ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 156, 2002

Dorota Witkowska, Iwona Staniec***

DYCHOTOMICZNA KLASYFIKACJA KREDYTOBIORCÓW PRZY UŻYCIU WIELOWYMIAROWEJ ANALIZY DYSKRYMINACYJNEJ

Streszczenie. W artykule przedstawiono możliwości wykorzystania analizy dyskryminacyjnej do klasyfikacji wniosków kredytowych, a właściwie wspomagania procesu decyzyjnego inspektorów kredytowych. Rezultaty badań empirycznych pokazują, że wielowymiarowa analiza dyskryminacyjna może być wykorzystywana do tego celu.

Słowa kluczowe: analiza dyskryminacyjna, dychotomiczna klasyfikacja, ryzyko kredytowe.

I. WPROWADZENIE

Idea wykorzystania metod ilościowych do oceny przedsiębiorstwa nie jest nowa, bowiem od roku 1932, kiedy P. Fitzpatrick opublikował swoją pracę, obserwowany jest znaczny rozwój badań, zarówno teoretycznych, jak i empirycznych, dotyczących prognozowania upadłości. Badania te koncentrują się zarówno na poszukiwaniu skutecznych metod prognozowania, jak i najlepszego zbioru wskaźników prognozujących.

Do roku 1980 dominowała wprowadzona przez E. I. Altmana (w 1968 r.) metoda analizy dyskryminacyjnej. Praca Altmana zapoczątkowała istotny postęp w badaniu wiarygodności przedsiębiorstw poprzez zastosowanie wielowymiarowej liniowej analizy dyskryminacyjnej. W latach 80. wprowadzono metody analizy logistycznej, a od początku lat dziewięćdziesiątych zaczęto stosować sztuczne sieci neuronowe do prognozowania upadłości przedsiębiorstw.

* Dr hab., Zakład Metod Ilościowych w Zarządzaniu Instytutu Zarządzania Politechniki Łódzkiej.

** Dr inż., Zakład Metod Ilościowych w Zarządzaniu Instytutu Zarządzania Politechniki Łódzkiej.

Do znaczących badań w dziedzinie wypłacalności przedsiębiorstw można zaliczyć prace: E. I. Altmana, W. H. Beavera, R. H. Fishera, J. J. Fritz Patricka, P. Weibla, J. Beermana, R. Ch. Moyera, M. Tamariego, K. Y. Tama, M. Kianga, R. Hena, M. Schumanna oraz M. Feidickera.

W celu minimalizowania ryzyka, standaryzowania i obiektywizowania procesu związanego z udzieleniem kredytu, banki na świecie doskonalą metody oceny zdolności kredytowej wykorzystując nowości naukowe, techniczne oraz własne doświadczenia wynikające z wieloletniej praktyki.

Celem prezentowanego artykułu jest zbadanie możliwości wykorzystania wielowymiarowej analizy dyskryminacyjnej do klasyfikacji kredytobiorców.

II. STRUKTURA ZBIORU DANYCH

W badaniach wykorzystano informacje dotyczące 110 przedsiębiorstw ubiegających się w latach 1994–1998 o przyznanie kredytu w jednym z banków w regionie łódzkim. Na ich podstawie obliczono podstawowe wskaźniki finansowe, które stanowiły pierwotną bazę danych w eksperymentach dotyczących dychotomicznej klasyfikacji firm za pomocą liniowej i logistycznej funkcji dyskryminacji. Każde przedsiębiorstwo zostało opisane przez wskaźniki finansowe wyznaczone na podstawie sprawozdań finansowych dla roku poprzedzającego składanie wniosku kredytowego (t–1) oraz dla okresu bieżącego t (tzn. od początku roku do momentu ubiegania się o kredyt) oraz przez dwie cechy jakościowe, tj. rodzaj branży oraz decyzję odnośnie do przyznania (lub nie) kredytu. Zatem podstawowy zbiór danych wykorzystywanych w eksperymentach zawierał 50 zmiennych dla każdego ze 110 przedsiębiorstw, które zostały zdefiniowane następująco:

x_1 i x_{25} – wskaźnik rentowności aktywów (ROA) (w %) odpowiednio w okresie t–1 i t,

x_2 i x_{26} – wskaźnik rentowności kapitału własnego (ROE) (w %) odpowiednio w okresie t–1 i t,

x_3 i x_{27} – wskaźnik rentowności sprzedaży (ROS) (w %) odpowiednio w okresie t–1 i t,

x_4 i x_{28} – wskaźnik rentowności brutto (w %) odpowiednio w okresie t–1 i t,

x_5 i x_{29} – wskaźnik rentowności netto (w %) odpowiednio w okresie t–1 i t,

x_6 i x_{30} – wskaźnik płynności bieżącej odpowiednio w okresie t–1 i t,

x_7 i x_{31} – wskaźnik krótkoterminowej płynności finansowej (szybkiej) odpowiednio w okresie t–1 i t,

x_8 i x_{32} – wskaźnik długoterminowej płynności finansowej odpowiednio w okresie t–1 i t,
x_9 i x_{33} – wskaźnik rotacji należności w dniach odpowiednio w okresie t–1 i t,
x_{10} i x_{34} – wskaźnik rotacji zapasów w dniach odpowiednio w okresie t–1 i t,
x_{11} i x_{35} – wskaźnik produktywności aktywów odpowiednio w okresie t–1 i t,
x_{12} i x_{36} – wskaźnik poziomu kosztów odpowiednio w okresie t–1 i t,
x_{13} i x_{37} – okres płacenia zobowiązań w dniach odpowiednio w okresie t–1 i t,
x_{14} i x_{38} – wskaźnik rotacji majątku trwałego odpowiednio w okresie t–1 i t,
x_{15} i x_{39} – wskaźnik rotacji majątku obrotowego odpowiednio w okresie t–1 i t,
x_{16} i x_{40} – wskaźnik ryzyka aktywów odpowiednio w okresie t–1 i t,
x_{17} i x_{41} – wskaźnik ogólnego zadłużenia odpowiednio w okresie t–1 i t,
x_{18} i x_{42} – wskaźnik pokrycia majątku trwałego kapitałem stałym odpowiednio w okresie t–1 i t,
x_{19} i x_{43} – wskaźnik długu (dźwignia finansowa) odpowiednio w okresie t–1 i t,
x_{20} i x_{44} – wskaźnik zadłużenia kapitału własnego odpowiednio w okresie t–1 i t,
x_{21} i x_{45} – wskaźnik pokrycia obsługi długu odpowiednio w okresie t–1 i t,
x_{22} i x_{46} – wskaźnik pokrycia odsetek odpowiednio w okresie t–1 i t,
x_{23} i x_{47} – wskaźnik zadłużenia środków trwałych odpowiednio w okresie t–1 i t,
x_{24} i x_{48} – stopa zadłużenia odpowiednio w okresie t–1 i t,
x_{49} – rodzaj branży,
x_{50} – decyzja kredytowa (wiarygodny lub niewiarygodny klient).

Oprócz prezentowanego zbioru zmiennych eksperymenty numeryczne przeprowadzono wykorzystując w tym celu przyrosty wskaźników finansowych wyrażających różnicę między wartościami wskaźników w okresach t i t–1. W ten sposób powstała nowa baza danych, którą uzupełniono zmiennymi x_{49} i x_{50} zawierającymi 26 zmiennych.

W zadaniach klasyfikacji jednym z ważniejszych problemów jest dobór zmiennych diagnostycznych. Wykorzystano różne zestawy zmiennych diagnostycznych otrzymane przy użyciu następujących metod: analizy macierzy współczynników korelacji (metody Nowaka i Hellwiga) oraz algorytmu genetycznego.

Redukcję liczby zmiennych przeprowadza się również korzystając z preprocessingu bazy danych, tworząc w ten sposób swego rodzaju zmienne sztuczne o dużej zawartości informacyjnej i wzajemnie nie skorelowane. W badaniach wykorzystano następujące metody preprocessingu zmiennych w postaci wskaźników finansowych x_{it} oraz ich przyrostów:

- analizę czynnikową,
- analizę głównych składowych.

Eksperymenty przeprowadzono dla zmiennych zdefiniowanych w postaci:

- wskaźników finansowych x_{it} (pierwotna baza danych),
- z przyrostów wskaźników finansowych Δx_{it},
- zmiennych po preprocessingu przeprowadzonego za pomocą analizy czynnikowej,
- zmiennych po preprocessingu przeprowadzonego za pomocą analizy głównych składowych.

Dokonując klasyfikacji ze wzorcem, istnieje konieczność podziału bazy danych na zbiór uczący (treningowy) i testowy. Strukturę obu zbiorów podano w tabl. 1.

Tablica 1

Struktura zbioru danych

	Niewiarygodnych	Wiarygodnych
Próba treningowa (ucząca)	45	45
Próba testowa	10	10

Źródło: opracowanie własne.

III. MIERNIKI JAKOŚCI KLASYFIKACJI

W przypadku, gdy znane są wzorce grupowania, można wyznaczyć błędy klasyfikacji, wynikające ze złego zakwalifikowania obiektów badania. Dokonując klasyfikacji dychotomicznej, wyróżnia się błędy pierwszego i drugiego rodzaju.

W naszym przypadku z błędem pierwszego rodzaju mamy do czynienia wtedy, gdy odrzucony przez bank klient zostanie zaklasyfikowany jako wiarygodny. Opierając się wówczas wyłącznie na takiej decyzji, bank narażony zostałby na ryzyko nieterminowej spłaty udzielonego kredytu wraz z odsetkami.

$$E_1 = \frac{N_1}{Z} \times 100\% \tag{1}$$

gdzie:

N_1 – liczba negatywnie rozpatrzonych przez bank wniosków kredytowych zaklasyfikowanych za pomocą wybranej metody grupowania jako wnioski, na podstawie których należy udzielić kredytu,

Z – liczba wszystkich negatywnie rozpatrzonych przez bank wniosków kredytowych znajdujących się w zbiorze uczącym (testującym).

Błąd drugiego rodzaju powstaje wówczas, gdy pozytywnie rozpatrzone przez bank przedsiębiorstwo zostanie w wyniku analizy dyskryminacyjnej rozpoznane jako niewiarygodne. W tym przypadku bank traci możliwe do uzyskania dochody w postaci odsetek od kredytu.

$$E_2 = \frac{N_2}{D} \times 100\% \tag{2}$$

gdzie:

N_2 – liczba pozytywnie rozpatrzonych przez bank wniosków kredytowych zaklasyfikowanych w wyniku analizy dyskryminacyjnej jako przedsiębiorstwa niewiarygodne,

D – liczba wszystkich pozytywnie rozpatrzonych przez bank wniosków kredytowych w zbiorze uczącym (testującym).

Ogólny błąd klasyfikacji zdefiniować można w postaci:

$$E = \frac{N_1 + N_2}{Z + D} \times 100\% \tag{3}$$

Ogólny błąd klasyfikacji (3) określa, jaka część rozpatrywanych w procedurze klasyfikacyjnej obiektów została niepoprawnie rozpoznana.

IV. METODY DOBORU ZMIENNYCH DIAGNOSTYCZNYCH

Dobór cech diagnostycznych należy do zadań szczególnie ważnych, jako że w znacznym stopniu zależą od niego ostateczne wyniki badania. Zestaw cech diagnostycznych powinien być tak sporządzony, by w sposób możliwie pełny charakteryzował najważniejsze aspekty badanego zjawiska. Wybór cech odbywa się przez przetwarzanie i analizę informacji statystycznych za pomocą odpowiednich procedur formalnych. Podstawą do wyboru cech diagnostycznych jest tzw. wstępna lista cech zaproponowana przez badacza na podstawie ogólnej znajomości zjawiska[1].

Dobór cech diagnostycznych można podzielić na dwie grupy kryteriów: merytoryczne i statystyczne[2]. Kryterium merytoryczne jest oceną jakościową

[1] Por. B. Podolec, K. Zając (1978), s. 20.

[2] Por. W. Dębski (1994).

i może być przeprowadzone m. in. na podstawie metody delfickiej lub tzw. burzy mózgów. Kryterium statystyczne oparte jest na miernikach ilościowych, które wyznaczane są za pomocą formalnych procedur.

W wielu badaniach ekonomicznych istnieje potrzeba redukowania liczby zmiennych opisujących badany wycinek rzeczywistości. Potrzeba ta może wynikać z faktu posiadania mało licznej próby i jednakowo dużej liczby szacowanych parametrów lub występowania zmiennych powielających tę samą informację. Przeprowadzana redukcja musi odpowiadać pewnym wymaganiom, aby uzyskany opis nie fałszował rzeczywistości. Do tego celu powinno się wykorzystać odpowiednie metody, których zastosowanie umożliwia uzyskanie zestawu zmiennych charakteryzujących w sposób możliwie pełny badane jednostki, a przy tym tworzących zespół jak najmniej liczny. Podane wymagania są spełnione wtedy, gdy zmienne diagnostyczne posiadają następujące własności:

– są nieskorelowane lub co najwyżej słabo skorelowane między sobą;

– są silnie skorelowane ze zmiennymi nie wchodzącymi do zespołu diagnostycznego;

– posiadają zdolność dyskryminacji badanych jednostek, tj. charakteryzują się wysoką zmiennością wśród wszystkich jednostek zbioru, a niską wśród jednostek wydzielonych grup;

– nie ulegają wpływom zewnętrznym.

Podstawowe metody doboru zmiennych diagnostycznych to:

– metoda analizy macierzy współczynników korelacji (tzw. metoda Nowaka i Hellwiga);

– metoda algorytmu genetycznego.

Tablica 2

Specyfikacja zmiennych diagnostycznych dla danych w formie wskaźników finansowych

Specyfikacja zmiennych		Zestaw zmiennych diagnostycznych
Metoda Nowaka	wszystkie	x_{25}, x_{40}, x_{42}
	podzielne	x_{25}, x_{40}
	wszystkie + rodzaj branży	x_{25}, x_{40}, x_{42}, x_{49}
	podzielne + rodzaj branży	x_{25}, x_{40}, x_{49}
Metoda Hellwiga	wszystkie	x_2, x_3, x_4, x_6, x_7, x_8, x_{14}, x_{18}, x_{19}, x_{22}, x_{24}, x_{25}, x_{27}, x_{30}, x_{33}, x_{35}, x_{38}, x_{47}
	wszystkie podzielne	x_2, x_3, x_4, x_6, x_{19}, x_{22}, x_{24}, x_{25}, x_{27}, x_{30}, x_{33}
	centralne	x_3, x_4, x_7, x_8, x_{18}, x_{19}, x_{24}, x_{25}, x_{27}, x_{30}, x_{33}, x_{35}, x_{47}
	centralne podzielne	x_3, x_4, x_{19}, x_{24}, x_{25}, x_{27}, x_{30}, x_{33}
	izolowane	x_2, x_{14}, x_{22}, x_6, x_{38}
	izolowane podzielne	x_2, x_{22}, x_6

Tablica 2 (cd.)

Specyfikacja zmiennych	Zestaw zmiennych diagnostycznych
Algorytm genetyczny	x_1, x_3, x_6, x_7, x_9, x_{10}, x_{14}, x_{18}, x_{25}, x_{28}, x_{29}, x_{30}, x_{32}, x_{42}, x_{49}
Zmienne zaproponowane przez J. Gajdkę i D. Stos [1996]	x_{35}; x_{25}; x_{27}; x_{48} $x_{51} = \frac{\text{zobowiązania krótkoterminowe}}{\text{koszt wytworzenia produkcji sprzedanej}} \times 360$

Źródło: opracowanie własne.

Analizując tabl. 2, można zauważyć, że niemal wszystkie zestawy zmiennych zawierają wskaźnik rentowności aktywów (x_{25}), który nie występuje jedynie w zestawie izolowanych i izolowanych podzielnych zmiennych wyznaczonych metodą Hellwiga.

Tablica 3

Specyfikacja zmiennych diagnostycznych dla danych w postaci przyrostów wskaźników finansowych

Specyfikacja zmiennych		Zestaw zmiennych diagnostycznych
Metoda Nowaka	wszystkie	Δx_1, Δx_{20}
	wszystkie + rodzaj branży	Δx_1, Δx_{20}, x_{49}
Metoda Hellwiga	wszystkie	Δx_1, Δx_5, Δx_6, Δx_8, Δx_9, Δx_{15}, Δx_{14}, Δx_{16}, Δx_{18}, Δx_{21}, Δx_{22}, Δx_{24}
	wszystkie podzielne	Δx_1, Δx_5, Δx_6, Δx_8, Δx_{16}, Δx_{21}, Δx_{18}, Δx_{22}, Δx_{24}
	centralne	Δx_1, Δx_5, Δx_6, Δx_8, Δx_9, Δx_{15}, Δx_{16}, Δx_{21}
	centralne podzielne	Δx_1, Δx_5, Δx_6, Δx_8, Δx_{16}, Δx_{21}
	izolowane	Δx_{14}, Δx_{18}, Δx_{22}, Δx_{24}
	izolowane podzielne	Δx_{18}, Δx_{22}, Δx_{24}
Algorytm genetyczny		Δx_1, Δx_2, Δx_3, Δx_6, Δx_7, Δx_8, Δx_9, Δx_{10}, Δx_{13}, Δx_{18}, Δx_{21}, x_{49}

Źródło: opracowanie własne.

W przypadku zmiennych zdefiniowanych w postaci przyrostów wskaźników finansowych we wszystkich zestawach zmiennych (wykluczając zmienne izolowane i izolowane podzielne) występuje przyrost wskaźnika rentowności aktywów.

Zastosowane metody preprocessingu to:
- analiza głównych składowych,
- analiza czynnikowa.

Tablica 4

Specyfikacja zmiennych diagnostycznych otrzymanych po preprocessingu danych w formie wskaźników i ich przyrostów

Specyfikacja zmiennych	Liczba zmiennych		Procent objaśnianej zmienności pierwotnego zestawu danych	
	wskaźniki	przyrosty	wskaźniki	przyrosty
Analiza czynnikowa	7	10	–	–
Analiza głównych składowych	2	2	96,02%	94,85%
	3	5	98,77%	99,71%
	8	8	99,8%	100%

Źródło: opracowanie własne.

Należy zauważyć, iż każdy z zaproponowanych zestawów zmiennych objaśnia w ponad 94,85% zmienność pierwotnego zestawu danych. Przedstawiono w tabl. 2, 3 i 4 zestawy zmiennych diagnostycznych wykorzystywane w dalszych analizach.

V. WIELOWYMIAROWA ANALIZA DYSKRYMINACYJNA

Budowę funkcji dyskryminacji należy poprzedzić wielowymiarowymi analizami zmiennych diagnostycznych. Podstawowe założenia, które należy zweryfikować przed przeprowadzeniem wielowymiarowej analizy dyskryminacyjnej, to[3]:

- rozkład normalny[4],
- podzielność zmiennych,
- równość macierzy kowariancji.

Rozkład normalny. Zakłada się, że zmienne dyskryminacyjne reprezentują wielowymiarowy rozkład normalny. Dotychczasowe badania z użyciem wielowymiarowej funkcji dyskryminacji potwierdzają, że jest ona dobrym

[3] Podobnie sformułowano założenia w pracach M. Krzyśko (1990), s. 19 i *Statistica*^{TM} *PL* (1997), s. 3069. Jak twierdzą C. Domański, M. Misztal (1998), s. 96, „liniowa funkcja dyskryminacji jest optymalna przy spełnieniu obu tych założeń [w niniejszej pracy założenia 1 i 3], jednak często jest ona wykorzystywana z dobrym rezultatem nawet, kiedy żadne z tych założeń nie jest spełnione. Wynika to z faktu, że liniowa funkcja dyskryminacji jest odporna na te założenia".

[4] Dotychczasowe badania z użyciem wielowymiarowej funkcji dyskryminacji potwierdzają, że jest ona dobrym klasyfikatorem mimo naruszenia tego założenia. Por. D. Morrison (1990), s. 347; *Machine Learning, Neural and Statistical Classification*, (1993), s. 22; A. Sokołowski (1999), s. 40.

klasyfikatorem mimo naruszenia tego założenia. Do weryfikacji założenia o wielowymiarowym normalnym rozkładzie używa się testów normalności wielowymiarowego rozkładu normalnego, np.: testu Kołmogorowa–Smirnowa, Shapiro–Wilka lub testu zgodności Hellwiga.

Podzielność zmiennych. Podzielność zmiennych przejawia się w systematycznej różnicy wartości średnich między grupami. Do wyeliminowania zmiennych niepodzielnych korzysta się z testu *U*–Manna–Whitney'a (jest to wielowymiarowa odmiana jednowymiarowego testu *t*-Studenta).

Równość macierzy kowariancji. Zakłada się, że macierze kowariancji zmiennych diagnostycznych są równe w grupach. Badania empiryczne wykazują, że można pominąć to założenie. Poza tym wielowymiarowy test M. Boxa na równość kowariancji jest szczególnie wrażliwy na odchylenia od wielowymiarowego rozkładu normalnego. Ito i Schull (1964) zbadali zachowanie rozkładów, gdy macierze kowariancji są różne i pokazali, że przy dużych liczebnościach niejednakowe macierze kowariancji nie mają wpływu na prawdopodobieństwo błędu pierwszego rodzaju oraz moc testu[5].

Wielowymiarowa analiza dyskryminacyjna jest metodą klasyfikacji danego obiektu O_i ze zbioru Ω do jednej z wcześniej ustalonych klas[6]. Zakwalifikowanie obiektu O_i opisanego przez zmienne zawarte w wektorze $\mathbf{x}_i$ dokonuje się na podstawie wartości funkcji dyskryminacyjnej D, którą wyznacza się następująco:

$$D(\mathbf{x}_i) = a_0 + a_1 x_{i1} + a_2 x_{i2} + \ldots + a_k x_{ip} = a_0 + \mathbf{a}^T \mathbf{x}_i \tag{4}$$

gdzie:

$\mathbf{a}^T = [a_1, a_2, \ldots, a_p]$ – wektor współczynników dyskryminacyjnych;

a_0 – wartość krytyczna,

$\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ – wektor zmiennych diagnostycznych (zmiennych dyskryminacyjnych) dla i-tego obiektu.

Liniowa funkcja dyskryminacji opisuje hiperpłaszczyznę rozdzielającą zbiory obiektów w ten sposób, aby je jak najlepiej odseparować. Zatem powstaniu liniowej funkcji dyskryminacji dla dwóch grup towarzyszy założenie, że dwie niezależne próby o liczebności n_1 (liczba elementów klasy K_1) i n_2 (liczba elementów klasy K_2) pochodzą z p-wymiarowych rozkładów normalnych o wektorach wartości oczekiwanych odpowiednio równych μ_1 i μ_2 oraz takiej samej macierzy kowariancji Σ. Dobrze zdefiniowana funkcja dyskryminacyjna uwzględnia wzajemne powiązania pomiędzy różnymi zmiennymi diagnostycznymi, przez co może dostarczać dodatkowych informacji.

[5] Por. K. Ito, W. J. Schull (1964), s. 71–82.

[6] Por. I. Staniec, D. Witkowska (1998), s. 541–546.

Wektor parametrów funkcji dyskryminacyjnej wyznacza się ze wzoru:

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1) \tag{5}$$

gdzie:

$\mathbf{S}$ – macierz kowariancji,

$\bar{\mathbf{x}}_1$; $\bar{\mathbf{x}}_2$ – wektory przeciętnych wartości zmiennych niezależnych w klasie pierwszej i drugiej.

Jeżeli wariancje obserwowanych zmiennych są identyczne, to elementy wektora parametrów funkcji dyskryminacji przedstawiają udział poszczególnych zmiennych dyskryminacyjnych. W przeciwnym przypadku porównywalność współczynników funkcji dyskryminacji uzyskuje się dzieląc każdy z nich przez odchylenie standardowe odpowiedniej zmiennej[7].

Przeciętne wartości funkcji dyskryminacyjnej wynoszą:

– dla klasy pierwszej:

$$\bar{y}_1 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_1 \tag{6}$$

– dla klasy drugiej:

$$\bar{y}_2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_2 \tag{7}$$

Wartością krytyczną jest liczba wyznaczona na podstawie reguły:

$$a_0 = -\left\lfloor \alpha(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_1 + (1-\alpha)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_2 \right\rfloor \tag{8}$$

gdzie:

α – prawdopodobieństwo wystąpienia elementów klasy pierwszej[8],

$1-\alpha$ – prawdopodobieństwo wystąpienia elementów klasy drugiej.

Regułę klasyfikującą można przedstawić w postaci jednej statystyki[9]:

$$D(\mathbf{x}_1) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_i - \alpha(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_1 - (1-\alpha)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\bar{\mathbf{x}}_2 \tag{9}$$

Obserwację o wektorze zmiennych $\mathbf{x}_i$ należy zaklasyfikować do klasy pierwszej (K_1) jeżeli $D(\mathbf{x}_i) \geqslant 0$, a do klasy drugiej (K_2), jeżeli $D(\mathbf{x}_i) < 0$.

W przypadku dyskryminacji logistycznej przyjmuje się, że prawdopodobieństwo przynależności obiektu o wektorze cech $\mathbf{x}_i$ do klasy K_s jest wartością dystrybuanty rozkładu logistycznego (L):

[7] Por. D. Morrison (1990), s. 343.

[8] Wielu autorów, m. in. K. Jajuga (1993), E. Gatnar (1998), D. Morrison (1990), przyjmuję automatycznie $\alpha = 0{,}5$.

[9] Statystyka (6) jest nazywana statystyką klasyfikacyjną Walda-Andersona, gdyż jako pierwszy podał ją A. Wald (1944), s. 145–162, a jako pierwszy podał jej własności T. W. Anderson (1958), s. 31–50.

Tablica 5

Porównanie błędów klasyfikacji przy użyciu wielowymiarowej analizy dyskryminacyjnej dla różnych zestawów zmiennych diagnostycznych

Funkcje dyskryminacyjne	Specyfikacja zmiennych		Zmienne diagnostyczne										
			zmienne zdefiniowane przez J. Gajdkę i D. Stos	metoda Nowaka	metoda Hellwiga						preprocessing		algorytm genetyczny
					wszystkie	wszystkie podzielne	centralne	centralne podzielne	izolowane	izolowane podzielne	analiza czynnikowa	analiza głównych składowych	
Liniowa	wskaźniki		x_{25}, x_{48}	x_{25}, x_{40}	x_{25}	x_{25}, x_4	x_{25}	x_{25}	x_{26}, x_{14}	x_{26}	C6, C7, C4	PC4, PC1	–
	błędy	E	20%	20%	20%	20%	20%	15%	45%	35%	25%	60%	–
		E_1	0%	0%	0%	0%	0%	0%	10%	10%	10%	20%	–
		E_2	40%	40%	40%	40%	40%	30%	80%	60%	40%	100%	
	przyrosty		–	Δx_1, Δx_{20}	Δx_{16}, Δx_5, Δx_{18}	Δx_8, Δx_5	Δx_{16}, Δx_{21}, Δx_9	$\Delta x16$, Δx_{21}	Δx_{18}	Δx_{18}	C3, C8	PC8, PC1	–
	błędy	E		35%	35%	40%	40%	40%	45%	35%	55%	40%	–
		E_1		10%	0%	10%	20%	20%	10%	10%	30%	0%	–
		E_2		60%	70%	70%	60%	60%	80%	60%	80%	80%	
Logistyczna	wskaźniki		–	x_{25}, x_{42}	x_{25}, x_{27}, x_{26}	x_{25}, x_{27}, x_{26}, x_{24}	x_{30}	x_{25}, x_{24}, x_3	x_{26}, x_2	x_{26}	C6, C4, C5,	PC1	x_{25}, x_1, x_{29}, x_{28}
	błędy	E		5%	30%	40%	20%	10%	30%	25%	15%	35%	5,56%
		E_1		0%	50%	40%	20%	10%	10%	10%	20%	20%	11,11%
		E_2		10%	10%	40%	20%	10%	50%	40%	10%	50%	0%
	przyrosty		–	Δx_1	Δx_{21}, Δx_1, Δx_5	Δx_{24}, Δx_1, Δx_{16}, Δx_{18}	Δx_1, Δx_{16}	Δx_1, Δx_{16}	Δx_{24}	Δx_{24}	C1, C3, C8	PC7, PC3	Δx_1, Δx_3, Δx_2, Δx_{25}
	błędy	E		45%	45%	55%	40%	40%	50%	30%	45%	35%	50%
		E_1		20%	30%	30%	20%	10%	20%	0%	10%	40%	30%
		E_2		70%	60%	80%	60%	70%	80%	60%	80%	30%	70%

Źródło: opracowanie własne.

Uwaga: W tablicy w zestawach zmiennych dyskryminacyjnych podano zmienne, które mają istotny wpływ na zdolności dyskryminacyjne danego modelu (uporządkowane według istotności). Podane w wierszach błędy są najniższymi błędami klasyfikacji, jakie uzyskano przy podanym zestawie zmiennych dla próby testowej.

$$P(K_s/\mathbf{x}_i) = L(a_0 + \mathbf{a}^T\mathbf{x}_i) \tag{10}$$

gdzie przyjęto założenie o liniowości logarytmu ilorazu wiarygodności.

W przypadku klasyfikacji dychotomicznej model dyskryminacji logistycznej jest równoważny modelowi regresji logistycznej, który jest postaci:

$$P(\mathbf{x}_i) = \frac{1}{1 + e^{-(a_0 + \mathbf{a}^T\mathbf{x}_i)}} \tag{11}$$

Parametry równania (11) szacuje się metodą największej wiarygodności. Uzyskane oceny równania logistycznego można interpretować następująco:

– jeżeli $a_j > 0$, to czynnik opisywany przez zmienną x_j działa stymulująco na prawdopodobieństwo wystąpienia badanego zjawiska;

– jeżeli $a_j < 0$, to czynnik opisywany przez zmienną x_j działa limitująco na prawdopodobieństwo wystąpienia badanego zjawiska;

– jeżeli $a_j = 0$, to czynnik opisywany przez zmienną x_j nie wpływa na prawdopodobieństwo wystąpienia badanego zjawiska.

VI. WYNIKI EKSPERYMENTÓW NUMERYCZNYCH

Celem badań była klasyfikacja klientów banku za pomocą liniowej i logistycznej funkcji dyskryminacyjnej. Eksperymenty zostały przeprowadzone dla 39 zestawów zmiennych diagnostycznych, przedstawionych w tabl. 2–4. Jakość klasyfikacji oceniono na podstawie błędów (1)–(3), których wartości dla 20-elementowego zbioru testującego zamieszczono w tab. 5. Podano w niej również symbole zmiennych statystycznie istotnych.

VII. WNIOSKI KOŃCOWE

Na podstawie przeprowadzonej analizy empirycznej można sądzić, że logistyczna funkcja dyskryminacji jest przy dychotomicznej klasyfikacji klientów banku na klasy: wiarygodnych i niewiarygodnych kredytobiorców sprawnym instrumentem.

Przy budowie modeli wykorzystano zmienne zdefiniowane jako wskaźniki finansowe oraz ich przyrosty. Modele zbudowane przy użyciu zmiennych w formie przyrostów ogólnie nie radzą sobie z rozpoznawaniem kredytobiorców, bowiem niezależnie od zestawu zmiennych wejściowych i modelu wykorzystywanego do klasyfikacji odsetek poprawnie rozpoznanych przedsiębiorstw dla zmiennych w postaci wskaźników jest większy niż dla zmiennych w postaci przyrostów wskaźników.

W przypadku liniowej funkcji dyskryminacji najlepsze wyniki klasyfikacji otrzymano dla podzielnych zmiennych centralnych wybranych metodą Hellwiga. Ogólny błąd klasyfikacji wynosi 15%, a błąd pierwszego rodzaju 0% i błąd drugiego rodzaju 30%. Dla logistycznej funkcji dyskryminacji najlepsze wyniki klasyfikacji uzyskano dobierając zmienne diagnostyczne metodą Nowaka. W tym przypadku ogólny błąd klasyfikacji wynosi 5%, a błąd pierwszego rodzaju 0% i błąd drugiego rodzaju 10%. Nieznacznie gorsze wyniki uzyskano stosując algorytm genetyczny do wyboru zmiennych: ogólny błąd klasyfikacji wynosi 5,56%, a błąd pierwszego rodzaju 11%, błąd drugiego rodzaju 0%.

Najwyższe błędy klasyfikacji zaobserwowano dla funkcji liniowej, w której zmiennymi diagnostycznymi były zmienne skonstruowane za pomocą analizy głównych składowych ($E=60\%$, $E_1=20\%$, $E_2=100\%$ dla zmiennych w postaci wskaźników x_i oraz $E=40\%$, $E_1=0\%$, $E_2=80\%$ dla zmiennych Δx_i. Stosują analizę czynnikową dla zmiennych w postaci wskaźników finansowych uzyskano ogólne błędy klasyfikacji równe 25% i 15% odpowiednio dla funkcji liniowej i logistycznej.

Na podstawie wyników badań empirycznych można sądzić, że największą siłę dyskryminacyjną mają zmienne: x_{25} – wskaźnik ROA w okresie t, Δx_1 – przyrost wskaźnika ROA, Δx_{16} – przyrost wskaźnika ryzyka aktywów oraz Δx_{18} – przyrost wskaźnika pokrycia majątku trwałego kapitałem stałym. Jak pokazano, ostateczne wyniki klasyfikacji zależą od zmiennych wykorzystywanych do budowy modeli i jest to element wyraźnie wpływający na efektywność metod klasyfikacji.

LITERATURA

Altman E. I. (1968), *Financial Ratios Discriminant Analysis and the Prediction of Corporate Bankruptcy*, „Journal of Finanse", **23**, 589–609.

Anderson T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York–London.

Azoff E. M. (1995), *Neural Network Time Series Forecasting of Financial Markets*, John Wiley & Sons Ltd., Chichester.

Dębski W. (1994), *Ryzyko bankowe*, „Bank i Kredyt", **10**, 5–10.

Domański C., Misztal M. (1998), *Zastosowanie wybranych metod dyskryminacji do wspomagania diagnozy i określania ryzyka operacyjnego u pacjentów z chorobą wieńcową*, [w:] *Modelowanie preferencji, a ryzyko '98* red. T. Trzaskalik, Katowice, s. 93–106.

Fritzpatrik P. (1932), *A Comparison of the Ratios of Successful Industrial Enterprises with These of Falied Companies*, The accountants Publishing Company.

Gajdka J., Stos D. (1996), *Wykorzystanie analizy dyskryminacyjnej w ocenie kondycji finansowej przedsiębiorstw*, [w:] *Restrukturyzacja w procesie przekształceń i rozwoju przedsiębiorstw*, red. R. Borowiecki, Akademia Ekonomiczna, Towarzystwo Naukowe Organizacji i Kierownictwa, Kraków, s. 56–65.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.
Gwiazda T. D. (1998), *Algorytmy genetyczne. Zastosowanie w finansach*, Wyższa Szkoła Przedsiębiorczości i Zarządzania im. L. Koźmińskiego, Warszawa.
Ito K., Schull W. J. (1964), *On the Robustness of the T_0^2 Test in Multivariate Analysis of Variance when Variance-Covariance Matrices Are Not Equal*, „Biometrika", **51**, 71–82.
Jajuga K. (1993), *Statystyczna analiza wielowymiarowa*, Biblioteka ekonometryczna, PWN, Warszawa.
Kolonko J. (1980), *Analiza dyskryminacyjna i jej zastosowania w ekonomii*, PWN, Warszawa.
Krzyśko M. (1990), *Analiza dyskryminacyjna*, WNT, Warszawa.
Machine Learning, Neural and Statistical Classification, (1993), Comparative Testing of Statistical and Logical Learning.
Morrison D. (1990), *Statystyczna analiza wielowymiarowa*, PWN, Warszawa.
Podolec B, Zając K. (1978), *Ekonometryczne metody ustalania regionów konsumpcji*, PWE, Warszawa.
Refenes Apostolos-Paul. (1994), *Neural Networks in the Capital Markets*, John Wiley & Sons Ltd., Chichester.
Sokołowski A. (1999), *Analizy wielowymiarowe*, Materiał kursowy StatSoft Polska, 6–7 maja, Kraków.
Staniec I., Witkowska D. (1998), *Analiza dyskryminacyjna w klasyfikacji wniosków kredytowych*, Materiały z V Międzynarodowej Konferencji Naukowej „Zarządzanie Organizacjami Gospodarczymi", red. J. Lewandowski, Łódź, 541–546.
Statistica™ PL, (1997), t. 3, StatSoft.
Wald A. (1944), *On statistical problem arising in the classification of an individual into one of two groups*, Annals of Mathematical Satistics, **15**, 145–162.

Dorota Witkowska, Iwona Staniec

DISCRIMINANT ANALYSIS TO CREDIT GRANTING PROCEDURE

(Summary)

The paper deals with the problem whether and to what extent multivariate linear discriminant analysis (MDA) are suitable for the credit investigation of companies. Sometimes in cases of credit evaluation, formalised methods aiming at the objectification and rationalisation of that operation are made use of. More often than not, statistical methods serve as formalised methods, but methods of pattern recognition are also employed. So far, the statistical method of the MDA has frequently and successfully been used for the purpose of credit evaluation. 110 data records, each of which represents the annual financial statements of – a company formed the basis of the inquiry. The annual financial statements analysed were prepared in accordance with the regulation of the GUS (Central Statistical Office).