

TAKSONOMIA 14

Klasyfikacja i analiza danych – teoria i zastosowania

Małgorzata Gliwa

Akademia Ekonomiczna w Katowicach

**WYKORZYSTANIE KRYTERIUM
KOŁMOGOROWA-SMIRNOWA PRZY BUDOWIE
DRZEWA KLASYFIKACYJNEGO
DLA ZBIORU OBIEKTÓW SYMBOLICZNYCH**

1. Wstęp

Podziału zbioru obiektów symbolicznych na jednorodne klasy można dokonać na wiele sposobów, jednak proces ten wygodnie jest przedstawić w postaci drzewa.

Efektywność tworzenia drzew klasyfikacyjnych zależy od wyboru sposobu podziału zbiorów obiektów w węzłach drzewa [Gatnar 1998]. Wybieramy tę cechę, która daje dwa najbardziej różniące się podzbiory. Takiego wyboru dokonać można na podstawie pewnej miary jakości podziału. W literaturze można spotkać wiele propozycji miar jakości podziału. Dwie najczęściej stosowane to funkcja entropii oraz wskaźnik zróżnicowania Giniego. Jednak żadna z tych miar nie wykorzystuje wiedzy o uporządkowaniu zbioru wartości cechy [Gatnar 1998].

Celem artykułu jest zaprezentowanie kryterium Kołmogorowa-Smirnowa (K-S) jako propozycji miary jakości podziału zbioru obiektów symbolicznych w węzłach drzewa. Jest to rozszerzenie wprowadzonej przez Friedmana (1977) metryki opartej na odległości Kołmogorowa-Smirnowa. K-S zakłada uporządkowanie zbioru wartości cechy [Mballo, Diday 2004].

2. Porządkowanie danych

Jak wspomniano, K-S wymaga, aby zbiór wartości cechy był uporządkowany, zatem kryterium to stosować można w odniesieniu do określonego charakteru

zmiennych. W artykule rozważono przypadek zmiennych przedziałowych oraz zmiennych wyrażonych w postaci listy kategorii z wagami [Diday 2002], a także omówiono sposób ich porządkowania.

2.1. Porządkowanie danych przedziałowych

Niech Ω będzie zbiorem wszystkich obserwacji, a α i β takimi dwiema liczbami rzeczywistymi, że dla każdego $\omega \in \Omega$, $Y(\omega) = [\alpha, \beta]$. Zbiór wszystkich przedziałów domkniętych oznaczmy przez $\mathfrak{S}$. Wówczas dla dowolnego przedziału $p \in \mathfrak{S}$, $p = [d_p, g_p]$, d_p oznacza dolną, a g_p górną granicę przedziału.

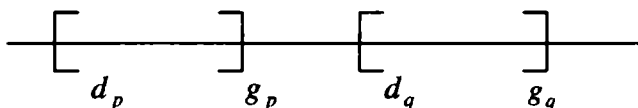
Podczas porządkowania danych przedziałowych rozważane są dwie możliwości, a mianowicie: przedziały są rozłączne lub mają część wspólną [Mballo, Diday 2004; 2005].

Przedziały rozłączne

Przez $D \subseteq \mathfrak{S}$ oznaczmy zbiór wszystkich przedziałów domkniętych rozłącznych. Niech $p, q \in D$. Na zbiorze D ustalamy relację mniejszości R w następujący sposób:

$$pRq \Leftrightarrow g_p < d_q, \quad (1)$$

co prezentuje następujący schemat:



Relacja określona w powyższy sposób definiuje porządek na zbiorze D .

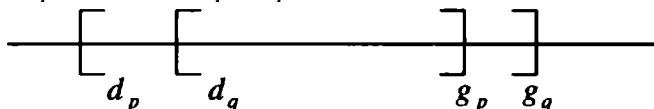
Przedziały o części wspólnej

W przypadku przedziałów mających część wspólną można mówić o uporządkowaniu względem dolnych lub górnych granic przedziałów. Dodatkowo należy rozważyć sytuację, kiedy odpowiednio dolne lub górne granice przedziałów są sobie równe.

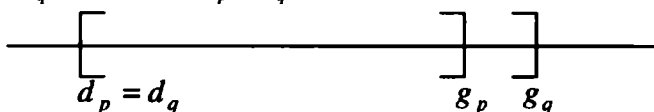
Porządkowanie względem dolnych granic przedziałów

Na zbiorze $\mathfrak{S}$ ustalamy relację mniejszości R w następujący sposób:

$$\text{– jeśli } d_p \neq d_q, \text{ to } pRq \Leftrightarrow d_p < d_q \quad (2)$$



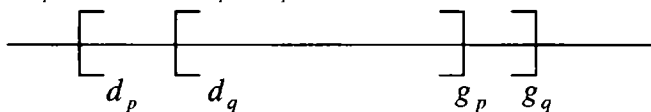
$$\text{– jeśli } d_p = d_q, \text{ to } pRq \Leftrightarrow g_p < g_q \quad (3)$$



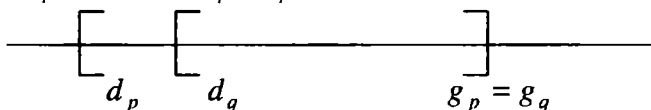
Porządkowanie względem górnych granic przedziałów

Porządkowanie względem górnych granic przedziałów odbywa się w analogiczny sposób. Na zbiorze $\mathfrak{I}$ ustalamy następującą relację mniejszości R :

- jeśli $g_p \neq g_q$, to $pRq \Leftrightarrow g_p < g_q$ (4)



- jeśli $g_p = g_q$, to $pRq \Leftrightarrow d_p < d_q$ (5)



Wszystkie zdefiniowane relacje są przechodnie i przeciwzwrotne, zatem określają porządek na zbiorze $\mathfrak{I}$.

2.2. Porządkowanie danych wyrażonych w postaci listy kategorii z wagami

Dane wyrażone w postaci listy kategorii z wagami porządkować można na wiele sposobów. Do najpopularniejszych zaliczamy porządkowanie za pomocą [Mballo, Diday 2004]:

- średniej, dla każdej obserwacji cechy obliczamy średnią z wszystkich wag,
- mediany, dla każdej obserwacji cechy wyznaczmy medianę wszystkich wag,
- odchylenia standardowego, dla każdej obserwacji cechy obliczamy odchylenie standardowe z wszystkich wag.

Wyznaczone na podstawie jednego z tych sposobów wartości porządkujemy w naturalny sposób. Jeżeli jednak wystąpi sytuacja, w której dwie lub więcej z obliczonych wartości są takie same, to kolejno wyznaczamy następną z miar i wprowadzamy porządek leksykograficzny [Mballo, Diday 2004].

3. Kryterium Kołmogorowa-Smirnowa

Po uporządkowaniu zbioru wartości cechy można wprowadzić kryterium Kołmogorowa-Smirnowa (K-S). W tym celu rozważamy zbiór uczący zawierający obiekty symboliczne $o_1, \dots, o_n$, które należą do jednej z m określonych *a priori* klas $K_1, \dots, K_m$. Klasy te grupujemy do dwóch tzw. superklas S_1 oraz S_2 . Mamy $2^{m-1} - 1$ możliwości podziału. Na przykład:

- dla dwóch *a priori* zdefiniowanych klas K_1, K_2 , istnieje tylko jedna możliwość wyznaczenia superklas ($2^{2-1} - 1 = 1$): $S_1 = \{K_1\}$, $S_2 = \{K_2\}$,

– dla trzech *a priori* zdefiniowanych klas K_1, K_2, K_3 , istnieją trzy możliwości wyznaczenia superklas ($2^{3-1} - 1 = 3$):

$$S_1 = \{K_1\}, S_2 = \{K_2, K_3\} \text{ lub } S_1 = \{K_1, K_2\}, S_2 = \{K_3\} \text{ lub } S_1 = \{K_1, K_3\}, S_2 = \{K_2\}.$$

Ponieważ podział zbioru obiektów symbolicznych następuje na podstawie wartości cech obiektów, w związku z tym w korzeniu drzewa, a następnie w każdym węźle, dla wszystkich wartości cech wyznaczamy dystrybuanty empiryczne według wzoru (6) [Mballo, Diday 2004; 2005]:

$$\hat{F}_i(p) = \frac{\text{card}\left(\left\{p \in D_{x_j} : p \leq q\right\} \cap \left\{p \in D_{x_j} : p \in S_i\right\}\right)}{\text{card}\left(\left\{p \in D_{x_j} : p \in S_i\right\}\right)}, \quad i=1, 2, \quad (6)$$

gdzie: $D_{x_j} \subseteq \mathfrak{S}$ to zbiór wszystkich wartości zmiennej objaśniającej X_j ,

S_i to jedna z superklas.

Kryterium Kołmogorowa-Smirnowa wyznaczamy jako różnicę dystrybuant wyznaczonych odpowiednio dla S_1 i S_2 [Mballo, Diday 2004; 2005]:

$$KS(p) = \sup_p \left| \hat{F}_1(p) - \hat{F}_2(p) \right|. \quad (7)$$

Należy dodać, że grupowanie k klas ($k \leq m$) do dwóch superklas S_1 i S_2 odbywa się w każdym węźle.

Po wskazaniu wszystkich możliwych układów superklas proces porządkowania wartości cech oraz podział zbioru obiektów symbolicznych (OS) odbywa się według algorytmu:

1. Dla każdego układu superklas uporządkuj zbiór wartości cech.
2. Dla każdej wartości cechy z każdego układu wyznacz K-S.
3. Porównaj obliczone wartości K-S.
4. Wybierz tę cechę, dla której K-S jest największe.
5. Podziel zbiór OS względem wybranej cechy.
6. Kroki 1-5 powtarzaj, aż do zakończenia budowy drzewa.

4. Przykład wykorzystania K-S

Rozważamy 10 obiektów symbolicznych, które opisują państwa rozwinięte lub rozwijające się z 5 kontynentów. Państwa scharakteryzowane zostały przez 5 zmiennych objaśniających:

- X_1 – ludność (w tys.),
- X_2 – roczny przyrost naturalny (w %),
- X_3 – powierzchnia (w tys. km²),

– X_4 – średnia długość życia (w latach),

– X_5 – odsetek analfabetów (w %).

Zmienną określającą przynależność obiektu do klasy jest stan rozwoju ekonomicznego kraju: 0 – kraje rozwinięte, 1 – kraje rozwijające się. Obiekty te zostały przedstawione w tab. 1. Na przykład obiekt symboliczny AME0 oznacza kraje rozwinięte Ameryki, a EUR1 – rozwijające się państwa europejskie.

Tabela 1. Tabela obiektów symbolicznych

Nazwa	X_1	X_2	X_3	X_4	X_5	Klasa
AME0	[15211;283230]	[0,9;1,8]	[757;9976]	[67,2;78,5]	[0;14,6]	0
AME1	[2576;25662]	[0,8;2,6]	[11;1285]	[64;76]	[4,3;38,7]	1
EUR0	[4469;82017]	[0;0,5]	[41;547]	[77,2;79,3]	[0;0]	0
EUR1	[1988;145491]	[-0,4;0,2]	[20;17075]	[66,1;78,2]	[0;0]	1
OCE0	[814;19138]	[0,9;1,2]	[18;7682]	[68,4;78,7]	[0;9,2]	0
OCE1	[159;4809]	[0;2,7]	[3;462]	[55,6;70,5]	[21;52]	1
AFR0	[1161;67884]	[0,8;1,9]	[2;1221]	[56,7;70,7]	[15,4;63,9]	0
AFR1	[1757;35119]	[0,9;2,9]	[196;945]	[40,6;52,3]	[15,3;72,3]	1
AZJ0	[4913;1275130]	[0,3;2,9]	[20;9597]	[65,1;80,5]	[1;23,7]	0
AZJ1	[16189;1008940]	[1,4;2,6]	[185;3287]	[62,3;71,9]	[4,8;54,6]	1

Źródło: <http://www.ceremade.dauphine.fr/%7Etuati/pays1.htm>.

Dla każdej zmiennej porządkujemy zbiór wartości cechy. Występujące w każdej komórce tabeli nawiasy to klasa, z której pierwotnie pochodziła cecha charakteryzująca obiekt. Otrzymane wartości przedstawia tab. 2.

Tabela 2. Wartości cech uporządkowane względem lewych granic przedziałów

X_1	X_2	X_3	X_4	X_5
[159;4809] (1)	[-0,4;0,2] (1)	[2;1221] (0)	[40,6;52,3] (1)	[0;0] (0)
[814;19138] (0)	[0;0,5] (0)	[3;462] (1)	[55,6;70,5] (1)	[0;0] (1)
[1161;67884] (0)	[0;2,7] (1)	[11;1285] (1)	[56,7;70,7] (0)	[0;9,2] (0)
[1757;35119] (1)	[0,3;2,9] (0)	[18;7682] (0)	[62,3;71,9] (1)	[0;14,6] (0)
[1988;145491] (1)	[0,8;1,9] (0)	[20;9597] (0)	[64;76] (1)	[1;23,7] (0)
[2576;25662] (1)	[0,8;2,6] (1)	[20;17075] (1)	[65,1;80,5] (0)	[4,3;38,7] (1)
[4469;82017] (0)	[0,9;1,2] (0)	[41;547] (0)	[66,1;78,2] (1)	[4,8;54,6] (1)
[4913;1275130] (0)	[0,9;1,8] (0)	[185;3287] (1)	[67,2;78,5] (0)	[15,3;72,3] (1)
[15211;283230] (0)	[0,9;2,9] (1)	[196;945] (1)	[68,4;78,7] (0)	[15,4;63,9] (0)
[16189;1008940] (1)	[1,4;2,6] (1)	[757;9976] (0)	[77,2;79,3] (0)	[21;52] (1)

Źródło: opracowanie własne na podstawie tab. 1.

W kolejnym kroku sprawdzamy, jakie są możliwości utworzenia superklas. Ponieważ są dwie klasy zdefiniowane *a priori*, jest tylko jedna możliwość utworzenia superklas: $S_1 = \{0\}$, $S_2 = \{1\}$. Następnie ze wzoru (7) dla każdej cechy obliczamy wartości K-S. Ostateczne wyniki przedstawione zostały w tab. 3.

Tabela 3. Wartości K-S

X_1	X_2	X_3	X_4	X_5
1/5	1/5	1/5	1/5	1/5
0	0	0	2/5	0
1/5	1/5	1/5	1/5	1/5
0	0	0	2/5	2/5
1/5	1/5	1/5	3/5	3/5
2/5	0	0	2/5	2/5
1/5	1/5	1/5	3/5	1/5
0	2/5	0	2/5	0
1/5	1/5	1/5	1/5	1/5

Źródło: obliczenia własne.

W następnym kroku następuje podział zbioru OS na dwa podzbiory względem cechy, dla której wartość K-S jest największa (zob. tab. 3). Proces ten następuje po uprzedniej identyfikacji obiektów symbolicznych na podstawie zmiennej X_5 . Wyniki prezentuje tab. 4.

Tabela 4. OS podzielone względem odpowiedniej cechy

Wyszczególnienie	X_1	X_2	X_3	X_4	X_5
EUR0	[4469;82017]	[0;0,5]	[41;547]	[77,2;79,3]	[0;0]
EUR1	[1988;145491]	[-0,4;0,2]	[20;17075]	[66,1;78,2]	[0;0]
OCE0	[814;19138]	[0,9;1,2]	[18;7682]	[68,4;78,7]	[0;9,2]
AME0	[15211;283230]	[0,9;1,8]	[757;9976]	[67,2;78,5]	[0;14,6]
AZJ0	[4913;1275130]	[0,3;2,9]	[20;9597]	[65,1;80,5]	[1;23,7]
AME1	[2576;25662]	[0,8;2,6]	[11;1285]	[64;76]	[4,3;38,7]
AZJ1	[16189;1008940]	[1,4;2,6]	[185;3287]	[62,3;71,9]	[4,8;54,6]
AFR1	[1757;35119]	[0,9;2,9]	[196;945]	[40,6;52,3]	[15,3;72,3]
AFR0	[1161;67884]	[0,8;1,9]	[2;1221]	[56,7;70,7]	[15,4;63,9]
OCE1	[159;4809]	[0;2,7]	[3;462]	[55,6;70,5]	[21;52]

Źródło: opracowanie własne.

Pierwsze podziały drzewa przeprowadzono również względem innych wartości cech, dla których K-S wynosiło 3/5 (zob. tab. 3), i ostatecznie otrzymano te same rezultaty.

Tabela 5. Uporządkowane wartości cech dla lewego węzła drzewa

X_1	X_2	X_3	X_4	X_5
[814;19138] (0)	[-0,4;0,2] (1)	[18;7682] (0)	[65,1;80,5] (0)	[0;0] (0)
[1988;145491] (1)	[0;0,5] (0)	[20;9597] (0)	[66,1;78,2] (1)	[0;0] (1)
[4469;82017] (0)	[0,3;2,9] (0)	[20;17075] (1)	[67,2;78,5] (0)	[0;9,2] (0)
[4913;1275130] (0)	[0,9;1,2] (0)	[41;547] (0)	[68,4;78,7] (0)	[0;14,6] (0)
[15211;283230] (0)	[0,9;1,8] (0)	[757;9976] (0)	[77,2;79,3] (0)	[1;23,7] (0)

Źródło: opracowanie własne.

Określanie superklas, porządkowanie wartości cech oraz obliczanie K-S powtórzo dla lewego i prawego węzła drzewa. Wyniki przedstawiono w tab. 5-8.

Tabela 6. Uporządkowane wartości cech dla prawego węzła drzewa

X_1	X_2	X_3	X_4	X_5
[159;4809] (1)	[0;2,7] (1)	[2;1221] (0)	[40,6;52,3](1)	[4,3;38,7] (1)
[1161;67884] (0)	[0,8;1,9] (0)	[3;462] (1)	[55,6;70,5](1)	[4,8;54,6] (1)
[1757;35119] (1)	[0,8;2,6] (1)	[11;1285] (1)	[56,7;70,7](0)	[15,3;72,3](1)
[15211;283230] (0)	[0,9;2,9] (1)	[185;3287](1)	[62,3;71,9 (1)	[15,4;63,9](0)
[16189;1008940](1)	[1,4;2,6] (1)	[196;945] (1)	[64;76] (1)	[21;52] (1)

Źródło: opracowanie własne.

Tabela 7. Wartości K-S dla tabeli 5

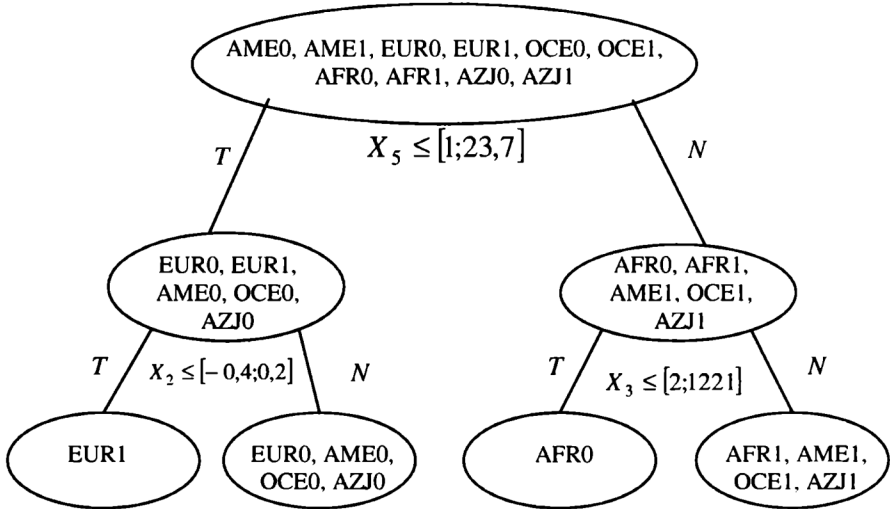
X_1	X_2	X_3	X_4	X_5
$\frac{1}{4}$	1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\frac{3}{4}$	$\frac{3}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{3}{4}$
$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
0	0	0	0	0

Źródło: obliczenia własne.

Tabela 8. Wartości K-S dla tabeli 6

X_1	X_2	X_3	X_4	X_5
$\frac{1}{4}$	$\frac{1}{4}$	1	$\frac{1}{4}$	$\frac{1}{4}$
$\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{2}{4}$	$\frac{2}{4}$
$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{3}{4}$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
0	0	0	0	0

Źródło: obliczenia własne.



Rys. 1. Drzewo klasyfikacyjne
Źródło: opracowanie własne.

Z przedstawionych tabel wynika, że dla zarówno lewego, jak i prawego węzła drzewa największa wartość K-S wynosi 1 i odnosi się odpowiednio do cech

$X_2 = [-0,4; 0,2]$ i $X_3 = [2; 1221]$. W kolejnym kroku, po przeprowadzeniu identyfikacji obiektów symbolicznych na podstawie powyższych wartości cech, a następnie podziale zbioru obiektów, otrzymano homogeniczne podzbiory. W związku z powyższym na tym etapie budowa drzewa zostaje zakończona. Otrzymane drzewo klasyfikacyjne przedstawiono na rys. 1. Na podstawie zaprezentowanego drzewa można przedstawić cztery reguły klasyfikacji. Na przykład, przemieszczając się wzdłuż kolejnych krawędzi drzewa do trzeciego liścia drzewa, otrzymamy następującą regułę klasyfikacji:

jeśli $X_3(z) > [1; 23, 7] \wedge X_3(z) \leq [2; 1221]$, to $Y(z) = 0$,

gdzie: z to obserwacja, którą chcemy zaklasyfikować, 0 – klasa zdefiniowana *a priori*. Jak widać, z zostanie zaklasyfikowane do grupy państw rozwiniętych.

5. Podsumowanie

Zaprezentowane w artykule kryterium Kołmogorowa-Smirnowa to narzędzie, na podstawie którego dokonywany jest wybór sposobu podziału zbioru obiektów symbolicznych w węzłach drzewa. Zaletą tego kryterium jest fakt, iż w przeciwieństwie do popularnych miar jakości podziału uwzględnia ono wiedzę o uporządkowaniu zbioru wartości cechy. Niestety, miara ta nie jest niezależna od charakteru zmiennych. Wobec tego otwarty staje się problem rozważenia stosowności tej miary w odniesieniu do innych typów danych.

Literatura

- Diday E. (2002), *An Introduction to Symbolic Data Analysis and the SODAS Software*, „The Electronic Journal of Symbolic Data Analysis”.
- Friedman J. H., *A Recursive Partitioning Decision Rule for Non Parametric Classification*, „IEEE Transactions on Computers” C-26, s. 404-408.
- Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa, s. 163-176.
- Mballo C., Diday E. (2005), *Decision Trees on Interval Valued Variables*, „The Electronic Journal of Symbolic Data Analysis” t. 3, nr 1, czerwiec.
- Mballo C., Diday E. (2004), *Kolmogorov-Smirnov for Decision Tree on Interval and Histogram Variables*, [w:] D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (red.), *Classification, Clustering and Data Mining Applications*, Springer, Chicago.

USING OF THE KOLMOGOROV-SMIRNOV CRITERION WHILE BUILDING CLASSIFICATION TREE FOR SET OF SYMBOLIC OBJECTS

Summary

If we are interested in dividing symbolic objects' into groups, as regards to specific variable values, it would be more comfortable to present this process as a decision tree. To build decision tree, we need a criterion in order to split the symbolic objects' set in a nodes. In this article the Kolmogorov-Smirnov criterion is presented, as a splitting criterion of symbolic objects' set. For this criterion, the values taken by the explanatory variables have to be ordered. We considered different possible orders of interval and histogram variables. At the end we show an example to illustrate this approach.