ZASTOSOWANIA MATEMATYKI
V (1960)

B. KOPOCIŃSKI (Wrocław)

DYSKRYMINACJA ZA POMOCĄ DENDRYTÓW

1. Bardzo częstym zagadnieniem występującym w badaniach przyrodniczych jest podział zbioru punktów na części. Podziału tego można dokonać za pomocą dendrytu wrocławskiego [1], który jest najkrótszą linią łamaną łączącą wszystkie punkty swymi odcinkami. Podział na K części powstaje przez usunięcie w dendrycie $K-1$ dowolnych odcinków. Główną zaletą dendrytu w ogólności jest to, że pozwala on narysować na płaszczyźnie ważniejsze zależności między punktami z wielowymiarowej przestrzeni. Dendryt nie daje natomiast pełnego obrazu przestrzeni, w której znajdują się badane punkty, a zatem wnioskowanie na podstawie dendrytu oparte jest na mniejszej ilości informacji o tych punktach niż ta, którą mamy w pełnej tablicy Czekanowskiego wszystkich odległości między punktami.

Inną poważną wadą dendrytu jest to, że dorzucony do niego jeden punkt zmienia na ogół połączenia dendrytowe istniejące w starym dendrycie.

Ponadto w zastosowaniach bardzo często odległości między punktami są zmiennymi losowymi, a więc połączenia dendrytowe mogą być silnie zależne od przypadku.

2. Niedogodności te można pokonać za pomocą dendrytów wyższych stopni. Wyżej opisany dendryt nazwijmy *dendrytem* I *stopnia*. W tablicy Czekanowskiego w miejsce tych odległości, które weszły do połączeń w dendrycie I stopnia, wstawiamy ∞ i z takiej tablicy rysujemy dendryt, który nazywamy *dendrytem* II *stopnia*. Następnie wstawiamy ∞ w miejsce odległości użytych do dendrytu II stopnia, rysujemy dendryt III stopnia itd.

Postępowanie kończymy z chwilą wyczerpania wszystkich skończonych odległości.

Uwaga. Może się zdarzyć, że dendryt pewnego stopnia nie będzie spójny, tzn. otrzymamy dwie lub więcej grup punktów, między którymi nie będzie skończonych połączeń. W tym przypadku przyjmujemy między tymi grupami połączenie odcinkiem o długości ∞. Zauważmy, że dendryty

wszystkich stopni wyczerpują całą tablicę Czekanowskiego, a więc dla każdej pary punktów istnieje dendryt, w którym te punkty są połączone.

3. O punkcie mówimy, że *należy do grupy punktów*, o ile odległości od niego do punktów tej grupy są krótsze niż odległości od pozostałych punktów. Ustawmy odległości punktu od pozostałych punktów w ciąg rosnący. Punkt tym silniej należy do grupy, im więcej jest na początku tego ciągu odległości łączących go ze swoją grupą, a im mniej ich jest na końcu tego ciągu. Uwagi te znajdują niezwykle proste odzwierciedlenie w dendrytach wyższych stopni.

Jeżeli punkt należy do grupy punktów, to w pierwszych dendrytach łączy się z punktami tej grupy, a nie łączy się z nimi w ostatnich. Nie będziemy określać gdzie kończą się „pierwsze" i zaczynają „ostatnie" dendryty, zajmiemy się natomiast połączeniami dendrytowymi. Niech dendryt łączy dwie grupy punktów; przez 1 oznaczamy połączenie punktu z jego grupą, a przez 0 połączenie punktu z grupą przeciwną. Do każdego punktu możemy zbudować ciąg złożony z zer i jedynek; na j-tym miejscu wstawiamy 0 lub 1 zależnie od tego, czy w dendrycie j-tego rzędu punkt ten łączy się z grupą przeciwną, czy ze swoją.

Nasuwa się prosta charakterystyka S należenia punktu do grupy. Oznaczając przez k liczbę jedynek na początku a przez l liczbę zer na końcu wyżej opisanego ciągu, określamy S jako

$$S = k+l.$$

Zakładając hipotezę, że nie ma grup, i przyjmując wartość połączenia za zmienną losową podpadającą pod schemat urnowy (czyli stosując randomizację), możemy obliczyć prawdopodobieństwa

$$P(S = N) = \sum_{i=\max(0,N-n_2+1)}^{\min(n_1-2,N)} \frac{\binom{n_1-1}{i+1}\binom{n_2}{N-i+1}}{\binom{n_1+n_2-1}{i+1}\binom{n_1+n_2-i-2}{N-i+1}}$$

dla $N < n_1+n_2-3$,

$$P(S = n_1+n_2-1) = \frac{1}{\binom{n_1+n_2-1}{n_1-1}}$$

oraz

$$P(S > S_0) = 1-\sum_{N=0}^{S_0} P(S = N),$$

gdzie n_1 jest liczbą punktów zbioru, do którego należy badany punkt, a n_2 liczbą pozostałych punktów.

Przyjmując poziom istotności α możemy znaleźć takie S_α, że $P(S > S_\alpha) < \alpha$. Mówimy wtedy: gdy $S > S_\alpha$, to *punkt należy do grupy na poziomie istotności* α.

Moc tego testu jest jednak mała. Przyczyną tego jest wada dendrytu polegająca na łączeniu przeciwstawnych grup, chociaż mogą one być znacznie od siebie oddalone. Tym samym dla pewnej pary punktów już w pierwszych połączeniach dendrytowych znajdzie się połączenie mające wartość zero. Mimo to przedstawiona tu charakterystyka stanowi niezwykle prostą i łatwą ilustrację przeprowadzonego podziału.

4. Przykład zastosowania. Mamy 20 chorych bądź na epilepsję, bądź na histerię. U chorych tych obserwowano 17 cech alternatywnych (posiada = 1, nie posiada = 0), oraz notowano wiek (cecha 1) i stan cywilny (cecha 2). Dane przedstawia tablica 1. Tablicę Czekanowskiego (tablica 2) obliczamy z cech unormowanych na 0 i 1 ($x_i = 0$, $\sigma_i = 1$), za odległość d_{ij} od i-tej do j-tej osoby przyjmujemy

$$d_{ij} = d_{ji} = \sum_{k=1}^{19} |x_{kj} - x_{ki}| .$$

Otrzymuje się tutaj 12 dendrytów, ostatnim jest dendryt XII stopnia, który zawiera tylko jedno połączenie skończone.

TABLICA 1

Osoby	Cechy																		Diagnoza
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	18	19	23	24	
1	24	1		1				1		1	1	1	1	1		1		1	1	histeria
2	64	1			1	1	1	1		1					1					epilepsja
3	35		1		1	1			1								1			,,
4	39	1		1	1				1	1					1					,,
5	24			1	1	1			1								1			,,
6	16			1			1	1	1	1							1			,,
7	34	3		1	1	1				1	1						1			,,
8	28	1					1		1	1	1									,,
9	19				1		1		1					1	1		1			,,
10	19	1		1	1	1				1	1					1			1	histeria
11	54	3					1		1	1							1			epilepsja
12	16,5		1			1				1	1	1	1	1		1		1	1	histeria
13	17									1	1	1		1					1	,,
14	17									1	1	1				1		1	1	,,
15	41	1			1	1	1		1	1	1		1	1		1			1	,,
16	41	1			1					1	1		1	1	1	1		1	1	,,
17	45	1			1		1	1			1		1	1	1	1			1	,,
18	21		1			1		1	1	1	1	1	1			1		1	1	,,
19	33									1	1			1		1			1	,,
20	40	1				1				1	1	1	1			1		1	1	,,

TABLICA 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0																			
2	29,88	0																		
3	34,77	20,41	0																	
4	25,03	13,49	15,29	0																
5	28,55	21,07	6,22	11,60	0															
6	22,75	19,09	18,07	14,43	11,84	0														
7	23,96	18,38	14,29	12,93	10,37	17,21	0													
8	21,63	16,89	17,52	11,95	16,40	11,02	14,91	0												
9	33,71	17,07	11,04	12,13	9,18	11,66	21,72	13,00	0											
10	15,46	20,20	20,47	14,74	14,25	19,95	9,66	15,35	22,27	0										
11	30,65	15,71	16,19	14,24	16,86	11,48	14,70	9,01	13,46	23,50	0									
12	10,89	32,46	25,48	31,41	31,19	30,29	26,33	23,65	32,84	16,67	32,67	0								
13	12,45	25,41	23,94	20,29	22,08	19,28	19,23	12,53	21,72	13,57	21,56	11,12	0							
14	10,37	27,49	26,02	22,37	24,26	21,36	21,31	14,61	23,80	11,65	23,55	9,04	6,28	0						
15	18,93	19,60	21,43	19,14	22,10	24,74	19,23	13,73	22,71	13,02	19,75	16,93	18,03	20,31	0					
16	12,94	21,57	27,84	16,91	28,51	31,33	21,62	20,32	24,68	15,41	26,34	14,96	16,40	14,32	10,61	0				
17	17,86	17,01	28,03	22,10	28,61	27,71	26,80	21,30	20,67	20,60	26,40	24,51	21,59	23,86	11,60	9,55	0			
18	12,31	29,61	23,17	29,10	26,88	24,40	28,05	21,35	30,99	18,84	30,38	6,93	17,95	11,67	18,82	20,87	25,80	0		
19	12,86	23,38	21,92	18,27	22,13	20,95	17,20	11,67	22,92	10,77	19,52	13,14	6,02	8,30	12,01	10,38	15,57	19,05	0	
20	10,44	24,05	25,49	23,19	26,15	28,99	19,26	17,97	30,90	13,05	[illegible]	[illegible]	[illegible]	[illegible]	12,68	10,71	20,27	10,16	12,25	0

TABLICA 3

Numer osoby	Ciąg połączeń																		S
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	19
2	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	12
3	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	19
4	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	17
5	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	15
6	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	19
7	0	1	1	1	1	0	1	1	1	0	0	0	0	1	0	0	0	0	5
8	1	1	0	1	0	1	1	0	0	1	1	0	1	0	0	0	0	0	8
9	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	17
10	0	1	1	1	0	0	1	0	1	1	1	1	1	0	0	1	0	0	3
11	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	19
12	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	17
13	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	1	0	0	0	7
14	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	11
15	1	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	13
16	1	1	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	15
17	1	1	1	0	1	1	0	1	0	1	0	1	1	1	0	0	0	0	0	8
18	1	1	1	1	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0	15
19	1	1	1	1	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	9
20	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	17

Rysunek 1 pokazuje pierwsze trzy dendryty. Na rysunku grupę histeryków wyróżniono przez zakreskowanie odpowiednich kółek dendrytów. Jest to diagnoza lekarzy (1). Aby zobaczyć, jak silnie rozdzielone są obie grupy, weźmy ciągi wartości połączeń i obliczmy prawdopodobieństwa. Mamy

dla histeryka:

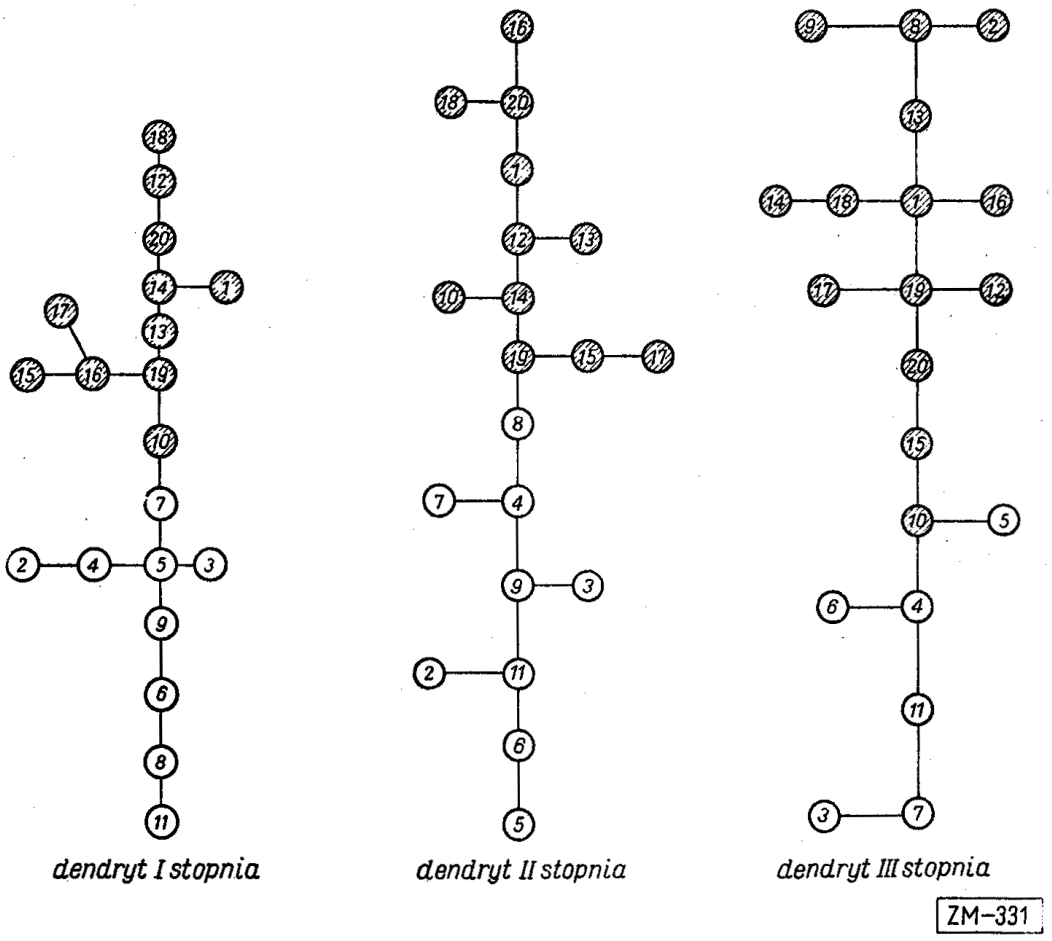
$$P(S > 3) = 0{,}1564,$$
$$P(S > 4) = 0{,}0835,$$
$$P(S > 5) = 0{,}0432,$$
$$P(S > 7) = 0{,}0087$$

dla epileptyka:

$$P(S > 1) = 0{,}7426,$$
$$P(S > 3) = 0{,}1674,$$
$$P(S > 4) = 0{,}0937,$$
$$P(S > 5) = 0{,}0492.$$

Wskaźnik S wskazuje, że większość osób bardzo silnie należy do swoich grup. Wątpliwości dotyczyć mogą jedynie osoby nr 10. Gdybyśmy jednak

(1) Źródłem materiałów jest Institut National d'Étude du Travail et d'Orientation Professionnelle, Paris.

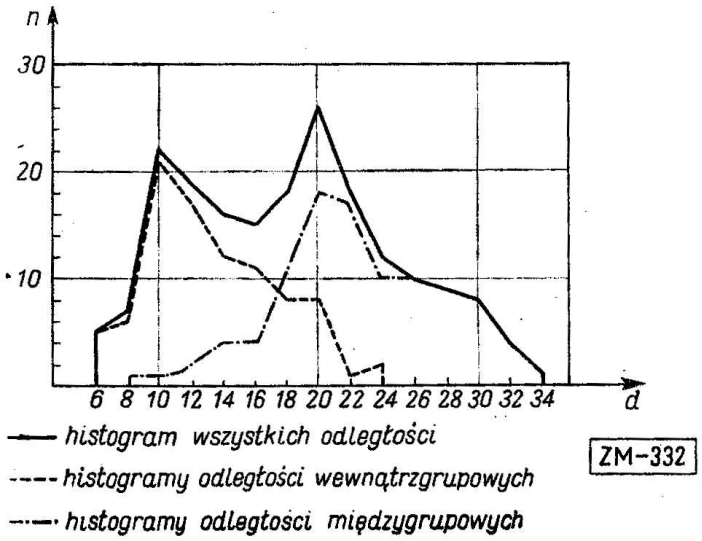


Rys. 1

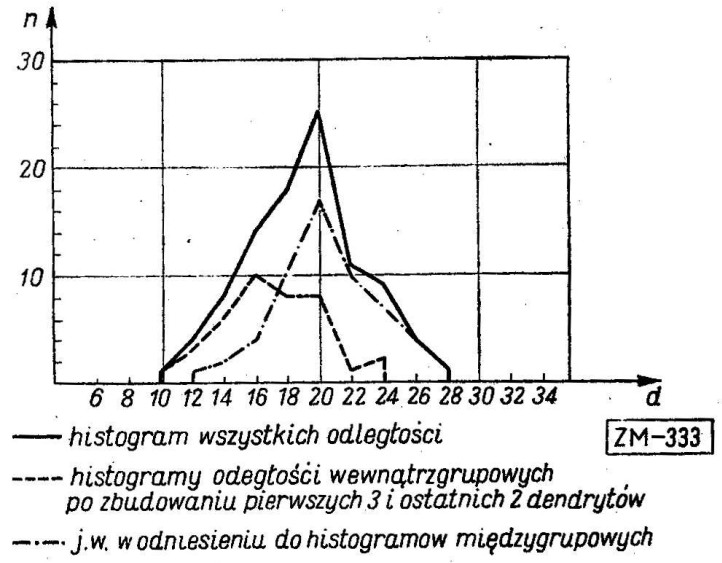
zaliczyli ją do grupy przeciwnej, wątpliwości byłyby jeszcze większe, $P(S > 1) = 0{,}7426$.

5. Opisany tu sposób postępowania pomimo swej prostoty zawiera pewne niedogodności. Przede wszystkim należy tu rysować dendryty wszystkich stopni, mimo że dendryty środkowe nie odgrywają na ogół poważniejszej roli, zawierają bowiem one długości średnie, które mogą być połączeniami zarówno międzygrupowymi, jak i wewnątrzgrupowymi i trudno na ich podstawie sądzić o podziale. Ponadto trudno obliczyć, ile informacji o podziale zawierają już narysowane dendryty. Niedogodności te można usunąć zastępując dendryty ostatnich stopni przez dendryty niepodobieństw, które buduje się z najdłuższych połączeń. W tej sytuacji połączenie obrazuje przechodzenie z jednej grupy do drugiej. Ilość dendrytów zawierających informacje o podziale możemy również wyznaczyć analizując histogram odległości. Przykład histogramu pokazano na rysunku 2, gdzie d jest odległością punktów, szerokość klasy wynosi 2, a częstość n umieszczono każdorazowo nad prawym kresem klasy. Histo-

gram odległości (rys. 2) jest sumą dwóch histogramów: odległości wewnątrzgrupowych i odległości międzygrupowych. Jeżeli grupy wzajemnie się przenikają, to te dwa histogramy pokrywają się. Inaczej jest, gdy w zbiorze punktów istnieją wyraźne grupy; w tym przypadku odległości międzygrupowe są większe od odległości wewnątrzgrupowych, oba te rozkłady są bardziej odległe i wtedy łączny rozkład (który możemy obserwować) jest nieregularny.



Rys. 2



Rys. 3

Tak więc nieregularność histogramu mówi o istniejącej w rozkładzie informacji o podziale na grupy. Budujemy teraz dendryty najkrótsze i najdłuższe, które wybierają najkrótsze i najdłuższe odległości ze zbiorów wszystkich odległości. Pozostające odległości zawierają coraz mniej

informacji o podziale na grupy, a ich histogram staje się coraz regularniejszy. Rysunek 3 przedstawia histogramy odległości pozostałych po zbudowaniu trzech pierwszych i dwóch ostatnich dendrytów. Wydaje się, że nie warto wyczerpywać pozostałych dendrytów.

Praca cytowana

[1] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus, S. Zubrzycki, *Sur la liaison et la division des points d'un ensemble fini*, Coll. Math. 3-4 (1951), str. 282-285. Tychże autorów: *Taksonomia wrocławska*, Przegl. Antrop. 27 (1951), str. 193-211.

Praca wpłynęła 12. 8. 1959

Б. КОПОЦИНСКИ (Вроцлав)

ДИСКРИМИНАЦИЯ ПРИ ПОМОЩИ ДЕНДРИТОВ

РЕЗЮМЕ

Данная работа дает возможность более полного использования таблицы Чекановского, чем в случае обычной дискриминации при помощи дендрита. Достигается это введением дендритов высших степеней. Дендрит высшей степени вычеркивается с таблицы Чекановского, в которою на место уже использованных расстояний вставляется ∞.

После выполнения распределения для данной точки составляется последовательность, членами которой являются значения последовательных соединений этой точки в дендритах.

Принимаем:

1 — значение соединения точки со своей группой,

0 — значение соединения точки с группой остальных точек.

Индекс S определяется общей суммой числа единиц в начале ряда и числа нулей в конце его. Вероятность $P(S > S_0)$ показывает, как сильно данная точка принадлежит к своей группе. Предлагается заменять дендриты найвысших степеней на дендриты длинейших соединений.

Индекс S обозначает число дендритов, требующихся для дискриминации. Число это также можно получить, анализируя гистограммы расстояния, принимая, что нерегулярность гистограммы информирует о распределении на группы.

B. KOPOCIŃSKI (Wrocław)

DISCRIMINATION BY MEANS OF DENDRITES

SUMMARY

The present paper shows that it is possible to make a fuller use of Czekanowski's table than is usually the case in discriminating by means of a dendrite. This is obtained by introducing dendrites of higher degrees. A dendrite of higher degree is drawn from Czekanowski's table, where in place of distances already used we put ∞.

Having made the division for a given point we form a sequence whose terms are the values of the successive connections of that point in dendrites.

We assume that

1 is the value of the connection of the point with its group,

0 is the value of the connection of the point with the group of the remaining points.

The index S is defined as the sum of the number of unities at the beginning of the sequence and the number of zeros at the end of the sequence. The probability $\Pr(S > S_0)$ shows how strongly a given point belongs to its group. It is suggested that the dendrites of the highest degrees should be replaced by the dendrites of the longest connections.

The index S determines the number of dendrites necessary for the discrimination. That number can also be determined by analysing the histogram of distances if we assume that the irregularity of the histogram informs of the division into groups.