: K – liczba kombinacji do przeanalizowania, N – maksymalna możliwa liczba cech w kombinacji wybranych cech M, N'– liczba wszystkich analizowanych cech, L – liczba najlepszych rozwiązań (zestawów cech) na etapie realizacji algorytmu (*number of solutions per stage*). Opracowano algorytmy poszukiwania cech tworzących zbiór M, które pozwalają uniknąć rozważania wszystkich możliwych kombinacji cech. Aby zmniejszyć liczbę rozważanych kombinacji cech, możemy: (i) zmniejszyć liczbę „najlepszych" rozwiązań na liście L; (ii) zmniejszyć liczbę cech N' przed zastosowaniem algorytmu; (iii) zwiększyć wielkość różnicy wartości cechy między obiektami, aby zostały uznane za różniące się znacznie pod względem tej cechy (domyślnie jakakolwiek różnica wartości oznacza różnicę między obiektami); (iv) zastosować postępowanie szybkiego wyszukiwania (*quick search*), polegającego na tym, że najpierw są testowane wszystkie cechy (zmienne) – każda osobno, ale potem nie są jak w podstawowym algorytmie rozważane ich kombinacje z wszystkimi cechami (zmiennymi) N', lecz tylko z cechami z listy L „najlepszych" rozwiązań [Strackeljan, Behr, Dedro 2003]. Należy jednak pamiętać, że takie postępowanie może mieć negatywny wpływ na jakość otrzymanego rozwiązania.

7. Przykład zastosowania algorytmu heurystycznego doboru cech

Prace prowadzone w Katedrze Informatyki Gospodarczej i Analiz Ekonomicznych WNE UW dotyczyły grupowania przedsiębiorstw w celu przeprowadzania ich analiz pod względem różnic kondycji ekonomiczno-finansowej i majątkowej. Poszukiwany był zestaw cech, który pozwoliłby wyodrębnić grupy przedsiębiorstw o istotnym zróżnicowaniu kondycji między grupami i jak największym podobieństwie w ramach grup. Analizą objęto spółki rynku podstawowego Warszawskiej Giełdy Papierów Wartościowych z sektorów: przemysłu oraz handlu i usług. Analiza spółek giełdowych objęła lata 1997-2001. Za zmienne, charakteryzujące przedsiębiorstwa pod względem kondycji ekonomicznej, przyjęto: (a) wielkości z bilansu, rachunku zysków i strat, rachunku przepływu środków pieniężnych, (b) wskaźniki kondycji majątkowej i finansowej: wskaźniki zyskowności, tj. marżę zysku brutto ze sprzedaży, marżę zysku operacyjnego, marżę zysku brutto, marżę zysku netto, stopę zwrotu z kapitału własnego, stopę zwrotu z aktywów; wskaźniki płynności, tj. kapitał pracujący, wskaźnik płynności bieżącej, wskaźnik płynności szybkiej, wskaźnik podwyższonej płynności; wskaźniki aktywności, czyli rotację należności, rotację zapasów, cykl operacyjny, rotację zobowiązań, cykl konwersji gotówki, rotację aktywów obrotowych, rotację aktywów; wskaźniki zadłużenia, tj. wskaźnik pokrycia ma-

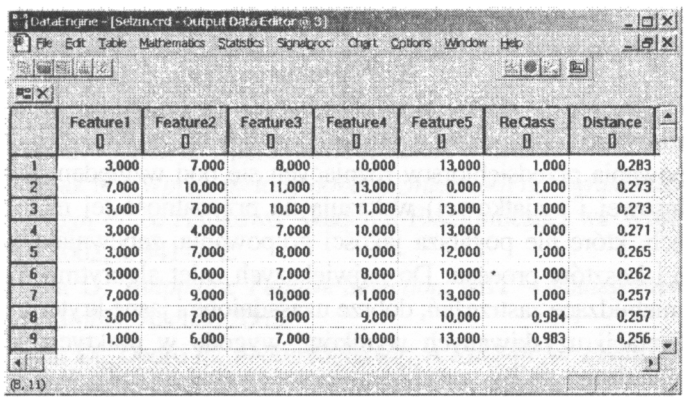
jątku, stopę zadłużenia, wskaźnik dźwigni finansowej. Przedstawiony zbiór cech jest dość liczny; wiele cech powiela informacje „niesione" przez inne cechy. Posługując się algorytmem heurystycznego doboru cech, postanowiono wyselekcjonować te cechy, które odznaczają się wysoką zdolnością dyskryminacji i które będą dobrze różnicować przedsiębiorstwa bez nadmiernej utraty informacji o ich kondycji ekonomicznej. W celu przeprowadzania obliczeń, sporządzania tabel i wykresów na potrzeby analizy wykorzystano narzędzie *Feature Selector* (*plug in*) programu *DataEngine* niemieckiej firmy Management Intelligenter Technologien GmbH z Aachen. Na potrzeby wyboru cech opracowano odpowiedni schemat przetwarzania (rys. 1).



Rys. 1. Konfigurowanie schematu przetwarzania za pomocą graficznego makrojęzyka *DataEngine* (ilustracja zestawiania bloków funkcjonalnych)
Źródło: opracowanie własne z użyciem programu *DataEngine*.

Posługując się programem *DataEngine* i przyjmując różne założenia dotyczące maksymalnej liczby cech, które mogą być zastosowane do analizy przedsiębiorstw, liczby grup, na które zamierzamy podzielić przedsiębiorstwa, liczby „najlepszych" rozwiązań na liście *L* (*number of solutions per stage*), ustaleń o tym, jaka wielkość różnicy cechy oznacza istotną różnicę między przedsiębiorstwami, zastosowania lub odrzucenia opcji *quick search*, otrzymywaliśmy kolejne zestawy cech do badań kondycji przedsiębiorstw. Analiza tych zestawów cech wskazała, że w rozpatrywanym przypadku kondycji ekonomicznej przedsiębiorstw często wybierane były jako „najlepsze" podobne zestawy cech. Podstawową różnicą było ich umiejscowienie (kolejność) na liście najlepszych rozwiązań. Zestawy cech są przedstawiane na liście, na której każdy zestaw jest umieszczony w jednym wierszu (rys. 2.). Zestawy na liście są uporządkowane od najlepszego do najgorszego, zgodnie z wielkościami *ReClass* i *Distance*, podawanymi dla każdego z zestawów. Pewną niedogodnością wykorzystywanego programu jest to, że są podawane

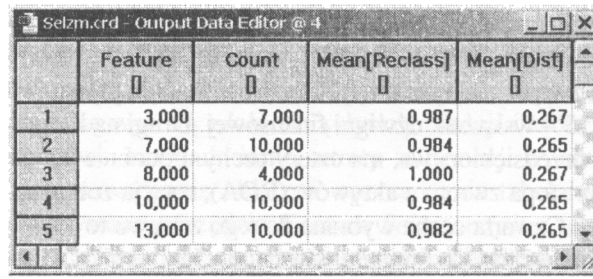
nie nazwy zmiennych, ale przypisane im numery (kody). W przedstawionym na rys. 2 „najlepszym" zestawie pięciu cech znalazły się następujące: 3 – zobowiązania długoterminowe, 7 – stopa zwrotu z aktywów (ROA), 8 – wskaźnik płynności bieżącej, 10 – rotacja zobowiązań, 13 – wskaźnik dźwigni finansowej. Drugi na liście zestaw cech, najlepiej różnicujących przedsiębiorstwa, ma cztery cechy. Podobnie jak w pierwszym zestawie, znalazły się tu: stopa zwrotu z aktywów (ROA), rotacja zobowiązań oraz wskaźnik dźwigni finansowej. Czwarta cecha wybrana do tego zestawu to 11 – wskaźnik pokrycia majątku (relacja zobowiązań do kapitału własnego). Jak wynika z rys. 2, aż siedem zestawów cech osiągnęło jednakową wielkość wskaźnika *ReClass* na poziomie 1. Różnią się one wielkościami wskaźnika *Distance*, który zadecydował o kolejności tych zestawów na liście najlepszych rozwiązań.



	Feature1	Feature2	Feature3	Feature4	Feature5	ReClass	Distance
	[]	[]	[]	[]	[]	[]	[]
1	3,000	7,000	8,000	10,000	13,000	1,000	0,283
2	7,000	10,000	11,000	13,000	0,000	1,000	0,273
3	3,000	7,000	10,000	11,000	13,000	1,000	0,273
4	3,000	4,000	7,000	10,000	13,000	1,000	0,271
5	3,000	7,000	8,000	10,000	12,000	1,000	0,265
6	3,000	6,000	7,000	8,000	10,000	1,000	0,262
7	7,000	9,000	10,000	11,000	13,000	1,000	0,257
8	3,000	7,000	8,000	9,000	10,000	0,984	0,257
9	1,000	6,000	7,000	10,000	13,000	0,983	0,256

Rys. 2. Proponowane zestawy cech wyselekcjonowane na podstawie kryteriów *ReClass* i *Distance*
Źródło: opracowanie własne z użyciem programu *DataEngine*.

Gdy zestawy cech nie różnią się obydwoma wskaźnikami, za „lepszy" zostaje uznany zestaw z mniejszą liczbą cech. Interesujące może być prześledzenie, jak często pojawiały się poszczególne cechy na liście najlepszych zestawów, tj. w różnych zestawach uznawanych za najlepsze. W omawianych powyżej wynikach lista najlepszych liczyła 20 zestawów, a za maksymalną liczbę cech w zestawie przyjęto 5 cech. Mogliśmy stwierdzić, że wśród cech z najlepszej listy najczęściej we wszystkich zestawach pojawiały się: stopa zwrotu z aktywów, rotacja zobowiązań oraz wskaźnik dźwigni finansowej (każda cecha w 10 zestawach na 20 zestawów na liście). Cecha „zobowiązania długoterminowe" wystąpiła w 7 zestawach z listy 20, a cecha „wskaźnik płynności bieżącej" tylko w 4 zestawach. Możemy porównać średnie wielkości wskaźników *ReClass* i *Distance* zestawów cech, na których pojawiała się dana cecha – por. rys. 3. Dane dotyczące częstości pojawiania się cech w zestawach najlepiej różnicujących przedsiębiorstwa, a także średnie wielkości *ReClass* i *Distance* dla zestawów, w których pojawiają się poszczególne cechy, dostarczają cennych wskazówek o zdolności dyskryminacyjnej poszczególnych cech.


Selzm.crd - Output Data Editor @ 4

	Feature []	Count []	Mean[Reclass] []	Mean[Dist] []
1	3,000	7,000	0,987	0,267
2	7,000	10,000	0,984	0,265
3	8,000	4,000	1,000	0,267
4	10,000	10,000	0,984	0,265
5	13,000	10,000	0,982	0,265

Rys. 3. Częstości (*Count*) pojawiania się poszczególnych cech na liście „najlepszych" zestawów cech oraz średnie wielkości *ReClass* i *Distance* zestawów zawierających wybrane cechy
Źródło: opracowanie własne z użyciem programu *DataEngine*.

8. Wnioski

Próby zastosowania opisywanej w artykule metody heurystycznego doboru cech na potrzeby grupowania przedsiębiorstw różniących się pod względem kondycji ekonomicznej (finansowej i majątkowej) wskazują na przydatność tej metody do takiego wybierania cech, które nie pogarsza jakości grupowania, zapewniając akceptowany poziom czasu i kosztów procesu. Do największych zalet algorytmu *FeatureSelector* zaliczyłabym niebudzące zastrzeżeń, dobrze uzasadnione i jasne kryteria doboru cech – łatwość interpretacji uzyskiwanych wyników, wygodę w praktycznych zastosowaniach. Przeprowadzone próby weryfikacji zastosowania metody w różnych zbiorach danych potwierdzają jej zalety. Za podstawową wadę algorytmu uznałabym długi czas obliczeń w bardzo dużej liczbie cech, sięgającej kilkuset – z taką liczbą cech mieliśmy do czynienia, analizując zbiór danych, dotyczących polskich gospodarstw domowych, w którym liczba cech przekraczała 450.

Proces doboru cech nie może być całkowicie zautomatyzowany: wymaga podjęcia decyzji dotyczących np. maksymalnej liczby cech w wybranym zbiorze przed rozpoczęciem wykorzystywania algorytmu, wskazania liczby grup podziału obiektów, określenia liczby dopuszczalnych rozwiązań (zestawów cech) rozpatrywanych w kolejnych krokach algorytmu. Na potrzeby niektórych analiz celowe jest, aby zestawy wybranych cech były modyfikowane „ręcznie", np. pewna zmienna może wnosić niemal tę samą informację, jaką wnosi inna zmienna, lecz z punktu widzenia analizy może nie być obojętne, którą wybierzemy. Nie jest możliwy wybór zestawu cech, który byłby „dobry", niezależnie od metody grupowania. Dlatego wybierając zmienne na potrzeby grupowania, powinniśmy zastanowić się, jaką metodę grupowania obiektów będziemy stosować. Na przykład określony zestaw cech może umożliwić satysfakcjonujące na potrzeby analiz pogrupowanie obiektów, gdy jest możliwy ich podział za pomocą liniowej funkcji dyskryminacji, podczas gdy inny zestaw cech może spowodować, że podział obiektów pozostanie poza możliwościami tej metody grupowania i będzie wymagać np. zastosowania sieci neuronowej.

Literatura

Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2001.

Kim Y.S., Street W.N., Menczer F., *Feature Selection in Data Mining*, [w:] *Data Mining. Opportunities and Challenges*, red. J. Wang, IRM Press – Idea Group Inc. 2003.

Lasek M., *Data Mining. Zastosowania w analizach i ocenach klientów bankowych*, Oficyna Wydawnicza „Zarządzanie i finanse", Biblioteka Menedżera i Bankowca, Warszawa 2002.

Lasek M., Nowak E., Pęczkowski M., *Analiza atrakcyjności turystycznej powiatów województwa świętokrzyskiego za pomocą metody Promethee*, „Turyzm" 2003, nr 13/1, (artykuł w dwóch wersjach językowych: polskiej i angielskiej).

Metody analiz porównawczych kondycji ekonomicznej przedsiębiorstw, red. M. Lasek, Wyd. Nowy Dziennik, Warszawa 2002.

Piramuthu S., *Evaluating Feature Selection Methods for Learning in Data Mining Applications*, European Journal of Operational Research, w druku.

Rozmus D., Wójciak M., *Wpływ metody doboru zmiennych na wyniki klasyfikacji sektora produkcyjnego*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 942, AE, Wrocław 2002, s. 341-352.

Strackeljan J., Behr D., *Automatic Feature Selection*, MIT – Management Intelligenter Technologien GmbH, Aachen 2002.

Strackeljan J., Behr D., Dedro F., *FeatureSelector: A Plug-In for Feature Selection with DataEngine*, International Data Analysis Symposium, Aachen 2003.

FEATURE SELECTION IN CLUSTERING – APPLICATION OF THE FEATURE SELECTOR HEURISTIC METHOD

Summary

The aim of this paper is to discuss problems concerning feature selection for grouping objects and to propose heuristic method for determining such features which can discriminate objects as best as possible. In the proposed method we start with single feature to which next features are added until obtaining such good possibilities of clustering with assumed criteria of quality, so that no further improvement of quality of clustering is possible. In the paper, algorithm of the method has been described, the strengths and weaknesses of the algorithm are examined and discussed. The method is illustrated with an example of feature selection for clustering of industrial enterprises in order to analyse and compare their financial standing.