PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Barbara Pawełek, Józef Pociecha, Jadwiga Kostrzewska, Mateusz Baryła, Artur Lipieta

Uniwersytet Ekonomiczny w Krakowie
e-mails: {barbara.pawelek, jozef.pociecha, jadwiga.kostrzewska, mateusz.baryla, artur.lipieta}@uek.krakow.pl

PROBLEM WARTOŚCI ODSTAJĄCYCH W PROGNOZOWANIU ZAGROŻENIA UPADŁOŚCIĄ PRZEDSIĘBIORSTW (NA PRZYKŁADZIE PRZETWÓRSTWA PRZEMYSŁOWEGO W POLSCE)[1]

PROBLEM OF OUTLIERS IN CORPORATE BANKRUPTCY PREDICTION (CASE OF MANUFACTURING COMPANIES IN POLAND)

DOI: 10.15611/pn.2016.426.15

Streszczenie: W pracach z zakresu prognozowania zagrożenia upadłością przedsiębiorstw można znaleźć rozważania dotyczące problemu wartości odstających. Propozycje rozwiązania tego problemu wahają się od ignorowania go, przez zamianę lub usunięcie wartości odstających, do stosowania metod odpornych na występowanie wartości odstających. W badaniach empirycznych pojawiają się zatem wątpliwości dotyczące wyboru poprawnego podejścia do problemu wartości odstających. Celem artykułu jest przedstawienie wyników badań empirycznych nad przydatnością wybranych metod wykrywania wartości odstających w prognozowaniu zagrożenia upadłością przedsiębiorstw. W badaniu rozważono różne metody wykrywania wartości odstających. Do oceny skuteczności klasyfikacyjnej wybranych metod prognozowania zagrożenia upadłością przedsiębiorstw na podstawie próby testowej wykorzystano mierniki: skuteczności ogólnej, wrażliwości i specyficzności. Badaniem objęte zostały przedsiębiorstwa przetwórstwa przemysłowego w Polsce.

Słowa kluczowe: wartości odstające, zagrożenie upadłością, prognozowanie.

Summary: The results of financial condition analysis are used, among other things, in the research on bankruptcy prediction of companies. The assessment of financial data quality involves also the detection of outliers. In the literature on bankruptcy prediction one can find deliberations on the problem of outliers. The proposals for solving this problem range from not taking any actions, through replacing or removing the outliers, to applying robust methods. Therefore, in the empirical research, some doubts concerning the choice of an

[1] Publikacja została dofinansowana ze środków przyznanych Wydziałowi Zarządzania Uniwersytetu Ekonomicznego w Krakowie, w ramach dotacji na utrzymanie potencjału badawczego.

appropriate approach to the outliers appear. The aim of the article is to present the outcomes of empirical research on the usefulness of selected techniques for identifying outliers in bankruptcy forecasting. In the study, both one-dimensional (Tukey's criterion) and multi-dimensional (projection depth function) procedures of outliers detection were considered. So as to assess the classification accuracy of chosen bankruptcy prediction methods for a test set, total accuracy, sensitivity and specificity measures were used. The analysis was based on data concerning manufacturing companies in Poland.

Keywords: outliers, bankruptcy, forecasting.

1. Wstęp

Przewidywanie pogorszenia się sytuacji finansowej przedsiębiorstwa jest ważnym zagadnieniem w naukach społeczno-ekonomicznych. Metody prognozowania zagrożenia upadłością cieszą się niesłabnącym zainteresowaniem naukowców, praktyków gospodarczych i instytucji finansowych. Podstawą badań są zbiory przedsiębiorstw zawierające jednostki, które ogłosiły upadłość, i obiekty, które nie ogłosiły upadłości w rozważanym okresie. Jedno z dwóch podejść stosowanych w tego typu badaniach polega na zbudowaniu zbioru zbilansowanego z wykorzystaniem metody dobierania obiektów parami lub losowania niezależnego [Baryła, Pawełek, Pociecha 2015]. Drugie podejście w prognozowaniu zagrożenia upadłością przedsiębiorstw bazuje na zbiorach niezbilansowanych. W przypadku tego podejścia, częściej niż w sytuacji analizowania zbiorów zbilansowanych, występuje problem niskiej skuteczności klasyfikacyjnej bankrutów w ramach rozważanych metod prognozowania upadłości przedsiębiorstw[2].

W pracy postawiono tezę, iż na skuteczność klasyfikacyjną bankrutów ma wpływ występowanie obiektów nietypowych wśród przedsiębiorstw zdrowych. Przez nietypowe przedsiębiorstwo zdrowe autorzy rozumieją obiekt o odstających wartościach wskaźników finansowych[3]. Występowanie tego typu obiektów w zbiorach, będących podstawą budowy modeli i reguł klasyfikacyjnych, utrudnia uzyskanie skutecznego narzędzia służącego do przewidywania zagrożenia upadłością przedsiębiorstw. Warto zatem rozważyć problem wartości odstających przy prognozowaniu zagrożenia upadłością przedsiębiorstw na podstawie zbiorów o strukturze bankrutów i nie-bankrutów zbliżonej do występującej w realnej gospodarce[4].

[2] Jako przyczynę tego zjawiska wskazuje się przede wszystkim mały udział bankrutów w badanych zbiorach.

[3] Przedsiębiorstwa tak zdefiniowane mogą charakteryzować się zarówno bardzo dobrą sytuacją finansową, jak i słabą sytuacją finansową, zbliżoną pod względem wielu wskaźników do sytuacji przyszłych bankrutów.

[4] W pracach z zakresu prognozowania zagrożenia upadłością przedsiębiorstw można znaleźć rozważania dotyczące występowania w danych wartości odstających. Propozycje rozwiązania tego problemu wahają się od ignorowania [Spicka 2013], przez zamianę lub usunięcie wartości odstają-

W prezentowanej pracy przyjęto, że obserwacja odstająca to taka, która wydaje się znacznie różnić od innych elementów zbiorowości, w której występuje [Barnett, Lewis 1994]. W literaturze przedmiotu można znaleźć różne klasyfikacje metod wykrywania wartości odstających. Jeden z podziałów rozróżnia metody: jednowymiarowe [Tukey 1977] i wielowymiarowe [Zuo, Serfling 2000].

Celem pracy jest przedstawienie wyników badań empirycznych nad przydatnością wybranych metod wykrywania wartości odstających w prognozowaniu zagrożenia upadłością przedsiębiorstw na podstawie niezbilansowanego zbioru obiektów.

Sformułowano następujące pytania badawcze:

- Czy wykrywanie obiektów nietypowych wśród przedsiębiorstw zdrowych w niezbilansowanym zbiorze obiektów sprzyja poprawie skuteczności klasyfikacyjnej metod prognozowania zagrożenia upadłością przedsiębiorstw?
- Czy wybór między podejściem jednowymiarowym a podejściem wielowymiarowym w wykrywaniu wartości odstających ma wpływ na poprawę skuteczności klasyfikacyjnej rozważanych metod?
- Czy usunięcie nietypowych przedsiębiorstw zdrowych ze zbioru uczącego wpływa na wybór finalnego zbioru wskaźników finansowych w rozważanych metodach?

W literaturze przedmiotu można znaleźć prace dotyczące prognozowania zagrożenia upadłością przedsiębiorstw na podstawie próby niezbilansowanej. W jednej z prac przeglądowych [García, Marqués, Sánchez 2015] po analizie ponad 140 prac z lat 2000–2013 wskazano cechy charakterystyczne tego typu badań5. Zaprezentowane w dalszej części pracy badania empiryczne mają cechy charakterystyczne dla badań publikowanych w rozważanym zakresie. Do przeprowadzenia obliczeń i prezentacji wyników wykorzystano środowisko R oraz programy Stata, Statistica i Excel.

2. Dane i procedura badawcza

W badaniach wykorzystano dwa niezbilansowane zbiory obiektów: zbiór S_1 (służący do prognozowania na rok przed upadłością) oraz zbiór S_2 (służący do prognozowania na dwa lata przed upadłością). Każdy zbiór zawierał 5435 przedsiębiorstw z sektora

cych [Pawełek, Kostrzewska, Lipieta 2015], do stosowania metod odpornych. W badaniach empirycznych pojawiają się zatem wątpliwości dotyczące wyboru poprawnego podejścia do problemu wartości odstających. Wykrywać wartości odstające czy ich nie wykrywać? Jeśli wykrywać, to w jaki sposób i co zrobić z wiedzą o wartościach odstających?

[5] Podstawą rozważań są dane rzeczywiste dotyczące określonych gospodarek państw (65% prac); analizy są oparte na jednej bazie danych (69%); bazy zawierają do 1000 obiektów (54%); stosowany jest podział na zbiór uczący i testowy (35%); podział jest w stosunku 80/20 (w kolejności 70/30); wykorzystuje się następujące miary skuteczności klasyfikacyjnej: skuteczność ogólna (88%), błąd I i II rodzaju (41%), miernik AUC (10%), koszt (5%); nie stosuje się testów statystycznych (68%).

przetwórstwa przemysłowego działających w Polsce. Około 0,9% przedsiębiorstw stanowili bankruci z lat 2007–2010. Każde przedsiębiorstwo było opisane przez 32 wskaźniki finansowe podzielone na grupy wskaźników[6]: płynności (4), zadłużenia (10), rentowności (7) i sprawności działania (11).

Badaniu poddano dwie klasyczne metody prognozowania zagrożenia upadłością przedsiębiorstw, a mianowicie: model logitowy i drzewo klasyfikacyjne[7].

Analizowane zbiory przedsiębiorstw S_1 i S_2 zostały 30 razy losowo podzielone na zbiór uczący i testowy w stosunku 80/20. W otrzymanych 60 parach zbiorów uczących i testowych zachowano stosunek między przedsiębiorstwami zdrowymi a bankrutami, występujący w zbiorze wejściowym.

W celu wskazania nietypowych przedsiębiorstw zdrowych, w badaniu wykorzystano jednowymiarową metodę wykrywania wartości odstających opartą na kryterium Tukeya oraz wielowymiarową metodę opartą na funkcji głębi projekcyjnej.

Procedura oparta na kryterium Tukeya [Tukey 1977] miała następujące etapy:

- Dla każdego wskaźnika finansowego, w każdym zbiorze S_1 lub S_2 i dla każdego podziału na część uczącą i testową, obliczono kwartyle pierwszy i trzeci oraz odchylenie kwartylowe. W analizie wykorzystano wartości wskaźników finansowych przedsiębiorstw zdrowych przydzielonych do zbioru uczącego.
- Za wartości odstające uznano wartości spoza przedziału: $\langle Q_1 - 1{,}5Q,\ Q_3 + 1{,}5Q\rangle$, gdzie Q oznacza odchylenie kwartylowe.
- Przedsiębiorstwo zdrowe uznano za nietypowe, jeżeli przynajmniej jedna z wartości wskaźników finansowych została uznana za odstającą.

Procedura oparta na funkcji głębi projekcyjnej[8] przebiegała następująco:

- Obliczenia z wykorzystaniem funkcji głębi projekcyjnej wykonano osobno dla każdego zbioru S_1 lub S_2 i dla każdego podziału na część uczącą i testową.
- W przypadku zastosowania funkcji głębi za nietypowe przedsiębiorstwa zdrowe uznano 10% spośród wszystkich przedsiębiorstw zdrowych w danym zbiorze uczącym, które leżały najdalej od wielowymiarowego centrum wyznaczonego dla przedsiębiorstw zdrowych w rozważanym zbiorze uczącym.

[6] Dane finansowe dotyczące lat 2005–2009 zostały pobrane z serwisu Emerging Markets Information Service.

[7] Redukcję zbioru wskaźników finansowych prowadzono w modelu logitowym z wykorzystaniem analizy krokowej wstecznej, zaś w przypadku drzewa klasyfikacyjnego zgodnie z algorytmem CART.

[8] Koncepcja głębi danych to zagadnienie nieparametrycznej, odpornej wielowymiarowej analizy statystycznej, rozwijane w ramach eksploracyjnej analizy danych [Kosiorowski 2012]. Umożliwia ona określenie liniowego porządku wielowymiarowych obserwacji z wykorzystaniem wielowymiarowej mediany, definiowanej jako wielowymiarowe centrum zbioru obserwacji [Zuo, Serfling 2000]. Istnieje wiele propozycji funkcji, zwanych funkcjami głębi (np. euklidesowa funkcja głębi, głębia Mahalanobisa, głębia Tukeya, głębia projekcyjna, głębia Studenta), przyporządkowujących każdej obserwacji pochodzącej z pewnego rozkładu dodatnią liczbę będącą miarą jej odstawania od centrum, ze względu na ten rozkład.

Po zastosowaniu opisanych powyżej metod wykrywania wartości odstających zbudowano 120 dodatkowych zbiorów uczących[9].

Do oceny skuteczności klasyfikacyjnej rozważanych metod wykorzystano: sprawność ogólną (procent przedsiębiorstw, które zostały dobrze zaklasyfikowane), wrażliwość (procent bankrutów, którzy zostali dobrze zaklasyfikowani) i specyficzność (procent zdrowych przedsiębiorstw, które zostały dobrze zaklasyfikowane).

3. Wyniki badań empirycznych

W celu uzyskania odpowiedzi na pierwsze pytanie badawcze dokonano porównania skuteczności klasyfikacyjnej rozważanych metod prognozowania zagrożenia upadłością przedsiębiorstw uzyskanej na zbiorze testowym po zbudowaniu modelu lub reguł decyzyjnych na podstawie zbioru uczącego oczyszczonego z nietypowych przedsiębiorstw zdrowych (zbiory: S_1_T, S_1_G, S_2_T, S_2_G) i nieoczyszczonego zbioru uczącego (zbiory: S_1 i S_2). Wyniki obliczeń[10] zamieszczono w tab. 1.

Tabela 1. Porównanie skuteczności klasyfikacyjnej wybranych metod prognostycznych opartych na zbiorze uczącym zawierającym lub niezawierającym wartości odstających

Metoda prognostyczna	Wyprzedzenie upadłości	Relacja	Liczba wystąpień relacji w zbiorze testowym		
			skuteczność ogólna	wrażliwość	specyficzność
Model logitowy	1 rok	$S_1 < S_1\_T$	0	**30**	0
		$S_1 < S_1\_G$	0	**23**	0
	2 lata	$S_2 < S_2\_T$	0	**28**	0
		$S_2 < S_2\_G$	0	**18**	0
Drzewo klasyfikacyjne	1 rok	$S_1 < S_1\_T$	0	10	0
		$S_1 < S_1\_G$	3	2	3
	2 lata	$S_2 < S_2\_T$	**24**	5	**24**
		$S_2 < S_2\_G$	11	9	11

Relacja np. 1: $S_1 < S_1\_T$ oznacza, że wartość miary skuteczności klasyfikacyjnej na zbiorze testowym była większa, gdy metoda bazowała na zbiorze uczącym S_1_T niż w przypadku zbioru S_1. Maksymalna wartość w komórce = 30.

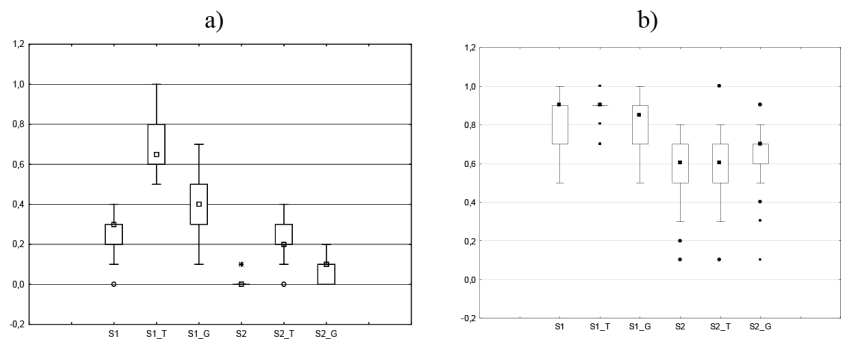
Źródło: obliczenia własne.

[9] Powstało 60 zbiorów uczących po usunięciu nietypowych przedsiębiorstw zdrowych wskazanych przy użyciu metody opartej na kryterium Tukeya (S_1_T i S_2_T) oraz 60 zbiorów uczących po usunięciu nietypowych przedsiębiorstw zdrowych wskazanych przy użyciu metody opartej na funkcji głębi projekcyjnej (S_1_G i S_2_G).

[10] Jeżeli skuteczność klasyfikacyjna metody prognostycznej zbudowanej na podstawie oczyszczonego zbioru uczącego była większa niż w przypadku wykorzystania nieoczyszczonego zbioru uczącego, to metodzie wykrywania wartości odstających dla danego podziału przypisano liczbę 1. W przeciwnym przypadku rozważanej metodzie wykrywania wartości odstających przypisano liczbę 0. Skuteczność klasyfikacyjną oceniano na podstawie wartości trzech mierników. Ze względu na dany miernik skuteczności klasyfikacyjnej rozważana metoda wykrywania wartości odstających, po zsumowaniu liczb przypisanych w kolejnych podziałach zbioru na część uczącą i testową, mogła uzyskać maksymalnie liczbę 30. Wartości większe od 15 (zaznaczone w tab. 1 pogrubioną czcionką) oznaczają, że zastosowanie danej metody wykrywania wartości odstających prowadziło częściej do poprawy skuteczności klasyfikacyjnej, mierzonej określonym miernikiem, niż do pogorszenia analizowanej skuteczności.

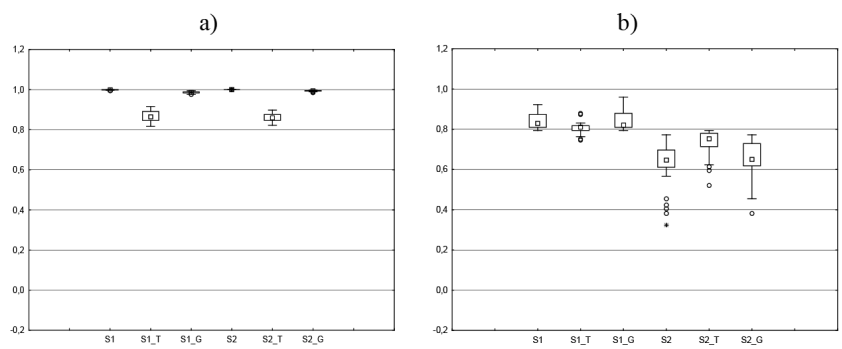
W większości przypadków uzyskano wzrost wartości miernika wrażliwości dla modelu logitowego po zastosowaniu obu analizowanych metod wykrywania wartości odstających. Wyniki uzyskane dla drzewa klasyfikacyjnego potwierdzają panujące przekonanie o odporności tej metody na wartości odstające i przydatności tej techniki w prognozowaniu zagrożenia upadłością przedsiębiorstw w oparciu o niezbilansowane zbiory obiektów. Warto jednak zauważyć, że zastosowanie rozważanej jednowymiarowej metody wykrywania wartości odstających w większości podziałów wpłynęło na poprawę skuteczności klasyfikacyjnej drzewa, mierzonej skutecznością ogólną i miernikiem specyficzności na dwa lata przed upadłością.

Podejmując próbę odpowiedzi na drugie pytanie badawcze sporządzono wykresy ramka-wąsy rozkładów empirycznych rozważanych mierników skuteczności klasyfikacyjnej (rys. 1, 2)[11].



Rys. 1. Miernik wrażliwości a) modelu logitowego i b) drzewa klasyfikacyjnego w zbiorze testowym

Źródło: opracowanie własne.



Rys. 2. Miernik specyficzności a) modelu logitowego i b) drzewa klasyfikacyjnego w zbiorze testowym

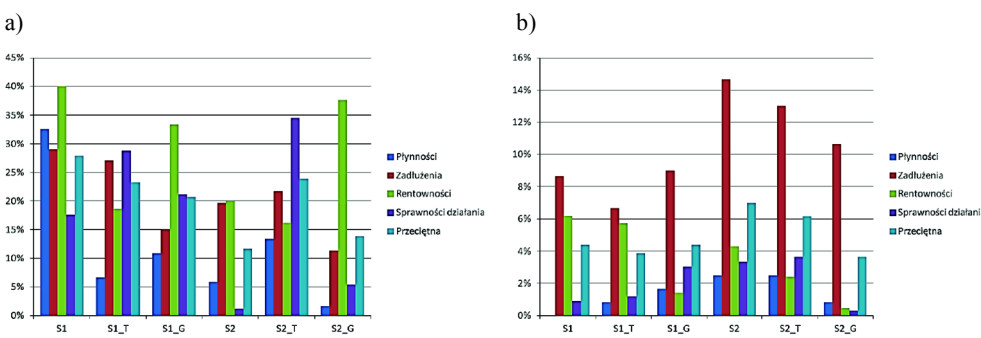
Źródło: opracowanie własne.

[11] W prezentacji wyników pominięto wykresy dla miernika skuteczności ogólnej ze względu na ich podobieństwo do wykresów uzyskanych dla miernika specyficzności.

Analizując rys. 1, można zauważyć, że wartości mediany miernika wrażliwości dla modelu logitowego opartego na oczyszczonym zbiorze uczącym są większe niż wartości mediany tego miernika dla modelu oszacowanego na nieoczyszczonym zbiorze uczącym. Wartości mediany po zastosowaniu jednowymiarowej metody wykrywania wartości odstających są zdecydowanie większe niż po zastosowaniu metody wielowymiarowej. W przypadku drzewa klasyfikacyjnego zastosowanie jednowymiarowej metody wykrywania wartości odstających na zbiorze S_1 nie poprawiło wartości mediany miernika wrażliwości, ale wpłynęło na zmniejszenie zakresu wartości tego miernika. Z kolei zastosowanie funkcji głębi projekcyjnej na zbiorze S_2 wpłynęło zarówno na poprawę wartości mediany miernika wrażliwości, jak i na zmniejszenie zakresu zmienności tego miernika.

Interesujących wniosków dostarcza analiza wykresów sporządzonych dla miernika specyficzności (rys. 2). W przypadku modelu logitowego obserwuje się znaczne pogorszenie skuteczności klasyfikacyjnej przedsiębiorstw zdrowych po zastosowaniu jednowymiarowej metody wykrywania wartości odstających. Natomiast niewielkie pogorszenie wystąpiło po zastosowaniu funkcji głębi projekcyjnej. W przypadku drzewa klasyfikacyjnego można zauważyć zarówno spadek dla zbioru S_1, jak i wzrost dla zbioru S_2 wartości mediany miernika specyficzności po zastosowaniu metody jednowymiarowej. Przyjęcie podejścia wielowymiarowego w wykrywaniu wartości odstających nie wpłynęło znacząco na skuteczność klasyfikacyjną przedsiębiorstw zdrowych.

Odpowiedź na trzecie pytanie badawcze otrzymano po wyznaczeniu stopnia wykorzystania przez grupy wskaźników możliwości pozostania wśród zmiennych objaśniających lub decyzyjnych[12]. Na rysunku 3 przedstawiono wyniki obliczeń (w %) dla poszczególnych grup wskaźników wraz z przeciętną obliczoną dla wskaźników bez podziału na grupy.



Rys. 3. Grupy wskaźników finansowych w a) modelu logitowym i b) drzewie klasyfikacyjnym

Źródło: opracowanie własne.

[12] Rozważano grupy wskaźników zamiast pojedynczych wskaźników z uwagi na występującą na ogół korelację liniową między zmiennymi z danej grupy wskaźników.

Przyglądając się rys. 3a, można zauważyć, że w przypadku zbioru S_1 ważną rolę pełniły wskaźniki z grupy rentowności, płynności i zadłużenia, jeśli model logitowy był szacowany na podstawie nieoczyszczonego zbioru uczącego. Po oczyszczeniu zbioru uczącego metodą jednowymiarową na znaczeniu zyskały wskaźniki z grupy sprawności działania i zadłużenia. W przypadku oszacowania modelu logitowego na podstawie oczyszczonego zbioru uczącego z wykorzystaniem wielowymiarowej metody wykrywania wartości odstających ważną rolę w modelu pełniły wskaźniki rentowności i sprawności działania. W sytuacji rozważania zbioru S_2 znaczny udział wśród zmiennych objaśniających modelu logitowego szacowanego na podstawie nieoczyszczonego zbioru uczącego miały wskaźniki rentowności i zadłużenia. Oczyszczenie zbioru uczącego z nietypowych przedsiębiorstw zdrowych z wykorzystaniem kryterium Tukeya doprowadziło do wzmocnienia roli wskaźników sprawności działania. Po zastosowaniu funkcji głębi projekcyjnej w modelu logitowym główną rolę pełniły wskaźniki rentowności.

W większości przypadków drzewo klasyfikacyjne było oparte głównie na wskaźnikach zadłużenia (rys. 3b). Tylko w przypadku prognozowania zagrożenia upadłością na rok przed upadłością ważną rolę pełniły także wskaźniki rentowności. Po zastosowaniu funkcji głębi projekcyjnej w przypadku zbioru S_1 znaczenie wskaźników rentowności było zdecydowanie mniejsze niż w przypadku drzewa zbudowanego na podstawie nieoczyszczonego zbioru uczącego.

4. Zakończenie

Głównym wnioskiem z przeprowadzonych badań jest stwierdzenie, że wykrywanie i usuwanie nietypowych przedsiębiorstw zdrowych z niezbilansowanych zbiorów uczących może sprzyjać poprawie skuteczności klasyfikacyjnej metod prognozowania zagrożenia upadłością przedsiębiorstw na zbiorach testowych.

Wykrywanie nietypowych przedsiębiorstw zdrowych w niezbilansowanych zbiorach uczących metodami jednowymiarowymi może sprzyjać większej poprawie skuteczności klasyfikacyjnej metod prognozowania zagrożenia upadłością niż metodami wielowymiarowymi, w przypadku miernika wrażliwości dla modelu logitowego oraz miernika specyficzności dla drzewa klasyfikacyjnego na dwa lata przed upadłością. Natomiast wyższość metody wielowymiarowej może ujawnić się w przypadku skuteczności klasyfikacyjnej mierzonej miernikiem wrażliwości dla drzewa klasyfikacyjnego na dwa lata przed upadłością.

Wykrywanie i usuwanie nietypowych przedsiębiorstw zdrowych z niezbilansowanego zbioru uczącego wpływa na wybór finalnego zbioru wskaźników finansowych, służących do prognozowania zagrożenia upadłością przedsiębiorstw.

W dalszych badaniach autorzy planują: włączyć do analizy inne metody wykrywania wartości odstających; w weryfikacji uzyskanych wyników rozważyć inne podejścia, np. V-krzyżowy sprawdzian; zwiększyć liczbę podziałów na część uczą-

cą i testową; powtórzyć badania dla innych podziałów, np. 70/30 i 60/40; objąć badaniami także inne metody prognozowania zagrożenia upadłością przedsiębiorstw.

Literatura

Barnett V., Lewis T., 1994, *Outliers in Statistical Data*, John Wiley & Sons, Chichester.

Baryła M., Pawełek B., Pociecha J., 2015, *Selection of balanced structure samples in corporate bankruptcy prediction*, [w:] W. Adalbert, H. Kestler (red.), *Conference: ECDA Conference 2014 Bremen: Analysis of Large and Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer (in press).

García V., Marqués A.I., Sánchez S.S., 2015, *An insight into the experimental design for credit risk and corporate bankruptcy prediction systems*, Journal of Intelligent Information Systems, vol. 44, no. 1, s. 159–189, DOI 10.1007/s10844-014-0333-4.

Kosiorowski D., 2012, *Statystyczne funkcje głębi w odpornej analizie ekonomicznej*, Seria specjalna: Monografie, nr 208, Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, Kraków.

Pawełek B., Kostrzewska J., Lipieta A., 2015, *The problem of outliers in the research on the financial standing of construction enterprises in Poland*, [w:] M. Papież, S. Śmiech (red.), *Proceedings of the 9th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-economic Phenomena*, Foundation of the Cracow University of Economics, Cracow.

Spicka J., 2013, *The financial condition of the construction companies before bankruptcy*, European Journal of Business and Management, vol. 5, no. 23, s. 160–169.

Tukey J.W., 1977, *Exploratory Data Analysis*, Addison-Wesley, Reading.

Zuo Y. Serfling R., 2000, *General Notions of Statistical Depth Functions*, The Annals of Statistics, vol. 28, no. 2, s. 461–482.