PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

**Jerzy Korzeniewski**
Uniwersytet Łódzki

# BADANIE ODPORNOŚCI METODY HINOV NA BŁĘDNIE ZADANĄ LICZBĘ SKUPIEŃ W ZBIORZE DANYCH

**Streszczenie:** Metoda HINoV służąca do wybierania zmiennych w analizie skupień jest popularna i jest jedną z najlepszych [Steinley, Brusco 2008]. Nieznana jest jednak efektywność tej metody wtedy, gdy liczba skupień w zbiorze danych jest błędnie zadana. Taka sytuacja jest powszechna, gdyż indeksy wyznaczające liczbę skupień mają na ogół charakter optymalizacyjny dla przyjętej metody grupowania i popełniają dość duże błędy. W artykule zbadana jest odporność metody na kilku tysiącach zbiorów danych wygenerowanych w postaci mieszanin rozkładów normalnych. Dobór liczb skupień, liczby zmiennych istotnych i maskujących, stopnia zachodzenia skupień na siebie, rozkładów zmiennych maskujących jest taki sam jak w eksperymencie symulacyjnym Steinleya i Brusco [2008].

## 1. Wstęp

Przystępując do badania zbioru danych, bardzo często spotykamy się z problemem występowania zmiennych nieistotnych, zwanych inaczej zanieczyszczającymi lub maskującymi. W kontekście analizy skupień niezwykle ważnym zagadnieniem jest wybór zmiennych istotnych dla struktury skupień zbioru danych. Jeśli taki wybór zostanie dokonany błędnie lub zbiór zmiennych nie zostanie w ogóle oczyszczony ze zmiennych nieistotnych, to fakt ten na ogół wpływa bardzo negatywnie na wnioski uzyskane z analizy skupień.

W swojej publikacji przedstawiającej metodę HINoV Carmone, Kara, Maxwell [1999] wymienili kilka znanych wówczas metod wybierania zmiennych opracowanych pod kątem analizy skupień. Podali krótką charakterystykę każdej metody i jej wady. Omówione zostały-metody takich autorów, jak: De Soete (1986, 1988), Fowlkes, Gnanadesikan, Kettenring (1987), Donoghue (1995), De Sarbo i in. (1984), van Buuren i in. (1989), Duffy i in. (1991), Hartigan (1972). Wady, jakie wytknięto tym metodom, dotyczą tego, że metody praktycznie nie dadzą się zastosować do dużych zbiorów danych, pojawiają się trudności z obserwacjami nietypowymi, trudno uzyskać dobre rozwiązania globalne, brak jest spójnych wniosków i istnieje częściowy subiektywizm. Artykuł Carmonego i in. [1999] można uważać za przełomowy w omawianym temacie, ponieważ później nikt raczej nie wracał do metod wcześniejszych.

Zaproponowanych zostało natomiast kilka nowych metod. Wśród nich wymienić należy trzy podejścia modelowe:

- metodę wyrazistości cech (*feature saliency method*) [Law, Jain, Figueiredo 2003];
- metodę opartą na wielkości rozproszenia *(scatter separability method)* [Dy, Brodley 2000];
- metodę opartą na wyborze właściwego modelu *(model selection method)* [Raftery, Dean 2006];

oraz pięć niemodelowych:

- metodę opartą na grupowaniu obiektów na wybranych podzbiorach zmiennych (COSA ) [Friedman, Meulman 2004];
- metodę opartą na badaniu entropii (*entropy based*) [Dash, Liu 2000];
- metodę kolejnych rzutowań (*projection pursuit*) [Montanari, Lizzani 2001];
- metodę VS-KM opartą na grupowaniu metodą *k*-średnich (*VS-KM method*) [Brusco, Cradit 2001].
- metodę opartą na grupowaniu metodą *k*-średnich zmodyfikowaną indeksem skupialności (*relative clusterability weighting method*) [Steinley, Brusco 2008].

Dwie ostatnie metody są modyfikacjami metody HINoV.

W roku 2008 Steinley i Brusco przeprowadzili obszerny eksperyment symulacyjny, w którym dokonali porównania ośmiu metod służących do wybierania zmiennych w analizie skupień. Wybrali oni trzy podejścia modelowe i pięć niemodelowych. Podstawą porównania metod były trzy wskaźniki: pamięć (*recall*), precyzja (*precision*) oraz asymptotyczna odzyskiwalność poprawnego przypisania obserwacji do skupień (*ARI asymptotic recovery*). **Pamięć** to stosunek liczby wybranych zmiennych istotnych do liczby wszystkich zmiennych istotnych. **Precyzja** – stosunek liczby wybranych zmiennych istotnych do liczby wszystkich wybranych zmiennych.

**Tabela 1.** Jakość ośmiu metod mierzona za pomocą trzech kryteriów

| Metoda | Pamięć | Precyzja | ARI |
|---|---|---|---|
| $\mathbf{M_1}$ Law i in. [2004] | 0,911 | 0,813 | 0,6070 |
| $\mathbf{M_2}$ Raftery, Dean [2006] | 0,326 | 0,552 | 0,4334 |
| $\mathbf{M_3}$ Dy, Brodley [2000] | 0,506 | 0,612 | 0,3864 |
| $\mathbf{M_4}$ Friedman, Meulman [2004] | 0,660 | 0,743 | 0,7211 |
| $\mathbf{M_5}$ Montanari, Lizzani [2001] | 0,732 | 0,699 | 0,7082 |
| $\mathbf{M_6}$ HINoV , Carmone i in. [1999] | 0,926 | 0,951 | 0,8514 |
| $\mathbf{M_7}$ VS-KM Brusco, Cradit [2001] | 0,983 | 0,841 | 0,8507 |
| $\mathbf{M_8}$ Steinley, Brusco [2007] | 0,893 | 0,964 | 0,8611 |

Źródło: [Steinley, Brusco 2008].

Spośród zbadanych metod najlepiej wypadły dwie modyfikacje metody HINoV, tj. metoda VS-KM, Brusco, Cradit [2001], metoda wykorzystująca indeks skupialności [Steinley, Brusco 2007] oraz metoda HINoV. Różnice pomiędzy tymi

trzema metodami są tak marginalne, że trudno którąkolwiek z nich uważać za lepszą od innych.

Z takimi warunkami (por. opis eksperymentu), w jakich badane były metody, w praktyce nie spotykamy się prawie nigdy, gdyż np. nigdy nie mamy pewności co do tego, jaka jest faktyczna liczba skupień w zbiorze danych. Przedmiotem tego artykułu jest badanie efektywności metody HINoV w warunkach błędnie zadanej liczby skupień w zbiorze danych. Taka sytuacja w praktyce zdarza się często, gdyż indeksy liczby skupień w zbiorze danych mają różne i często błędne wskazania (por. [Najman, Najman 2005]). Indeksy te przy małej liczbie skupień mylą się standardowo o jedno skupienie.

## 2. Charakterystyka metody HINoV

Dla każdej zmiennej *v* przeprowadzamy 50 razy grupowanie metodą *k*-średnich i zapamiętujemy to grupowanie, które miało najmniejszą sumę kwadratów odchyleń obserwacji od środków skupień, tzn.

$$SSE(v)=\sum_{k=1}^{K}\sum_{i\in C_k}\left(x_{iv}-\overline{x}_{iv}\right)^2 \text{ , gdzie } \overline{x}_{iv}=\frac{1}{N_k}\sum_{i\in C_k}x_{iv} .$$

Następnie obliczamy poprawiony indeks Randa dla każdej pary zmiennych

$$ARI(u,v)=\frac{\binom{n}{2}(t_1+t_2)-\left[(t_1+t_3)(t_1+t_4)+(t_3+t_2)(t_4+t_2)\right]}{\binom{n}{2}^2-\left[(t_1+t_3)(t_1+t_4)+(t_3+t_2)(t_4+t_2)\right]},$$

gdzie: $t_1$ – liczba par obserwacji należących do tego samego skupienia dla obu zmiennych;

$t_2$ – liczba par obserwacji należących do innych skupień dla obu zmiennych;

$t_3$ – liczba par obserwacji należących do tego samego skupienia dla zmiennej *u* i do różnych skupień dla zmiennej *v*;

$t_4$ – liczba symetryczna do $t_3$.

Wartość indeksu $ARI(u,v)$ interpretuje się jako miarę podobieństwa dwóch podziałów/grupowań tego samego zbioru obserwacji. Następnie dla każdej zmiennej obliczamy sumę

$$TOPRI(u)=\sum_{u\neq v}ARI(u,v),$$

którą możemy interpretować jako miarę siły związku podziałów zbioru, w których brane były pod uwagę wartości tej zmiennej ze wszystkimi podziałami, w których nie były one uwzględniane. Zmienne o największych wartościach *TOPRI*(*u*) mają najsilniejszy związek ze strukturą skupień. Ostatnim etapem jest podzielenie wszystkich zmiennych na dwa zbiory: zmiennych istotnych i maskujących. W tym celu można wykorzystać kryterium największego skoku wskaźnika *TOPRI*(*u*), tzn. po uporządkowaniu wartości tego wskaźnika malejąco obliczać ilorazy

$$RR(s) = \frac{\left(TOPRI\left(k(s)\right) - TOPRI\left(k(s+1)\right)\right)}{\left(TOPRI\left(k(s-1)\right) - TOPRI\left(k(s)\right)\right)}$$

i wybierać *s* początkowych zmiennych do największej wartości *RR*(*s*)

Taka metoda ma wady. Po pierwsze, metoda ta z góry zakłada, że odrzucimy jakieś zmienne, co jest założeniem dość daleko idącym. Po drugie, iloraz jest bardzo nieodporny na wartości mianownika bliskie zeru, które mogą powodować fałszywe podziały zbioru wszystkich zmiennych. Trudno jednak w takim eksperymencie symulacyjnym zastosować jakiś inny sposób, np. oryginalnie polecany przez twórców metody sposób podziału wszystkich zmiennych opierający się na wizualnym badaniu wykresu osypiska.

## 3. Eksperyment badawczy

W celu zachowania ogólnej porównywalności badania przeprowadzono eksperyment na wzór eksperymentu Steinleya i Brusco [2008]. Poszerzono nawet nieco zakres eksperymentu, dopuszczając zbiory z 3 skupieniami i maskujące rozkłady równomierne. Wszystkie zbiory składały się z 200 elementów, różniły się między sobą następującymi czynnikami:

- Pierwszy czynnik, liczba skupień, był równy 3, 4, 6 lub 8.
- Drugi czynnik, liczebności skupień, miał trzy warianty: (a) równe liczebności wszystkich skupień; (b) 10% obserwacji i (c) 60% obserwacji w jednym skupieniu, a pozostałe skupienia były równoliczne.
- Trzeci czynnik, liczba zmiennych istotnych, był równy 2, 4 lub 6.
- Czwarty czynnik, prawdopodobieństwo „zachodzenia na siebie" (*overlap*) skupień na każdej ze zmiennych istotnych, miał pięć wariantów – 0, 0.1, 0.2, 0.3, 0.4. Separowalność skupień była typu „łańcuchowego", tj. na każdym wymiarze było $k - 1$ par skupień zachodzących na siebie w takim samym stopniu równym *overlap* ($k$ – liczba skupień).
- Piąty czynnik, siła korelacji wewnątrz skupień, miał dwa warianty: (a) macierz kowariancji w każdym skupieniu była macierzą jednostkową; (b) w każdym skupieniu była taka sama macierz kowariancji z jedynkami na przekątnej, poza przekątną zaś liczba wylosowana z odcinka [0.3; 0.8].

- Szósty czynnik, liczba zmiennych maskujących, przyjmował wartości 2, 4 lub 6.
- Siódmy czynnik, rozkład zmiennych maskujących, miał siedem wariantów: (a) wszystkie zmienne zostały niezależnie wygenerowane z rozkładu skośnego jednomodalnego (rozkład gamma z jednym stopniem swobody dla licznika i z jednym dla mianownika); (b) wszystkie zmienne zostały niezależnie wygenerowane z rozkładu normalnego ze średnią zero i jednostkową wariancją; (c) wszystkie zmienne zostały niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0.25 poza przekątną; (d) wszystkie zmienne zostały niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0.5 poza przekątną; (e) wszystkie zmienne zostały niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0.75 poza przekątną; (f) wszystkie zmienne zostały wygenerowane z niezależnych rozkładów normalnych ze średnimi równymi zeru i wariancjami losowanymi z odcinka [1; 20]; (g) wszystkie zmienne zostały niezależnie wygenerowane z rozkładów równomiernych na odcinku [1; 20]. Każdy układ parametrów został powtórzony dwukrotnie, co razem dało liczbę 15 120 zbiorów.

Biorąc pod uwagę niewielką liczbę skupień występującą w eksperymencie, przyjęto założenie o tym, że błąd popełniany przy określaniu liczby skupień może być również niewielki, tj. równy 1 lub2. Wobec tego każdy zbiór poddany został metodzie HINoV dla prawidłowo zadanej liczby skupień, dla liczby skupień większej od niej o 1 oraz większej o 2. Dla każdego zbioru zanotowano pamięć i precyzję metody, następnie wielkości te wykorzystywane były do obliczania średniej arytmetycznej pamięci i precyzji dla określonych grup zbiorów bądź wszystkich zbiorów.

## 4. Wyniki i wnioski

Ogólnie (por. tab. 2) nie ma dużych spadków efektywności mierzonej pamięcią i precyzją. Różnice pamięci i precyzji są rzędu 2-3%. Ciekawszych spostrzeżeń można dokonać, analizując szczegółowo określone przypadki zbiorów. Na przykład okazało się, że wyniki nie zależą od stopnia pokrywania się skupień (czynnik czwarty) oraz od siły korelacji wewnątrz skupień (czynnik piąty). Zależą z kolei, w znacznym stopniu, od typu rozkładu nieistotnego, stąd też w tab. 2 wyszczególniono wyniki dla różnych typów rozkładów maskujących. Dla niektórych rozkładów metoda spisuje się o wiele lepiej (np. dla rozkładu gamma) niż dla innych (najgorzej dla równomiernego). Wyniki zależą również w dużym stopniu od liczby skupień (por. tab. 3). Dla liczby skupień równej 8 sfałszowanie jej do 9, lub nawet 10, oczywiście nie ma takich konsekwencji jak powiększenie dwóch skupień do 3 lub 4.

**Tabela 2.** Średnia pamięć i precyzja dla różnych typów rozkładów nieistotnych dla poprawnej liczby skupień (*lsk*), za dużej o jeden (*lsk* +1) i za dużej o dwa (*lsk* + 2)

| Typ rozkładu nieistotnego | *lsk* | | *lsk* + 1 | | *lsk* + 2 | |
|---|---|---|---|---|---|---|
| | pamięć | precyzja | pamięć | precyzja | pamięć | precyzja |
| a | 0,917 | 0,872 | 0,934 | 0,920 | 0,932 | 0,883 |
| b | 0,930 | 0,878 | 0,914 | 0,860 | 0,886 | 0,824 |
| c | 0,915 | 0,871 | 0,882 | 0,831 | 0,866 | 0,819 |
| d | 0,824 | 0,802 | 0,819 | 0,809 | 0,824 | 0,802 |
| e | 0,660 | 0,666 | 0,681 | 0,642 | 0,706 | 0,662 |
| f | 0,909 | 0,892 | 0,908 | 0,852 | 0,887 | 0,835 |
| g | 0,858 | 0,831 | 0,798 | 0,744 | 0,759 | 0,693 |
| Średnio | 0,859 | 0,830 | 0,848 | 0,808 | 0,837 | 0,788 |

Źródło: obliczenia własne.

Ciekawe jest to, że gdy rozkładem maskującym jest rozkład beta (przypadek a), to efektywność metody wzrasta przy zbyt dużej, błędnej liczbie skupień. Wydaje się, że rezultat ten można logicznie wytłumaczyć w następujący sposób. Metoda HINoV tak dobrze wykrywa skośne rozkłady maskujące, że fałszywe zwiększenie liczby skupień tylko poprawia efektywność – pamiętać należy o tym, że jest to typowa metoda *wrapper*, tj. „owinięta" wokół pewnej metody grupowania danych. Wobec tego uaktywniać się mogą cechy przyjętej metody grupowania, w tym przypadku metody *k*-średnich, np. to, że dla większej fałszywej liczby skupień na zmiennej o rozkładzie beta większość obserwacji jest przypisana do jednego z większej liczby skupień, co powoduje większą, łatwiejszą do wykrycia deformację. Fakt ten, tj. poprawa efektywności tylko dla jednego typu rozkładu maskującego, niestety bardzo deformuje wnioski wynikające z porównania średnich wskazań pamięci i precyzji, ponieważ przy porównywaniu średnich zacierają się wyraźniejsze spadki efektywności dla innych typów rozkładów maskujących.

**Tabela 3.** Średnia pamięć i precyzja dla zbiorów o małej liczbie skupień (równej 3), dla różnych typów rozkładów nieistotnych, dla poprawnej liczby skupień (*lsk* = 3), za dużej o jeden (*lsk* + 1) i za dużej o dwa (*lsk* + 2)

| Typ rozkładu nieistotnego | *lsk* = 3 | | *lsk* + 1 | | *lsk* + 2 | |
|---|---|---|---|---|---|---|
| | pamięć | precyzja | pamięć | precyzja | pamięć | precyzja |
| a | 0,802 | 0,750 | 0,922 | 0,840 | 0,915 | 0,771 |
| b | 0,937 | 0,862 | 0,939 | 0,792 | 0,910 | 0,740 |
| c | 0,891 | 0,825 | 0,848 | 0,771 | 0,799 | 0,711 |
| d | 0,674 | 0,615 | 0,611 | 0,626 | 0,586 | 0,588 |
| e | 0,457 | 0,419 | 0,394 | 0,366 | 0,365 | 0,350 |
| f | 0,904 | 0,936 | 0,889 | 0,836 | 0,860 | 0,786 |
| g | 0,798 | 0,755 | 0,723 | 0,701 | 0,689 | 0,678 |
| Średnio | 0,78 | 0,737 | 0,761 | 0,705 | 0,732 | 0,661 |

Źródło: obliczenia własne.

Podsumowując przeprowadzony eksperyment, można stwierdzić, że metoda HINoV raczej nie jest odporna na błędnie (przy błędzie proporcjonalnym do liczby skupień) zadaną liczbę skupień.

## Literatura

Brusco M., Cradit D., A *variable-selection heuristic for k-means clustering*, „Psychometrika" 2001 no 66.

Carmone F.J. Jr., Kara A., Maxwell S., *HINoV: a new model to improve market segment definition by identifying noisy variables*, „Journal of Marketing Research" 1999 vol. 36.

Dash M., Liu H., *Feature Selection for Clustering* , Proceedings of Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD), 2000.

Dy J., Brodley C., *Feature Subset Selection and Order Identification for Unsupervised Learning*, Proc. 17th International Conf. on Machine Learning, 2000.

Friedman J., Meulman J., *Clustering objects on subsets of attributes*, „Journal of the Royal Statistical Society", Series B 66, 2004.

Law M., Jain A., Figueiredo M., *Feature Selection in Mixture-Based Clustering* , IEEE Transactions on Pattren Analysis and Machine Inteligence, 2003.

Montanari A., Lizzani L., *A projection pursuit approach to variable selection*, „Computational Statistics and Data Analysis" 2001 vol. 35(4).

Najman K., Najman K., *Analityczne metody ustalania liczby skupień*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław 2005.

Raftery A.E., Dean N., *Variable Selection for Model Based Clustering,* JASA 101, 2006.

Steinley D., Brusco M., *A new variable weighting and selection procedure for k-means cluster analysis*, „Multivariate Behavioral Research" 2008 no 43.

Steinley D., Brusco M., *Selection of variables in cluster analysis: an empirical comparison of eight procedures*, „Psychometrika" 2008 no 73.

## INVESTIGATING THE ROBUSTNESS OF HINOV TO WRONGLY PREDETERMINED NUMBER OF CLUSTERS

**Summary:** The HINoV method for choosing variables in the context of cluster analysis is a very popular one and one of the best [Steinley, Brusco 2008]. However, the efficiency of this method to the wrongly predetermined number of clusters remains an uninvestigated problem. The situation in which we cannot have precise knowledge about the number of clusters in a data set is very common since the indices most of which are of optimizing nature usually go wrong. In the paper, the robustness of HINoV is investigated in a broad simulation experiment on thousands of data sets in the form of the mixture of normal distributions. The organization of the experiment with respect to the number of variables, clusters, distributions etc. follows the experiment conducted by Steinley and Brusco [2008].