PROGNOZOWANIE BANKRUCTWA ZA POMOCĄ KLASYFIKATORÓW ROZMYTYCH REALIZUJĄCYCH IDEĘ MAKSYMALNEGO MARGINESU

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 13(19)

Damian Gąska

Uniwersytet Ekonomiczny we Wrocławiu

ISSN 1644-6739
e-ISSN 2449-9765

DOI: 10.15611/sps.2015.13.06

Streszczenie: Artykuł poświęcony jest zagadnieniu prognozowania bankructwa. Koncepcją poddaną analizie jest wykorzystanie w tym celu metody klasyfikatorów rozmytych maksymalnego marginesu (*Maximum Margin Fuzzy Classifiers* – MMCF). Praca zawiera zwięzłą charakterystykę podejść stosowanych do prognozowania bankructwa. Przedstawiono najważniejsze teoretyczne aspekty MMFC. W części badawczej zaprezentowano wyniki analizy porównawczej, ukazując MMFC na tle tradycyjnie stosowanych metod. Badanie dotyczyły przedsiębiorstw notowanych na Giełdzie Papierów Wartościowych w Warszawie.

Słowa kluczowe: prognozowanie bankructwa, uczenie pod nadzorem, klasyfikatory rozmyte maksymalnego marginesu.

1. Wstęp

W artykule dokonano analizy klasyfikatorów rozmytych maksymalnego marginesu (ang. *Maximum Margin Fuzzy Classifiers* – MMFC [Abe 2010a]) jako metody prognozowania bankructwa.

Metoda ta łączy w sobie cechy charakterystyczne dla uczenia pod nadzorem i wnioskowania na gruncie logiki rozmytej. Dzieli przy tym pewne własności z metodą wektorów nośnych SVM (ang. *Support Vector Machines*).

Uczenie pod nadzorem obejmuje pewną grupę metod łączącą klasyczne, statystyczne metody klasyfikacji, takie jak liniowa analiza dyskryminacyjna, z algorytmami uczenia opartymi na technikach sztucznej inteligencji. Wykorzystanie tego typu metod ma w prognozowaniu bankructwa długą tradycję.

W części badawczej MMFC zostały porównane z metodami klasyfikacji tradycyjnie stosowanymi w prognozowaniu bankructwa. Dane dotyczyły polskich spółek notowanych na Giełdzie Papierów Wartościowych – GPW.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

2. Metody prognozowania bankructwa

Poniżej krótko scharakteryzowano ideę uczenia pod nadzorem oraz metodę SVM stanowiącą jego szczególny przypadek. Skupiono się przy tym na tych cechach metody, które dziedziczy po niej MMFC.

Uczenie pod nadzorem (ang. *supervised learning*) obejmuje szeroki zakres metod klasyfikacji. W zadaniu prognozowania bankructwa każda obserwacja $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ jest wektorem cech opisujących spółkę. W praktyce są to najczęściej pewne liczbowe charakterystyki analizowanego przedsiębiorstwa, np. wskaźniki finansowe, pochodzące z pewnego (ustalonego) okresu w przeszłości. Y oznacza dychotomiczną etykietę klasy. Umownie przyjmiemy, że $Y = 1$ oznacza spółkę zagrożoną upadłością, zaś $Y = -1$ przedsiębiorstwo o dobrej kondycji finansowej.

Celem uczenia jest wyznaczenie funkcji $d : \mathrm{X} \to \{-1,1\}$, czyli klasyfikatora. $\mathrm{X} \subseteq \mathbb{R}^m$ oznacza przestrzeń, z której pochodzą obserwacje $\mathbf{X}$. Przy pomocy d można wówczas klasyfikować nowe obserwacje. Klasyfikator konstruuje się w oparciu o dane.

Dane wejściowe (uczące) do budowy klasyfikatora tworzą próbę uczącą $\mathrm{L}_n = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)\}$. O istocie uczenia pod nadzorem stanowi obecność zmiennej objaśnianej Y w próbie. W metodach uczenia bez nadzoru (jak np. analiza skupień) zmienna objaśniana nie występuje.

Metodą uczenia pod nadzorem, która zyskała w ostatnich latach dużą popularność w wielu zastosowaniach, jest metoda wektorów nośnych (podpierających), w literaturze znana jako SVM (zob. [Vapnik 1995]).

Pełniejszy opis samej metody i jej wykorzystania w prognozowaniu bankructwa można znaleźć w [Gąska 2013]. Tutaj zostały wyróżnione te cechy metody SVM, które są potrzebne do wprowadzenia MMFC. Są to zasada maksymalizacji marginesu oraz wykorzystanie funkcji jądrowej.

Zasada maksymalizacji marginesu oznacza, że poszukuje się takiej hiperpłaszczyzny separującej klasy, której położenie byłoby optymalne w tym sensie, że jej odległość od najbliższej obserwacji z próby uczącej (tzw. margines) byłaby największa.

W wariancie uogólnionym metody SVM obserwacje przekształca się do przestrzeni o wyższym wymiarze za pomocą pewnego operatora φ i rozwiązuje problem optymalizacyjny w tej nowej przestrzeni.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

Ze względu na to, że problem optymalizacyjny zdefiniowany w metodzie SVM wyraża się w terminach iloczynów skalarnych $< \mathbf{X}_i, \mathbf{X}_j >$, w przestrzeni wyższego wymiaru problem będzie zależał od obserwacji poprzez iloczyny $< \varphi(\mathbf{X}_i), \varphi(\mathbf{X}_j) >$. Twierdzenie Mercera (zob. np. [Krzyśko i in. 2008; Burges 1998]) orzeka, że dla pewnych rodzajów funkcji jądrowych $K(\mathbf{u}, \mathbf{v})$ istnieje operator φ taki, że funkcje te dają się wyrazić jako iloczyny skalarne $< \varphi(\mathbf{u}), \varphi(\mathbf{v}) >$. Pozwala to na rozważanie nieliniowej wersji metody *SVM* bez podawania jawnej postaci przekształcenia φ. Wystarcza bowiem obranie odpowiedniej funkcji jądrowej $K(\cdot, \cdot)$, spełniającej założenia twierdzenia Mercera, i przyjęcie $< \varphi(\mathbf{X}_i), \varphi(\mathbf{X}_j) >:= K(\mathbf{X}_i, \mathbf{X}_j)$. Rozumowanie to znane jest w literaturze (zob. [Abe 2010a; Hastie i in. 2009]) jako „sztuczka z funkcją jądrową" (ang. *kernel trick*).

W praktycznych zastosowaniach stosowane są głównie dwie funkcje jądrowe:
funkcja jądrowa Gaussa:

$$K(\mathbf{u}, \mathbf{v}) = \exp\left(-\tau \| \mathbf{u} - \mathbf{v} \|^2\right), \tag{1}$$

funkcja jądrowa wielomianowa stopnia q:

$$K(\mathbf{u}, \mathbf{v}) = (1 + \mathbf{u}^T \mathbf{v})^q, \; q \in \{1, 2, \ldots\}. \tag{2}$$

3. Klasyfikatory rozmyte maksymalnego marginesu (*MMFC*) i ich związki z SVM

Metoda klasyfikatorów rozmytych maksymalnego marginesu (zob. [Abe 2010b]) z SVM dzieli takie własności jak uczenie pod nadzorem, zasada maksymalizacji marginesu i *kernel trick*, zachowując przy tym możliwość interpretacji w charakterze wnioskowania rozmytego.

Podobnie jak w przypadku SVM, metodę można rozważać w wersji rozszerzonej, odwzorowując nieliniowo przestrzeń obserwacji do przestrzeni o wyższym wymiarze za pomocą pewnego operatora φ. Dla każdej z klas definiuje się wówczas regułę rozmytą (zob. [Abe 2010b; Rutkowski 2005]):

$$\mathrm{R}^{(y)}: \text{JEŻELI } \varphi(\mathbf{X}) \text{ jest } C_y \text{ TO } \mathbf{X} \text{ należy do klasy } Y = y,$$

gdzie $y \in \{-1, 1\}$ i jak dotychczas „1" oznacza bankruta, a „–1" spółkę finansowo zdrową. C_y oznacza wektor średni klasy y:

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

$$C_y = \frac{1}{N(y)} \sum_{i=1}^{N(y)} \varphi(\mathbf{X}_i^{(y)}),$$

gdzie $N(y)$ jest liczbą elementów z klasy y w próbie uczącej, a $\mathbf{X}_i^{(y)}$ i-tym elementem tak określonego podzbioru. Dla każdej z dwu określonych wyżej klas definiuje się jej funkcję przynależności w rozszerzonej przestrzeni $\varphi(\mathrm{X})$:

$$\mu_{\varphi,y}(\mathbf{X}) = \exp(-h^2_{\varphi,y}(\mathbf{X})), \tag{3}$$

gdzie $h^2_{\varphi,y}(\mathbf{X}) = \dfrac{d^2_{\varphi,y}(\mathbf{X})}{\alpha_y}$ oraz

$$d^2_{\varphi,y}(\mathbf{X}) = (\varphi(\mathbf{X}) - C_y)^T Q^+_{\varphi,y} (\varphi(\mathbf{X}) - C_y), \tag{4}$$

zaś $Q^+_{\varphi,y}$ to tzw. uogólniona macierz odwrotna (ang. *pseudo-inverse matrix* – zob. np. [Abe 2010b]) do macierzy kowariancji elementów z klasy $Y = y$ w rozszerzonej przestrzeni.

Dodatnie współczynniki α_y, występujące w (3), pełnią rolę parametrów. Przy $\alpha_y = 1$ funkcja h w (3) jest standardową odległością Mahalanobisa $\varphi(\mathbf{X})$ od C_y.

Na podstawie uzyskanej reguły rozmytej konstruuje się klasyfikator. Nową obserwację przyporządkowuje się do tej z dwu klas, dla której odpowiednia funkcja przynależności (3) ma wyższą wartość. W tym sensie jest to metoda uczenia pod nadzorem, dająca na wyjściu standardową, dwuwartościową funkcję dyskryminacyjną.

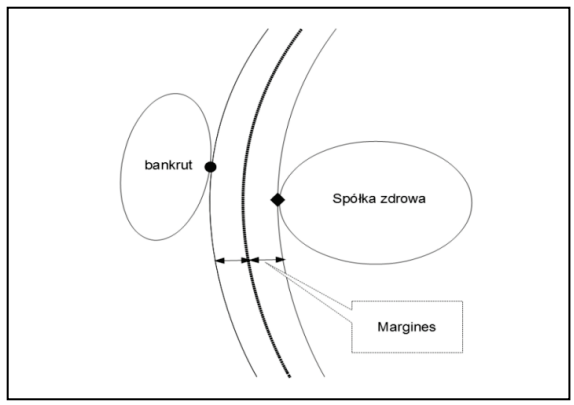
Podobnie jak w klasyfikacji SVM, nie chcemy podawać φ w sposób jawny, ale wykorzystać *kernel trick*. W tym celu należy pokazać, że problem można przeformułować do postaci, w której zależność od obserwacji będzie się wyrażać w terminach ich iloczynów skalarnych. Pozwoli to na podstawienie: $<\varphi(\mathbf{X}_i), \varphi(\mathbf{X}_j)> := K(\mathbf{X}_i, \mathbf{X}_j)$. W [Abe 2010a] pokazano jak – po serii przekształceń – można przeformułować pod tym kątem (4). Ze względu na to, że zapis ostatecznej postaci funkcji staje się skomplikowany, nie został tutaj przytoczony.

Analogicznie jak dla SVM, aby zwiększyć zdolność do prawidłowej klasyfikacji nowych obserwacji, maksymalizuje się margines oddzielający klasy. W tym przypadku odbywa się to poprzez odpowiednie do-

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

pasowanie parametrów α_y, które są odpowiedzialne za nachylenie funkcji przynależności (3). Idea ta została zilustrowana na rys. 1 i 2.



Rys. 1. Maksymalizacja marginesu

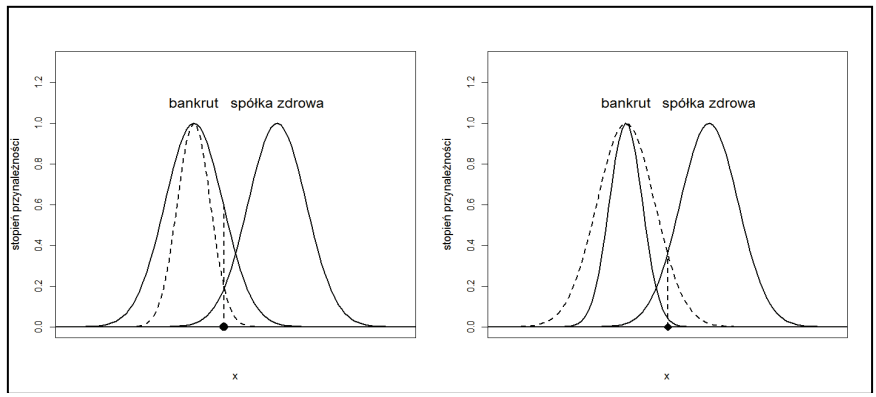
Źródło: opracowanie własne na podstawie [Abe 2010b].

Symbole ● i ♦ oznaczają obserwacje z próby uczącej, kolejno z klasy bankrutów i spółek zdrowych, których odległość od klasy przeciwnej jest możliwie najmniejsza. Granica decyzyjna leży wówczas pomiędzy cieńszymi krzywymi zaznaczonymi na rys. 1.

W lewej części rys. 2 widać, że jeżeli obserwacja ● należy do klasy bankrutów, to jest prawidłowo sklasyfikowana, gdyż wartość funkcji przynależności jest wyższa dla tej klasy. Nie ulegnie to zmianie do momentu, w którym nie zwiększymy nachylenia funkcji przynależności aż do przekroczenia poziomu zaznaczonego krzywą przerywaną.

Przez analogię – obserwacja należąca do klasy spółek zdrowych ♦, zaznaczona w prawej części rys. 2, nie zostanie błędnie sklasyfikowana, dopóki nie zmniejszy się nachylenia funkcji przynależności dla klasy bankrutów poniżej poziomu zaznaczonego krzywą przerywaną. Biorąc pod uwagę wszystkie obserwacje, możemy wyznaczyć dolną i górną granicę dla α_1, w obrębie których nie dojdzie do wzrostu liczby błędnie klasyfikowanych obserwacji.

Maksymalizację marginesu uzyskamy zatem, jeżeli za α_1 przyjmiemy wartość w środku przedziału o krańcach wyznaczonych przez wartości skrajne. Mając wyznaczone α_1, analogiczne rozumowanie przeprowadza się dla $\alpha_{(-1)}$. Pełny opis algorytmu wyznaczania α_y optymalnych pod kątem maksymalizacji marginesu można znaleźć w [Abe 2010b].



Rys. 2. Górna granica dla α_1 (klasa bankrutów), przy której nie dochodzi do błędnej klasyfikacji – lewa część. Dolna granica dla α_1 przy której nie dochodzi do błędniej klasyfikacji – część prawa.

Źródło: opracowanie własne na podstawie [Abe 2010b].

4. Wyniki badań

Celem przeprowadzonych badań było porównanie różnych metod klasyfikacji w zastosowaniu do prognozowania bankructwa w warunkach rynku krajowego, ze szczególnym uwzględnieniem metody MMFC.

Zwróćmy uwagę, że – przy przyjętych założeniach – ostatecznym rozstrzygnięciem w klasyfikacji metodą MMFC jest dychotomiczna decyzja o przynależności obserwacji do jednej z dwóch grup. Jest to zatem decyzja wyrażająca się w klasycznej, zero-jedynkowej logice. Możliwość rozpatrywania wyniku w dziedzinie zbiorów rozmytych (poprzez analizę wartości odpowiednich funkcji przynależności) stanowi pewną dodatkową możliwość interpretacyjną (zob. [Korol 2013; Nogueira 2005; Korol 2011; Devulapalli i Vadlamani 2009]). Z tego powodu MMFC zostały porównane z innymi metodami, w których reguła dyskryminacyjna wyznaczana jest w oparciu o próbę uczącą (są to zatem metody uczenia pod nadzorem). Pełna lista przetestowanych metod przedstawia się następująco:

- klasyfikacja z zastosowaniem modelu regresji logistycznej (zob. np. [Ćwik i Koronacki 2008; Hastie i in. 2009; Krzyśko i in. 2008; Ostasiewicz 2012]),
- las losowy (zob. [Breiman 2001]),
- metoda *k*-najbliższych sąsiadów (ang. *k-Nearest Neightbours*, *K-NN*, np. [Krzyśko i in. 2008; Ćwik i Koronacki 2008; Hastie i in. 2009]),

- SVM z funkcją jądrową Gaussa (1) oraz wielomianową (2),
- MMFC z funkcją jądrową Gaussa oraz wielomianową.

Wykorzystane dane pochodzą z rocznych sprawozdań finansowych spółek notowanych na GPW w latach 2008–2014 (do marca). Podstawą badań była analiza rocznych sprawozdań za rok poprzedzający o dwa lata datę upadłości, co uznaje się za rozsądny kompromis pomiędzy oczekiwaną skutecznością predykcji – z jednej strony, a praktyczną przydatnością klasyfikatora – z drugiej (zob. [Prusak 2005]).

Z analiz wyłączone zostały firmy z sektora finansowego, ze względu na ich odmienną charakterystykę. Ostatecznie wyłoniono 94 spółki, w tym 37, które ogłosiły upadłość we wspomnianym okresie. Próbę konstruowano w ten sposób, by – w miarę możliwości – spółki „zdrowe" były zbliżone do poszczególnych bankrutów pod względem sektora działalności oraz wielkości przedsiębiorstwa.

Jako cechy przyjęto zestaw wskaźników finansowych. Ich wykorzystanie do oceny kondycji spółek i prognozowania bankructwa ma długą tradycję i jest stosowane w tym kontekście od lat 60. ubiegłego wieku (zob. np. [Beaver 1966]). Istotną przesłanką takiego doboru cech była również dostępność danych statystycznych. Zestaw cech przedstawiono w tabeli 1.

Tabela 1. Wybrane wskaźniki finansowe

Wskaźniki rentowności	X_1	zysk brutto ze sprzedaży/aktywa ogółem
	X_2	zysk netto/aktywa ogółem
	X_3	zysk brutto/aktywa ogółem
	X_4	zysk z działalności operacyjnej/ przychody netto ze sprzedaży
Wskaźniki płynności	X_5	aktywa obrotowe bez krótkoterminowych rozliczeń międzyokreswych/zobowiązania krótkoterminowe
	X_6	aktywa obrotowe bez krótkoterminowych rozliczeń międzyokresowych – zapasy/zobowiązania krótkoterminowe
	X_7	kapitał obrotowy/aktywa ogółem
	X_8	inwestycje krótkoterminowe/zobowiązania krótkoterminowe
Wskaźniki zadłużenia	X_9	zobowiązania krótkoterminowe/aktywa ogółem
	X_{10}	zobowiązania ogółem/aktywa ogółem
	X_{11}	kapitał własny/zobowiązania ogółem
	X_{12}	(kapitał własny + zobowiązania długoterminowe)/aktywa trwałe
	X_{13}	(zysk netto + amortyzacja)/zobowiązania ogółem
	X_{14}	zysk brutto/zobowiązania krótkoterminowe
Wskaźniki sprawności	X_{15}	koszty operacyjne(bez pozostałych kosztów operacyjnych)/zobowiązania krótkoterminowe
	X_{16}	przychody ze sprzedaży/suma bilansowa
	X_{17}	przychody ze sprzedaży/należności krótkoterminowe

Źródło: opracowanie własne.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

Zgodnie z przyjętymi założeniami badawczymi, uzyskane wyniki miały umożliwić porównanie różnych metod klasyfikacji w zagadnieniu prognozowania bankructwa. W szczególności oznacza to, że każda metoda powinna wykorzystywać ten sam bazowy zbiór danych.

Chęć rzetelnego porównania zobowiązuje ponadto do przetestowania metod przy różnych wariantach przetwarzania wstępnego danych i/lub wyboru cech, tak aby każda miała szansę być zastosowana w warunkach najbardziej dla niej sprzyjających. Wyciągnięte w ten sposób wnioski będą bowiem bardziej przydatne z punktu widzenia praktycznych użytkowników modeli prognozowania bankructwa. Można się np. spodziewać, że w klasyfikacji metodą najbliższych sąsiadów najlepszą zdolność dyskryminacyjną uzyskamy (zob. [Krzyśko i in. 2008]) po standaryzacji danych. Z kolei w algorytmie lasów losowych zwykle nie ma potrzeby wstępnego przetwarzania lub stosowania wyboru cech. Takie przypuszczenia można formułować na wstępie, jednak nie można mieć tutaj pewności, dlatego skuteczność wszystkich metod sprawdzono przy różnych wariantach przetwarzania wstępnego i/lub wyboru cech. Miało to na celu zagwarantowanie, że wyniki nie będą zaburzone wyborem badacza co do techniki przetwarzania danych optymalnej dla danej metody.

Ostatecznego porównania dokonuje się zatem dla wyników uzyskanych dla każdej metody w najbardziej sprzyjającym dla niej wariancie, mając jednak w pamięci, że oryginalny, bazowy zbiór danych uczących był wspólny dla wszystkich metod.

Tabela 2. Warianty wyboru cech i przetwarzania wstępnego danych

Techniki wyboru cech	
Nazwa/Oznaczenie	Opis
W1	Wybór cech wykorzystujący zestawienie metody *bootstrap* i procedury wstecznej eliminacji (ang. *backward eliminaction*) – szczegóły opisane w dalszej części paragrafu.
W2	Wybór cech najsilniej dyskryminujących na podstawie testu Wilcoxona-Manna-Whitney'a, eliminacja cech skorelowanych liniowo.
Techniki przetwarzania danych	
Standaryzacja	Przekształcenie polegające na odjęciu średniej oraz podzieleniu przez odchylenie standardowe.
PCM	Wykorzystanie metody składowych głównych (PCM, *principal components method*, zob. np. [Hastie i in. 2009; Krzyśko i in. 2008; Ćwik i Koronacki 2008]). Redukcja do składowych tłumaczących 90% zmienności danych.

Źródło: opracowanie własne.

Uwzględniono przy tym pewne kombinacje technik przetwarzania wstępnego z metodami wyboru cech, mając jednocześnie na uwadze, że dana technika wyboru cech może teoretycznie dawać w rezultacie inny ich podzbiór na danych oryginalnych, a inny, jeśli dane podda się wcześniej standaryzacji. Oczywiście nie stosowano wszystkich możliwych kombinacji. Na przykład metoda składowych głównych sama w sobie umożliwia redukcję wymiaru danych i nie łączono jej dodatkowo z wyborem cech. Pełną listę zastosowanych kombinacji wskazano w tabeli 4 z wynikami.

Wybór cech W1 opierał się na wykorzystaniu 100 bootstrapowych replikacji próby uczącej. Następnie dla każdej replikacji wybierany był podzbiór cech metodą krokową eliminacji wstecznej dla modelu logitowego (ang. *backward elimination* zob. [Agresti 2002]).

Procedura wstecznej eliminacji rozpoczyna się od modelu pełnego – zawierającego wszystkie cechy. Następnie sekwencyjnie eliminuje się poszczególne składniki. W kolejnych krokach usuwa się z modelu tę zmienną, której usunięcie ma najmniejszy wpływ na redukcję wartości miary dopasowania modelu (w tym przypadku było to kryterium Akaikego – AIC). Eliminację powtarza się dopóki jakość dopasowania nie spadnie w sposób istotny. Procedurę eliminacji wstecznej przeprowadza się dla każdej replikacji bootstrapowej próby uczącej, uzyskując w ten sposób 100 zestawów cech, przy czym w *b*-tym zestawie znajdują się cechy, które pozostałyby w modelu po wykonaniu wstecznej eliminacji dla *b*-tej próby bootstrapowej.

W kolejnym etapie wybierana była cecha najczęściej występująca w uzyskanych zestawach. Wybór powtarzano, wybierając kolejno cechy, które pozostawały najczęściej, biorąc jednak pod uwagę tylko te replikacje, w których do modelu zakwalifikowano cechy wybrane w poprzednich krokach itd.

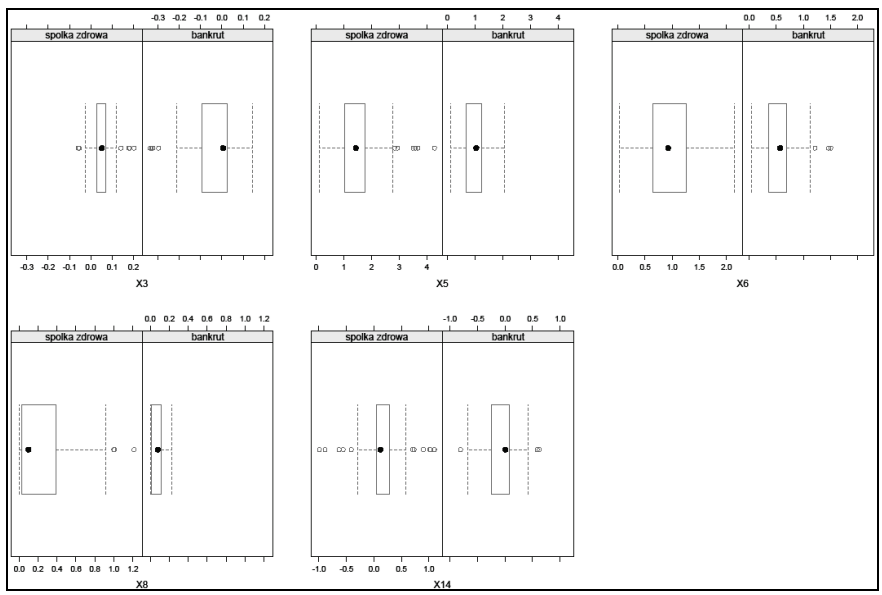
Procedurę kończono, gdy w kolejnej iteracji algorytmu miałaby zostać wybrana cecha, której współczynnik określający frakcję tych replikacji, dla których cecha nie była wykluczana w procedurze *backward elimination*, był poniżej 60% liczby wszystkich replikacji.

W wyniku zastosowania tej metody na pełnym zbiorze danych (procedurę wyboru cech powtarzano w kolejnych przebiegach metody *cross-validation*, zgodnie ze schematem opisanym niżej, dla różnych podzbiorów uczących), wybrano następujące cechy: X_8, X_{12}, X_{14} oraz X_{16}.

W ramach wyboru cech W2 wśród atrybutów, których moduł współczynnika korelacji Pearsona przekraczał wartość 0,7, wybierany był atrybut najsilniej dyskryminujący na podstawie testu Wilcoxona-

-Manna-Whitney'a. Przy czym ograniczono się do cech istotnie dyskryminujących, tj. na poziomie istotności $\alpha = 0,05$ w teście WMW.

W tej metodzie – przy pełnej i nie poddanej przetwarzaniu próbie uczącej – wybrane zostały cechy: X_3, X_5, X_6 oraz X_{14}. Na rys. 3 zilustrowano rozkłady wyselekcjonowanych w ten sposób cech w grupach spółek zdrowych i bankrutów.



Rys. 3. Wykresy pudełkowe dla wybranych cech w grupach spółek zdrowych i bankrutów

Źródło: opracowanie własne.

Przy ocenie wyników klasyfikacji wykorzystana została metoda 5-krotnego sprawdzania krzyżowego (ang. *5-fold cross-validation*, zob. [Kohavi 1995]), nazywana dalej CV. W metodzie tej próba ucząca dzielona jest losowo na pięć możliwie równych części. Cztery z nich służą do konstrukcji klasyfikatora, a pozostała stanowi próbę testową. Procedurę uczenia i testowania powtarza się 5-krotnie przy zmieniającym się podzbiorze testowym, następnie uśrednia się uzyskane wyniki.

Zaletą metody CV jest fakt, że pozwala jednocześnie uwzględnić dwa trudne do pogodzenia wymagania. Po pierwsze do uzyskiwania oszacowań wskaźników skuteczności klasyfikacji wykorzystywane są wszystkie dostępne obserwacje (w przeciwieństwie do metody podzia-

łu próby na dane konstrukcyjne i testowe, gdzie część obserwacji zostaje zupełnie wykluczona z budowy funkcji klasyfikacyjnych).

Po drugie zachowana jest przy tym reguła, że w żadnym momencie funkcja klasyfikacyjna nie jest testowana na obserwacji, która była wykorzystana do jej konstrukcji.

Metoda CV pozwala również na porównywanie wskaźników skuteczności klasyfikacji pomiędzy wynikami uzyskanymi dla różnych wariantów przetwarzania wstępnego danych. Pod warunkiem jednak, że w każdym przypadku zachowana jest ta sama liczba obserwacji, oraz każdej obserwacji w zbiorze przetworzonym można przypisać dokładnie jedną obserwację z oryginalnej próby. Zwróćmy uwagę, że opisane wcześniej techniki przetwarzania wstępnego danych pozwalają na zachowanie tych założeń.

W analizie wykorzystano wariant metody CV z próbkowaniem warstwowym (ang. *stratified sampling*). Oznacza to, że losowe podziały próby były generowane w ten sposób, by, w miarę możliwości, zachować oryginalne proporcje klas.

Zgodnie z zaleceniami literaturowymi (zob. [Hastie i in. 2009]), w przypadku metod obejmujących wieloetapową budowę klasyfikatora, te kroki, w których analizowane są etykiety klasy Y powinny być powtarzane osobno dla każdego podziału danych w metodzie CV. Takimi etapami mogą być np. wybór cech, redukcja wymiaru, czy ustalanie wartości parametrów poprzez kalibrację. Niekiedy wymaga to dodatkowej, wewnętrznej pętli sprawdzania krzyżowego (tzw. *internal cross-validation loop*), osobno dla każdego podziału generowanego w nadrzędnym przebiegu metody.

Zanim dokona się podziałów w metodzie sprawdzania krzyżowego dopuszczalne są jedynie takie etapy konstrukcji klasyfikatora, w których etykiety klas „nie są widoczne", co obejmuje np. normalizację czy standaryzację cech oraz różne techniki uczenia bez nadzoru – np. metodę składowych głównych.

Odstąpienie od tej zasady może prowadzić do błędnych wniosków. Uzasadnienie takiego postępowania, a także przykład nieprawidłowego oraz poprawnego wnioskowania o jakości klasyfikacji w oparciu o metodę *cross-validation* można znaleźć w [Hastie i in. 2009].

Praktyczną stronę przeprowadzonych badań ilustruje tabela 3. W kolumnie drugiej ujęto to, co w danej metodzie stanowi o istocie uczenia pod nadzorem.

Porównując – po kolei dla każdej obserwacji – jej prawdziwą klasę ze wskazaniem klasyfikatora (uzyskanym, gdy obserwacja weszła

Tabela 3. Podsumowanie metod

Metoda	Zasadnicza idea uczenia	Parametry dodatkowe – ustalane z góry lub w wyniku kalibracji (ang. *internal CV*)
Klasyfikacja z wykorzystaniem modelu regresji logistycznej	Estymacja współczynników modelu metodą NW.	
Las losowy	Wyznaczanie podziałów w wierzchołkach budowanych iteracyjnie drzew decyzyjnych.	Liczba konstruowanych drzew klasyfikacyjnych była ustalona na 500, co jest wartością domyślną biblioteki randomForest pakietu statystycznego R.
K-NN	Metoda wymaga „zapamiętania" całej próby uczącej, co umożliwia wyznaczenie najbliższych sąsiadów nowych obserwacji	Liczba k najbliższych sąsiadów. (wyznaczana w wewnętrznej pętli *5-fold CV*)
SVM	Wyznaczenie wartości parametrów modelu w wyniku rozwiązania nieliniowego problemu optymalizacyjnego (zob. np. [Gąska 2013; Burges 1998]).	Parametry funkcji jądrowych (wyznaczane w wewnętrznej pętli *5-fold CV*)
MMFC	Wyznaczenie wektorów średnich oraz macierzy kowariancji koniecznych do obliczenia odległości Mahalanobisa. Dopasowanie parametrów α_y funkcji przynależności (3) stanowiło również integralną część uczenia.	Parametry funkcji jądrowych (wyznaczane w wewnętrznej pętli *5-fold CV*)

Źródło: opracowanie własne.

do podpróby testowej w metodzie *CV*), można wyznaczyć następującą tabelę kontyngencji:

	$\mathbf{Y} = -1$	$\mathbf{Y} = 1$
$d(\mathbf{X}) = -1$	TN	FN
$d(\mathbf{X}) = 1$	FP	TP

gdzie: TN (ang. *True Negatives*) oznacza liczbę spółek zdrowych właściwie sklasyfikowanych, FN (ang. *False Negatives*) – liczbę bankrutów błędnie zaklasyfikowanych do klasy spółek zdrowych, FP (ang. *False Positives*) – liczbę spółek zdrowych zaklasyfikowanych do bankrutów, zaś TP (ang. *True Positives*) – liczbę bankrutów właściwie zaklasyfikowanych.

Dla klasyfikatora d możemy wówczas oszacować dokładność klasyfikacji (*ACC*, *Accuracy*), tzn. wartość prawdopodobieństwa $ACC = P(d(\mathbf{X}) = Y)$, jako

$$\widehat{ACC} = (TP + TN) / (TP + TN + FP + FN) . \quad (5)$$

Ponadto istotnym kryterium w ocenie klasyfikacji powinna być właściwa identyfikacja spółek zagrożonych bankructwem. Przy czym pożądane byłoby – po pierwsze – by klasyfikator właściwie prognozował bankructwo w grupie spółek, które rzeczywiście upadną, co sprowadza się do maksymalizacji prawdopodobieństwa warunkowego $TPR = P(d(\mathbf{X}) = 1 \mid Y = 1)$. TPR oznacza *True Positive Rate* i jest też czasem określane jako „czułość" klasyfikatora (ang. *sensitivity*, *recall*, zob. [Krzyśko i in. 2008]). Naturalnym, częstościowym estymatorem tej wielkości jest $\widehat{TPR} = TP / (FN + TP)$.

Po drugie chcielibyśmy, by wśród spółek zaklasyfikowanych do bankrutów były przedsiębiorstwa faktycznie upadłe. Oznacza to maksymalizację prawdopodobieństwa $PPV = P(Y = 1 \mid d(\mathbf{X}) = 1)$. *PPV* oznacza *Positive Predicitive Value*, co jest też nazywane precyzją (ang. *precision*) klasyfikacji i jest szacowane jako $\widehat{PPV} = TP / (FP + TP)$.

W zaprezentowanych wynikach podajemy ich wspólną miarę, bilansującą przeciwstawne *TPR* i *PPV*. Jest to tzw. *F-score* i jest zdefiniowany jako średnia harmoniczna

$$F_{score} = 2(PPV \cdot TPR) / (PPV + TPR). \quad (6)$$

Wartości F_{score} mieszczą się w przedziale [0, 1], przy czym jakość predykcji jest tym większa im wartości bliższe są 1.

Kolejnym badanym wskaźnikiem był współczynnik korelacji Matthewsa wyrażający się wzorem

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (7)$$

Określa on korelację pomiędzy klasą prognozowaną a obserwowaną. Przyjmuje wartości z przedziału $[-1,1]$, gdzie wartość „– 1" oznacza całkowitą niezgodność, „0" – zbieżność nie lepszą niż losowa, a „1" – identyczność.

W tabeli 4 zaprezentowano wyniki. W zamiarze redukcji wariancji estymatorów uzyskanych metodą *CV* procedurę sprawdzania krzyżowego powtórzono 1000-krotnie, a wartości uśredniono.

W większości przypadków metody wykazały w miarę wysoką zdolność dyskryminacyjną. Spoglądając na wszystkie trzy kryteria jakości klasyfikacji (5), (6) i (7), można zauważyć, że największą sku-

Tabela 4. Skuteczność klasyfikacji

Wariant	ACC	F_{score}	MCC
Dane oryginalne			
Logit	66%	53%	25%
Las losowy	74%	60%	40%
K-NN	61%	17%	3%
SVM z funkcją jądrową Gaussa	76%	63%	45%
SVM z funkcją jądrową wielomianową	71%	54%	33%
MMFC z funkcją jądrową Gaussa	65%	53%	22%
MMFC funkcją z funkcją jądrową wielomianową	69%	33%	28%
Dane standaryzowane			
Logit	66%	53%	24%
Las losowy	72%	58%	38%
K-NN	72%	53%	35%
SVM z funkcją jądrową Gaussa	74%	63%	43%
SVM z funkcją jądrową wielomianową	67%	50%	24%
MMFC z funkcją jądrową Gaussa	72%	62%	40%
MMFC funkcją z funkcją jądrową wielomianową	68%	32%	23%
Wybór cech metodą bootstrap (W1)			
Logit	69%	56%	31%
Las losowy	67%	49%	24%
K-NN	64%	13%	0%
SVM z funkcją jądrową Gaussa	64%	34%	12%
SVM z funkcją jądrową wielomianową	63%	50%	19%
MMFC z funkcją jądrową Gaussa	67%	54%	27%
MMFC funkcją z funkcją jądrową wielomianową	67%	39%	19%
Wybór cech, test Wilcoxona-Manna-Whitney'a (W2)			
Logit	72%	59%	38%
Las losowy	71%	58%	35%
K-NN	70%	44%	28%
SVM z funkcją jądrową Gaussa	69%	44%	26%
SVM z funkcją jądrową wielomianową	70%	48%	29%
MMFC z funkcją jądrową Gaussa	70%	64%	39%
MMFC funkcją z funkcją jądrową wielomianową	69%	59%	33%
Standaryzacja + wybór cech (W2)			
Logit	72%	58%	37%
Las losowy	71%	57%	34%
K-NN	72%	46%	35%
SVM z funkcją jądrową Gaussa	69%	44%	26%
SVM z funkcją jądrową wielomianową	67%	43%	25%
MMFC z funkcją jądrową Gaussa	72%	66%	42%
MMFC funkcją z funkcją jądrową wielomianową	71%	56%	34%
Metoda składowych głównych (PCM)			
Logit	62%	33%	8%
Las losowy	64%	41%	14%
K-NN	62%	16%	2%
SVM z funkcją jądrową Gaussa	65%	38%	14%
SVM z funkcją jądrową wielomianową	67%	50%	24%
MMFC z funkcją jądrową Gaussa	66%	35%	17%
MMFC funkcją z funkcją jądrową wielomianową	64%	42%	14%

Źródło: opracowanie własne.

teczność uzyskano dla metody SVM z funkcją jądrową Gaussa i danych oryginalnych. Wielomianowy wariant SVM okazał się prawie zawsze słabiej dyskryminujący od gaussowskiego odpowiednika.

Model regresji logistycznej dawał dobre wyniki pod warunkiem zastosowania odpowiednich technik wyboru cech.

Zgodnie z oczekiwaniami metoda najbliższych sąsiadów okazała się skuteczna jedynie przy standaryzacji danych. W innych wariantach uzyskiwana korelacja między klasą prognozowaną a prawdziwą (MCC, zgodnie z (7)) była bliska 0.

Najbardziej uniwersalnymi klasyfikatorami okazały się las losowy oraz MMFC z funkcją jądrową Gaussa, dając w miarę dobre wyniki klasyfikacji we wszystkich wariantach przetwarzania danych. Podobnie jak w przypadku SVM wyższe wartości wskaźników skuteczności klasyfikacji dla MMFC uzyskiwano z zastosowaniem gaussowskiej funkcji jądrowej. Metoda MMFC w najlepszym dla siebie wariancie – uwzględniającym standaryzację i eliminację cech skorelowanych – dała wyniki zbliżone do najlepszych, uzyskanych dla SVM i lasu losowego.

Redukcja wymiaru metodą składowych głównych nie przyczyniła się do poprawy wyników klasyfikacji.

W odniesieniu do podanych wcześniej wyników, tam gdzie to było możliwe, podano parametry identyfikujące postać klasyfikatora uzyskaną w wariancie najkorzystniejszym dla danej metody (najpierw wybrano wariant, przy którym osiągnięto najlepsze wyniki, następnie proces uczenia powtórzono dla wszystkich dostępnych danych w najbardziej korzystnym wariancie).

Logit

Najkorzystniejszym wariantem okazał się wybór cech z pomocą testu Wilcoxona-Manna-Whitney'a połączony z eliminacją cech wysoko skorelowanych (ACC = 72%, F_{score} = 59%, MCC = 38%, zgodnie z wzorami (5), (6), (7)).

Ostatecznie model przedstawiał się następująco

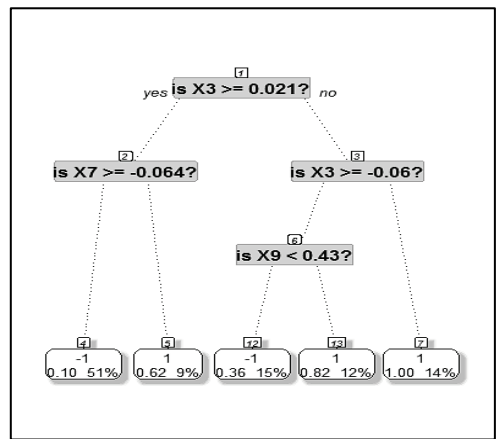
$$P(\mathrm{Y}=1 \mid \mathbf{X}) = \frac{\exp(1{,}56 - 21{,}7X_3 - 0{,}18X_5 - 1{,}37X_6 - 3.54X_8 + 2.8X_{14})}{1 + \exp(1{,}56 - 21{,}7X_3 - 0{,}18X_5 - 1{,}37X_6 - 3.54X_8 + 2.8X_{14})}.$$

Las losowy

Zgodnie z oczekiwaniami metoda sprawdziła się najlepiej na oryginalnych (nieprzetworzonych) danych (ACC = 74%, F_{score} = 60%, MCC = 40%). Ze względu na to, że u jej podstaw stoi „głosowanie"

wielu drzew decyzyjnych, dla których zestaw cech do podziału był dobierany losowo, uzyskany klasyfikator trudno przedstawić w prostej analitycznej postaci.

Interesującym wynikiem z punktu widzenia analityka badającego kondycję finansową spółek jest pojedyncze drzewo zbudowane w oparciu o wszystkie cechy. Zostało ono przedstawione na rys. 4.



Rys. 4. Drzewo decyzyjne. W liściach (węzłach końcowych, zaznaczonych na biało) wskazano dominującą klasę (1 lub –1), podano frakcję spółek zdrowych (po lewej stronie) oraz procent obserwacji z próby trafiającej do liścia (z prawej strony)

Źródło: opracowanie własne.

K-NN

Metoda najbliższych sąsiadów osiąga najlepsze wyniki przy standaryzacji danych (*ACC* = 72%). W wyniku kalibracji optymalną wartość parametru *k* oszacowano na 8 (analizowano zakres od 1 do 25).

SVM

Najwyższą skuteczność klasyfikacji dla tej metody uzyskano na danych oryginalnych w wariancie z funkcją jądrową Gaussa (*ACC* = 76%, F_{score} = 63%, *MCC* = 45%, zgodnie ze wzorami (5), (6), (7)). Uzyskano następującą optymalną postać dla tej funkcji $K(\mathbf{u},\mathbf{v}) = \exp\left(-0,069 \parallel \mathbf{u}-\mathbf{v} \parallel^2\right)$.

MMFC

Najlepszy wynik (*ACC* = 72%, F_{score} = 66%, *MCC* = 42%) uzyskano dla wersji z funkcją jądrową Gaussa postaci $K(\mathbf{u},\mathbf{v}) = \exp\left(-2.29 \parallel \mathbf{u}-\mathbf{v} \parallel^2\right)$.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

Parametry $\alpha_{(-1)}$ i α_1 funkcji przynależności (3) wyniosły w tym wariancie odpowiednio 8,99 i 37,6. Opcją najkorzystniejszą dla tej metody było połączenie standaryzacji i eliminacji cech metodą W2 (zob. tabela 2).

5 Podsumowanie

Zbadane w tej pracy klasyfikatory rozmyte maksymalnego marginesu MMFC cechują się skutecznością predykcyjną porównywalną z metodami uczenia pod nadzorem tradycyjnie wykorzystywanymi w prognozowaniu bankructwa oraz w innych zagadnieniach klasyfikacji.

Ze względu na to, że przedstawione wyniki wpisane są w kontekst konkretnego zbioru danych, stanowiącego podstawę do budowy modeli klasyfikacyjnych, nie pozwalają na jednoznaczne rozstrzygnięcie, czy metoda MMFC może być skutecznym środkiem do predykcji bankructwa dla polskich przedsiębiorstw. Sugerują jednak, że może to być narzędzie potencjalnie konkurencyjne wobec innych technik. Szczególnie w wariancie wykorzystującym funkcję jądrową Gaussa widoczna jest wysoka skuteczność klasyfikacji na tle modeli referencyjnych.

Metoda MMFC może być postrzegana jako wnioskowanie rozmyte, co jest jej pewną zaletą interpretacyjną. Niemniej, zawężając zagadnienie do prognozowania bankructwa, konieczność wykorzystania rozumowań rozmytych może budzić wątpliwości.

Należy przy tym zwrócić uwagę, że zastosowanie wielu metod opartych na statystycznym uczeniu pod nadzorem nie musi ograniczać się do binarnych predykcji, lecz oferuje możliwość estymacji prawdopodobieństwa bankructwa.

Prawdopodobieństwo bankructwa wyraża stopień niepewności co do tego, czy spółka upadnie i – jako takie – jest pożądaną informacją z punktu widzenia użytkowników modeli predykcyjnych. Można uzasadnić stanowisko (zob. [Ostasiewicz 2003]), że teoria zbiorów rozmytych opisuje raczej nieostrość (ang. *vagueness*) niż niepewność (ang. *uncertanity*) i – z tego punktu widzenia – jej stosowalność w prognozowaniu bankructwa jest ograniczona.

Literatura

Abe S., *Kernel-based methods*, [w:] *Support Vector Machines for Pattern Classification*. Springer, 2010a, rozdz. 6.

Abe S., *Maximum-Margin Fuzzy Classifiers*, [w:] *Support Vector Machines for Pattern Classification*. Springer, 2010b, rozdz. 10.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 13(19)

Agresti A., *Categorical Data Analysis*. John Wiley & Sons, 2002.

Beaver H. W., *Financial Ratios As Predictors of Failure*, „Journal of Accounting Research" 1966, s. 71-102..

Breiman L., *Random Forests*, Machine Learning, 2001.

Burges C., *A tutorial on support vector machines for pattern recognition*, Data Mining and Knowledge Discovery, February 1998, s. 121-167.

Ćwik J., Koronacki J., *Statystyczne systemy uczące się*, EXIT, Warszawa 2008.

Devulapalli K.C., Vadlamani R., *Feature selection and fuzzy rule-based classifier applied to bankruptcy prediction in bank*, „Int. J. of Information and Decision Sciences" 2009, nr 1, s. 343-365.

Gąska D., *Zastosowanie metody SVM do oceny ryzyka bankructwa i prognozowania upadłości przedsiębiorstw*, „Śląski Przegląd Statystyczny" 2013, nr 11, s. 289-310.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Kohavi R., *A study of cross-validation and bootstrap for accuracy,* [w:] *14th Intl. Joint Conf. Art. Int.*, 1995, s. 1137-1143.

Korol T., *Multi-Criteria Early Warning System Multi-Criteria Early Warning System*, „International Research Journal of Finance and Economics" 2011.

Korol T., *Nowe podejście do analizy wskaźnikowej w przedsiębiorstwie*, Wolters Kluwer Polska Warszawa, Polska, 2013.

Korol T., *Prognozowanie upadłości firm przy wykorzystaniu miękkich technik obliczeniowych*, Finansowy Kwartalnik Internetowy „e-Finanse" 2010.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M., *Systemy uczące się*, WNT Warszawa 2008.

Nogueira R., Vieira M. S., Sousa M. C. J., *The prediction of bankruptcy using fuzzy classifiers,* [w:] *2005 ICSC Congres*, 2005.

Ostasiewicz W., *Certainty And Uncertainty Versus Precision And Vagueness*, „Badania Operacyjne i Decyzje" 2003, s. 139-148.

Ostasiewicz W., *Myślenie statystyczne*. Wolters Kluwer Polska, 2012.

Prusak B., *Nowoczesne metody prognozowania zagrożenia finansowego przesiębiorstw*, Difin Warszawa 2005.

Rutkowski L., *Metody rozpoznawania wiedzy z wykorzystaniem zbiorów rozmytych typu 1*, [w:] tegoż, *Metody i techniki sztucznej inteligencji*, PWN Warszawa, 2005, rozdz. 4, s. 52-131.

Vapnik V. V., *The Nature of Statistical Learning Theory*, New York, 1995.

BANKRUPTCY PREDICTION WITH MAXIMUM MARGIN FUZZY CLASSIFIERS

Summary: The paper is devoted to the bankruptcy prediction problem. Analyzed concept is the usage of Maximum Margin Fuzzy Classifiers. The article gives a brief overview of approaches used for the purpose of bankruptcy prediction. The most important theoretical aspects of *MMFC* method are presented. The final part contains results and conclusions of a study on real-world data regarding Warsaw Stock Exchange companies.

Keywords: bankruptcy prediction, supervised learning, maximum-margin fuzzy classifiers.