

O PEWNYM WYNIKU ORZEKAJĄCYM  
NIEMOŻLIWOŚĆ PRZEPROWADZENIA  
ANALIZY SKUPIEŃ  
*Krótki dowód z komentarzem*

Stanisław Ambroszkiewicz i Jacek Koronacki

Instytut Podstaw Informatyki PAN

Zakopane, Grudzień 2010

- Uwagi wstępne - sformułowanie problemu
- Twierdzenie o niemożliwości
- Dowód
- Parę słów komentarza

Zastanawiając się nad właściwym zrozumieniem zadania analizy skupień, widzianym w całej jego ogólności (m.in. bez zakładania a priori liczby skupień),

Jon Kleinberg tak pisze w szeroko cytowanej pracy pt. *An Impossibility Theorem for Clustering* (w: S. Becker, S. Thrun, K. Obermayer (red.), *Advances in Neural Information Processing Systems* (2002), str. 446-453):

there has been relatively little work aimed at reasoning about clustering independently of any particular algorithm, objective function, or generative data model.

I proponuje ogólne podejście aksjomatyczne:

Niech  $S$  będzie zbiorem, na którym dokonywane mają być partycje.

Jest to dowolny zbiór skończony, powiedzmy o liczności  $n$ .

Odstęp  $d$  (*distance* w oryginale) niech będzie funkcją na iloczynie kartezyjskim  $S \times S$  o wartościach rzeczywistych, nieujemną i taką, że

$d(i; j) = 0$  wtedy i tylko wtedy, gdy  $i = j$  oraz  $d(i; j) = d(j; i)$ .

Niech  $f$  będzie funkcją przypisującą każdej funkcji odstępów  $d$  odpowiadającą jej partycję (funkcją klasteryzacyjną), spełniającą następujące warunki:

- 1. Niezmienniczość względem skali: Dla każdej funkcji odstępów  $d$  i każdego  $c > 0$ ,  $f(d) = f(cd)$ .
- 2. Bogactwo możliwych partycji:  $\text{Range}(f)$  równy jest zbiorowi wszystkich partycji zbioru  $S$ .
- Niech  $l$  będzie partycją zbioru  $S$ , zaś  $d_1$  i  $d_2$  dwoma odstępami na  $S$ . Mówimy, że  $d_2$  jest  $l$ -transformacją odstępów  $d_1$ , jeżeli
  - (a) dla wszystkich  $i, j$  należących do tego samego skupienia partycji  $l$ ,  $d_2(i, j) \leq d_1(i, j)$ ;
  - (b) dla wszystkich  $i, j$  należących do różnych skupień partycji  $l$ ,  $d_2(i, j) \geq d_1(i, j)$ .
- 3. Zgodność: Niech  $d_1$  i  $d_2$  będą dwoma odstępami. Jeżeli  $f(d_1) = l$  i  $d_2$  jest  $l$ -transformacją odstępów  $d_1$ , to  $f(d_2) = l$ .

**Dla żadnego  $n \geq 2$  nie istnieje funkcja klasteryzacyjna  $f$ , spełniająca warunki 1-3 (niezmienniczości, bogactwa i zgodności).**

Założmy, że dla odstępów  $d_1$  mamy  $f(d_1) = \{S\}$  (tzn. partycją zbioru  $S$  jest jedno skupienie  $S$ ; odstęp  $d_1$  istnieje na mocy w-ku 2).

Pokażemy, że dla dowolnego odstępów  $d_2 : f(d_1) = f(d_2)$ , czyli że w-k 2 nie może zachodzić.

Tym sposobem dowód będzie zakończony.

Niech  $c_1 = \min\{d_1(i, j) :$

$i, j$  są różnymi elementami należącymi do  $S$   
(jest to minimalna odległość w  $d_1$ ).

Niech  $C_1 = \max\{d_2(i, j) :$

$i, j$  są różnymi elementami należącymi do  $S$   
(jest to maksymalna odległość w  $d_2$ ).

Weźmy takie  $c > 0$ , że  $C_1/c < c_1$ .

Zdefiniujmy taką nową funkcję odstępów  $d'$ , że  $d'(i, j) = d_2(i, j)/c$ .

Ponieważ  $C_1$  jest największą wartością  $d_2$ , oraz  $C_1/c < c_1$ , więc również  $d'(i, j) < c_1$ . Wartość  $c_1$  jest najmniejsza dla  $d_1$ , więc również  $d'(i, j) < d_1(i, j)$  dla każdych różnych  $i, j$  należących do tego samego skupienia zbioru  $S$ .

Na mocy w-ku 3  $d'$  jest  $S$ -transformacją odstępów  $d_1$ , czyli  $f(d') = f(d_1)$ . Z kolei  $d'$  jest przeskalowaniem  $d_2$ , a więc na mocy w-ku 1,  $f(d') = f(d_2)$ .

Ostatecznie  $f(d_1) = f(d_2)$ , co kończy dowód.

Analogiczny dowód można by oprzeć wychodząc od takiego odstępu  $d_1$ , że  $f(d_1) = I = \{\{s\} \in S\}$ , (tzn. partycja  $I$  rozbija  $S$  na "singeltony"). Znowu łatwo pokazać, że dla dowodnego odstępu  $d_2$  otrzymujemy  $f(d_1) = f(d_2)$ .

Do wykazania tej tezy należy zwiększać wartości odstępu  $d_1$  definiując nowy odstęp  $d'$  (na mocy (3) prowadząc do otrzymania tej samej wartości funkcji  $f$  dla  $d'$  co dla  $d_1$ ), będący zarazem przeskalowanym odstępem  $d_2$ . Wówczas znowu otrzymamy  $f(d_1) = f(d') = f(d_2)$ .

- Oryginalny dowód Kleinberga, jakkolwiek dowodzący więcej niż potrzeba, jest długi i dość złożony. Na co wskazuje możliwość przeprowadzenia dowodu tak bardzo prostego, jak pokazany, jeśli nie na złe skomponowanie ogólnych aksjomatów analizy skupień?
- W dowodzie pokazaliśmy, że warunek bogactwa nie może być spełniony, jeśli spełnione są pozostałe dwa warunki.
- Ale to nie znaczy, że warunek zgodności albo niezmiennoczości ma sens! W niezmienionej, niby ogólnej postaci, każdy z wprowadzonych aksjomatów jest bez sensu z punktu widzenia **rzeczywistego** problemu analizy skupień. I tylko oraz aż tyle.
- Ważny cel, jaki przyświecał Kleinbergowi (i innym), nadal oczekuje właściwego sformułowania i rozwiązania. Naszym zdaniem, kluczem jest uwzględnienie potrzeby kompromisu między dopasowaniem do danych modelu uwzględniającego istnienie skupień oraz złożonością tego modelu.

**DZIĘKUJEMY PAŃSTWU ZA UWAGĘ**