

317/2007

Raport Badawczy
Research Report

RB/75/2007

**Podział niejednorodnych zbiorów
danych na podzbiory jednorodne z
wykorzystaniem cech najbardziej
informatywnych**

**M. Bereziński, M. Inkielman,
D. Wagner**

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Pracowni zgłaszający pracę:
Prof. dr inż. Roman Kulikowski

Warszawa 2007

PODZIAŁ NIEJEDNORODNYCH ZBIORÓW DANYCH NA PODZBIORY JEDNORODNE Z WYKORZYSTANIEM CECH NAJBARDZIEJ INFORMATYWNYCH

Mirosław Bereziński¹, Michał Inkielman², Dariusz Wagner³

Instytut Badań Systemowych PAN

01-447 Warszawa, ul. Newelska 6

e-mail: ¹Mirosław.Berezinski@ibspan.waw.pl

²Michał.Inkielman@ibspan.waw.pl

³Dariusz.Wagner@ibspan.waw.pl

Streszczenie: Punktem wyjścia do rozważań jest stwierdzenie, że dekompozycja wielowymiarowych zbiorowości – z którymi ma się do czynienia w ocenach jakości procesu edukacyjnego – na podzbiory jednorodne, nie musi być wykonywana w oparciu o pełen zbiór cech, charakteryzujących elementy zbiorowości (studentów lub nauczycieli). Może być przeprowadzona w oparciu o podzbiór cech najbardziej informatywnych. W pracy zaproponowano ogólną, unormowaną miarę stopnia informatywności cechy i pokazano, że istniejące dotychczas miary są jej szczególnymi przypadkami. Szczegółowo rozpatrzono różne miary informatywności oparte na pojęciu ilości informacji w sensie Shannona. Przedstawiono metodę identyfikacji podzbioru cech najbardziej informatywnych spośród z góry ustalonego zbioru cech. Zaproponowano metodę dekompozycji niejednorodnej zbiorowości na części jednorodne w oparciu o zespół cech najbardziej informatywnych. Sformułowano kryterium minimalizujące ryzyko błędu decyzyjnego, przy przydzielaniu studentów bądź nauczycieli do już istniejących grup jednorodnych. Przedstawiono metodę weryfikacji hipotezy o jednorodności, sformuowaną w kategoriach ilości informacji.

Słowa kluczowe: proces edukacyjny, ewaluacja, model matematyczny, analiza danych wielowymiarowych, informatywność cech, ilość informacji, entropia

1. Wprowadzenie

Ważnym czynnikiem kształtowania jakości edukacji w szkolnictwie wyższym jest systematyczne sprawdzanie i ocenianie osiągnięć studentów oraz nauczycieli akademickich, a następnie wykorzystywanie tej wiedzy do kształtowania jakości szkoły, a zwłaszcza realizowanego w niej procesu edukacyjnego. Są to zagadnienia związane z tzw. pomiarem dydaktycznym. W teorii pomiaru dydaktycznego proces oceniania nazywa się też ewaluacją (Niemierko 1975, 1990, 1991).

O wyborze postaci matematycznego modelu obiektu rzeczywistego decyduje się, przede wszystkim, na podstawie analizy systemowego sformułowania zadania oraz analizy zbioru danych, które mają być użyte do szacowania warto-

ści parametrów modelu (Kacprzyński 1974; Bubnicki 1974, 2005; Gutenbaum 2003). Zakłada się, że dane są zawsze obciążone jakąś dozą niepewności, przy czym w tradycyjnym nurcie teorii identyfikacji przyjmuje się, że jest to niepewność probabilistyczno-statystyczna, nazywana też stochastycznością. Matematycznym modelem wielkości nieznannej jest w tym przypadku zmienna losowa o znanym (rachunek prawdopodobieństwa) lub nieznanym (statystyka matematyczna) rozkładzie prawdopodobieństwa.

Spójrzmy teraz na zadanie podziału zbiorowości studentów (nauczycieli) na grupy jednorodne z punktu widzenia teorii informacji. Podejście to daje możliwość sprowadzenia zadania podziału do zadania identyfikacji zbioru najbardziej informatywnych cech i sformułowania go w kategoriach pojęciowych entropii i informacji. Metody wykorzystujące miary entropii i informacji są szczególnie wygodne wtedy, gdy wśród tych charakterystyk studentów (nauczycieli), które przyjmujemy za podstawę podziału ich zbiorowości na grupy jednorodne, dominują cechy jakościowe. Jest oczywiste, że nie wszystkie cechy są jednakowo istotne. Intuicja podpowiada, że za istotne należy uznać te, które są najbardziej informatywne, z punktu widzenia zadania podziału zbiorowości studentów (nauczycieli) na grupy jednorodne, a więc te, których znajomość pozwoli z dużo większą wiarygodnością niż znajomość innych wnioskować o wartościach pozostałych cech. Powstaje więc konieczność identyfikacji cech istotnych.

Przyjmijmy, że wszystkie cechy są wyrażone w jednostkach nominalnych. Jest to równoważne stwierdzeniu, że są one jakościowo nieuporządkowane. Przyjęcie tego założenia pozwala oceniać stopień informatywności cech za pomocą ilości informacji Shannona (Shannon 1949; Brillouin 1969; Nowakowski i Sobczak 1970).

2. Wybór cech istotnych (najbardziej informatywnych)

2.1. Uwagi wstępne

Probabilistyczno-statystyczne miary zależności między cechami dają najlepsze wyniki wtedy, gdy wartości wszystkich rozpatrywanych cech ilościowych bądź jakościowych są wyrażone w jednostkach odpowiadających im skal nominalnych. Skale te można konstruować zarówno dla cech ilościowych (np. numer studenta na liście, oceny z poszczególnych przedmiotów, ocena średnia semestralna itp.), jak i jakościowych (student zdolny, pracowity, mało aktywny itp.), przy czym istotą każdej skali nominalnej jest gradacja stopnia natężenia przejawiania się danej cechy. Należy pamiętać, że każda skala nominalna ma własność symetryczności i przechodności. Skale te powinny być skonstruowane w taki sposób, żeby różnice między elementami należącymi do jednej i tej samej klasy były małe, a różnice między elementami należącymi do dwóch różnych klas – duże.

Probabilistyczno-statystyczne miary zależności, w większości oparte na statystyce chi-kwadrat, są dobrze omówione w literaturze (zob., np.: Hald 1952; Anderson 1952; Cramér 1945; Hellwig 1966; Fisz 1967; Neyman 1969; Barra 1982; Pacut 1985; Zieliński 1990; Rao 1994; Plucińska i Pluciński 2000; Nowak 2002), pomijamy więc ich przedstawienie. Skupimy się na rozpatrzeniu miar opartych na pojęciu ilości informacji.

2.2. Miary oparte na pojęciu ilości informacji

O ile pojęcie niezależności jest w rachunku prawdopodobieństwa i statystyce matematycznej rozumiane w sposób jednoznaczny, o tyle pojęcie zależności ma wiele znaczeń. Będziemy mówić, że dwie cechy są od siebie funkcyjnie zależne, jeśli znajomość wartości jednej z nich pozwala jednoznacznie określić wartość drugiej. W tym przypadku znajomość wartości jednej cechy całkowicie usuwa nieokreśloność co do wartości drugiej. Zmniejszenie nieokreśloności jest zawsze konsekwencją otrzymania określonej ilości informacji. Jak wiadomo (Ashby 1956; Beer 1959; Klir 1976; Beneš 1979; Bertalanfy 1984; Heller, Lubański i Ślaga 1992; Penrose 2005) informacja pociąga za sobą ograniczenie różnorodności stanów obiektu. Z kolei, ograniczenie różnorodności stanów jest jedną z głównych metod regulacji obiektów rzeczywistych, ponieważ im mniejsza jest różnorodność stanów obiektu, tym bardziej jest przewidywalne jego zachowanie. Wedle Ashby'ego (1956): „Ograniczenie różnorodności jest relacją między dwoma zbiorami, która ma miejsce wtedy, gdy różnorodność mająca miejsce w jednych warunkach jest mniejsza niż różnorodność w innych warunkach”. Z tego punktu widzenia identyfikację statystycznej zależności zmiennej losowej X_2 od zmiennej losowej X_1 można traktować jako określenie, w jaki sposób ograniczenie różnorodności wartości zmiennej X_1 wpływa na ograniczenie różnorodności wartości zmiennej Y w porównaniu z różnorodnością jaką miałyby zmienna X_2 , gdyby była rozpatrywana niezależnie od X_1 . Konstruując probabilistyczno-statystyczne miary zależności zmiennych losowych na ogół szuka się takich formuł matematycznych, które w przypadku niezależności zmiennych X_1 i X_2 będą przyjmowały wartość zero, zaś w przypadku ich funkcyjnej zależności – wartość jeden. Spełnienie tego drugiego warunku osiąga się za pomocą operacji normowania miary. Jeżeli więc przez $U(X_2)$ oznaczymy nieokreśloność zmiennej losowej X_2 rozpatrywanej niezależnie od X_1 , zaś przez $U(X_2 | X_1)$ nieokreśloność zmiennej losowej X_2 pod warunkiem, że są znane wartości zmiennej losowej X_1 , to staje się oczywiste, że ogólną postacią unormowanej miary wpływu stopnia różnorodności zmiennej X_1 na stopień różnorodności zmiennej X_2 jest

$$W(X_2 | X_1) = \frac{U(X_2) - U(X_2 | X_1)}{U(X_2)}. \quad (2.1)$$

W przypadku, gdy zmienne X_1 i X_2 są wzajemnie niezależne, wtedy znajomość wartości zmiennej X_1 nie zmniejsza nieokreśloności co do wartości zmiennej X_2 . Wobec tego $U(X_2 | X_1) = U(X_2)$ i $W(X_2 | X_1) = 0$. Natomiast w przypadku, gdy między zmiennymi X_1 i X_2 zachodzi zależność funkcyjna, wtedy $U(X_2 | X_1) = 0$ i $W(X_2 | X_1) = 1$. Miarę (2.1) można uważać za wskaźnik względnego zmniejszenia nieokreśloności zmiennej X_2 wskutek otrzymania informacji o zmiennej X_1 .

Wchodzą w grę różne sposoby definiowania i formalizacji pojęć nieokreśloności i informacji, a w ślad za tym różne możliwości określania miary (2.1). Załóżmy, że pojęcia te są określone w sensie nadanym im przez Shannona (zob., np.: Shannon 1949; Kullback 1959; Jagłom i Jagłom 1973; Sobczak 1981). Jeżeli więc zmienne losowe X_1 i X_2 są dyskretne i pierwsza z nich przyjmuje wartości x_i z prawdopodobieństwami p_i ($i = 1, 2, \dots, I$), a druga wartości y_j z prawdopodobieństwami q_j ($j = 1, 2, \dots, J$), to entropie tych zmiennych wyrażają się wzorami

$$H(X_1) = - \sum_{i=1}^I p_i \log p_i, \quad (2.2a)$$

$$H(X_2) = - \sum_{j=1}^J q_j \log q_j. \quad (2.2b)$$

Entropia łącznej zmiennej losowej (X_1, X_2) , której realizacjami są pary liczb (x_i, x_j) o prawdopodobieństwach p_{ij} ($i = 1, 2, \dots, I; j = 1, 2, \dots, J$) jest określona wzorem

$$H(X_1, X_2) = - \sum_{i=1}^I \sum_{j=1}^J p_{ij} \log p_{ij}. \quad (2.3)$$

Jak wiadomo, znając prawdopodobieństwa łączne p_{ij} zmiennej losowej dwuwymiarowej można wyznaczyć prawdopodobieństwa warunkowe

$$p_{j|i} = \frac{p_{ij}}{p_i}, \quad (i = 1, 2, \dots, I; j = 1, 2, \dots, J). \quad (2.4)$$

Na podstawie prawdopodobieństwa warunkowego definiuje się entropię warunkową. Na przykład, entropia warunkowa zmiennej losowej X_2 przy warunku, że $X_1 = x_{1i}$ ($i = 1, 2, \dots, I$) jest zdefiniowana następująco

$$H(X_2 | X_1) = - \sum_{j=1}^J p_{j|i} \log p_{j|i}. \quad (2.5)$$

Jeżeli więc w formule (2.1) uwzględnimy zależności (2.2b) i (2.5), to formuła ta przyjmie postać

$$R(X_2 | X_1) = \frac{H(X_2) - H(X_2 | X_1)}{H(X_2)}. \quad (2.6)$$

Rozpatrzmy różnicę stojącą w liczniku tego wyrażenia i przyjmijmy oznaczenie

$$I(X_1, X_2) = H(X_2) - H(X_2 | X_1). \quad (2.7)$$

Wielkość $I(X_1, X_2)$ nazywa się ilością informacji w sensie Shannona. Pozwala ona zapisać formułę (2.6) w postaci

$$R(X_2 | X_1) = \frac{I(X_2 | X_1)}{H(X_2)}. \quad (2.8)$$

Wielkość $R(X_2 | X_1)$ jest więc unormowanym współczynnikiem ilości informacji. Współczynnik ten ma następujące własności:

1. $0 \leq R(X_2, X_1) \leq 1$, przy czym $R(X_2 | X_1) = 0$ jeśli zmienne X_1 i X_2 są niezależne, oraz $R(X_2 | X_1) = 1$, jeśli zachodzi między nimi ścisła zależność funkcyjna.
2. Wartość $R(X_2 | X_1)$ nie zależy od zmiany kolejności położenia rozpatrywanych elementów na odpowiedniej skali, tzn. od zamiany miejscami wierszy bądź kolumn macierzy danych.

Entropia shannonowska nie jest jedyną możliwą reprezentacją nieokreśloności. Załóżmy, jak poprzednio, że zmienne losowe X_1 i X_2 są dyskretne, przy czym pierwsza z nich przyjmuje wartości x_{1i} z prawdopodobieństwami p_i ($i = 1, 2, \dots, I$), druga zaś wartości x_{2j} z prawdopodobieństwami q_j ($j = 1, 2, \dots, J$). Niech $f(q)$ będzie funkcją postaci (2.1) określoną na przedziale $0 \leq q \leq 1$, malejącą w tym przedziale i spełniającą warunek $f(1) = 0$. Załóżmy,

że jeśli została zaobserwowana ta wartość zmiennej X_2 , której prawdopodobieństwo wystąpienia jest równe q , to wielkość $f(q)$ pozwala oszacować średnią ilość informacji otrzymanej w wyniku obserwacji. Średnia ta wyraża się następująco

$$\sum_{j=1}^J q_j f(q_j). \quad (2.9)$$

Zauważmy, że jeżeli prawdopodobieństwo którejkolwiek z obserwacji jest równe 1 (obserwacja pewna), to średnia (2.9) jest równa zero. Dla obliczenia wartości tej średniej trzeba w każdym konkretnym przypadku zastąpić w formule (2.9) prawdopodobieństwa teoretyczne ich wartościami empirycznymi p_j ($j = 1, 2, \dots, m$). Wtedy nieokreśloności występujące w wyrażeniu (2.1) będą wyrażały się – odpowiednio – formułami:

$$U(X_2) = \sum_j p_j f(p_j), \quad (2.10)$$

oraz

$$U(X_2 | X_1) = \sum_{i=1}^I p_i \sum_{j=1}^J p_{ji} g(p_{ji}). \quad (2.11)$$

Ponieważ liczba funkcji f jest praktycznie nieograniczona, więc nieograniczona jest również liczba miar postaci (2.11). Rozpatrzmy, dla przykładu, trzy rodzaje funkcji f .

W przypadku, gdy $f(q) = -\log q$, otrzymuje się unormowany współczynnik ilości informacji $R(X_2 | X_1)$, określony wzorem (2.8).

Jeżeli $g(q) = 1 - q$, to miara (2.1) przyjmuje postać

$$\tau(X_2 | X_1) = \frac{\sum_{i=1}^I p_i \sum_{j=1}^J p_{ji}^2 - \sum_{j=1}^J p_j^2}{1 - \sum_{j=1}^J p_j^2}. \quad (2.12)$$

Jest to tzw. współczynnik Wallace'a (Wallace 1983). Ma on następujące własności:

1. $0 \leq \tau(X_2 | X_1) \leq 1$, przy czym $\tau(X_2 | X_1) = 0$, jeśli zmienne X i Y są niezależne, oraz $\tau(X_2 | X_1) = 1$, jeśli między zmiennymi zachodzi ścisła zależność funkcyjna.
2. Wartość współczynnika $\tau(X_2 | X_1)$ nie zmienia się, gdy zmienia się kolejność położenia rozpatrywanych elementów na odpowiedniej skali, tzn. gdy dokona się przestawienia wierszy bądź kolumn macierzy danych.

Współczynnik Wallace'a jest naturalną realizacją zasady proporcjonalnej predykcji, sformułowanej przez Goodmana i Kruskala (Goodman i Kruskal 1954, 1959, 1963, 1967). Stwierdza ona, że miarą zależności zmiennych losowych X_2 i X_1 powinno być względne zmniejszenie prawdopodobieństwa błędu predykcji zmiennej X_2 , pod warunkiem, że zmienna X_1 jest znana, w porównaniu z prawdopodobieństwem błędu predykcji X_2 bez znajomości X_1 . Istota metody Goodmana-Kruskala polega na analizie tablicy zależności w taki sposób, że prowadzący analizę przewiduje prawdopodobieństwo tego, iż

$$\frac{P_{j|i}^q}{\sum_{j=1}^J P_{j|i}^q} \quad (2.13)$$

spośród rozpatrywanych obiektów będzie należało do kategorii j z punktu widzenia cechy reprezentowanej przez zmienną losową X_2 , jeśli wiadomo, że pod względem cechy reprezentowanej przez zmienną losową X_1 należą one do kategorii i . W tym przypadku, jeśli nie wiadomo do jakiej kategorii należy obiekt względem cechy X_1 , to prawdopodobieństwo jego przynależności do kategorii j względem cechy X_2 wynosi

$$\frac{P_j^q}{\sum_j P_j^q}, \quad (2.14)$$

gdzie q ($q \geq 0$) jest dowolną stałą. Jeśli nie wiadomo, do której kategorii należy obiekt względem cechy X_1 , to oczekiwana część błędnych przewidywań wynosi

$$1 - \frac{P_j^{q+1}}{\sum_j P_j^q}. \quad (2.15)$$

Różnica między tym wyrażeniem i średnią ważoną liczbą błędnych predykcji, gdy znana jest kategoryzacja względem cechy X_1 , rozpatrywana w stosunku do liczby błędnych predykcji, gdy nie jest znana kategoryzacja względem cechy X_1 , przedstawia sobą – w istocie – procent wielkości objaśnianej wariacji.

W szczególnym przypadku, gdy $q = 1$, otrzymuje się współczynnik Wallace'a. Współczynnik ten wyraża fakt, że zmniejszaniem się błędów predykcji rządzi zasada proporcjonalności (tzn. jeśli, na przykład, $\tau = 0.50$, to przyjmuje się, że znajomość kategoryzacji obiektów względem cechy X_1 sprawia, że liczba błędów zmniejsza się dwukrotnie). Tym samym ocenia się możliwość poprawnego zaszeregowania obiektu do kategorii j ($j = 1, 2, \dots, m$) względem cechy X_2 , jeśli wiadomo, że należy do kategorii i ($i = 1, 2, \dots, n$) względem cechy X_1 , przy czym równocześnie ocenia się siłę zależności między cechami X_1 i X_2 . Wszystko to zwiększa heurystyczną wartość procedury pomiaru siły zależności między cechami, jako elementu analizy statystycznej.

Jeśli przyjąć, że $q = \infty$, to otrzyma się miarę zależności znaną jako współczynnik Goodmana-Kruskala (Goodman i Kruskal 1954, 1959, 1963, 1967):

$$\lambda(X_2 | X_1) = \frac{\sum_{i=1}^n p_i \max_j p_{ji} - \max_j p_{.j}}{1 - \max_j p_{.j}}, \quad (2.16)$$

gdzie $\max_j p_{.j}$ jest prawdopodobieństwem odpowiadającym wartości modalnej rozkładu brzegowego cechy X_2 , zaś $\max_j p_{ji}$ jest prawdopodobieństwem wartości modalnej rozkładu brzegowego cechy X_2 pod warunkiem, że cecha X_1 przyjęła wartość i .

Współczynnik Goodmana-Kruskala też można otrzymać jako szczególny przypadek miary (2.1), gdy

$$f(q) = \begin{cases} 1, & \text{gdy } 0 \leq q \leq q_0, \\ 0, & \text{gdy } q_0 \leq q < 1. \end{cases} \quad (2.17)$$

Rozpatrywanie współczynnika $\lambda(X_2 | X_1)$ w ramach koncepcji Goodmana-Kruskala polega więc na szacowaniu wartości względnego błędu prognozy, że obiekt należący do danej kategorii względem cechy X_1 , należy pod wzglę-

dem cechy X_2 do tej kategorii j , której odpowiada największe prawdopodobieństwo. W ten sposób minimalizuje się liczbę błędnych prognoz.

Zaletą współczynnika Goodmana jest możliwość szybkiego przeprowadzenia obliczeń, ale poważną wadą jest to, że z faktu, iż $\lambda(X_2 | X_1) = 0$ nie wynika, że cechy reprezentowane przez zmienne X_1 i X_2 są statystycznie niezależne, lecz jedynie to, że wartości modalne tych cech są takie same. Zwróćmy też uwagę, że współczynniki $\tau(X_2 | X_1)$ i $\lambda(X_2 | X_1)$ są stosunkami prawdopodobieństwa, co pozwala interpretować je na gruncie klasycznej statystyki matematycznej.

Mimo specyficznych odrębności wszystkie trzy współczynniki – $R(X_2 | X_1)$, $\tau(X_2 | X_1)$ i $\lambda(X_2 | X_1)$ – mają ważną wspólną własność. Jest nią asymetria, czyli fakt, że $R(X_2 | X_1) \neq R(X_1 | X_2)$, $\tau(X_2 | X_1) \neq \tau(X_1 | X_2)$ i $\lambda(X_2 | X_1) \neq \lambda(X_1 | X_2)$. Asymetria oznacza, że zależność między cechą X_2 , traktowaną jako zmienna zależna (objaśniana) i cechą X_1 , traktowaną jako zmienna niezależna (objaśniająca) musi być rozpatrywana jako zależność probabilistyczno-statystyczna, a miary zależności muszą być interpretowane jako wskaźniki względnego zmniejszenia nieokreśloności wiedzy o X_2 pod warunkiem znajomości X_1 .

W sytuacjach, kiedy kierunek zmienności X_2 w zależności od zmian X_1 nie jest oczywisty, Řehak i Řehakova (1973) proponują korzystać z następujących symetrycznych postaci współczynników $R(X_2 | X_1)$, $\tau(X_2 | X_1)$ i $\lambda(X_2 | X_1)$:

$$R(X_2, X_1) = \frac{I(X_2 | X_1)}{\frac{1}{2}[H(X_1) + H(X_2)]}, \quad (2.18)$$

$$\tau(X_2 | X_1) = \frac{\frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J \left[(p_{ij} - p_i \cdot p_j)^2 \frac{p_i + p_j}{p_i \cdot p_j} \right]}{1 - \frac{1}{2} \left(\sum_{i=1}^I p_i^2 + \sum_{j=1}^J p_j^2 \right)}. \quad (2.19)$$

$$\lambda(X_2, X_1) = \frac{\frac{1}{2} \left[\sum_{i=1}^I \max_j p_{ij} + \sum_{j=1}^J \max_i p_{ij} - (\max_i p_i + \max_j p_j) \right]}{1 - \frac{1}{2} (\max_i p_i + \max_j p_j)}. \quad (2.20)$$

Zalety i wady tych współczynników są takie same, jak ich postaci asymetrycznych.

Ilość informacji w sensie Shannona jest wielkością symetryczną względem zmiennych X_1 i X_2 . Podobnie jak statystyka chi-kwadrat wyraża ona jedynie fakt występowanie lub niewystępowania statystycznej zależności między tymi zmiennymi. W przypadku, kiedy nie ma konieczności podziału zmiennych na objaśniane i objaśniające, informacyjną miarą ich zależności może być współczynnik Rajskiego (1964)

$$R(X_1, X_2) = \frac{I(X_1, X_2)}{H(X_1, X_2)}, \quad (2.21)$$

gdzie $H(X_1, X_2)$ jest entropią łączną cech X_1 i X_2 . Współczynnik ten ma następujące własności:

1. $0 \leq R(X_1, X_2) \leq 1$.
2. $R(X_1, X_2) = 0$ wtedy i tylko wtedy, gdy cechy X i Y są niezależne.
3. $R(X_1, X_2) = 1$ wtedy i tylko wtedy, gdy między zmiennymi zachodzi ścisła zależność funkcyjna. Zależność ta ma miejsce wtedy, gdy w każdym wierszu i w każdej kolumnie macierzy zależności występuje jeden i tylko jeden element różny od zera. Jeśli jest to element stojący na przecięciu wiersza i ($i = 1, 2, \dots, I$) oraz kolumny j ($j = 1, 2, \dots, J$), to zachodzi równość $p_{ij} = p_{i.} = p_{.j}$. Wobec tego w tym przypadku ilość informacji w sensie Shannona wynosi $H(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2) = H(X_1, X_2)$. Ostatecznie więc $R(X_1, X_2) = 1$.

Zauważmy jeszcze jedną ważną własność współczynnika $R(X_1, X_2)$. Z klasycznej statystyki matematycznej wiadomo, że jeśli współczynnik korelacji zmiennych X_1 i X_2 jest równy zeru, to z tego nie wynika, że cechy te są niezależne. W przypadku współczynnika $R(X_1, X_2)$ jest inaczej: z zerowej wartości tego współczynnika wynika niezależność zmiennych losowych.

Współczynnik $R(X_1, X_2)$, podobnie jak inne informacyjne miary zależności oparte na pojęciu łącznej ilości informacji $I(X_1, X_2)$, może być użyty do oceny zależności danych ilościowych bądź jakościowych, podzielonych na klasy wyznaczone przez progi zmienności wartości zmiennych losowych X_1 i X_2 .

Taką kategoryzację przedstawia się zazwyczaj w postaci tablic dwudzielnych (zob., np.: Fisz 1967; Hellwig 1980).

Na podstawie wielkości $I(X_1, X_2)$ można konstruować rozmaite miary, wyrażające siłę wpływu zmiennej X_1 na zmienną X_2 . Przykładem może być współczynnik

$$\Gamma(X_2, X_1) = \frac{I(X_1, X_2)}{H(X_1)}. \quad (2.22)$$

Można go zapisać w postaci

$$\Gamma(X_2, X_1) = \frac{H(X_2) - H(X_2 | X_1)}{H(X_1)}. \quad (2.23)$$

Współczynnik ten objaśnia, jaką część entropii cechy X_2 określa entropia cechy X_1 . Zwróćmy uwagę, że mimo formalnej identyczności formuł (2.6) i (2.23) sens każdej z nich jest zupełnie inny. Przede wszystkim zauważmy, że współczynnik (2.6), którego równoważną formą jest (2.8), można przedstawić w postaci

$$R(X_2 | X_1) = \frac{I(X_2 | X_1)}{H(X_2)} = \Gamma(X_2, X_1) \frac{H(X_1)}{H(X_2)}. \quad (2.24)$$

Formuła ta stwierdza, że wielkość względnego zmniejszenia nieokreśloności zmiennej X_2 , spowodowanego oddziaływaniem zmiennej X_1 , jest proporcjonalna do względnej entropii zmiennej X_1 . Istnieje więc związek między współczynnikiem $R(X_2 | X_1)$, jako wskaźnikiem względnego zmniejszenia entropii zmiennej X_2 pod warunkiem, że jest znane X_1 , a $R(X_2 | X_1)$ jako wskaźnikiem wpływu zmiennej X_1 na zmienną X_2 .

Podobnie jak w przypadku współczynników $R(X_2 | X_1)$ i $\tau(X_2 | X_1)$ wymieńmy podstawowe własności współczynnika $\Gamma(X_2, X_1)$. Oto one:

1. Jeśli zmienne X_1 i X_2 są niezależne, to $\Gamma(X_2, X_1) = 0$.
2. Jeśli między cechami X_1 i X_2 zachodzi jednoznaczna zależność funkcyjna, to $\Gamma(X_2, X_1) = 1$.
3. Ponieważ $I(X_1, X_2) \leq \min[H(X_1), H(X_2)]$, więc jeśli nieokreśloność własna zmiennej X_1 jest większa niż bezwarunkowa

entropia zmiennej X_2 , tzn. jeśli $H(X_1) > H(X_2)$, to niemożliwy jest pełny wpływ cechy X_1 na cechę X_2 .

4. Współczynnik $I(X_2, X_1)$ może być uważany za częściowy wskaźnik informatywności zmiennej X_1 . Własność ta może być wykorzystana przy rozwiązywaniu takich zadań związanych z klasyfikacją i podziałem, jak identyfikacja najbardziej informatywnych cech, konstruowanie najbardziej efektywnej reguły decyzyjnej itp.

Jeżeli rozpatrywane zbiorowości są odpowiednio duże, to statystyki odpowiadające omówionym wyżej informacyjnym miarom zależności mają podobne własności asymptotyczne. Na przykład, Khan (1973) udowodnił, że jeśli jest prawdziwa hipoteza o niezależności X_1 i X_2 , to statystyka $2nI$ ma rozkład asymptotycznie chi-kwadrat o $(r-1)(s-1)$ stopniach swobody, natomiast w przypadku prawdziwości hipotezy alternatywnej statystyka ta ma asymptotycznie niecentralny rozkład χ^2 (chi-kwadrat) o $(r-1)(s-1)$ stopniach swobody i parametrze niecentralności $\eta^2 = 2nI$. Wartość oczekiwana niecentralnego rozkładu χ^2 wynosi $(r-1)(s-1) + \lambda^2$, zaś wariancja jest równa $2[(r-1)(s-1) + 2\lambda^2]$. Przy dużej niecentralności dobrą aproksymacją niecentralnego rozkładu χ^2 jest rozkład normalny o takich samych parametrach. Wykorzystując ten fakt można przedstawić współczynnik $R(X_2 | X_1)$ w takiej postaci:

$$R(X_2 | X_1) = \frac{\chi^2}{\eta^2 - 2nH(X_2 | X_1)}. \quad (2.25)$$

Khan i Ali (1973) udowodnili też, że podobne własności asymptotyczne mają statystyki odpowiadające współczynnikom $\lambda(X_2 | X_1)$ i $\tau(X_2 | X_1)$.

Wykorzystując powyższe własności można w następujący sposób weryfikować poziom istotności informacyjnej miary zależności zmiennych X_1 i X_2 :

1. Znając licznosc n badanej zbiorowości, określa się liczbę stopni swobody i ustala się poziom istotności.
2. Na podstawie tablicy rozkładu χ^2 określa się wartość krytyczną

$$I_0 = \frac{\chi^2}{2n}, \text{ z którą porównuje się wartość } I(X_1, X_2).$$

3. Jeśli $I_0 \geq I(X_1, X_2)$, to przyjmuje się, że na przyjętym poziomie istotności hipoteza o niezależności zmiennych X_1 i X_2 jest prawdziwa.
4. Jeśli $I_0 < I(X_1, X_2)$, to przyjmuje się, że na przyjętym poziomie istotności hipoteza o niezależności cech X_1 i X_2 jest nieprawdziwa.

Zwróćmy uwagę, że wykorzystując koncepcję informacyjnej miary zależności można skonstruować wskaźniki dotyczące zależności między zmiennymi wyrażonymi w jednostkach jednej i tej samej skali, jak i między zmiennymi, którym odpowiadają różne skale. Zwróćmy też uwagę, że jeśli zmienna X_1 jest określona w skali nominalnej, a cecha X_2 w skali przedziałowej, to pojęcia nieokreśloności (entropii), nieokreśloności warunkowej (entropii warunkowej) oraz informacji bezpośrednio wiążą się z tradycyjnymi pojęciami statystycznymi, takimi jak różnorodność, zmienność czy wariancja. Ponieważ przy pomiarach w skali przedziałowej nie otrzymujemy samej tylko informacji o kategorii zmiennej X_2 , ale także ocenia się w kategoriach ilościowych stopień odmienności od innych obiektów należących do danej kategorii, więc wygodnie jest posługiwać się następującą definicją nieokreśloności:

$$U(X_2) = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y})^2, \quad (2.26)$$

$$U(X_2 | X_1) = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y}_i)^2, \quad (2.27)$$

gdzie

$$\bar{y} = \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^J y_{ij}, \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^J y_{ij}. \quad (2.28)$$

W tym przypadku informacyjna miara zależności (2.1) przyjmie następującą postać

$$\eta^2 = \frac{\sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y})^2 - \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y})^2}. \quad (2.29)$$

Po uwzględnieniu formuł (2.28), wprowadzeniu oznaczenia

$$F = \frac{\frac{1}{r-1} \sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2}{\frac{1}{n-r} \sum_{i=1}^I \sum_{j=1}^J (\bar{y}_{ij} - \bar{y})^2} \quad (2.30)$$

i wykonaniu odpowiednich przekształceń formuła (2.29) przyjmie postać

$$\eta^2 = \frac{(r-1)F}{(r-1)F + n - r}. \quad (2.31)$$

Jak wiadomo (Okta 1966), w przypadku gdy punkty odpowiadające parom (x, \bar{y}_x) , gdzie \bar{y}_x jest średnią warunkową zmiennej X_2 przy ustalonej wartości cechy X_1 , układają się wzdłuż pewnej krzywej, nie można mówić o korelacji. Miarą współzależności cech X_1 i X_2 jest współczynnik $\eta(X_2, X_1)$ nazywany stosunkiem korelacyjnym zmiennej X_2 względem zmiennej X_1 . Jego kwadrat jest definiowany jako stosunek sumy kwadratów odchyłeń średnich warunkowych od średniej ogólnej do sumy kwadratów odchyłeń y (Okta 1966), tzn.

$$\eta^2(X_2, X_1) = \frac{\sum_{i=1}^I (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y})^2}. \quad (2.32)$$

Oczywiście, $0 \leq \eta^2(X_2, X_1) \leq 1$. Wartość współczynnika $\eta^2(X_2, X_1)$ jest zawsze większa lub równa wartości współczynnika korelacji, przy czym równość ma miejsce jedynie wtedy, gdy regresja jest prostoliniowa. Stosunek korelacyjny przyjmuje wartość zero wtedy, gdy średnie warunkowe \bar{y}_i leżą dokładnie na prostej równoległej do prostej do osi x , a więc gdy nie zależą od x . Poza tym, jeśli między cechami X_1 i X_2 zachodzi związek funkcyjny wyrażony za pomocą odpowiedniej formuły matematycznej, to współczynnik korelacji jest równy jedności. Stosunek korelacyjny można uważać za miarę tendencji skupiania się punktów wzdłuż krzywej regresji. Im bardziej te punkty przylegają do krzywej regresji, tym kwadrat stosunku korelacyjnego jest bliższy jedności. Jak widać, stosunek korelacyjny $\eta^2(X_2, X_1)$ jest więc częściowo zgodny z koncepcją informacyjnego podejścia do konstruowania miar zależności.

2.3. Podział zbiorowości studentów na grupy jednorodne – ujęcie teoriiinformacyjne

Rozpatrzmy teraz zadanie podziału n -osobowej zbiorowości studentów (nauczycieli) S na grupy jednorodne, wykorzystując miary oparte na pojęciach entropii i informacji. Jak zobaczymy, podejście to pozwoli powiązać zadanie identyfikacji najbardziej informatywnych cech z zadaniem podziału i rozwiązywać je łącznie. Utrzymujemy założenie, że rozpatrywane cechy są wielkościami skategoryzowanymi. Chcemy oszacować stopień informatywności każdej z cech za pomocą ilości informacji Shannona. Jak już powiedzieliśmy, miara ta pozwala ocenić, o ile zmniejszy się nieokreśloność naszej wiedzy o wielkości x_i , jeśli jest znana wartość x_j . Najbardziej informatywną będzie więc ta cecha, której wpływ na daną cechę jest największy, tzn. dla której wyrażenie

$$\sum_{\substack{i=1 \\ i \neq j}}^I \Gamma_{ij} = \frac{1}{H_j} \sum_{\substack{i=1 \\ i \neq j}}^I I_{ij} \quad (2.33)$$

przyjmie wartość maksymalną, gdzie (zob. wzór (2.22))

$$\Gamma_{ij} = \frac{I_{ij}}{H_j}, \quad (i = 1, 2, \dots, I; j = 1, 2, \dots, J; i \neq j). \quad (2.34)$$

Oczywiście, równoważnikiem tego zadania jest maksymalizacja wyrażenia

$$M_j = \sum_{\substack{i=1 \\ i \neq j}}^I I_{ij}, \quad (j = 1, 2, \dots, J). \quad (2.35)$$

Przyjmijmy, że każdy ze studentów jest scharakteryzowany za pomocą m cech binarnych. Jeśli liczba studentów posiadających cechę i jest równa n_i , to entropia zbiorowości S wynosi

$$H(S) = m n \log n - \sum_{i=1}^m [n_i \log n_i + (n - n_i) \log(n - n_i)]. \quad (2.36)$$

Podstawa logarytmu może być dowolna, ale w przypadku cech binarnych wygodnie jest przyjmując podstawę równą 2.

Rozpatrzmy cechę X_j i podzielmy zbiorowość studentów na dwie części, S_1 i S_2 ($S = S_1 \cup S_2$, $S_1 \cap S_2 = \emptyset$), zaliczając do S_1 tych wszystkich studen-

tów, którzy posiadają cechę X_j , a do S_2 wszystkich pozostałych. Entropia zmniejszy się o wielkość

$$\Delta H(S, S_1, S_2) = H(S) - H(S_1) - H(S_2), \quad (2.37)$$

gdzie

$$H(Z) = \sum_{\substack{i=1 \\ i \neq j}}^J p_{i|j} \log p_{i|j}, \quad (Z = S_1, S_2). \quad (2.38)$$

Uwzględniając ten fakt można przedstawić wyrażenie (2.37) w postaci

$$\Delta H(S, S_1, S_2) = \sum_{\substack{i=1 \\ i \neq j}}^{m-1} [H_i - H(i|j)] = \sum_{\substack{i=1 \\ i \neq j}}^{m-1} I_{ij}, \quad (2.39)$$

a biorąc pod uwagę (2.35) otrzymamy

$$\Delta H(S, S_1, S_2) = M_j, \quad (j = 1, 2, \dots, J). \quad (2.40)$$

A zatem, zmniejszenie entropii przy podziale zbiorowości studentów na grupy wedle posiadania cechy X_j jest równe stopniowi informatywności tej cechy. Zwróćmy uwagę, że zależność (5.40) będzie prawdziwa również wtedy, gdy podziału zbiorowości studentów dokona się nie wedle cechy dwuwartościowej, lecz wielowartościowej (wiele alternatyw).

Jeżeli obliczymy wartość M_j dla każdej cechy j ($j = 1, 2, \dots, J$), to na pierwszym kroku postępowania za najbardziej informatywną trzeba uznać tę cechę, dla której M_j przyjmuje wartość maksymalną. Na następnych krokach należy dokonywać podziału zawsze względem tej cechy, dla której $\Delta H = \max_j \Delta H_j$. Procedurę podziału należy zatrzymać wtedy, gdy zostanie osiągnięty z góry określony stopień jednorodności otrzymanych podzbiorów.

Jednorodność grup może być oceniana w różny sposób. Möller i Capecchi (1975) zaproponowali następujące postępowanie. Niech $r_j(k)$ będzie liczbą równą modalnej częstości występowania cechy X_j ($j = 1, 2, \dots, J$) w klasie k ($k = 1, 2, \dots, K$), tzn. równą największej liczbie studentów charakteryzujących się posiadaniem jednej z gradacji cechy X_j wśród wszystkich studentów należących do klasy k , bliskich sobie w sensie tej cechy. Estymatorem prawdopo-

dobieństwa zaliczenia studenta wedle cechy X_j do kategorii różnej od modalnej jest

$$p_j(\varepsilon) = 1 - \frac{1}{n} \sum_{k=1}^K r_j(k), \quad (j = 1, 2, \dots, m). \quad (2.41)$$

Uśredniając to prawdopodobieństwo po zbiorze wszystkich cech otrzymujemy ocenę średniego prawdopodobieństwa błędnego podziału

$$p(\varepsilon) = \frac{1}{m} \sum_j p_j(\varepsilon), \quad (j = 1, 2, \dots, m). \quad (2.42)$$

Möller i Capecchi (1975) proponują, aby za miarę jednorodności przyjąć

$$h(\varepsilon) = 1 - p(\varepsilon). \quad (2.43)$$

Warunkiem zatrzymania procedury podziału jest spełnienie nierówności

$$h(\varepsilon) \geq h^*, \quad (2.44)$$

gdzie h^* jest pewną z góry ustaloną wartością stopnia jednorodności.

Oczywiście, można oceniać nie tylko stopień jednorodności całej rozpatrywanej zbiorowości studentów, ale także stopień jednorodności grup k ($k = 1, 2, \dots, K$), otrzymywanych na poszczególnych krokach procedury. W tym przypadku formuła (2.43) przyjmie postać

$$h(k) = 1 - \frac{1}{m} \sum_{j=1}^m \left(1 - \frac{r_j(k)}{n_k} \right), \quad (2.45)$$

przy czym $0 \leq h(k) \leq 1$, n_k - ogólna liczba elementów w grupie k ($k = 1, 2, \dots, K$). Procedurę podziału należy zakończyć, gdy wartość $h(k)$ przekroczy z góry ustaloną wartość progową.

Podział zbiorowości studentów (nauczycieli) na grupy jednorodne według określonego zbioru najbardziej informatywnych cech ma sens jedynie wtedy, gdy liczba cech opisujących każdą z otrzymanych grup jest mniejsza od w-sciowej liczby cech, charakteryzujących całą zbiorowość. Dotychczas przedstawiliśmy jedynie charakterystykę tych grup. Teraz trzeba określić, których studentów należących do rozpatrywanej zbiorowości S , należy zaliczyć do których grup. Można to zrobić w następujący sposób.

Przypuśćmy, że w wyniku zastosowania opisanej wyżej procedury zbiorowość studentów została podzielona na L grup. Wobec tego, każdemu studentowi (nauczycielowi) można przyporządkować L liczb nieujemnych $\mu(N_{l|s})$, ($l = 1, 2, \dots, L$), wyrażających prawdopodobieństwo tego, że student s należy do grupy l . Zachodzi równość $\sum_{l=1}^L \mu(N_{l|s}) = 1$. Dla każdej grupy można oszacować empirycznie wartość prawdopodobieństwa tego, że student (nauczyciel), któremu przyporządkowano l -tą wartość cechy x_r , należy do tej grupy. Prawdopodobieństwo to wyraża się wzorem

$$p_k(x_r^l) = \frac{n_k(x_r^l) + 1}{n_k(x_r) + l(x_r)}, \quad (2.46)$$

gdzie $n_k(x_r^l)$ - liczba elementów w grupie k , którym przyporządkowano l -tą wartość cechy x_r ; $n_k(x_r)$ - liczba elementów w grupie k , którym przyporządkowano jakąkolwiek wartość cechy x_r ; $l(x_r)$ - liczba stopni gradacji cechy x_r . Prawdopodobieństwo tego, że dany element będzie należał do klasy k ze względu na wszystkie rozpatrywane cechy jest równe iloczynowi wszystkich prawdopodobieństw $p_k(x_r^l)$, tzn.

$$\prod_r p_k(x_r^l) = \prod_r \frac{n_k(x_r^l) + 1}{n_k(x_r) + l(x_r)}. \quad (2.47)$$

Estymatorem prawdopodobieństwa grupy L_k jest oczywiście iloraz

$$p(L_k) = \frac{n_k}{n}, \quad (k = 1, 2, \dots, K). \quad (2.48)$$

Korzystając ze wzoru Bayesa otrzymuje się prawdopodobieństwo przynależności studenta (nauczyciela) s_i do grupy k

$$p(L_k | s_i) = \frac{p(L_k) p_k(s_i)}{\sum_{k=1}^K p(L_k) p_k(s_i)}, \quad (2.49)$$

gdzie $p_k(S_i) = \prod_r p_k(x_r^i)$, przy czym $p_k(x_r^i)$ jest określone formułą (2.46),

przy uwzględnieniu grupy, do której należy element s_i z punktu widzenia cechy x_r .

Zwróćmy uwagę, że powyższa metoda podziału zbiorowości na grupy jednorodnie (w oparciu o pojęcia entropii i informacji) ma konstrukcję hierarchiczną. Wyniki podziału mogą być przedstawione graficznie w postaci struktury wielopoziomowej, przy czym w odniesieniu do każdego poziomu można podać wartość entropii $\sum_{k=1}^K H(L_k)$ lub wartość ΔH , o jaką zmniejszy się entropia, obliczoną wedle formuły (2.37). Ponadto każdemu poziomowi można przyporządkować liczbę wyrażającą miarę jednorodności (zob. formuły (2.43), (2.44)), nazwę cechy, wedle której dokonuje się podziału itp.

3. Zadanie przydziału studentów (nauczycieli) do istniejących już grup jednorodnych

Proces oceny studentów (nauczycieli) nie zawsze wiąże się z koniecznością podziału ich zbiorowości na grupy jednorodne. Bardzo często powstaje konieczność rozwiązania innego zadania: istnieje już podział na grupy jednorodne i zachodzi potrzeba wskazania, do której z nich powinno się zaliczyć nowego studenta (nauczyciela).

Aby rozwiązać to zadanie trzeba wprowadzić pewną funkcję decyzyjną, za pomocą której można będzie z wystarczająco dużym stopniem wiarygodności rozróżniać elementy (studentów lub nauczycieli) należące do poszczególnych grup oraz wskazywać, do której z nich należy zaliczyć nowego (studenta lub nauczyciela). Identyfikacji grup można dokonać za pomocą hierarchicznej procedury, którą przedstawiliśmy w poprzednim punkcie. Rolę reguły decyzyjnej może pełnić formuła (2.49) – posługując się nią można określić prawdopodobieństwo przynależności nowego studenta lub nauczyciela do określonej grupy.

Ważna jest przy tym ocena błędu, jaki można popełnić rozwiązując to zadanie, polegającego na zaliczeniu nowej osoby do pewnej grupy, podczas gdy w rzeczywistości powinna być zaliczona do innej. Jeśli przyjmiemy, że cechy identyfikujące poszczególne grupy są niezależne, to jako miarę błędu można przyjąć stosunek prawdopodobieństw (2.49), czyli wielkość

$$\omega_{k,k+1} = \frac{p(L_k | s)}{p(L_{k+1} | s)}. \quad (3.1)$$

Uwzględnivszy wyrażenie (2.49) można tej formule nadać taką formę przybliżoną

$$\omega_{k,k+1} \approx \frac{p(L_k)}{p(L_{k+1})} \cdot \frac{p_k(s)}{p(s)}, \quad (3.2)$$

gdzie $p_k(s) = \prod_j p_k[r_j(k)]$.

Rozpatrzmy kryterium zgodności chi-kwadrat (χ^2). Niech α będzie prawdopodobieństwem popełnienia błędu polegającego na zaliczeniu studenta (nauczyciela) do grupy L_k , podczas gdy należało go zaliczyć do grupy L_{k+1} , zaś β - prawdopodobieństwem zdarzenia przeciwnego. Jeżeli

$$\frac{\alpha}{1-\beta} < \omega_{k,k+1} < \frac{1-\alpha}{\beta}, \quad (3.3)$$

to grupy L_k i L_{k+1} trzeba uznać za nierozróżnialne z punktu widzenia cech x_1, x_2, \dots, x_m i nowy student (nauczyciel) może być zaliczony do dowolnej z nich.

Zauważmy, że formuła (3.3) może być dodatkowym kryterium zatrzymania procesu podziału. Ciąg cech istotnych (informatywnych), który otrzymuje się w toku realizacji procedury podziału, jest ustalany zgodnie z kryterium (3.3).

4. Weryfikacja hipotezy o nieinformatywności cech

W przypadku stwierdzenia, że rozpatrywana zbiorowość studentów (nauczycieli) jest niejednorodna, zachodzi konieczność podzielenia jej zbioru jednorodny w sensie z góry określonego zespołu cech. Jeśli liczba m (tj. liczba cech) jest duża, to obliczenia komplikują się. Sensowny jest więc pomysł, by podziału na zbioru jednorodny dokonywać nie wedle pełnego zbioru cech, ale wedle l ($l \ll m$) cech najbardziej informatywnych w warunkach rozpatrywanego zadania. Wybór l cech najbardziej informatywnych jest równoważny odrzuceniu t cech, których informatywność jest mniejsza od pewnej z góry ustalonej wartości progowej ($m = l + t$).

Rozpatrzmy dwie m -wymiarowe zmienne losowe normalne, $X = (X_1, X_2, \dots, X_m)$ i $Y = (Y_1, Y_2, \dots, Y_m)$. Niech $E(X) = (\theta_{11}, \theta_{12}, \dots, \theta_{1m}) = \Theta_1$ będzie wartością oczekiwaną pierwszej z nich, a $E(Y) = (\theta_{21}, \theta_{22}, \dots, \theta_{2m}) = \Theta_2$ – drugiej. Załóżmy, że obie zmienne mają taką samą macierz kowariancji V .

Oznaczmy zbiór wartości j przez J^* i niech J_l będzie dowolnym podzbiorem zbioru J^* . Dopełnieniem zbioru J_l do zbioru J^* jest podzbiór J_{m-l} , tzn. J_l . Wobec tego, zbiorem cech nieinformatywnych J_l będzie taki zbiór cech, dla którego różnice $\theta_{1j} - \theta_{2j} = 0$, tzn.

$$J_l = \{j : \theta_{1j} - \theta_{2j} = 0\}. \quad (4.1)$$

Zbiór cech J_l , którego elementom odpowiadają różnice $\theta_{1j} - \theta_{2j} \neq 0$, będziemy uważać za informatywny. Warunek odpowiadający informatywnemu zbiorowi cech wygląda tak:

$$J_l = \{j : \theta_{1j} - \theta_{2j} \neq 0\}. \quad (4.2)$$

Ponieważ każdemu podzbiorowi J_l i J_l odpowiadają zmienne losowe X i Y , więc można je oznaczyć symbolami X_l , Y_l oraz X_l , Y_l . Wartości oczekiwane tych zmiennych oznaczymy symbolami θ_l i θ_l . Macierz kowariancji ma postać:

$$V = \begin{pmatrix} V_{ll} & V_{ll} \\ V_{ll} & V_{ll} \end{pmatrix}. \quad (4.3)$$

Uznanie rozpatrywanego zespołu l cech J_l za nieinformatywny odbywa się na podstawie weryfikacji hipotezy

$$H_0 : \theta_l' = \theta_l, \quad (4.4)$$

przy hipotezie alternatywnej (zespół cech jest informatywny)

$$H_1 : \theta_l' \neq \theta_l, \quad (4.5)$$

gdzie θ_l' i θ_l są wartościami oczekiwanymi l -wymiarowych zmiennych losowych X_l i Y_l .

Hipoteza (4.4) może być bezpośrednio sprawdzona za pomocą kryterium stosunku prawdopodobieństwa (3.1). Przy przyjętych przez nas oznaczeniach kryterium to przyjmuje postać:

$$\lambda(J_l) = \frac{\max_{\theta} L(H_1)}{\max_{\theta} L(H_0)}, \quad (4.6)$$

czyli

$$\lambda(J_1) = \frac{(2\pi)^{-\frac{l(n_1+n_2)}{2}} |\hat{V}_0^l|^{-\frac{l(n_1+n_2)}{2}} \exp\left[-\frac{l(n_1+n_2)}{2}\right]}{(2\pi)^{-\frac{l(n_1+n_2)}{2}} |\hat{V}_1^l|^{-\frac{l(n_1+n_2)}{2}} \exp\left[-\frac{l(n_1+n_2)}{2}\right]} \quad (4.7)$$

Stąd, po uproszczeniu,

$$\lambda(J_1) = \left(\frac{|\hat{V}_1^l|}{|\hat{V}_0^l|} \right)^{\frac{n_1+n_2}{2}}, \quad (4.8)$$

gdzie n_1 i n_2 - liczby obserwacji w macierzach X_1 i Y_1 odpowiednio; $|\hat{V}_0^l|$ - wyznacznik macierzy kowariancji z próby (tzn. oszacowania \hat{V}_0^l macierzy kowariancji V_{II} , obliczonej w warunkach prawdziwości hipotezy zerowej H_0 ; $|\hat{V}_1^l|$ - wyznacznik macierzy kowariancji z próby (tzn. oszacowania \hat{V}_1^l macierzy kowariancji V_{II} , obliczonej w warunkach prawdziwości hipotezy alternatywnej H_1).

Hipoteza zerowa (4.4) zostaje odrzucona, jeśli znaleziona wartość $\lambda(J_1)$ jest mniejsza od wartości dopuszczalnej λ_α przy poziomie istotności α . W tym przypadku rozpatrywany zbiór cech J_1 należy uznać za informatywny zgodnie z hipotezą (4.5). Jeśli jednak hipoteza zerowa nie przeczy danym empirycznym, to rozpatrywany zbiór cech J_1 należy uznać na nieinformatywny i można go wykluczyć z dalszej analizy.

Sformułowanej wyżej zadanie można nieco rozszerzyć, uogólniając jego warunki na z góry ustalony zbiór cech, który trzeba poddać sprawdzeniu na informatywność. W tym przypadku otwiera się zamancziwaja wcześniej perspektywa, tzn. do przeprowadzenia obserwacji statystycznych, ograniczyć zbiór danych tylko do zbioru cech informatywnych. Inaczej mówiąc, mamy tu na uwadze celowo ukierunkowane zbieranie danych, potrzebne do badania informacji w duchu teorii optymalnego planowania eksperymentu (Kacprzyński 1974).

Złożoność praktycznego rozwiązania takiego zadania polega na tym, że ocena informatywności przy posługiwaniu się formułą (4.6) powinna być

otrzymana bez dokładnej znajomości wielkości θ_{1j} i θ_{2j} . Dlatego, dysponując jedynie dostatecznie grubymi ocenami wartości oczekiwanych możemy wnioskować ze znanym stopniem ostrożności.

W ogólnym przypadku, gdy brak jest wcześniej określonego zbioru cech, rozważania sprowadzają się do tego, żeby prawdopodobieństwo pojawienia się błędnego wniosku o informatywności cech było możliwie jak najmniejsze. Jeśli oznaczyć przez Φ_I klasę wszystkich zbiorów J_I i każdemu z tych zbiorów przyporządkować indeks s ($s \in S$), to każdemu zbiorowi cech $J_{I_s} \in \Phi_I$ można przyporządkować liczbę

$$\lambda(J_{I_s}) = \frac{\max_K L(H_0^{I_s})}{\max_K L(H_1^{I_s})}, \quad (4.9)$$

gdzie $H_0^{I_s} : J_{I_s} = J_0^I$.

A zatem, stosunek prawdopodobieństwa (4.9) można rozpatrywać jako funkcję określoną na elementach J_{I_s} klasy J_I . Zbiór cech, dla którego ta funkcja przyjmuje wartość maksymalną, można uważać za najbliższy J_0^I . Przy tym, jeśli znalezione maksimum jest mniejsze od λ_α , to klasę Φ_I na poziomie istotności α rozpatruje się jako niezawierającą nieinformatywnych podzbiorów cech. Jeśli jednak maksimum funkcji (4.9) jest większe lub równe λ_α , to zbiór cech J_{I_s} można uważać za nieinformatywny i nie włączać go do planowanego badania.

5. Metodyka obliczeń

Weryfikacja hipotezy zerowej $H_0 : \theta^t = \theta$ dla dwóch m -wymiarowych zmiennych losowych X i Y , przyjmujących wartości $X_k = (x_{k1}, x_{k2}, \dots, x_{km})$, gdzie $k = 1, 2, \dots, n_1$, jest wektorem obserwacji zmiennej losowej X , a $Y_k = (y_{k1}, y_{k2}, \dots, y_{km})$, gdzie $k = 1, 2, \dots, n_2$, jest wektorem obserwacji zmiennej losowej Y za pomocą kryterium stosunku prawdopodobieństw jest związane z koniecznością znalezienia ocen \hat{V}_0 i \hat{V}_1 macierzy kowariancji V . Każdą z tych ocen oblicza się przy założeniu, że jest spełniona określona hipoteza: ocenie \hat{V}_0 odpowiada niesprzeczność hipotezy zerowej H_0 , a ocenie \hat{V}_1 - prawdziwość hipotezy alternatywnej H_1 .

Ocena \hat{V} macierzy kowariancji V jest macierzą ocen wariancji dyskretnych zmiennych losowych i odpowiadających im ocen kowariancji. Dlatego,

jeśli oznaczymy elementy macierzy \hat{V}_0 przez $\hat{\sigma}^0$, to można zapisać następującą równość:

$$\hat{\Sigma}_0 = \begin{pmatrix} \hat{\sigma}_{11}^0 & \hat{\sigma}_{12}^0 & \dots & \hat{\sigma}_{1m}^0 \\ \hat{\sigma}_{21}^0 & \hat{\sigma}_{22}^0 & \dots & \hat{\sigma}_{2m}^0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{m1}^0 & \hat{\sigma}_{m2}^0 & \dots & \hat{\sigma}_{mm}^0 \end{pmatrix}, \quad (5.1)$$

gdzie elementy diagonalne $\hat{\sigma}_{ii}^0$ ($i = 1, 2, \dots, m$) są ocenami jednowymiarowych zmiennych losowych, z których składają się wektory X i Y , zaś elementy niediagonalne $\hat{\sigma}_{ij}^0$ ($i, j = 1, 2, \dots, m; i \neq j$) są ocenami kowariancji.

Elementy diagonalne $\hat{\sigma}_{ii}^0$ ($i = 1, 2, \dots, m$) oblicza się wedle następującej formuły:

$$\hat{\sigma}_{ii}^0 = \frac{1}{n_1 + n_2 - 2} \left[\sum_{k=1}^{n_1} x_{ki}^2 + \sum_{k=1}^{n_2} y_{ki}^2 - \frac{1}{n_1 + n_2} \left(\sum_{k=1}^{n_1} x_{ki} + \sum_{k=1}^{n_2} y_{ki} \right)^2 \right]. \quad (5.2)$$

Elementy niediagonalne $\hat{\sigma}_{ij}^0$ ($i, j = 1, 2, \dots, m; i \neq j$) oblicza się wedle następującej formuły:

$$\hat{\sigma}_{ij}^0 = \frac{1}{n_1 + n_2 - 2} \cdot \left[\sum_{k=1}^{n_1} x_{ki} x_{kj} + \sum_{k=1}^{n_2} y_{ki} y_{kj} - \frac{1}{n_1 + n_2} \left(\sum_{k=1}^{n_1} x_{ki} + \sum_{k=1}^{n_2} y_{ki} \right) \left(\sum_{k=1}^{n_1} x_{kj} + \sum_{k=1}^{n_2} y_{kj} \right) \right] \quad (5.3)$$

Analogicznie, oznaczwszy elementy macierzy kowariancyjnej z próbki \hat{V}_1 przez $\hat{\sigma}^1$, mamy:

$$\hat{\Sigma}_1 = \begin{pmatrix} \hat{\sigma}_{11}^1 & \hat{\sigma}_{12}^1 & \dots & \hat{\sigma}_{1m}^1 \\ \hat{\sigma}_{21}^1 & \hat{\sigma}_{22}^1 & \dots & \hat{\sigma}_{2m}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_{m1}^1 & \hat{\sigma}_{m2}^1 & \dots & \hat{\sigma}_{mm}^1 \end{pmatrix}. \quad (5.4)$$

Elementy diagonalne $\hat{\sigma}_{ii}^1$ ($i = 1, 2, \dots, m$) oblicza się wedle następującej formuły:

$$\hat{\sigma}_{ii}^1 = \frac{1}{n_1 + n_2 - 2} \left[\sum_{k=1}^{n_1} x_{ki} + \sum_{k=1}^{n_2} y_{ki}^2 - \frac{1}{n_1} \left(\sum_{k=1}^{n_1} x_{ki} \right)^2 - \frac{1}{n_2} \left(\sum_{k=1}^{n_2} y_{ki} \right)^2 \right]. \quad (5.5)$$

Elementy niediagonalne $\hat{\sigma}_{ij}^1$ ($i, j = 1, 2, \dots, m; i \neq j$) oblicza się wedle następującej formuły:

$$\hat{\sigma}_{ij}^1 = \frac{1}{n_1 + n_2 - 2} \cdot \left[\sum_{k=1}^{n_1} x_{ki} x_{kj} + \sum_{k=1}^{n_2} y_{ki} y_{kj} - \frac{1}{n_1} \left(\sum_{k=1}^{n_1} x_{ki} \right) \left(\sum_{k=1}^{n_1} x_{kj} \right) - \frac{1}{n_2} \left(\sum_{k=1}^{n_2} y_{ki} \right) \left(\sum_{k=1}^{n_2} y_{kj} \right) \right] \quad (5.6)$$

Znalezione za pomocą formuł (5.2), (5.3) i (5.5), (5.6) wyznaczniki macierzy (5.1) i (5.2) wykorzystuje się do obliczenia wartości kryterium (4.9). W przypadku prawdziwości hipotezy o jednorodności obliczona wartość kryterium powinna być mniejsza lub równa wartości dopuszczalnej $\chi_{\alpha m}^2$. W przypadku, gdy obliczona wartość W jest większa od $\chi_{\alpha m}^2$ przyjmuje się hipotezę alternatywną.

Przyjęcie hipotezy zerowej usuwa konieczność zbioru cech informatywnych dla rozwiązywanego zadania. Z kolei odrzucenie hipotezy zerowej świadczy o niejednorodności zbiorowości statystycznej i dlatego jest sens dalej sprawdzać hipotezę o informatywności cech $H_0: J_I = J_I^0$ przy hipotezie alternatywnej $H_1: J_I \neq J_I^0$.

Sprawdzenia hipotezy $H_0: J_I = J_I^0$, że dany zbiór cech J_I jest nieinformatywny, można dokonać przez obliczenie wyznaczników dwóch macierzy $\hat{V}_0^{J_I}$ i $\hat{V}_1^{J_I}$. Macierz te otrzymuje się z macierzy \hat{V}_0 i \hat{V}_1 przez wykreślenie tych wierszy i kolumn, którym odpowiada $j \notin J_I$. Obliczone wartości wyznaczników wykorzystuje się potem dla sprawdzenia hipotezy zerowej

$$W_{J_I} = -z \ln \frac{|\hat{V}_1^{J_I}|}{|\hat{V}_0^{J_I}|}. \quad (5.7)$$

gdzie $z = n_1 + n_2 - \frac{l}{2} - 2$.

Hipotezę zerową odrzuca się, gdy ma miejsce nierówność

$$W_{J_l} > \chi_{\alpha, l}^2, \quad (5.8)$$

gdzie $\chi_{\alpha, l}^2$ jest dopuszczalną wartością χ^2 przy poziomie istotności α i l stopniach swobody. Jeśli jednak okaże się, że $W_{J_l} \leq \chi_{\alpha, l}^2$, to rozpatrywany zbiór cech J_l może być uznany za nieinformatywny i wyłączony z dalszej analizy.

Cechy nieinformatywne wyklucza się metodą znalezienia maksymalnej wartości stosunku prawdopodobieństwa. Proces ten istotnie upraszcza się, jeśli cechy są uporządkowane zgodnie z unormowaną różnicą

$$d_j = \frac{n_1 n_2}{n_1 + n_2} \frac{(\bar{x}_j - \bar{y}_j)^2}{\hat{\sigma}_j^2}, \quad (j = 1, 2, \dots, m). \quad (5.9)$$

W formule tej współczynnik $\frac{n_1 n_2}{n_1 + n_2}$ jest wielkością stałą dla każdego $j = 1, 2, \dots, m$. Dlatego unormowaną różnicę można oceniać w jednostkach wariancji, tzn.

$$d_j = \frac{(\bar{x}_j - \bar{y}_j)^2}{\hat{\sigma}_j^2}. \quad (5.10)$$

Jest zrozumiałe, że w procesie porządkowania cech nie dopuścimy do błędu, jeśli będziemy posługiwać się nie formułą (5.10), lecz wartościami statystyki t Studenta:

$$t = \frac{\bar{x}_j - \bar{y}_j}{m_d}, \quad (5.11)$$

gdzie m_d jest błędem losowym różnicy średnich, określonym tak:

$$m_d = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}. \quad (5.12)$$

Jeśli znalezione wartości t uporządkować w kolejności wzrostu wartości kryterium, $t_{j_1} \geq t_{j_2} \geq \dots \geq t_{j_m}$, to dowolna cecha z lewej strony szeregu wartości t będzie grała większą rolę w różnicach między grupami, na które jest rozbita zbiorowość, niż dowolna cecha z prawej. Dlatego dla wykrycia cech nieinformatywnych l można ograniczyć się do sprawdzenia hipotezy dotyczącej elementów stojących w prawej części rozpatrywanego szeregu.

Po uporządkowaniu całego zbioru cech procedura obliczeniowa sprowadza się do obliczenia wartości kryterium

$$\lambda_l^{n_1+n_2} = \frac{|\hat{V}_n^0| |\hat{V}_n^1|}{|\hat{V}_n^1| |\hat{V}_n^0|}, \quad (5.13)$$

gdzie $|\hat{V}_n^0|$ i $|\hat{V}_n^1|$ - wyznaczniki podmacierzy macierzy empirycznej V_0 , której elementy są obliczone przy założeniu o nieinformatywności zbioru cech J_1 i J_2 , natomiast $|\hat{V}_n^1|$ i $|\hat{V}_n^0|$ - wyznaczniki podmacierzy empirycznej V_1 , której elementy są obliczone przy założeniu o nieinformatywności zbioru cech J_2 i J_1 .

Jeśli dla danego l dopuścić, że J_1 nie zależy od J_2 , to macierze V_0 i V_1 w formule (4.3) będą składały się z elementów zerowych. W tym przypadku oceny V i w warunkach hipotezy zerowej H_0 , i w warunkach hipotezy alternatywnej H_1 będą miały postać:

$$\hat{V}_n^0 = \begin{pmatrix} \hat{V}_n^1 & 0 \\ 0 & \hat{V}_n^0 \end{pmatrix}, \quad (5.14)$$

$$\hat{V}_n^1 = \begin{pmatrix} \hat{V}_n^0 & 0 \\ 0 & \hat{V}_n^1 \end{pmatrix}. \quad (5.15)$$

Dla oceny informatywności zbioru cech J_1 i J_2 spośród wszystkich otrzymanych $m-1$ wartości kryterium (5.13) wybiera się maksymalną i wykonuje się procedurę wykreślenia elementów $j \notin J_1$ z macierzy \hat{V}_n^0 i \hat{V}_n^1 , a następnie stosuje się kryterium (5.7) i warunek (5.8).

4. Uwagi końcowe i wnioski

Ze statystycznego punktu widzenia proces grupowania ma sens tylko wtedy, gdy rozpatrywana zbiorowość jest niejednorodna z punktu widzenia pew-

nego z góry określonego zespołu cech. Dlatego zanim przystąpi się do dzielenia zbiorowości studentów na grupy, trzeba najpierw oszacować stopień jej niejednorodności. Odpowiada ta weryfikacji hipotezy o równości średnich wielowymiarowych. Jeśli hipoteza ta zostanie odrzucona, to trzeba niejednorodną zbiorowość studentów podzielić na grupy jednorodne. Powstaje wtedy pytanie, czy można tego dokonać postępując się mniejszą liczbą cech?

Najczęściej wyboru cech informatywnych dokonuje się za pomocą analizy czynnikowej, analizy składowych głównych lub analizy kanonicznej. W odniesieniu do każdej z nich opracowano wiele algorytmów numerycznych, które pozwalają – chociaż nie zawsze w sposób w pełni zadowalający – określić zbiór cech informatywnych. Istnieje przekonanie, że podstawowym kryterium informatywności cech jest wielkość strat ponoszonych przez decydena, który w oparciu o nie podejmuje konkretną decyzję, podczas gdy w rzeczywistości jego decyzja powinna być inna. Praktyczność tego poglądu jest jednak wątpliwa. Zła decyzja pociąga za sobą określone konsekwencje, których skutki – nie tylko w procesie edukacyjnym – mogą być nieodwracalne. Aby temu zapobiec trzeba szukać innych metod.

W pracy zwrócono uwagę, że dekompozycja wielowymiarowych zbiorowości studentów lub nauczycieli – z koniecznością rozwiązywania tego zadania spotykamy się przy ocenie (ewaluacji) jakości procesu edukacyjnego – na podzbiory jednorodne, nie musi być wykonywana w oparciu o pełen zbiór cech, charakteryzujących elementy zbiorowości (studentów lub nauczycieli). Może być przeprowadzona w oparciu o podzbiór cech najbardziej informatywnych. W pracy zaproponowano ogólną, unormowaną miarę stopnia informatywności cechy i pokazano, że istniejące dotychczas miary są jej szczególnymi przypadkami. Szczegółowo rozpatrzono różne miary informatywności oparte na pojęciu ilości informacji w sensie Shannona. Przedstawiono metodę identyfikacji podzbioru cech najbardziej informatywnych spośród z góry ustalonego zbioru cech. Zaproponowano metodę dekompozycji niejednorodnej zbiorowości na części jednorodne w oparciu o zespół cech najbardziej informatywnych. Sformułowano kryterium minimalizujące ryzyko błędu decyzyjnego, przy przydzielaniu studentów bądź nauczycieli do już istniejących grup jednorodnych. Przedstawiono metodę weryfikacji hipotezy o jednorodności, sformułowaną w kategoriach ilości informacji.

W ostatnich trzech dekadach w niemal wszystkich krajach nastąpiło załamanie badań nad zastosowaniami metod matematycznych jako środków wspomagających kształtowanie procesów edukacyjnych. Zostało to spowodowane ogólnościowym ruchem w zakresie reformowania szkolnictwa wszystkich szczebli. Doświadczenie pokazało, że w większości przypadków reformy te nie tylko nie przyniosły oczekiwanych rezultatów, ale – co gorsza – doprowadziły do rozspójnienia systemów edukacyjnych i zniszczenia wielu bardzo dobrych tradycji edukacyjnych. Immanentną właściwością procesów edukacyjnych jest wieloraka niepewność i nieokreśloność. Konieczne staje się więc nowe spojrzenie na procesy edukacyjne, przede wszystkim z punktu widzenia metod

analizy systemowej i współczesnych metod matematycznych, uwzględniające te właściwości procesów edukacyjnych.

Przedstawione w tej pracy metody i uzyskane wyniki są zgodne z tym zapotrzebowaniem. Obszarem ich potencjalnych zastosowań jest nie tylko szkolnictwo wyższe, ale także szkolnictwo niższe (podstawowe, gimnazjalne i licealne).

Literatura

- Anderson T.W., Darling D.A. (1952). Asymptotic theory of certain „Goodness of fit” criteria based on stochastic processes. *Annals of Mathematical Statistics*, 23, 193-212.
- Ashby W.R. (1956). An introduction to cybernetics. Chapman and Hall, London.
- Barra J.R. (1982). Matematyczne podstawy statystyki. Państwowe Wydawnictwo Naukowe, Warszawa.
- Beer S. (1959). Cybernetics and management. The English Universities Press, London.
- Beneš J. (1979). Teoria systemów. Państwowe Wydawnictwo Naukowe, Warszawa.
- Bertalanfy L., von (1984). Ogólna teoria systemów. Postawy rozwój, zastosowania. Państwowe Wydawnictwo Naukowe, Warszawa.
- Bolszew L.N. (1963). Asimptotičeski pirsonowskije priobrazowanija. *Teorija wierojatnostiej i jejo primienienija*, 8, 129-155.
- Brillouin L. (1969). Nauka a teoria informacji. Państwowe Wydawnictwo Naukowe, Warszawa.
- Bubnicki Z. (1974). Identyfikacja obiektów sterowania. Państwowe Wydawnictwo Naukowe, Warszawa.
- Bubnicki Z. (2005). Teoria i algorytmy sterowania. Wydawnictwo Naukowe PWN, Warszawa.
- Cramér H. (1945). Mathematical methods of statistics. Almqvist and Wiksells, Uppsala.
- Fisz M. (1967). Rachunek prawdopodobieństwa i statystyka matematyczna. Państwowe Wydawnictwo Naukowe, Warszawa.
- Goodman L.A., Kruskal W.H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 1732-1769.
- Goodman L.A., Kruskal W.H. (1959). Measures of association for cross classification. II. Further discussions and references. *Journal of the American Statistical Association*, 54, 123-163.
- Goodman L.A., Kruskal W.H. (1963). Measures of association for cross classification. III. Approximate sampling theory. *Journal of the American Statistical Association*, 58, 310-364.
- Goodman L.A., Kruskal W.H. (1954). Measures of association for cross classification. IV. Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67, 415-421.

- Gutenbaum J. (2003). Modelowanie matematyczne systemów. Akademicka Oficyna Wydawnicza EXIT, Warszawa.
- Hald A. (1952). Statistical theory with engineering applications, Chapman and Hall, London.
- Heller M., Lubański W., Ślaga S.W. (1992). Zagadnienia filozoficzne współczesnej nauki. Wstęp do filozofii przyrody. Akademia Teologii Katolickiej, Warszawa.
- Hellwig Z. (1980). Elementy rachunku prawdopodobieństwa i statystyki matematycznej. Państwowe Wydawnictwo Naukowe, Warszawa.
- Jagłom A.M., Jagłom I.M. (1973). Wierojatność i informacja. Nauka, Moskwa.
- Kacprzyński B. 1974). Planowanie eksperymentu. Podstawy matematyczne. Wydawnictwa Naukowo-Techniczne, Warszawa.
- Khan A.N., Ali S.M. (1973). A new coefficient of association. *Annals of the Institute of Statistical Mathematics*, 25, 41-50.
- Klir G.J., red. (1976). Ogólna teoria systemów. Wydawnictwa Naukowo-Techniczne, Warszawa.
- Kołmogorow A. N. (1933). Sulla determinazione empirico di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83-91.
- Koroluk W.S. (1959). Asimptotičeskiy analiz raspriedielenij maksimalnych ukłonenij w schiemie Bernulli. *Teorija wierojatnostiej i jejo primienienija*, 4, 369-397.
- Kullback S. (1959). Information theory and statistics. John Wiley, New York.
- Lebiediew N.N. (1957). Funkcje specjalne i ich zastosowania. Państwowe Wydawnictwo Naukowe, Warszawa.
- Möller F., Capecchi V. (1975). The role of entropy in nominal classification. W: H.H. Blalock i inni, Quantitative sociology. Academic Press, New York, 381-429.
- Neyman J. (1969). Zasady rachunku prawdopodobieństwa i statystyki matematycznej. Państwowe Wydawnictwo Naukowe, Warszawa.
- Niemierko B. (1975). Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe. Wydawnictwa Szkolne i Pedagogiczne, Warszawa.
- Niemierko B. (1990). Pomiar sprawdzający w dydaktyce. Państwowe Wydawnictwo Naukowe, Warszawa.
- Niemierko B. (1991). Między oceną szkolną a dydaktyką. Bliżej dydaktyki. Wydawnictwa Szkolne i Pedagogiczne, Warszawa.
- Nowak R. (2002). Statystyka dla fizyków. Wydawnictwo Naukowe PWN, Warszawa.
- Nowakowski J., Sobczak W. (1970). Teoria informacji. Wydawnictwa Naukowo-Techniczne, Warszawa.
- Oktaba W. (1966). Elementy statystyki matematycznej i metodyka doświadczalnictwa. Państwowe Wydawnictwo Naukowe, Warszawa.
- Pacut A. (1985). Prawdopodobieństwo. Teoria. Modelowanie probabilistyczne w technice. Wydawnictwa Naukowo-Techniczne, Warszawa.
- Penrose R. (2006). Droga do rzeczywistości. Prószyński i S-ka, Warszawa.

- Plucińska A., Pluciński E. (2000). Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne. Wydawnictwa Naukowo-Techniczne, Warszawa.
- Rajski C. (1964). On the normed information rate of discrete random variables. W: Transactions of the 3rd Prague Conference on Information Theory". *Probability and Mathematical Statistics*, 12, 583-585.
- Řehák J., Řeháková B. (1973). Měření statistické závislosti nominálních znaků. *Sociologický Časopis*, 4, 404-417.
- Rao C.R. (1994). Statystyka i prawda. Wydawnictwo Naukowe PWN, Warszawa.
- Shannon C.E. (1949). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379-432, 623-656.
- Sobczak W. (1981). Podstawy probabilistyczne teorii systemów informacyjnych. Wydawnictwa Naukowo-Techniczne, Warszawa.
- Wallace D.L. (1983). A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, 78, 569-574.
- Zieliński R. (1990). Siedem wykładów wprowadzających do statystyki matematycznej. Państwowe Wydawnictwo Naukowe PWN, Warszawa.

Spis treści

1. Wprowadzenie	1
2. Wybór cech istotnych (najbardziej informatywnych)	2
2.1. Uwagi wstępne	2
2.2. Miary oparte na pojęciu ilości informacji	3
2.3. Podział zbiorowości studentów na grupy jednorodnie – ujęcie teorio- informacyjne	14
3. Zadanie przydziału studentów (nauczycieli) do już istniejących grup jednorodnych	19
4. Weryfikacja hipotezy o niejednorodności cech	20
5. Metodyka obliczeń	23
6. Uwagi końcowe i wnioski	27

Literatura

