Barometr Regionalny
Nr 3(21) 2010

Modelowanie stanu małych i średnich przedsiębiorstw w warunkach niepewności

Andrzej Burda
Wyższa Szkoła Zarządzania i Administracji w Zamościu

Zdzisław S. Hippe
Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie

Streszczenie: *W artykule przedstawiono nową procedurę uczenia i testowania modeli opracowaną specjalnie dla danych niepewnych. Opiera się ona na kombinacji metody resubstytucji ze zmodyfikowanym paradygmatem uczenia i testowania, zwanym przez nas walidacją kolejkową. Opracowana procedura została sprawdzona na niepewnych danych, przekłamanych prawdopodobnie w procesie kreatywnej i agresywnej księgowości, odnoszących się do małych i średnich przedsiębiorstw pochodzących z regionu podkarpackiego. W trakcie tych badań zweryfikowano modele uczenia w formie drzew i reguł decyzji. Poprawność klasyfikacji obu typów modeli (drzewa i reguły) oszacowano w oparciu o wskaźnik błędu klasyfikacji. Potwierdzono, że błędy fałszywie pozytywnej klasyfikacji są znacznie większe niż klasyfikacji fałszywie negatywnej. Różnice odkryte za pomocą zaproponowanej procedury walidacji mogą być prawdopodobnie wykorzystane jako sygnał występowania przekłamań w danych poddawanych analizie.*

Wstęp

Obecnie o sile gospodarczej krajów wysoko rozwiniętych nie decydują duże koncerny, lecz małe i średnie przedsiębiorstwa, zwane dalej **MSP**, zatrudniające do 250 osób[1]. **MSP** stanowią 99,8% firm funkcjonujących w UE w sektorze pozafinansowym. Są one miejscem pracy dla ponad 67% wszystkich osób zatrudnionych w sektorze prywatnym [Schmiemann 2008]. Ich upadłość w każdym przypadku stanowi bardzo duże zagrożenie bezrobociem na obszarze funkcjonowania, a szczególnie w regionach słabo zurbanizowanych. Z tego względu poszukiwanie niezawodnych i efektywnych metod oceny ich stanu ma duże znaczenie nie tylko dla nich samych, ale spełnia również ważną funkcję społeczną.

1. Cel i zakres badań

Niepewność danych ekonomiczno-finansowych **MSP** wynika z wielu powodów. Jednym z ważniejszych może być umyślne działanie przedsiębiorcy w procesie tzw. kreatywnej księgowości [Nowak 1998, s. 89]. Pojęcie to odnosi się do praktyk księgowych, które co do zasady nie są sprzeczne z obowiązującym prawem, ale na pewno odbiegają od ducha tych zasad. Innymi słowy, termin ten zwykle odnosi się do systematycznego skrywania prawdziwego dochodu, majątku czy też kosztów prowadzenia działalności przedsiębiorstw, korporacji i innych organizacji. Dlatego wydaje się, że skutki kreatywnej księgowości mogą być jedną z najważniejszych przyczyn niezadowalających wyników oceny stanu małych i średnich przedsiębiorstw za pomocą modeli opisanych w dostępnej literaturze [Haider i in. 2007, Pongsatat i in. 2004, Kim i in. 2010], a budowanych na rzeczywistych danych. Potwierdzenie tych spostrzeżeń uzyskaliśmy w naszych wcześniejszych doświadczeniach z użyciem wybranych metod statystycznych i uczenia maszynowego [Burda 2009].

[1] W Polsce kryteria podziału przedsiębiorstw ze względu na ich wielkość reguluje ustawa z 2 lipca 2004 r. o swobodzie działalności gospodarczej (DzU z 2004 r., nr 173, poz. 1808 ze zmianami).

Podkreślić należy, że podobnie jak w wielu krajach, również w Polsce roczne sprawozdanie zgłaszane przez **MSP** służbom skarbowym zawiera jakościową deklarację jego właściciela, czy przedsiębiorstwo jest w stanie przeżycia bądź w stanie upadłości lub likwidacji. Bardziej szczegółowy audyt przedsiębiorstwa jest realizowany tylko raz na cztery lata, więc istnieje pewna luka pozwalająca na kreatywną księgowość. Dlatego w podjętych przez nas badaniach – przy użyciu metod nadzorowanego uczenia maszynowego – poddaliśmy wstępnej ocenie dane **MSP** służące do ich klasyfikacji (*bankrut*, *nie-bankrut*) poprzez specyficzną procedurę walidacji modeli uczonych na danych rzeczywistych. Metodologia walidacji, krótko opisana w następnej części, została sprawdzona na danych dotyczących małych i średnich przedsiębiorstw z regionu podkarpackiego. W trakcie tych badań zweryfikowano modele uczenia w formie drzew i reguł decyzji. Poprawność obu typów modeli (drzewa i reguły) oszacowano w oparciu o wskaźnik błędu klasyfikacji.

1.1. Opis kryteriów procedury walidacyjnej

Najważniejszym kryterium skuteczności metod indukcyjnych jest wskaźnik błędu. Jeśli liczba przypadków jest mniejsza niż 100, do oszacowania błędu modelu uczenia najczęściej ma zastosowanie metoda *leaving-one-out.* W metodzie tej liczba eksperymentów sekwencyjnego uczenia i testowania jest równa liczbie przypadków w zbiorze danych. Podczas *i*-tego eksperymentu *i*-ty przypadek usuwany jest z zestawu danych. Następnie zbiór reguł otrzymany za pomocą pozostałych przypadków zgodnie z zasadą indukcji służy do klasyfikacji pomijanego przypadku, a wynik klasyfikacji jest rejestrowany. Błąd klasyfikacji wyznacza się jako stosunek całkowitej liczby błędnych klasyfikacji do liczby przypadków w zbiorze.

Gdy liczba przypadków w zbiorze danych jest większa lub równa 100, najczęściej stosuje się metodę dziesięciokrotnej walidacji krzyżowej. Technika ta jest podobna do metody *leaving-one-out* i oparta jest na podobnym paradygmacie uczenia i testowania. Procedura rozpoczyna się od losowego sortowania wszystkich przypadków zbioru, a następnie zbiór wszystkich przypadków jest dzielony na 10 wzajemnie rozłącznych podzbiorów, w przybliżeniu równolicznych. Dla każdego *n*-tego podzbioru wszystkie pozostałe przypadki wykorzystywane są do uczenia, tj. do indukcji reguł, a podzbiór ten jest używany do testów. Metoda ta jest powszechnie stosowana ze względu na oszczędność czasu przy jednoczesnym braku istotnego wpływu na dokładność oceny błędu. Jest powszechnie akceptowana jako standardowy sposób walidacji systemów regułowych.

W przypadku dużych zbiorów danych (co najmniej 1000 przypadków) przyjmuje się paradygmat testowania znany pod nazwą *train-and-test*. Technika ta znana jest również jako *holdout*. Dwie trzecie przypadków powinno być wykorzystane do uczenia, a jedna trzecia do testowania. W jeszcze innym sposobie sprawdzania poprawności, resubstytucji, zakłada się, że można przyjąć te same dane do uczenia i testowania modelu. Ogólnie rzecz biorąc, można zauważyć, że oszacowany tą metodą wskaźnik błędu jest zwykle zbyt optymistyczny, jednak technika ta ma częste zastosowanie.

Głównym celem naszej pracy było opracowanie i przetestowanie nowej procedury walidacji niepewnych danych (czyli walidacji niepewnych modeli uczenia), która jest połączeniem techniki resubstytucji (stosowana w celu sprawdzenia *i*-tego modelu z wykorzystaniem danych z tego samego, *i*-tego roku) i walidacji opartej o zmodyfikowany *learn-and-test* paradygmat badań, w którym model uczenia z *i*-tego roku jest kolejno stosowany i oceniany jako model dla roku *i*-tego±1, *i*-tego±2, *i*-tego±3 itp. Modyfikacja paradygmatu badań opiera się w naszym przypadku na niezależnej walidacji (klasyfikacji) na wszystkich zbiorach danych z roku i, $i \pm 1$, $i \pm 2$, $i \pm 3$ itd. Taka procedura oceny została przez nas nazwana walidacją kolejkową (*queue validation*).

1.2. Zbiór badawczy

Obiekty to małe i średnie przedsiębiorstwa z woj. podkarpackiego zebrane w 7 podzbiorach danych dla lat 2000–2006, podzielone na dwie równoliczne kategorie: *bankrut* i *nie-bankrut*. Każdy obiekt został przedstawiony za pomocą 7 atrybutów opisujących:

- udział zapasów w aktywach ogółem,
- udział kapitału obrotowego w finansowaniu majątku ogółem,
- niedobór kapitału obrotowego netto,
- produktywność majątku,
- wynik finansowy brutto,
- stopa zmian sprzedaży,
- stopa zmian zatrudnienia.

W tabeli 1 przedstawiono taksonomię numeryczną badanych danych w kolejnych latach.

Tab. 1. Taksonomia numeryczna danych **MSP** w latach 2000–2006

Nazwa zbioru	Rok	Liczba przypadków		
		Razem	*nie-bankrut*	*bankrut*
SME_2000	2000	132	66	66
SME_2001	2001	150	75	75
SME_2002	2002	144	72	72
SME_2003	2003	130	65	65
SME_2004	2004	128	64	64
SME_2005	2005	132	66	66
SME_2006	2006	132	66	66
Łącznie		948	474	474

1.3. Opis badań

Modele uczenia zostały wygenerowane za pomocą dwóch metod uczenia maszynowego (*machine learning*, **ML**), czyli udoskonalonego algorytmu **ID3/C4.5** do tworzenie *quasi*-optymalnych drzew decyzji [Hippe i in. 2003] i algorytmu **NGTS** pierwotnie opisanego w: [Hippe 1999]. Kilka szczegółów dotyczących jego nowej wersji przedstawiono w sekcji 2.

2. Algorytm NGTS

Algorytm **NGTS** (uproszczony jego schemat pokazany jest na rys. 1) generuje w procesie uczenia zbiór reguł decyzji, począwszy od najbardziej ogólnych do bardziej szczegółowych. Istotą algorytmu jest użycie specyficznej formuły H służącej do walidacji i generowania kolejnych reguł:

[1] $$H = G + sqrt(A),$$

gdzie:
G (ogólność) – łączna liczba przykładów klasyfikowanych poprawnie oraz błędnie, podzielona przez liczbę wszystkich przykładów w tabeli decyzji;
A (dokładność) jest obliczana jako liczba poprawnie sklasyfikowanych przykładów podzielona przez całkowitą liczbę poprawnie i nieprawidłowo klasyfikowanych przykładów.

Algorytm **NGTS** startuje od utworzenia pustego zbioru reguł R. W każdym etapie postępowania pierwszy obiekt ze zbioru U, zwany u_1, jest używany do tworzenia zbioru U' stanów obiektu i zbioru D' decyzji dla tego obiektu. Dodatkowo, tworzony jest pusty zbiór W' do przechowywania akceptowalnych warunków reguł. Tak długo, jak W' nie jest równe U', wykonywane są następujące czynności:
– tworzony jest zbiór S', zawierający wszystkie możliwe kombinacje niewykorzystanych (jak dotąd) i niepustych stanów obiektu – do formalnego opisu stosuje się następujący wzór:

[2] $$S' := P\{U' \setminus W'\} \setminus \{\emptyset\},$$

gdzie P oznacza zbiór potęgowy będący zbiorem wszystkich podzbiorów tego argumentu (w tym zbiór pusty) – dlatego na koniec zbiór pusty należy usunąć;
– dla każdego elementu s'_i ze zbioru S' tworzony jest zbiór potencjalnych warunków reguł zgodnie z formułą:

[3] $$w_i := W' \cup s'_i;$$

następnie zbiór potencjalnych warunków reguł w_i podlega walidacji na podstawie wyliczonych wartości wcześniej opisanych parametrów G, A i H;
– jeżeli dokładność (A) potencjalnych reguł równa się 1, to bieżący zbiór w_i przyjmuje się do utworzenia warunków bieżącej reguły; reguła r jest tworzona i dodana do zbioru R, a w końcu ze zbioru U wszystkie obiekty występujące w nowej regule są usuwane, co opisuje poniższa formuła:

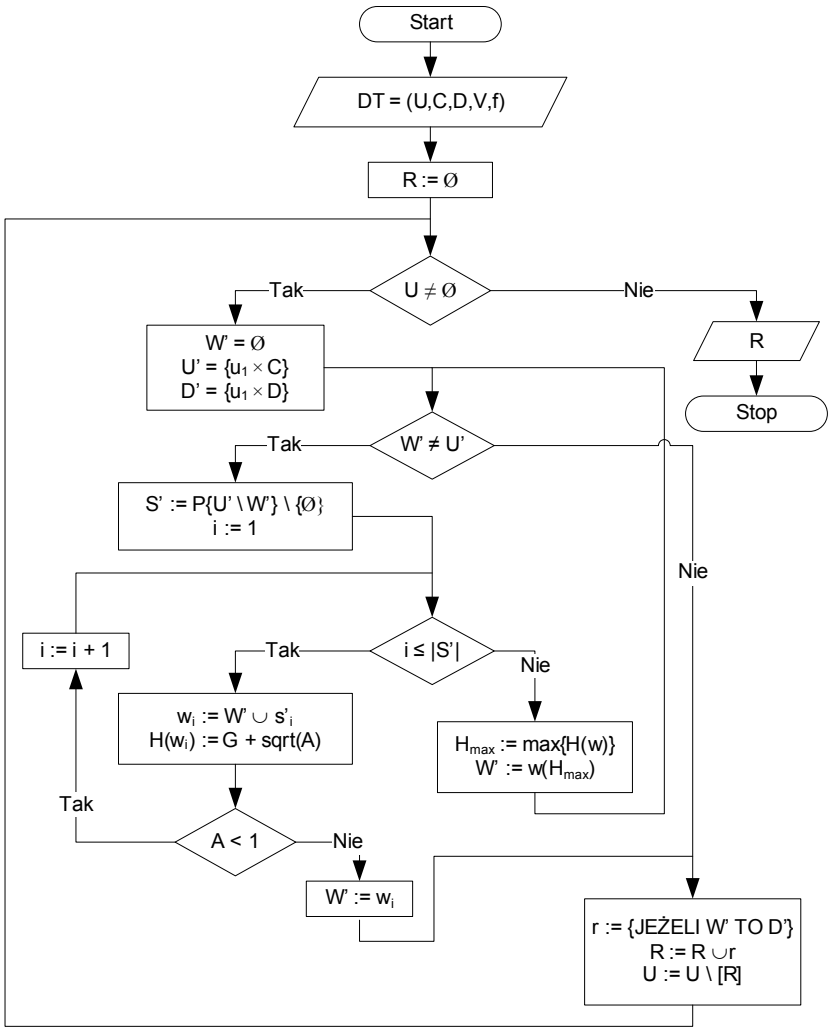
[4] $$U := U \setminus [R];$$

– jeżeli dokładność (A) potencjalnych reguł jest mniejsza od 1, to brany jest pod uwagę następny element $s'_i + 1$;
– jeżeli nie ma już więcej elementów w zbiorze S', to ze wszystkich egzaminowanych zbiorów warunków wybierany jest ten o maksymalnej wartości H do utworzenia nowej reguły, co można zapisać jako:

[5] $$H_{\max} := \max\{H(w)\},$$

[6] $$W' := w(H_{\max}).$$

Jeżeli żaden z testowanych zestawów warunków nie prowadzi do reguł mających $A = 1$, to W' wzrasta do U' i reguły muszą być zbudowane ze wszystkich warunków U', następnie dodawane są do R i wreszcie wszystkie użyte w regule obiekty są usuwane ze zbioru U. Przetwarzanie jest kontynuowane aż do momentu, gdy wejściowy zbiór obiektów U jest pusty. W rezultacie końcowym zwracany jest zbiór reguł R.



Objaśnienia:
DT – tablica decyzji (*Decision Table*), $DT = (U, C, D, V, f)$, gdzie:
U, C, D niepuste zbiory elementów, gdzie U jest zbiorem obiektów (przypadków), C jest zbiorem atrybutów opisujących, D jest zbiorem atrybutów decyzyjnych,
$C, D \subset A$, $C \cup D = A$, $C \cap D = \emptyset$ (A jest skończonym zbiorem atrybutów),
$V = \bigcup_{a \in A} V_a$ (V_a jest dziedziną atrybutu $a \in A$),

$f : U \times A \to V$ jest funkcją informacyjną, taką jak $\forall_{x \in U, a \in A}\, f(x, a) \in V_a$

Rys. 1. Schemat algorytmu **NGTS**

3. Rezultaty eksperymentu

Wyniki eksperymentów przedstawiono w tabeli 2 i 3. Etykiety wierszy oznaczono przez analogię do sposobu uczenia modelu (np. **ID3_2000** – model stworzony za pomocą algorytmu **ID3/C4.5** na podstawie danych zebranych w roku 2000). Kolumny zawierają błędy klasyfikacji badanych modeli na danych zebranych z lat 2000–2006.

Tab. 2. Walidacja niepewnych danych **MSP** (modele **ID3/C4.5**)

Model	Kategoria	Błąd klasyfikacji [%] w roku						
		2000	2001	2002	2003	2004	2005	2006
ID3_2000	*nie-bankrut*	**1,5**	10,7	19,4	12,3	10,9	9,1	9,1
	bankrut	**0,0**	52,0	55,6	53,8	25,0	33,3	27,3
ID3_2001	*nie-bankrut*	3,0	**1,3**	15,3	7,7	10,9	12,1	6,1
	bankrut	54,5	**4,0**	22,2	46,2	25,0	33,3	27,3
ID3_2002	*nie-bankrut*	4,5	8,0	**6,9**	3,1	4,7	3,0	6,1
	bankrut	36,4	44,0	**0,0**	38,5	0,0	16,7	27,3
ID3_2003	*nie-bankrut*	6,1	14,7	19,4	**3,1**	14,1	15,2	15,2
	bankrut	45,5	36,0	44,4	**0,0**	25,0	33,3	27,3
ID3_2004	*nie-bankrut*	4,5	6,7	12,5	3,1	**4,7**	3,0	6,1
	bankrut	36,4	40,0	11,1	30,8	**0,0**	16,7	27,3
ID3_2005	*nie-bankrut*	4,5	2,7	13,9	4,6	3,1	**3,0**	6,1
	bankrut	45,5	56,0	44,4	46,2	18,8	**0,0**	27,3
ID3_2006	*nie-bankrut*	7,6	10,7	20,8	10,8	9,4	10,6	**9,1**
	bankrut	45,5	36,0	44,4	38,5	18,8	16,7	**9,1**

Tab. 3. Walidacja niepewnych danych **MSP** (modele **GTS**)

Model	Kategoria	Błąd klasyfikacji [%] w roku						
		2000	2001	2002	2003	2004	2005	2006
NGTS_2000	*nie-bankrut*	**0,0**	12,0	5,6	6,2	3,1	9,1	4,5
	bankrut	**0,0**	28,0	22,2	30,8	12,5	50,0	36,4
NGTS_2001	*nie-bankrut*	10,6	**0,0**	11,1	9,2	10,9	9,1	9,1
	bankrut	36,4	**0,0**	44,4	30,8	25,0	16,7	27,3
NGTS_2002	*nie-bankrut*	6,1	10,7	**0,0**	9,2	7,8	10,6	13,6
	bankrut	54,5	36,0	**0,0**	46,2	18,8	33,3	45,5
NGTS_2003	*nie-bankrut*	10,6	4,0	13,9	**0,0**	12,5	12,1	10,6
	bankrut	36,4	16,0	22,2	**0,0**	18,8	50,0	45,5
NGTS_2004	*nie-bankrut*	7,6	20,0	6,9	7,7	**0,0**	10,6	9,1
	bankrut	18,2	20,0	11,1	15,4	**0,0**	33,3	27,3
NGTS_2005	*nie-bankrut*	10,6	8,0	9,7	21,5	6,3	**0,0**	13,6
	bankrut	54,5	32,0	55,6	30,8	25,0	**0,0**	9,1
NGTS_2006	*nie-bankrut*	7,6	10,7	16,7	10,8	9,4	7,6	**0,0**
	bankrut	9,1	20,0	44,4	30,8	25,0	50,0	**0,0**

4. Dyskusja wyników

Wszystkie wygenerowane modele uczenia oceniane metodą walidacji kolejkowej charakteryzują się zróżnicowaną wartością błędu prognozy w zależności od kategorii klasyfikowanych przypadków. Zdecydowana większość modeli z małą dokładnością klasyfikuje kategorię *bankrut*. Wartości błędów fałszywie pozytywnej prognozy oscylują w granicach 9,1–56% (z wyłączeniem modelu **ID3_2002** utworzonego na danych zebranych w 2004 r.). Błędy te są 2–5 razy większe niż błędy predykcji fałszywie negatywnej. Błędy prognozowania obiektów kategorii *nie-bankrut* są niskie i stabilne. Ich wartość średnia jest mniejsza niż 10%, a błędy w kolejnych latach nigdy nie przekroczyły 21%.

5. Wnioski

Przypuszczalnie głównym odkryciem przeprowadzonych badań jest to, że rozpoznanie stanu obiektu kategorii *bankrut* w analizie danych **MSP** jest znacznie trudniejsze niż kategorii *nie-bankrut*. Wartości błędów fałszywie pozytywnej klasyfikacji są znacznie większe niż błędów klasyfikacji fałszywie negatywnej. Zakłada się, że te różnice wykryte w procesie walidacji kolejkowej mogą być wykorzystane jako wskazówka wystąpienia nadużyć w procesie opracowania danych **MSP** do sprawozdań składanych do urzędów skarbowych.

W przyszłych badaniach chcemy zastosować inne metody sztucznej inteligencji do tworzenia modeli uczenia, lepiej dopasowane do interpretacji fałszywych danych oraz do wykrywania intencjonalnych przekłamań tych danych.

Podziękowania

Autorzy składają uprzejme podziękowania dr. Markowi Cierpiał-Wolanowi, dyrektorowi Urzędu Statystycznego w Rzeszowie, za udostępnienie danych niezbędnych do przeprowadzenia badań.

Literatura

Burda A (2009): *Wielokryterialna ocena modeli prognozowania stanu ekonomiczno-finansowego małych i średnich przedsiębiorstw*, „Barometr Regionalny", nr 1(15), s. 77–84.

Hippe Z.S. (1999): *Data Mining and Knowledge Discovery In Business: Past, Present, and Future*, [w:] W. Abramowicz, M. Orłowska (red.): *Business Information Systems'99*, Springer-Verlag, London, s. 158–169.

Hippe Z.S., Knap M. (2003): *Badania nad generowaniem pewnych oraz możliwych drzew decyzji*, „Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej, Technologie informacyjne", nr 1, Wyd. Wydziału ETI Politechniki Gdańskiej, Gdańsk, s. 189–194.

Haider S.A., Bukhari A.S. (2007): *Evaluating Financial Sector Firm's Creditworthiness for South-Asian Countries*, „Asian Journal of Information Technology", vol. 6, s. 329–341.

Kim H.S., Sohn S.Y. (2010): *Support Vector Machines for Default Prediction of SMEs Based on Technology Credit*, „European Journal of Operational Research", vol. 201, s. 838–846.

Nowak M. (1998): *Praktyczna ocena kondycji finansowej przedsiębiorstwa,* Fundacja Rozwoju Rachunkowości w Polsce, Warszawa.

Pongsatat S., Ramage J., Lawrence H. (2004): *Bankruptcy Prediction for Large and Small Firms in Asia: A Comparison of Ohlson and Altman*, „Journal of Accounting and Corporate Governance", vol. 2, s. 1–13.

Schmiemann M. (2008): *Enterprises by Size Class-Overview of SMEs in the EU*, http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-08-031/EN/KS-SF-08-031-EN.PDF.