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

Jan Meus, Justyna Stefaniak
Uniwersytet Jagielloński w Krakowie

ANALIZA I KLASYFIKACJA DANYCH W ASPEKCIE WSPÓŁCZYNNIKA *Z*

1. Wstęp

Wysoka złożoność zjawisk, którymi zajmujemy się coraz częściej, powoduje, że często badanie odosobnionych zmiennych nie prowadzi do konstruktywnych wniosków. Dotyczy to przede wszystkim zjawisk biologicznych, w których przedmiot badania – organizmy żywe – są opisane przez układy niezwykle skomplikowanych zmiennych wzajemnie oddziaływających na siebie. W efekcie wiele stanów charakteryzujących tak skomplikowane systemy traktujemy jako cechy jakościowe, co w analizie statystycznej prowadzi do badania zależności w tablicach kontyngencji [Bjømstad 1979; Kendall, Stuart 1979].

Takie tablice obrazują przejrzyście strukturę analizowanych danych. Jeżeli rozpatrujemy populację generalną pod kątem dwóch czynników, dzielących populacje odpowiednio na r (klasyfikacja A) i c-klas (klasyfikacja B), to w celach tablicy (tab. 1) znajdują się prawdopodobieństwa pochodzenia wylosowanych elementów z grup, które były sklasyfikowane jednocześnie przez oba czynniki. W związku z tym możemy łatwo policzyć prawdopodobieństwa brzegowe odpowiednich kategorii A_i ($p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$) czy kategorii B_j ($p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$).

Tabela 1. Tablica klasyfikacji podwójnej

B→ / A↓	B_1	B_2	...	B_c	Ogółem
A_1	p_{11}	p_{12}	...	p_{1c}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	...	p_{2c}	$p_{2\cdot}$
...	...	...	...	...	...
A_r	p_{r1}	p_{r2}	...	p_{rc}	$p_{r\cdot}$
Ogółem	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot c}$	1

Źródło: opracowanie własne.

Jednym z podstawowych zagadnień jest ustalenie oraz ocena wzajemnych związków (współzależności).

O ile pojęcie niezależności jest dość oczywiste, co możemy symbolicznie zapisać w formie warunku:

$$p_{ij} = p_{i.} \cdot p_{.j} \text{ dla wszystkich } i \in I,\ j \in J\ |^1,$$

o tyle pojęcie zależności nie jest precyzyjnie zdefiniowane. Dlatego też wprowadza się rozmaite indeksy (współczynniki), które mają mierzyć, wyznaczać poziom natężenia zależności.

Do takich tradycyjnych, znanych od dawna miar należą m.in. współczynnik Pearsona, Czuprowa, Cramera czy też współczynnik Steffensena [Goodman, Kruskal 1979].

Ze względu na asymetryczność problemów pojawiających się podczas analizy powstały też współczynniki, w których wiedzę o tym, która klasyfikacja ma poprzedzać chronologicznie, przyczynowo lub w jakikolwiek inny sposób „drugą" klasyfikację, wykorzystano do tworzenia takich miar. Poniższe wzory są przykładem tego typu indeksów. Opierają się one na metodach przewidywania:

- optymalnego – zaproponowana przez Goodmana [Goodman, Kruskal 1979]):

$$\lambda_a = \frac{\sum_{j=1}^{c} p_{mj} - p_{m.}}{1 - p_{m.}},\quad \lambda_b = \frac{\sum_{i=1}^{r} p_{im} - p_{.m}}{1 - p_{.m}},$$

gdzie $p_{.m} = \max_j p_{.j}$, $p_{m.} = \max_i p_{i.}$, $p_{im} = \max_j p_{ij}$ oraz $p_{mj} = \max_i p_{ij}$;

- proporcjonalnego – zaproponowanego przez Wallisa [Goodman, Kruskal 1979]):

$$\eta_a = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{p_{ij}^2}{p_{.j}} - \sum_{i=1}^{r} p_{i.}^2}{1 - \sum_{i=1}^{r} p_{i.}^2},\quad \eta_b = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{p_{ij}^2}{p_{i.}} - \sum_{i=1}^{r} p_{.j}^2}{1 - \sum_{i=1}^{r} p_{.j}^2}.$$

2. Współczynnik zależności Z i jego własności

Wychodząc z założenia asymetryczności problemu, autorzy zaproponowali nieco inny wzór, wykorzystujący wszystkie informacje, jakie czerpiemy z klasyfikacji przedstawionych w analizowanej tabeli. Dla dwóch zdarzeń A_i oraz B_j definiujemy stopień zależności między tymi zdarzeniami:

Definicja 1. Przez kwadratowy współczynnik zależności zdarzenia A_i od zdarzenia B_j będziemy rozumieli wartość poniższego wyrażenia

$$Z^2(A_i : B_j) = 1 - \left[P(B_j)\frac{1-P(A_i \mid B_j)}{P(\overline{A_i})} \cdot \frac{1-P(\overline{A_i} \mid B_j)}{P(A_i)} + P(\overline{B_j})\frac{1-P(A_i \mid \overline{B_j})}{P(\overline{A_i})} \cdot \frac{1-P(\overline{A_i} \mid \overline{B_j})}{P(A_i)} \right]. \quad (1)$$

Przyjęte oznaczenia mają sugerować, które z rozpatrywanych zdarzeń jest przyczyną, a które jej skutkiem (A_i – skutek, B_j – przyczyna).

[1] $I := \{1, 2, ..., r\}$, zaś $J := \{1, 2, ..., c\}$.

Definicja ta bazuje na stwierdzeniu, że zdarzenie A_i jest tym bardziej zależne od zdarzenia B_j, im więcej informacji przekazuje nam zajście zdarzenia B_j o możliwości pojawienia się zdarzenia A_i. Rozumiemy przez to, że jeżeli zdarzenie A_i jest zależne od zdarzenia B_j, to, jeśli mamy informację, że zaszło zdarzenie B_j lub zdarzenie przeciwne, to nasze przekonanie o zajściu A_i lub jego dopełnienia jest bardziej prawidłowe niż bez tej informacji.

Gdy zamienimy rolami A_i z B_j, otrzymamy definicję tzw. **kwadratowego współczynnika zależności zdarzenia B_j od zdarzenia A_i.**

Uwaga. $Z^2(A_i : B_j) = \dfrac{\left[P(A_i \cap B_j) - P(A_i).P(B_j)\right]^2}{P(A_i).P(\overline{A_i}).P(B_j).P(\overline{B_j})}$.

Z powyższego widać, że niezależnie od tego, który ze zbiorów przyjmiemy za skutek, a który za przyczynę, wartość kwadratowego współczynnika Z jest taka sama, czyli $Z^2(A_i : B_j) = Z^2(B_j : A_i)$.

Wprowadzając definicję kwadratowego współczynnika zależności dowolnego zdarzenia A_i zależnego od np. zdarzeń B_1 oraz B_2, bierzemy pod uwagę nie tylko wpływ zdarzeń B_1 i B_2, ale również dopełnienie ich sumy. W związku z tym zaproponowany wzór wygląda tak jak w definicji drugiej (pojawia się tu dodatkowy składnik).

Definicja 2. Kwadratowy współczynnik $Z^2(A_i : B_1, B_2)$ zależności pojedynczego zdarzenia A_i zależnego od zdarzeń B_1 oraz B_2 wyraża się następująco:

$$Z^2(A_i : B_1, B_2) = 1 - [P(B_1)\frac{1-P(A_i \mid B_1)}{P(\overline{A_i})} \cdot \frac{1-P(\overline{A_i} \mid B_1)}{P(A_i)} + P(B_2)\frac{1-P(A_i \mid B_2)}{P(\overline{A_i})} \cdot$$
$$\cdot \frac{1-P(\overline{A_i} \mid B_2)}{P(A_i)} + P(\overline{B_1 \cup B_2})\frac{1-P(A_i \mid \overline{B_1 \cup B_2})}{P(\overline{A_i})} \cdot \frac{1-P(\overline{A_i} \mid \overline{B_1 \cup B_2})}{P(A_i)}] . \quad (2)$$

Jeśli wprowadzamy wzór dla pary zdarzeń, np. A_1 i A_2, zależnych od jakiegoś pojedynczego zdarzenia B_j, uwzględniamy oprócz zdarzeń A_1 A_2, zdarzenie będące ich dopełnieniem, co we wzorze jest widoczne jako dodatkowy czynnik w iloczynach.

Definicja 3. Kwadratowy współczynnik dla zdarzeń A_1 oraz A_2 zależnych od zdarzenia B_j obliczamy ze wzoru:

$$Z^2(A_1, A_2 : B_j) = 1 - [P(B_j)\frac{1-P(A_1 \mid B_j)}{P(\overline{A_1})} \cdot \frac{1-P(A_2 \mid B_j)}{P(\overline{A_2})} \cdot \frac{1-P(\overline{A_1 \cup A_2} \mid B_j)}{P(A_1 \cup A_2)} +$$
$$+ P(\overline{B_j})\frac{1-P(A_1 \mid \overline{B_j})}{P(\overline{A_1})} \cdot \frac{1-P(A_2 \mid \overline{B_j})}{P(\overline{A_2})} \cdot \frac{1-P(\overline{A_1 \cup A_2} \mid \overline{B_j})}{P(A_1 \cup A_2)}] . \quad (3)$$

Wszystko to, co napisano do tej pory, można sprowadzić do następującej postaci, będącej wzorem na kwadratowy współczynnik zależności dowolnej k-elemeto-

wej rodziny $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ – zdarzeń zależnych od – n-elementowej rodziny $B_{j_1}, B_{j_2}, \ldots, B_{j_n}$ – zdarzeń.

Definicja 4. Niech $\{i_1, i_2, \ldots, i_k\} \subset I$ oraz $\{j_1, j_2, \ldots, j_n\} \subset J$, wówczas dla pewnych ustalonych $k \leq c$ oraz $n \leq r$

$$Z^2(A_{i_1}, A_{i_2}, \ldots, A_{i_k} : B_{j_1}, B_{j_2}, \ldots, B_{j_n}) = \begin{cases} 1 - \sum\limits_{j \in \{j_1, \ldots, j_{n+1}\}} \left[p_{.j} \prod\limits_{i \in \{i_1, \ldots, i_{k+1}\}} \dfrac{1 - \frac{p_{ij}}{p_{.j}}}{1 - p_{i.}} \right], & n < c, k \leq r \\ 1 - \sum\limits_{j=1}^{c} \left[p_{.j} \prod\limits_{i \in \{i_1, \ldots, i_{k+1}\}} \dfrac{1 - \frac{p_{ij}}{p_{.j}}}{1 - p_{i.}} \right], & n = c, k \leq r \end{cases} \tag{4}$$

Wzór na kwadratowy współczynnik zależności rodziny zbiorów $B_{j_1}, B_{j_2}, \ldots, B_{j_n}$ -kategorii od rodziny zbiorów $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$ -kategorii przedstawia definicja 5. Zamieniają się ze sobą zbiory, po których indeksujemy sumę i iloczyn oraz związane z nimi prawdopodobieństwa brzegowe.

Definicja 5. $Z^2(B_{j_1}, B_{j_2}, \ldots, B_{j_n} : A_{i_1}, A_{i_2}, \ldots, A_{i_k}) =$

$$= \begin{cases} 1 - \sum\limits_{i \in \{i_1, \ldots, i_{k+1}\}} \left[p_{i.} \prod\limits_{j \in \{j_1, \ldots, j_{n+1}\}} \dfrac{1 - \frac{p_{ij}}{p_{i.}}}{1 - p_{.j}} \right], & n < c, k \leq r \\ 1 - \sum\limits_{i \in \{i_1, \ldots, i_{k+1}\}} \left[p_{i.} \prod\limits_{j=1}^{c} \dfrac{1 - \frac{p_{ij}}{p_{i.}}}{1 - p_{.j}} \right], & n = c, k \leq r \end{cases} \tag{5}$$

Wprowadzone indeksy $k + 1$ i $n + 1$ odnoszą się do poszczególnych dopełnień, a prawdopodobieństwa z nimi związane mają postać:

- $p_{i_{k+1}.} = \sum\limits_{i \in I \setminus \{i_1, \ldots, i_{k+1}\}} p_{i.}$,
- $p_{.j_{n+1}} = \sum\limits_{j \in I \setminus \{j_1, \ldots, j_{n+1}\}} p_{.j}$,
- $p_{i_{k+1}, j_{n+1}} = \sum\limits_{i \in I \setminus \{i_1, \ldots, i_{k+1}\}} \sum\limits_{j \in I \setminus \{j_1, \ldots, j_{n+1}\}} p_{ij}$.

Twierdzenie. Współczynnik Z ma następujące własności:

(i) Współczynnik zależności Z rodziny A-klas od rodziny B-klas jest nieokreślony, jeżeli populacja leży w jednym wierszu, czyli jest jedna A-klasa.

(ii) W przeciwnym razie wartość współczynnika zależności Z jest liczbą z przedziału [0,1].

(iii) Jeżeli rodzina zbiorów A jest niezależna od rodziny zbiorów B, to wartość współczynnika zależności $Z = 0$.

(iv) Jeżeli podział na A-klasy (odpowiednio B-klasy) jest rozdrobnieniem podziału na B-klasy (odpowiednio A-klasy), to $Z(B{:}A) = 1$ (odpowiednio $Z(A{:}B) = 1$), szczególnie, jeżeli podział na A-klasy jest istotnym rozdrobnieniem podziału na B-klasy, tzn. podziały się nie pokrywają, to $Z(A{:}B) = 1$ oraz $Z(B{:}A) = 1$.

(v) Jeżeli mamy ustalone podziały na A- oraz B-klasy i zaczniemy rozdrabniać rodzinę, którą traktujemy jako „wyjaśniającą", to współczynnik Z wzrasta wraz ze wzrostem liczby podziałów.

Przedstawione twierdzenie oddaje istotę tego, co nazywamy asymetrycznością.

3. Przykład

Zbadajmy, czy w populacji, liczącej 400 pacjentów (wiemy, z jakiej warstwy społecznej się wywodzą i na co chorują), zachodzi zależność pomiędzy ich chorobą a statusem społecznym. Status społeczny można podzielić na klasę robotniczą lub inteligencję zamieszkującą wieś lub miasto. Brane są pod uwagę następujące choroby: choroba płuc, cukrzyca, nadciśnienie tętnicze, depresja oraz nadwaga. Liczba chorych sklasyfikowanych jednocześnie na podstawie obu cechy znajduje się w tab. 2.

Tabela 2. Choroba a pozycja społeczna

Choroba→		A_1	A_2	A_3	A_4	A_5	Ogółem
Pozycja społeczna↓		choroby płuc	cukrzyca	nadciśnienie tętnicze	depresja	nadwaga	
B_1	inteligencja – miasto	10	5	28	6	80	129
B_2	robotnicy – miasto	30	5	12	18	40	105
B_3	inteligencja – wieś	10	5	20	6	40	81
B_4	robotnicy – wieś	30	5	20	10	20	85
Ogółem		80	20	80	40	180	400

Dane w tabeli mają charakter umowny i służą jedynie do ilustracji przedstawianej teorii.

Źródło: opracowanie własne.

Jeżeli zastąpimy prawdopodobieństwa przez częstości w definicji współczynnika Z i obliczymy wartości empirycznego współczynnika Z dla takich danych, zauważymy, że:

1. Najbardziej zróżnicowane są choroby płuc $Z(A_1{:}B) = 0{,}2842$ oraz nadwaga $Z(A_5{:}B) = 0{,}2897$. Zajmowana pozycja społeczna nie ma żadnego wpływu na wystąpienie cukrzycy $Z(A_2{:}B) = 0{,}0427$.

2. Największa zależność, będąca zarazem największą wartością współczynnika Z otrzymaną dla powyższego przykładu, zachodzi pomiędzy występującą jednocześnie chorobą płuc i depresją a statusem społecznym $Z(A_1{,}A_4{:}B) = 0{,}3061$. Wartość współczynnika dotyczącego występującej jednocześnie choroby płuc i nadwagi jest mniejsza i wynosi $Z(A_1{,}A_5{:}B) = 0{,}2088$.

3. Największą wartość zależności stanu społecznego od przedstawionych chorób otrzymujemy, gdy $Z(B_1{:}A) = 0{,}2861$. Można zatem wyciągnąć wniosek, że przedstawiciele inteligencji mieszkający w mieście są bardziej narażeni na zachorowanie niż przedstawiciele innych warstw społecznych.

4. Zastosowanie

Metoda wyznaczania współczynników Z w Zakładzie Bioinformatyki i Telemedycyny CM UJ w Krakowie znalazła rozmaite zastosowania. Między innymi:

1. Wykorzystano ją do oceny zależności struktury białka od sekwencji aminokwasów. Zależność przestrzennej struktury białka od sekwencji aminokwasów w łańcuchu polipeptydowym jest jednym z dogmatów biologii molekularnej. W analizowanej tabeli znalazło się 146 940 kolumn (sekwencje tetrapeptydów) oraz 2397 wierszy (możliwe konformacje tetrapeptydów). Tabela 3 przedstawia 20 największych wartości współczynnika Z (w postaci logarytmicznej) danej sekwencji aminokwasów do struktury [Meus i in.].

Tabela 3. Zależności typu struktury i rodzaju sekwencji białkowej, obliczone na podstawie danych ze strukturalnej białkowej bazy danych PDB. Alternatywna metoda jest oparta na teorii informacji Shannona [Shannon 1948].

Współczynnik Z *E-001	Typ struktury	Rodzaj sekwencji	Numer pozycji w alternatywnej metodzie	Współczynnik Z *E-001	Typ struktury	Rodzaj sekwencji	Numer pozycji w alternatywnej metodzie
7.45	AEGD	GNES	7	4.95	AEED	GLRL	2
6.92	BEBF	ERSY		4.91	DAFA	DGPG	
6.58	AFFB	GFRN	10	4.91	AEDE	GIFR	
6.57	AEBB	GPVY		4.88	AGAE	IKYG	
6.18	AEGE	GIGH	5	4.85	CFAG	NTGG	
5.78	ABFF	GPHF		4.79	EBCB	ELPD	8
5.71	EAAD	KGGP		4.77	CADD	RGRC	
5.49	GAGD	FNAG		4.67	GGED	KGHH	
5.34	DBAA	NGGD		4.66	AGCE	GGDD	
5.19	GCFG	GDSG	1;2	4.57	DFDA	YNPV	
5.16	DDBG	SHHG		4.56	BFBE	PEPV	6
5.10	BBDF	KTRS		4.54	GFDD	GQTN	
5.09	GDAE	ESGH		4.51	FAEA	PGFG	
5.03	DBEB	AERS		4.47	AEGF	GCAQ	
5.02	BACE	GGAE	3	4.47	ADFA	GTQC	
4.95	CABF	AGPH					

W ostatniej (czwartej) kolumnie przedstawiono uzyskane pozycje zależności danej sekwencji od struktury w analogicznej analizie zależności, liczonej metodami informatycznymi.

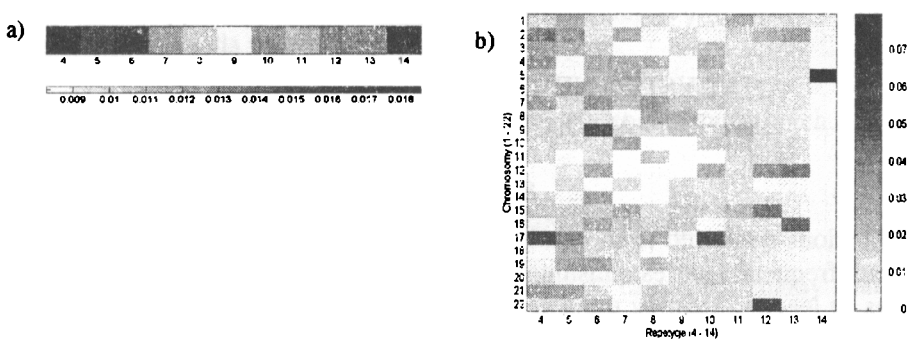
2. Metodę wykorzystano w analizie trinukleotydowych mikrosatelitów w genomie ludzkim na przykładzie tandemów nCAG, gdzie $n = 1, ..., 14$. Badania eksperymentalne wskazują, że tandemy nCAG, osiągające długość powyżej 10 powtó-

rzeń, występują stosunkowo rzadko i pojawiają się tylko w określonych partiach materiału genetycznego. Podobny wynik uzyskano po zastosowaniu metody Meusa w analizie tandemów *n*CAG w zsekwencjonowanym genomie ludzkim.

Tabela 4. Liczba tandemów mikrosatelitów (CAG) w chromosomach somatycznych człowieka

Wyszczególnienie	4CAG	5CAG	6CAG	7CAG	8CAG	9CAG	10CAG	11AG	12CAG	13CAG	14CAG
Chromosom01	198	37	21	7	3	2	2	2	0	0	0
Chromosom02	144	47	17	9	3	2	1	0	1	1	0
Chromosom03	143	25	10	5	3	2	0	1	0	0	0
Chromosom04	91	24	14	6	4	1	2	0	0	0	0
Chromosom05	114	26	7	1	3	1	0	0	0	0	1
Chromosom06	158	27	19	3	4	1	2	1	0	0	0
Chromosom07	117	23	5	6	0	2	0	0	0	0	0
Chromosom08	88	23	8	3	4	2	1	0	0	0	0
Chromosom09	77	16	15	2	1	0	0	1	0	0	0
Chromosom10	107	26	8	1	2	1	0	1	0	0	0
Chromosom11	108	25	13	4	4	1	1	1	0	0	0
Chromosom12	106	26	6	4	2	1	2	0	1	1	0
Chromosom13	47	13	5	1	1	0	0	0	0	0	0
Chromosom14	75	17	12	3	2	0	1	0	0	0	0
Chromosom15	81	15	5	5	1	0	0	1	1	0	0
Chromosom16	94	25	5	5	1	0	1	0	0	1	0
Chromosom17	79	30	11	2	4	1	3	1	0	0	0
Chromosom18	57	17	4	1	1	0	0	0	0	0	0
Chromosom19	52	16	2	1	0	1	0	0	0	0	0
Chromosom20	55	13	6	3	1	1	1	0	0	0	0
Chromosom21	38	4	4	1	0	0	0	0	0	0	0
Chromosom22	47	10	7	2	2	0	0	0	1	0	0

Źródło: dane zaczerpnięte z [Piwowar, Meus, Roterman 2004].



Rys. 1. Wartości współczynników *Z* obliczone dla tandemów o długości od 4 do 14 w związku z występowaniem a) w całym genomie, z wyłączeniem chromosomów płciowych; b) na konkretnych chromosomach somatycznych człowieka
Źródło: opracowanie własne.

Na rysunku 1 wartości współczynników *Z* są przedstawione za pomocą odcieni szarości. Im ciemniejszy jest kolor, tym wartości *Z* są większe.

Tandem 9CAG okazał się najrównomierniej rozproszonym tandemem w genomie z najniższą wartością współczynnika Z (rys. 1a). Tandemy o skrajnych wartościach współczynnika Z – 4CAG i 14CAG – charakteryzuje większe skoncentrowanie w pewnych partiach genomu, a więc ich rozproszenie jest małe. Poza tym bardziej szczegółowa analiza z rozbiciem genomu na chromosomy (rys. 1b) wykazała, że określone długości tandemów zależą od rodzaju chromosomu, na którym występują.

Podziękowanie

Autorzy dziękują dr hab. Irenie Roterman-Koniecznej oraz mgr Monice Piwowar za udostępnienie danych dotyczących struktur białkowych oraz danych genomowych.

Literatura

Bjørnstad J.F., *Inference Theory in Contingency Tables*, Statistical Research Report, Oslo 1979, s. 1-26.

Goodman L.A., Kruskal W.H., *Measures of Association for Cross Classifications*, Series in Statistics, Springer , New York 1979.

Kendall M.G., Stuart A., *The Advanced Theory of Statistics*, vol. 2, Griffin, London 1979.

Meus J., Bryliński M., Piwowar P., Wiśniowski Z., Stefaniak J., Piwowar M., Konieczny L., Roterman I., *A Tabular Approach to the Sequence-to-Structure Relation*, w druku.

Piwowar M., Meus J., Roterman I., *Analiza zawartości mikrosatelitów nCAG w genomach wybranych organizmów oraz w mRNA ludzkim*, SIIB, Kraków 2004.

Shannon C.E., *A Mathematical Theory of Communication*, „Bell System Technical Journal" 1948, 27, s. 379-423, 623-656.

ANALYSIS AND DATA CLASSIFICATION IN THE ASPECT OF COEFFICIENT Z

Summary

Publications dedicated to the analysis of contingency tables in statistical literature focus frequently on establishing statistical significance between investigated variables. It leads to treating them as single states, though these variables are usually defined as systems of events creating the the specific one. The reason of this approach is in the lack of method which would treat both variables as a system of two systems of disjoint events and would introduce the analysis of the "interior" structure of the data. In this paper a new asymmetrical measure of dependence has been introduced which bases on the above mentioned assumption.