PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 507 • 2018

Taksonomia 30
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu
e-mail: marcin.pelka@ue.wroc.pl

PODEJŚCIE WIELOMODELOWE ANALIZY DANYCH SYMBOLICZNYCH W OCENIE ZDOLNOŚCI KREDYTOWEJ OSÓB FIZYCZNYCH

ENSEMBLE LEARNING FOR SYMBOLIC DATA IN INDIVIDUAL CREDIT SCORING

DOI: 10.15611/pn.2018.507.20
JEL Classification: C53, C87, C39

Streszczenie: Ustawa prawo bankowe definiuje zdolność kredytową jako zdolność do spłaty zaciągniętego kredytu wraz z odsetkami w terminach określonych w umowie. Analiza i ocena zdolności kredytowej jest kluczowym zagadnieniem z punktu widzenia banku. W ocenie zdolności kredytowej istotne miejsce zajmują szeroko rozumiane metody analizy danych – w tym podejście wielomodelowe. Głównym celem artykułu jest zaprezentowanie zastosowania podejścia wielomodelowego danych symbolicznych w ocenie zdolności kredytowej osób fizycznych (na przykładzie dwóch zbiorów danych rzeczywistych) oraz porównanie wyników otrzymanych z zastosowaniem podejścia wielomodelowego z pojedynczym modelem oraz znaną przynależnością obiektów do klas. Otrzymane wyniki wskazują, że podejście wielomodelowe analizy danych symbolicznych może być użytecznym narzędziem w ocenie zdolności kredytowej osób fizycznych i pozwala otrzymać z reguły lepsze wyniki niż model pojedynczy.

Słowa kluczowe: zdolność kredytowa, analiza danych symbolicznych, podejście wielomodelowe.

Summary: Bank law defines credit ability as the ablitiy to repay credit and interest in terms that are in the credit agreement. Credit scoring is a key term for a bank. It should not be done only at the begininng but during the whole credit period. In credit scoring data analysis methods play a very important role. The main aim of the paper is to present the possibility of applying ensemble learning methods for symbolic data in credit scoring (on the basis of two real data sets). Results obtained from a single model are compared with the ensemble approach. The obtained results show that ensemble learning for symbolic data can be a useful tool for credit scoring and it allows to obtain better results than a single model.

Keywords: credit scoring, symbolic data analysis, ensemble approach.

1. Wstęp

Komisja Nadzoru Finansowego w raporcie o sytuacji banków w pierwszym półroczu 2017 r. wskazuje, że w analizowanym okresie utrzymywało się wysokie tempo wzrostu kredytów konsumpcyjnych (nastąpił przyrost o 5,4 mld zł w porównaniu z analogicznym okresem roku 2016) (zob. [Kotowicz (red.) 2017, s. 8]). Jednocześnie jakość portfela kredytowego nie uległa znacznemu pogorszeniu. Niemniej jednak odnotowano wzrost stanu kredytów zagrożonych o 0,8 mld zł oraz niewielki wzrost ich udziału w portfelu kredytowym. Dodatkowo zaobserwowano także zwiększenie stanu kredytów opóźnionych w spłacie powyżej 30 dni (o 0,7 mld zł), chociaż ich udział w portfelu kredytowym nie uległ znaczącej zmianie (por. [Kotowicz (red.) 2017, s. 8–9]). Warto dodać także, że kredyty konsumpcyjne stanowią ponad 60% zobowiązań polskich gospodarstw domowych [Kolasa 2017, s. 17].

Zgodnie z ustawą Prawo bankowe przez zdolność kredytową rozumie się zdolność do spłaty zaciągniętego kredytu wraz z odsetkami w terminach określonych w umowie [Ustawa z 29 sierpnia 1997, art. 70 pkt 1].

Jedocześnie ustawodawca wskazuje, że to na kredytobiorcy ciąży obowiązek dostarczenia danych i informacji niezbędnych do dokonania oceny jego zdolności kredytowej.

Analiza i ocena zdolności kredytowej jest kluczowym zagadnieniem z punktu widzenia banku. Ocena zdolności kredytowej nie jest dokonywana jednorazowo – wręcz przeciwnie, przez cały czas trwania umowy banki dokonują oceny zdolności kredytowej, aby umieć stwierdzić, czy kredytobiorca będzie w stanie spłacić całą kwotę w terminie zawartym w umowie.

Najczęściej stosowanymi przez banki metodami oceny zdolności kredytowej są: analiza ilościowa, jakościowa oraz punktowa. **Analiza ilościowa** polega na ustaleniu wysokości i stabilności dochodów, które mają zapewnić terminową spłatę kredytu wraz z odsetkami. Natomiast **analiza jakościowa** jest dokonywana na podstawie cech indywidualnych. Ocenie poddawane są m.in. cechy osobowe (wiek, stan cywilny, liczba osób w gospodarstwie domowym, status mieszkaniowy oraz majątkowy, staż pracy, wykształcenie itd.), historia współpracy z bankiem (historia rachunku, korzystanie z innych produktów banku, terminowość spłat innych zobowiązań), ryzyko transakcji kredytowej (kwota kredytu, długość okresu kredytowania, udział własny oraz zabezpieczenia). **Analiza punktowa** (*credit scoring*) polega na punktowej ocenie poszczególnych cech jakościowych i ilościowych kredytobiorcy. Poszczególnym wariantom cech przypisuje się określoną liczbę punktów (np. 10 za wskazanie trzech potencjalnych zabezpieczeń, 5 za dwa zabezpieczenia, 2 za jedno oraz 0 za niskie). W praktyce wykorzystuje się scoring aplikacyjny (użytkowy) oraz scoring behawioralny. **Scoring użytkowy** dotyczy nowych klientów i polega na analizie wniosku kredytowego. Karta scoringu statystycznego jest budowana na podstawie danych dotyczących klientów, którym bank udzielił kredytu w przeszłości. **Scoring behawioralny** dotyczy stałych klientów. Podstawą oceny nie jest wniosek

kredytowy, ale dotychczasowa współpraca klienta z bankiem. Scoring tego typu pozwala m.in. określić nowy limit kredytowy, zmodyfikować istniejący limit kredytu, udostępniać nowe produkty, przedłużać umowy itd.

W ocenie zdolności kredytowej z powodzeniem znajdują zastosowanie metody statystycznej analizy wielowymiarowej. Do najbardziej popularnych metod statystycznych stosowanych w rzeczywistym credit scoringu zalicza się (zob. [Wójciak 2007]):

1. Regresja logistyczna,
2. Regresja typu MARS,
3. Algorytm drzew klasyfikacyjnych,
4. Lasy losowe.

Głównym celem artykułu jest zaprezentowanie zastosowania podejścia wielomodelowego danych symbolicznych z wykorzystaniem drzew klasyfikacyjnych opartych na optymalnym podziale, jądrowej analizy dyskryminacyjnej oraz metody *k*-najbliższych sąsiadów danych symbolicznych w ocenie zdolności kredytowej osób fizycznych (na przykładzie dwóch zbiorów danych rzeczywistych – danych z banku BGŻ S.A. z 2004 r. oraz danych z banków niemieckich) i porównanie wyników otrzymanych z zastosowaniem podejścia wielomodelowego z pojedynczym modelem oraz znaną przynależnością obiektów do klas. Do obliczeń wykorzystano pakiety `symbolicDA`, `pROC` oraz autorskie skrypty w programie R.

2. Podejście wielomodelowe danych symbolicznych

Obiekty symboliczne, w przeciwieństwie do obiektów w ujęciu klasycznym, mogą być opisywane przez następujące rodzaje zmiennych [Bock, Diday (eds.) 2000, s. 2–3; Billard, Diday 2006, s. 7–30; Dudek 2013, s. 35–36; Diday, Noirhomme-Fraiture 2008, s. 10–19] (wizualizację przykładowych obiektów symbolicznych dla zbioru danych z BGŻ S.A. zaprezentowano na rys. 1): zmienne nominalne, porządkowe, przedziałowe oraz ilorazowe, zmienne interwałowe – czyli przedziały liczbowe, zmienne wielowariantowe – czyli listy kategorii lub wartości, zmienne wielowariantowe z wagami – czyli listy kategorii z wagami, zmienne histogramowe – czyli listy wartości z wagami.

Szerzej o obiektach i zmiennych symbolicznych, sposobach otrzymywania zmiennych symbolicznych z baz danych, różnicach i podobieństwach między obiektami symbolicznymi a klasycznymi piszą m.in. [Bock, Diday (eds.) 2000, s. 2–8; Dudek 2013, s. 42–43; 2004; Billard, Diday 2006, s. 7–66; Noirhomme-Fraiture, Brito 2011; Diday, Noirhomme-Fraiture 2008, s. 3–30].

Generalnie podejście wielomodelowe polega na łączeniu (agregacji) M modeli bazowych $D_1,\ldots,D_M$ w jeden model połączony (zagregowany) (por. [Kuncheva 2004]).

Głównym celem zastosowania podejścia wielomodelowego jest zmniejszenie błędu predykcji. Model zagregowany jest bowiem bardziej dokładny niż jakikolwiek z pojedynczych modeli, które go tworzą (zob. [Gatnar 2008, s. 62]). Jedną z bardziej znanych metod agregacji modeli bazowych jest metoda agregacji bootstrapowej, za-

proponowana przez Breimana w 1996 – metoda ta znana jest jako bagging (zob. [Gatnar 2008, s. 140; Breiman 1996, s. 123]). Realizuje on architekturę równoległą modeli zagregowanych.

Metoda bagging polega na utworzeniu modeli bazowych na podstawie prób uczących losowanych ze zwracaniem ze zbioru uczącego (próby te nazywa się próbami bootstrapowymi). Na podstawie każdej z prób budowany jest model z wykorzystaniem odpowiedniej metody. Ostatecznie modele bazowe łączone są za pomocą głosowania większościowego w przypadku zagadnień dyskryminacyjnych lub uśredniania wyników w przypadku analizy regresji (por. [Gatnar 2008, s. 140; Kuncheva 2004, s. 204]).

Z uwagi na fakt, że regresja logistyczna danych symbolicznych pozwala na zastosowanie jedynie zmiennych interwałowych (zob. np. [Pełka 2015]), w części empirycznej wykorzystane zostaną następujące metody analizy danych symbolicznych:

1. Drzewa klasyfikacyjne oparte na optymalnym podziale (DT)[1].
2. Jądrowa analiza dyskryminacyjna (KDA)[1].
3. Metoda *k*-najbliższych sąsiadów (kNN_SDA).

3. Charakterystyka zbiorów danych

W badaniu wykorzystano dwa zbiory danych. Pierwszy z nich zawiera informacje o kredytach konsumpcyjnych udzielonych oraz odrzuconych w 2004 r. przez Bank Gospodarki Żywnościowej S.A. Oddział w Kłodzku. Zbiór danych zawiera 80 obserwacji (obiektów symbolicznych pierwszego rzędu) i jest opisywany przez 14 zmiennych symbolicznych różnego typu (zob. tab. 1).

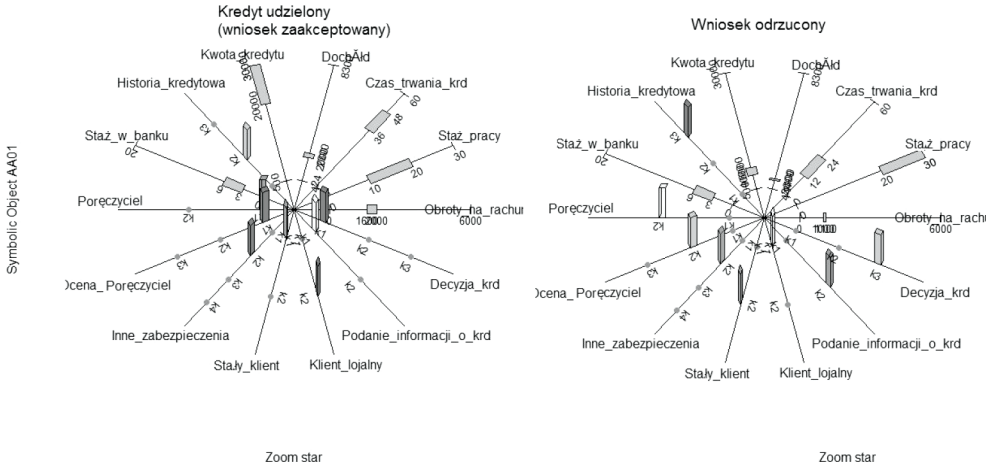
Tabela 1. Charakterystyka zmiennych dla pierwszego zbioru danych

Lp.	Nazwa zmiennej	Typ zmiennej	Lp.	Nazwa zmiennej	Typ zmiennej
1	średnie wpływy na rachunek bieżący	interwałowa	8	wskazanie poręczyciela	wielowariantowa
2	staż pracy kredytobiorcy	interwałowa	9	ocena poręczyciela	wielowariantowa
3	czas trwania kredytu	interwałowa	10	inne proponowane zabezpieczenia	wielowariantowa
4	dochody kredytobiorcy	interwałowa	11	ocena stałości klienta	wielowariantowa
5	wnioskowana kwota kredytu	interwałowa	12	ocena lojalności klienta	wielowariantowa
6	historia kredytowa	wielowariantowa	13	udzielona informacja kredytowa	wielowariantowa
7	staż klienta w banku BGŻ S.A.	interwałowa	14	przynależność do klasy 1 lub 2	nominalna

Źródło: opracowanie własne.

[1] Szerzej o drzewach klasyfikacyjnych opartych na optymalnym podziale oraz jądrowej analizie dyskryminacyjnej piszą m.in. [Gatnar, Walesiak (red.) 2011, s. 280–291; Dudek 2013, s. 143–168].

Na rysunku 1 zawarto dwa przykłady obiektów z klas: wniosek kredytowy zaakceptowany (kredyt udzielony – klasa 1) oraz wniosek kredytowy odrzucony (kredyt nieudzielony – klasa 2). W dalszej części zbiór ten będzie oznaczany jako BGŻ.



Rys. 1. Przykładowe obiekty z klas wniosek kredytowy zaakceptowany (kredyt udzielony) oraz wniosek kredytowy odrzucony (kredyt nieudzielony)

Źródło: opracowanie własne.

Tabela 2. Charakterystyka zmiennych dla drugiego zbioru danych

Lp.	Nazwa zmiennej	Typ zmiennej	Lp.	Nazwa zmiennej	Typ zmiennej
1	przynależność do klasy 1 (spłacony terminowo) lub klasy 2 (spłacony z problemami)	nominalna	10	poręczyciele	wielowariantowa
2	czas trwania kredytu	interwałowa	11	najbardziej wartościowe aktywa	wielowariantowa
3	informacja o innych kredytach	wielowariantowa	12	wiek	interwałowa
4	przeznaczenie	wielowariantowa	13	informacja o innych kredytach	wielowariantowa
5	wysokość kredytu	interwałowa	14	typ własności lokalu	wielowariantowa*
6	oszczędności	interwałowa	15	poprzednie kredyty	wielowariantowa*
7	zatrudnienie	interwałowa	16	charakterystyka zawodu i formy zatrudnienia	wielowariantowa
8	rata jako procent dochodów	interwałowa	17	obcokrajowiec	wielowariantowa*
9	płeć	wielowariantowa*			

* – zmienna wielowariantowa, której realizacją jest tylko jeden z jej wariantów.

Źródło: opracowanie własne.

Dla celów podejścia wielomodelowego zbiór ten podzielono na zbiór uczący o liczebności 53 obiektów oraz testowy o liczebności 27 obiektów.

Drugi zbiór danych zawiera informacje o tysiącu kredytobiorców niemieckich banków. Tablicę danych symbolicznych na potrzeby monografii przygotował dr hab. Andrzej Dudek, prof. UE (zob. [Dudek 2013, s. 162]). Zbiór danych zawiera obiekty pierwszego rzędu, które opisuje 17 zmiennych symbolicznych różnego typu (zob. tab. 2).

Dla celów podejścia wielomodelowego zbiór ten podzielono na zbiór uczący o liczebności 928 obiektów oraz testowy o liczebności 72 obiekty. W dalszej części zbiór ten będzie oznaczany jako GC.

4. Wyniki badań empirycznych

W tabeli 3 zestawiono parametry przyjęte w ramach każdego z podejść.

Tabela 3. Parametry przyjęte w ramach pojedynczych modeli oraz podejścia wielomodelowego

Kryteria	Jeden model			Podejście wielomodelowe		
metoda	DT	KDA	kNN_SDA	DT	KDA	kNN_SDA
parametry – zbiór BGŻ i GC	$n^* = 2$ $W^* = -1e10$	szerokość pasma $h = 1,2$ znormalizowana miara odległości Ichino-Yaguchi (U_3)	znormalizowana miara odległości Ichino-Yaguchi (U_3) liczba sąsiadów 10	$n^* = 2$ $W^* = -1e10$ 50 modeli	szerokość pasma $h \in [0,4; 2]$ różne miary odległości 50 modeli	znormalizowana miara odległości Ichino-Yaguchi liczba sąsiadów 15 50 modeli

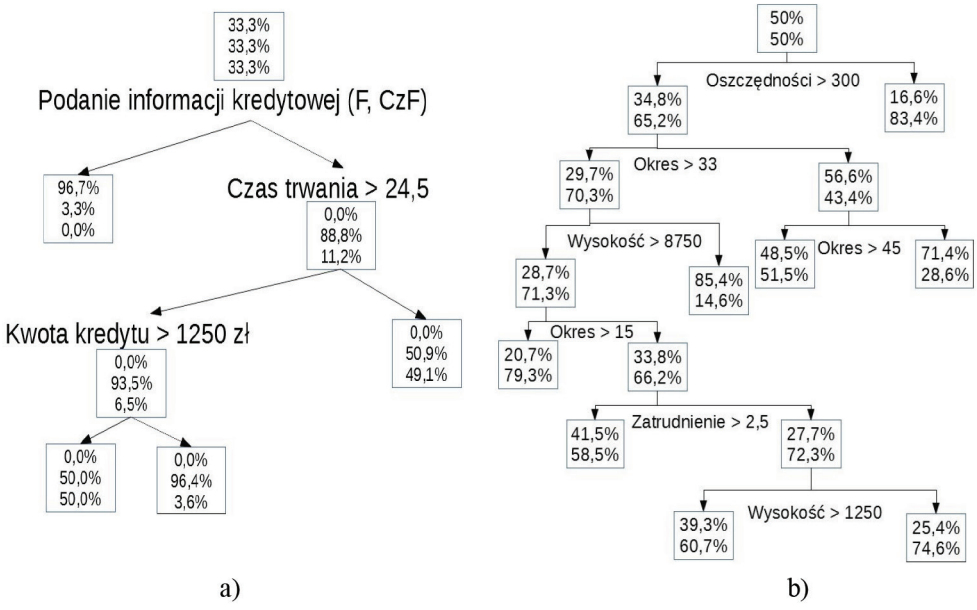
n^* – minimalna liczba obiektów w węźle; W^* – minimalna wartość funkcji-kryterium.

Źródło: opracowanie własne.

Na rysunku 2 w części a) zaprezentowano drzewo klasyfikacyjne otrzymane z zastosowaniem pojedynczego modelu dla zbioru danych BGŻ, a w części b) drzewo klasyfikacyjne otrzymane z zastosowaniem pojedynczego dla zbioru danych GC.

W wypadku pojedynczych drzew klasyfikacyjnych najistotniejszymi zmiennymi, które decydują o przynależności obiektów do klas w zbiorze GC, są oszczędności, okres kredytowania oraz zatrudnienie i wysokość kredytu. Natomiast w zbiorze BGŻ kluczowymi zmiennymi są podanie informacji kredytowej (udzielenie prawdziwej informacji we wniosku skutkowało przydzieleniem kredytu), czas trwania kredytu oraz kwota kredytu.

W tabeli 4 zestawiono błędy otrzymane dla pojedynczych modeli oraz podejścia wielomodelowego danych symbolicznych dla każdego zbioru danych.



Rys. 2. Drzewa klasyfikacyjne dla zbioru BGŻ (część a) oraz zbioru GC (część b)

Źródło: opracowanie własne z wykorzystaniem programu R.

Tabela 4. Błędy klasyfikacji dla każdego z modeli, zbioru danych

Zbiór danych	Model	Metoda		
		KDA	DT	kNN_SDA
BGŻ	pojedynczy	0,15	0,18	0,11
	zagregowany	0,10	0,097	0,098
GC	pojedynczy	0,12	0,069	0,16
	zagregowany	0,11	0,056	0,12

Źródło: opracowanie własne.

5. Podsumowanie

Zarówno pojedyncze modele, jak i podejście wielomodelowe danych symbolicznych mogą z powodzeniem być zastosowane w ocenie zdolności kredytowej osób fizycznych.

Podejście wielomodelowe pozwala zazwyczaj otrzymać dokładniejsze wyniki (obarczone mniejszym błędem) niż pojedyncze modele. W przypadku zbioru danych pochodzącego z banku BGŻ S.A. zarówno pojedyncze modele, jak i modele

zagregowane osiągnęły zbliżone wyniki. W przypadku danych dotyczących banków niemieckich nieco lepsze rezultaty osiągają drzewa klasyfikacyjne oparte na optymalnym podziale.

W wypadku pojedynczych drzew klasyfikacyjnych najistotniejszymi zmiennymi, które decydują o przynależności obiektów do klas w zbiorze danych pochodzących z rynku niemieckiego, są oszczędności, okres kredytowania oraz zatrudnienie i wysokość kredytu. Natomiast w zbiorze z banku BGŻ S.A. kluczowymi zmiennymi są podanie informacji kredytowej (udzielenie prawdziwej informacji we wniosku skutkowało przydzieleniem kredytu), czas trwania kredytu oraz kwota kredytu.

Literatura

Bock H.-H., Diday E. (eds.), 2000, *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from ComplexDdata*, Springer Verlag, Berlin–Heidelberg.

Billard L., Diday E., 2006, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

Breiman L., 1996, *Bagging predictors*, Machine Learning, vol. 24, s. 123–140.

Diday E., Noirhomme-Fraiture M., 2008, *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, Wiley, Chichester.

Dudek A., 2013, *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu, Wrocław.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., Walesiak M. (red.), 2011, *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, C.H. Beck, Warszawa.

Kolasa A., 2017, *Sytuacja finansowa sektora gospodarstw domowych w I kw. 2017*, Departament Analiz Ekonomicznych NBP, Warszawa, http://www.nbp.pl.

Kotowicz A. (red.), 2017, *Raport o sytuacji banków w I półroczu 2017*, Urząd Komisji Nadzoru Finansowego, Warszawa, https://www.knf.gov.pl/publikacje_i_opracowania.

Kuncheva L.I., 2004, *Combining Pattern Classifiers. Methods and Algorithms*, Wiley, New Jersey.

Noirhomme-Fraiture M., Brito P., 2011, *Far beyond the classical data models: Symbolic data analysis*, Statistical Analysis and Data Mining, vol. 4, iss. 2, s. 157–170.

Pełka M., 2015, *Regresja logistyczna dla danych symbolicznych interwałowych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Ekonometria, nr 2(48), s. 44–52.

Ustawa z 29 sierpnia 1997 r. Prawo bankowe, Dz.U. nr 140, poz. 939 ze zm.

Wójciak M., 2007, *Metody oceny ryzyka kredytowego*, PWE, Warszawa.