PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Mariola Chrzanowska
Wyższa Szkoła Ekonomii i Administracji w Kielcach
Dorota Witkowska
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

ZASTOSOWANIE WYBRANYCH METOD KLASYFIKACJI DO ROZPOZNAWANIA INDYWIDUALNYCH KREDYTOBIORCÓW

1. Wstęp

Przez ostatnie pół roku portfele kredytowe banków urosły o sumę ponad 18 mld zł. To tylko 6 mld zł mniej niż w całym 2005 r. Banki wciąż skutecznie zabiegają o nowych klientów – według NBP w pierwszym półroczu roku 2006 wartość pieniędzy zdeponowanych w bankach urosła o 8 mld zł[1]. Kwota zdeponowana w bankach – prawie 229 mld zł – jest wyższa niż kiedykolwiek. Z tych funduszy banki finansują wciąż rosnący popyt na kredyty. Równocześnie wraz ze zwiększającą się liczbą kredytów rośnie zagrożenie kredytami niespłacanymi w terminie. Duża liczba takich kredytów może spowodować utratę płynności finansowej banku. Zatem problem rozpoznawania klientów, którzy terminowo nie spłacają kredytu, jest istotny. Niniejszy artykuł zawiera wyniki aplikacji wybranych metod klasyfikacji do oceny indywidualnych kredytobiorców.

Celem badań jest zastosowanie regresji binarnej, bayesowskiej analizy dyskryminacyjnej oraz zagregowanych drzew klasyfikacyjnych do rozpoznawania stanu zadłużenia przez kredytobiorców indywidualnych. Efektywność wykorzystanych metod porównano za pomocą wybranych miar.

2. Opis danych

Baza danych obejmuje informacje o 2576 kredytobiorcach, z których 419 nie spłaca kredytu (termin spłaty kredytu dla większości klientów jeszcze nie upłynął).

[1] Dla porównania w całym roku 2005 bankom przybyło 10 mld zł depozytów (na podstawie [Samcik 2006]).

Każdy z klientów opisany został za pomocą następujących cech, oznaczonych symbolami C0-C11:

– C0 – grupa klienta, która została określona przez bank na podstawie stanu spłacalności kredytu (warianty cechy: sytuacja normalna, sytuacja wątpliwa, kredyt pod obserwacją, kredyt poniżej standardu; kredyt stracony),
– C1 – czas spłaty kredytu,
– C2 – wiek kredytobiorcy,
– C3 – waluta pożyczki (CHF, EUR, PLN, USD),
– C4 – płeć kredytobiorcy,
– C5 – oddział (miejscowość/miasto, w którym udzielono kredytu),
– C6 – aktualna stopa procentowa,
– C7 – poprzednia stopa procentowa,
– C8 – wartość kredytu (wyrażona w zł),
– C9 – aktualnie spłacona kwota,
– C10 – posiadane lokaty (warianty cechy: konto *a'vista*; do 1 miesiąca, od 1 do 3 miesięcy, od 3 do 6 miesięcy, do 1 roku, powyżej 1 roku),
– C11 – grupa ryzyka NBP.

Ze względu na przyjęty cel badania cechą grupującą jest C0, opisująca stan spłacalności kredytu. Wstępna analiza danych pozwoliła wykryć liniową zależność pomiędzy cechami C0 „grupa klienta" oraz C11: „grupa ryzyka NBP". Klasyfikacja prowadzona za pomocą zmiennej C11 była bezbłędna zarówno w zbiorze uczącym, jak i testowym. Dlatego w dalszym badaniach zdecydowano się jej nie uwzględniać. Całkowity brak zależności z cechą grupującą C0 wykazały zmienne: C5: „oddział" oraz C4: „płeć". Dobór zmiennych do modelu przeprowadzono na podstawie analizy zależności[2]. Ostatecznie do badań wybrano zmienne: C3: „waluta", C1: „czas spłaty" i C2: „wiek", które są silnie skorelowane ze zmienną grupującą i jednocześnie nie są skorelowane między sobą.

W badaniach uwzględniano jedynie dwa warianty cechy C0, tj. spłatę kredytu (sytuacja normalna) i kredyt stracony, ponieważ pozostałe dwa warianty (tj. kredyt pod obserwacją oraz kredyt poniżej standardu) dotyczyły jedynie pojedynczych obiektów. Uznano je zatem za obserwacje nietypowe i usunięto z próby. W celu poprawy efektywności klasyfikacji z dostępnej próby wyodrębniono dwie równoliczne grupy klientów spłacających i niespłacających terminowo kredyt. Następnie próbę podzielono na dwie części. Do estymacji modeli wylosowano 560 obserwacji, pozostałych 278 przypadków stanowiło zbiór testowy, zachowując identyczną strukturę obu podprób.

3. Zastosowane metody klasyfikacji

W badaniach zastosowano trzy grupy metod:

- regresję binarną,

[2] Miary współzależności stosowane dla zmiennych opartych na różnych skalach pomiaru opisane są m.in. w [Walesiak 1996, s. 60].

- bayesowską analizę dyskryminacyjną,
- zagregowane drzewa klasyfikacyjne.

Niech grupująca zmienna zero-jedynkowa będzie objaśniana za pomocą k binarnych zmiennych objaśniających. Wtedy model regresji ma postać[3]:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + \varepsilon_i . \quad (1)$$

Równanie (1) można zapisać jako:

$$y_i = \mathbf{b}^T \mathbf{x}_i + \varepsilon_i, \quad (2)$$

gdzie: y_i – i-ta obserwacja zmiennej grupującej; i = 1, ..., N, (y_i = 1, jeśli kredytobiorca spłaca kredyt, y_i = 0, gdy kredytobiorca nie spłaca kredytu), N – liczba obserwacji, $\mathbf{x}_i = \begin{bmatrix} x_{1i} & ... & x_{ki} \end{bmatrix}^T$ – wektor [k×1] obserwacji dotyczących k zmiennych diagnostycznych (charakterystyk) opisujących i-ty obiekt badania (np. kredytobiorcę), $\mathbf{b} = [\beta_0 \ \beta_1 ... \beta_k]^T$ – wektor [(k+1)×1] parametrów modelu, ε_i – składnik losowy. Uwzględnione w modelu cechy C1-C3 zostały przekształcone w zmienne binarnie dla każdego wariantu cechy (np. wiek od 36 do 45 lat zapisano jako: $x_{21} = 1; x_{22} = x_{23} = 0$).

Podejście bayesowskie umożliwia wykorzystanie informacji *a priori*, dotyczącej rozpatrywanego problemu. W bayesowskiej regule klasyfikacji decyzji o przydzieleniu danej obserwacji do jednej z kłas dokonuje się na podstawie realizacji pewnego (k-wymiarowego) wektora losowego **x**. Innymi słowy, bayesowska reguła klasyfikacyjna polega na przydziale k-wymiarowej obserwacji **x** do klasy C_j, dla której zachodzi[4]:

$$\sum_{j=1}^{K} l_{ij} p_j f(j \mid x) = \min_l \sum_{j=1}^{K} l_{lj} P(j \mid m), \quad (3)$$

gdzie: $f(j|x)$ – gęstość rozkładu wektora losowego **x**, pod warunkiem że występuje j-ty stan (j = 1, ..., m), p_j – prawdopodobieństwo *a priori*, że badana obserwacja należy do klasy C_j, $P(j|m)$ – prawdopodobieństwo *a posteriori* wyznaczane na podstawie wzoru:

$$P(j/m) = \frac{p_j f(x \mid j)}{\sum_{l=1}^{K} p_l f(x \mid l)}, j = 1, ..., m, \quad (4)$$

I_{ij} – funkcja straty, która w zagadnieniach klasyfikacji często przyjmuje wartości binarne i ma postać:

$$l_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}, \quad (5)$$

dla (4) funkcja dyskryminacyjna przyjmuje postać:

[3] Przedstawiony w pracy model jest nazwany zredukowanym modelem regresji binarnej, gdyż nie zawiera zmiennych interakcyjnych oraz iloczynów zmiennych oryginalnych. Więcej informacji znaleźć można w: [Gruszczyński 2001, s. 69]. Dokładny opis modelu regresji binarnej znaleźć można m.in. w pracy [Amundsen 1974].

[4] Na podstawie: [Jajuga 1998, s. 25-29].

$$g_i(x) = \ln\left(f(x|i)\right) + \ln(p_i). \quad (6)$$

Jeśli składowe wektora losowego **x** są niezależnymi zmiennymi losowymi dyskretnymi, z których każda przyjmuje odpowiednio cztery, trzy i dwie wartości (jak w tab. 1), to funkcję dyskryminacyjną $g_j(x)$ można zapisać wzorem:

$$g_j(x) = w_{10}^j + w_{11}^j x_1 + w_{12}^j x_1^2 + w_{13}^j x_1^3 + w_{20}^j + w_{21}^j x_2 + w_{30}^j + w_{31}^j x_3 + w_{32}^j x_3^2 + \ln p_j, \quad (7)$$

gdzie: w_{kl}^j – wartość współczynnika funkcji dyskryminacyjnej wyznaczonej dla *j*-tej klasy, stojącego przy *k*-tej zmiennej dla *l*-tego wariantu cechy.

Tabela 1. Kodowania wybranych zmiennych w analizie dyskryminacyjnej

Kodowanie	Zmienna x_k		
	czas spłaty x_1	wiek x_2	waluta x_3
$x_{k0} = 0$	1-5	do 35 lat	zł
$x_{k1} = 1$	6-10	36-45 lat	pozostałe
$x_{k2} = 2$	11-15	46-55 lat	
$x_{k3} = 3$		56 i więcej lat	

Źródło: opracowanie własne.

Zagregowane modele dyskryminacyjne powstają ze złożenia kilku modeli mających postać drzew klasyfikacyjnych. W literaturze znaleźć można dwa rodzaje agregacji:

- agregację bootstrapową znaną jako *bagging* (por. [Breiman 1996]),
- losowanie adaptacyjne, w literaturze określane jako *boosting* (por. [Freund, Schapire 1997]).

Obie metody opierają się na wielokrotnym losowaniu ze zwracaniem obiektów ze zbioru uczącego *S*. W ten sposób buduje się *N*-elementowe próby uczące U_1 U_2 ... U_v. Na podstawie każdej z tych prób konstruowany jest model dyskryminacyjny w postaci drzewa klasyfikacyjnego D_1, D_2, ...D_v. Opierając się na tych modelach, dokonuje się predykcji zmiennej zależnej *y* dla każdej obserwacji (obiektu **x**) ze zbioru uczącego. Na podstawie tych prognoz dokonuje się ostatecznej klasyfikacji obserwacji, stosując zasadę majoryzacji:

$$D_A(\mathbf{x}) = \arg\max_j \left\{ p\left(D(\mathbf{x}) = j\right) \right\}. \quad (8)$$

Wszystkie zbudowane drzewa łączone są w jeden zagregowany model. Jest to rodzaj „czarnej skrzynki", bo użytkownik nie zna reguł, za pomocą których dokonywana jest klasyfikacja (por. [Gatnar 2002, s. 217]).

4. Wyniki badań

Ocenę jakości klasyfikacji przeprowadzono, opierając się na następujących miernikach, takich jak:

• udział poprawnie rozpoznanych obiektów w każdej klasie:

$$U = \left(1 - \frac{M_j}{N_j}\right) \cdot 100\%, \tag{9}$$

$$N = \sum_{j=1}^{m} N_j; \tag{10}$$

• ogólny błąd klasyfikacji wyrażony wzorem (por. [Witkowska 2003, s. 436]):

$$E = \frac{\sum_{j=1}^{m} M_j}{\sum_{j=1}^{m} N_j} \cdot 100\% = \frac{\sum_{j=1}^{m} M_j}{N} \cdot 100\%; \tag{11}$$

• zliczeniowy R^2

$$R^2 = \frac{N - \sum_{j=1}^{m} M_j}{N} \cdot 100\% = 100 - E; \tag{12}$$

gdzie: M_j – liczba obiektów (kredytobiorców), które zostały błędnie rozpoznane jako należące do grupy j-tej , N_j – liczba obiektów należących do grupy j-tej, (dla $j = 1, 2$); N_1 jest liczbą klientów banku, którzy nie spłacają kredytu, a przez model zostali rozpoznani jako spłacający kredyt, N_2 jest liczbą klientów banku, którzy spłacają kredyt, a przez model zostali rozpoznani jako niespłacający kredytu.

W wyniku zastosowania dyskryminacji bayesowskiej otrzymano dwie funkcje[5]:

$$-g_1(x) = -3{,}29 - 2{,}88x_1 + 0{,}9x_1^2 - 2{,}93x_2 + 1{,}37x_2^2 - 0{,}37x_2^3 - 0{,}01x_3,$$

$$-g_2(x) = -4{,}33 - 1{,}91x_1 + 0{,}34x_1^2 - 1{,}75x_2 + 0{,}54x_2^2 - 0{,}16x_2^3 - 0{,}08x_3.$$

Konkretny obiekt przypisywany jest do tej grupy, dla której uzyskuje wyższą wartość funkcji klasyfikacyjnej. (por. [Jajuga 1999, s. 31]).

W celu budowy modelu regresji binarnej zmienne opisujące wiek, walutę oraz czas spłaty kredytu zakodowano zgodnie z zapisem z tab. 1. W wyniku estymacji KMNK otrzymano model regresji binarnej o postaci:

$$\hat{y}_i = 0{,}197 + 0{,}03x_{21i} + 0{,}01x_{22i} + 0{,}04x_{23i} + 0{,}246x_{3i} + 0{,}443x_{11i} + 0{,}402x_{12i}$$

$$t_\alpha \quad (3{,}97) \quad (0{,}04) \quad (0{,}23) \quad (0{,}06) \quad (2{,}58) \quad (8{,}01) \quad (8{,}81)$$

$$s_e = 0{,}459 \quad R^2 = 0{,}167 \quad \overline{R}^2 = 0{,}159 \quad n = 560.$$

Przyjęto, że i-ty obiekt zostaje przypisany do grupy spłacającej kredyt, jeśli $\hat{y}_i \geq 0{,}5$.

[5] Grupa g_1 oznacza klientów, którzy nie spłacają kredytu (grupę kredytów straconych).

5. Podsumowanie

Porównanie jakości klasyfikacji przeprowadzonej za pomocą poszczególnych metod, odpowiednio dla próby uczącej i testowej, przedstawiono w tab. 2 i 3.

Tabela 2. Ocena jakości klasyfikacji dla próby uczącej

Mierniki poprawnej klasyfikacji	Modele zagregowane		Analiza dyskryminacyjna	Regresja binarna
	bagging	*boosting*		
Udział poprawnie pogrupowanych kredytów niespłaconych U	75%	78,5%	82%	49%
Ogólny błąd klasyfikacji E	22%	15%	32%	28%
Zliczeniowy R^2	78%	85%	68%	72%

Źródło: obliczenia własne.

Biorąc pod uwagę straty ponoszone przez bank z powodu kredytów straconych, najlepszą jakość klasyfikacji uzyskano dzięki zastosowaniu bayesowskiej analizy dyskryminacyjnej. Model poprawnie rozpoznał 82% kredytów niespłaconych. Za pomocą algorytmu *boosting* poprawnie rozpoznanych zostało 78,5% kredytów niespłaconych. Najgorszą klasyfikację, zbliżoną do przypadkowej, uzyskano stosując regresję binarną.

Tabela 3. Ocena jakości klasyfikacji dla próby testowej

Mierniki poprawnej klasyfikacji	Modele zagregowane		Analiza dyskryminacyjna	Regresja binarna
	bagging	*boosting*		
Udział poprawnie pogrupowanych kredytów niespłaconych U	58,9%	65%	81%	53%
Ogólny błąd klasyfikacji E	21%	19%	54%	28%
Zliczeniowy R^2	78%	81%	46%	71%

Źródło: obliczenia własne.

Wyniki klasyfikacji dotyczące próby testowej są nieco gorsze niż w próbie uczącej, tym niemniej skuteczność klasyfikacji uzyskanej dzięki zastosowaniu poszczególnych metod została zachowana. Biorąc pod uwagę minimalizację strat związanych ze złymi kredytami, najlepszą metodą okazała się bayesowska analiza dyskryminacyjna, model bowiem poprawnie rozpoznał 80,6% kredytów niespłaconych. Najlepszą jakość klasyfikacji odnośnie do wszystkich obserwacji w zbiorze testowym uzyskano dla modelu zagregowanego za pomocą algorytmu *boosting*. Uzyskane wyniki są zgodne z prezentowanymi w literaturze (por. [Gatnar 2001, s. 121]). Rozpoznawanie kredytobiorców za pomocą modeli zagregowanych jest obarczone mniejszymi błędami niż w przypadku pojedynczych klasyfikatorów (zob. [Koronacki, Ćwik 2005]).

Literatura

Amundsen H.T. (1974), *Binary Variable Multiple Regressions,* „Scandinavian Journal of Statistic", s. 59-70.

Breuman L. (1996), *Bagging Predictors,* „Machine Learning" 24, s. 123-140.

Freund Y., Schapire R.E. (1997), *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, „Journal of Computer and System Sciences" 55, s. 119-139.

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.

Gatnar E. (2002), *Agregacja modeli dyskryminacyjnych*, [w:] K. Jajuga, M. Walesiak (red.), *Taksonomia 9. Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe AE we Wrocławiu nr 942, AE, Wrocław, s. 217-225.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, Seria Biblioteka Ekonometryczna, Wydawnictwo Naukowe PWN, Warszawa.

Gruszczyński M. (2001), *Modele i prognozy zmiennych jakościowych w finansach i bankowości*, Seria Monografie i Opracowania nr 490, Wydawnictwo SGH, Warszawa.

Jajuga K. (1998), *Metoda analizy dyskryminacyjnej w przypadku zmiennych dyskretnych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 765, *Ekonometria 1*, AE, Wrocław, s. 24-32.

Koronacki J., Ćwik J. (2005), *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, Warszawa.

Samcik M. (2006), *Banki coraz bardziej spragnione naszych lokat*, http://serwisy. gazeta.pl/wyborcza/20290,68586,3558989.html, 2006-08-18.

Walesiak M. (1996), *Metody analizy danych marketingowych*, Wydawnictwo Naukowe PWN, Warszawa.

Witkowska D. (2003), *Uwagi na temat konstrukcji i oceny sztucznych sieci neuronowych wykorzystywanych do dychotomicznej klasyfikacji kredytobiorców*, [w:] J. Lewandowski (red.), *Zarządzanie organizacjami gospodarczymi w przyszłości*, Wydawnictwo Politechniki Łódzkiej, Łódź, s. 430-437.

APPLICATION OF SELECTED METHODS TO THE INDIVIDUAL BORROWERS RECOGNITION

Summary

In the paper the results of application selected methods to the individual credit repayment are presented. These methods are: binary regression, Bayesian discriminant analysis and aggregated classification trees. The experiments are provided for actual data concerning 2547 borrowers. The training set contains 560 selected objects and testing set – 278 cases. The efficiency of employed methods is evaluated in terms of chosen measures.