

Zastosowanie systemu regułowego opartego o teorię zbiorów przybliżonych do budowy modelu prognostycznego określającego kondycję ekonomiczno-finansową

Jarosław Sachajko

1. Wstęp

Zastosowanie teorii zbiorów przybliżonych do budowy syntetycznego modelu prognostycznego umożliwiające przewidywanie kondycji ekonomiczno-finansowej przedsiębiorstw wynika z niezaprzeczalnych jej zalet takich jak:

- przekształcanie dużej liczby przypadków uczących na prosty model z uogólnioną wiedzą,
- opracowanie modelu który jest zbiorem łatwych do zrozumienia zasad, w związku z czym nie istnieje potrzeba ich interpretacji,
- każda reguła wynika ze zbioru określonych przypadków uczących,
- brak konieczności posiadania dodatkowych informacji o atrybutach,

powodują liczne implementacje teorii zbiorów przybliżonych do złożonych problemów w medycynie, technice, farmakologii, i ekonomii.

Teoria zbiorów przybliżonych należy do grona metod sztucznej inteligencji. Jej podstawy opracowane zostały w roku 1982 na Uniwersytecie Warszawskim przez profesora Zdzisława Pawlaka.

Porównanie dotychczasowych wyników badań wykorzystujących teorię zbiorów przybliżonych w badaniu kondycji ekonomiczno - finansowej jest bardzo trudne ze względu na znaczne różnice w założeniach badawczych, poczynając od kluczowej w analizie definicji kondycji (bankructwa), poprzez nieporównywalny zestaw danych oraz okres, z jakiego pochodzą, na zestawie atrybutów opisujący przypadki kończąc. Wspólnymi cechami analizowanych artykułów [1, 4, 8, 9] są dobre wyniki

predykcji bankructwa firm, oraz wykorzystanie jako zbioru uczącego i walidacyjnego firm notowanych na giełdach papierów wartościowych. Innowacyjnością w badaniach opisanych w niniejszym artykule jest wykorzystanie danych z grupy małych firm, nie notowanych na giełdzie papierów wartościowych.

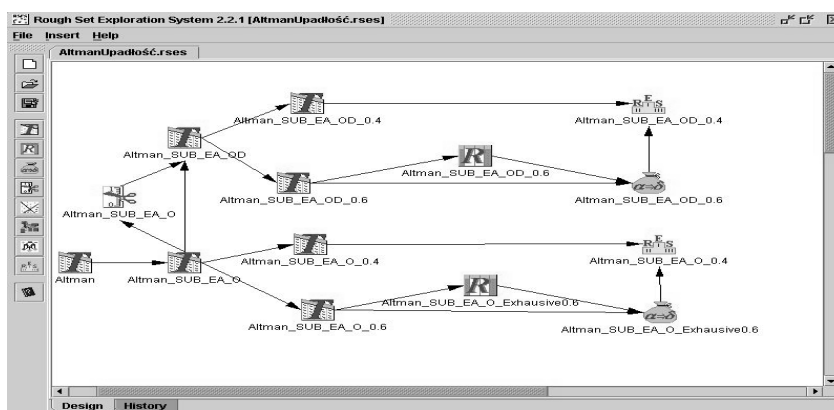
Artykuł ukazuje dobre wyniki stosowania proponowanej metody do budowy systemu prognozującego przyszłą sytuację ekonomiczno - finansową.

Drugi obszar w którym możliwe jest wykorzystanie teorii zbiorów przybliżonych to wspieranie działań doradców ekonomicznych. Dla tej grupy osób przygotowano odrębny, ograniczony liczebnościowo do 11 zbiorów reguł decyzyjnych. Limit liczby reguł wynika z możliwości analitycznych człowieka, który przy dużej liczbie reguł, nie byłby w stanie skorzystać z wiedzy w nich zawartej.

Do badań użyto oprogramowanie RSES (Rough Set Exploration System), stworzone przez grupę badawczą skupioną przy Zakładzie Logiki Instytutu Informatyki i Matematyki Uniwersytetu Warszawskiego, kierowaną przez prof. dr hab. A. Skowrona. Początki prac nad budową oprogramowania sięgają roku 1993, a jego najnowsza wersja 2.2.1 w znacznej części napisana w języku Java, została skompilowana w roku 2005¹. Obecnie system ten wykorzystując teorię zbiorów przybliżonych posiada bardzo szerokie możliwości badawcze takie jak:

- importowanie zestawu danych z posiadanych plików tekstowych,
- wstępną obróbkę danych pozwalającą na uzupełnienie brakujących danych, a następnie ich dyskretyzację,
- budowanie kilku typów klasyfikatorów z późniejszą możliwością oceny ich przydatności w analizie określonego zestawu danych.

Ponadto oprogramowanie to wyróżnia się przyjaznym, intuicyjnym interfejsem graficznym (Rys.1) oraz umożliwia wykonanie wielu eksperymentów w jednym oknie programu, a przez to znacznie upraszcza porównanie poszczególnych modeli.



Rys 1. Okno programu RSES 2.2.1.

¹ Oprogramowanie RSES 2.2.1 jest dostępne w Internecie na stronie <http://logic.mimuw.edu.pl/~rses>. Do celów niekomercyjnych (edukacyjnych, badawczych) oprogramowanie jest dostępne bez konieczności uiszczenia opłaty.

2. Podstawy teorii zbiorów przybliżonych

Reprezentację danych powinny charakteryzować: uniwersalność i efektywność. Takie cechy posiada tablicowy sposób reprezentacji danych. Tablica, której kolumny są etykietowane przez atrybuty, wiersze odpowiadają obiektom, a na przecięciu wiersza i kolumny znajduje się wartość odpowiedniego atrybutu, nosi nazwę tablicy informacyjnej, tablicy typu atrybut-wartość lub systemu informacyjnego [2, 4, 5, 7].

Tablica 1. Zapis systemu informacyjnego uniwersum [2].

$$\mathbf{y} = \begin{matrix} \text{1 rok} \\ \vdots \\ \text{2 rok} \\ \vdots \\ \vdots \\ \text{n rok} \end{matrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \mathbf{X} \mathbf{I} \mathbf{Z} = \begin{bmatrix} x_{101} & x_{102} & \dots & z_{01} & \dots & z_{0m} \\ x_{201} & x_{202} & \dots & z_{01} & \dots & z_{0m} \\ x_{301} & x_{302} & \dots & z_{01} & \dots & z_{0m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i01} & x_{i02} & \dots & z_{01} & \dots & z_{0m} \\ x_{111} & x_{112} & \dots & z_{11} & \dots & z_{1m} \\ x_{211} & x_{212} & \dots & z_{11} & \dots & z_{1m} \\ x_{311} & x_{312} & \dots & z_{11} & \dots & z_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i11} & x_{i12} & \dots & z_{11} & \dots & z_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1n-11} & x_{1n-12} & \dots & z_{n-11} & \dots & z_{n-1m} \\ x_{2n-11} & x_{2n-12} & \dots & z_{n-11} & \dots & z_{n-1m} \\ x_{3n-11} & x_{3n-12} & \dots & z_{n-11} & \dots & z_{n-1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{in-11} & x_{in-12} & \dots & z_{n-11} & \dots & z_{n-1m} \end{bmatrix}$$

U - zbiór przypadków zwany uniwersum,
 x_i - numer przypadku przypisanego chwili i , $i=1, \dots, I$
 q_n - n -ta zmienna charakteryzująca przypadek zwana dalej atrybutem $q_n \in Q$,
 V_{in} - wartość atrybutu q_n w stanie c_i , $V_{in} \in V_{q_n}$
 V_q - dziedzina atrybutu q .

Systemem informacyjnym nazywamy uporządkowaną czwórkę złożoną z niepustego skończonego zbioru zwanego uniwersum (U), niepustego skończonego zbioru atrybutów (Q), sumy dziedzin przyjmowanych przez atrybuty (V) oraz funkcji informacyjnej (f) odwzorowującej iloczyn kartezjański zbioru uniwersum (U) i zbioru atrybutów (Q) w sumę dziedzin atrybutów (V)

$$(1) \quad SI = \langle U, Q, V, f \rangle$$

Niech (1) będzie systemem informacyjnym oraz $P \subseteq Q$. Elementy $x, y \in U$ nazywamy P -nierozróżnialnymi w SI wtedy i tylko wtedy, gdy wartości funkcji f_x dla dowolnego atrybutu p i ustalonego x jest równa wartości funkcji f_y dla dowolnego atrybutu p i ustalonego y ze zbioru uniwersum dla każdego $p \in P$.

$$(2) \quad f_x(p) = f_y(p)$$

Dla dowolnego, niepustego $P \subseteq Q$ relacją P -nierozróżnialności w SI nazywamy relację \tilde{P} określoną w następujący sposób: dla każdego $x, y \in U$ $x \tilde{P} y$ wtedy i tylko wtedy, gdy $f_x(p) = f_y(p)$ dla każdego $p \in P$. Dla elementu $x \in U$ i relacji równoważności $\tilde{P} \subseteq U \times U$ odpowiednią klasę abstrakcji oznaczymy przez $[x]_{\tilde{P}}$, a ich rodzinę, zwaną ilorazem zbioru U przez relację \tilde{P} , oznaczymy przez P^* .

Niech będzie dana przestrzeń aproksymacji S składająca się z uporządkowanej pary uniwersum oraz relacji Q -nierozróżnialności

$$(3) \quad S = \langle U, Q \rangle$$

i dowolny zbiór $X \subseteq U$.

Klasy abstrakcji relacji Q będziemy nazywać Q -elementarnymi zbiorami w tej przestrzeni. W przypadku, gdy relacja nierozróżnialności Q jest ustalona, mówimy o zbiorach elementarnych (atomowych) w S .

Każdą, skończoną sumę zbiorów Q -elementarnych nazywamy zbiorem Q -dokładnym (Q -definiowalnym) w S . Każdy zbiór, nie będący sumą zbiorów Q -elementarnych nazywamy zbiorem Q -przybliżonym w S .

Q -dolną aproksymacją X w przestrzeni aproksymacji S nazywamy zbiór złożony ze wszystkich elementów x należących do zbioru uniwersum U takich że klasa abstrakcji $[X]_Q$ dla elementu x i relacji Q -nierozróżnialności zawiera się w zbiorze X :

$$(4) \quad \underline{Q}X = \{x \in U : [x]_Q \subseteq X\}$$

Q -górną aproksymacją X w przestrzeni aproksymacji S nazywamy zbiór złożony ze wszystkich elementów x ze zbioru uniwersum U takich, że część wspólna klasy abstrakcji $[X]_Q$ dla elementu x i relacji Q -nierozróżnialności Q i zbioru X nie jest zbiorem pustym:

$$(5) \quad \overline{Q}X = \{x \in U : [x]_Q \cap X \neq \emptyset\}$$

Dla dowolnego $X \subseteq U$, X jest Q -dokładny wtedy i tylko wtedy, gdy górna aproksymacja równa się dolnej aproksymacji $\underline{Q}X = \overline{Q}X$ natomiast zbiór X jest przybliżony, gdy $\underline{Q}X \neq \overline{Q}X$.

Q -dokładność aproksymacji X w przestrzeni aproksymacji S jest równa ilorazowi liczby elementów Q -dolnej aproksymacji X w S przez liczbę elementów Q -górną aproksymacji X w S .

$$(6) \quad \mu_Q(X) = \frac{\text{card}(\underline{Q}X)}{\text{card}(\overline{Q}X)}$$

nazywa się Q -dokładnością aproksymacji X w przestrzeni aproksymacji S . Oczywiście $0 \leq \mu_Q(X) \leq 1$ to $\mu_Q(X) = 1$

Niech

$$(7) \quad F = \{X_1, X_2, \dots, X_n\}$$

będzie rodziną podzbiorów U

Q -dolną aproksymacją F w S nazywamy zbiór złożony z Q -dolnych aproksymacji zbiorów X_1, X_2 do X_n

$$(8) \quad \underline{Q}F = \{\underline{Q}X_1, \underline{Q}X_2, \dots, \underline{Q}X_n\}$$

\overline{Q} -górną aproksymacją F w S nazywamy zbiór złożony z Q -górných aproksymacji zbiorów X_1, X_2 do X_n

$$(9) \quad \overline{Q}F = \{\overline{Q}X_1, \overline{Q}X_2, \dots, \overline{Q}X_n\}$$

Zbiór $Pos_Q(F) = \bigcup_{X_i \in F} \underline{Q}X_i$ - Q -pozytywny obszar

rodziny F nazywany sumę Q -dolnych aproksymacji zbiorów X_i z rodziny F .

Zbiór $Bn_Q(F) = \bigcup_{X_i \in F} Bn_Q X_i$ - Q -brzegowy obszar rodziny F nazywamy sumę Q -brzegowych obszarów zbioru X_i z rodziny F .

Zbiór $Neg_Q(F) = U - \bigcup_{X_i \in F} \overline{Q}X_i$ - Q -negatywny obszar rodziny F nazywamy różnicę uniwersum U i sumy wszystkich Q -górných aproksymacji zbiorów X_i z rodziny F .

Liczba

$$(10) \quad \gamma_Q(F) = \frac{\text{card}(Pos_Q(F))}{\text{card}(U)}$$

Q -jakość aproksymacji rodziny zbiorów F nazywamy iloraz liczby elementów Q -pozytywnego obszaru rodziny F przez liczbę elementów zbioru uniwersum U , natomiast

$$(11) \quad \beta_Q(F) = \frac{\text{card}(Pos_Q(F))}{\sum_{X_i \in F} \text{card}(\overline{Q}X_i)}$$

Q -dokładnością aproksymacji rodziny F nazywamy iloraz liczby elementów Q -pozytywnego obszaru rodziny F przez sumę liczby elementów Q -górných aproksymacji zbiorów X_i z rodziny F .

Liczby te spełniają warunek $0 \leq \beta_Q(F) \leq \gamma_Q(F) \leq 1$

Dla analizy danych podstawowe znaczenie ma pojęcie zależności między zbiorami atrybutów systemu informacyjnego. Badanie tej zależności ma na celu sprawdzenie czy dla jednoznacznego opisu elementów U potrzebna jest znajomość wartości wszystkich atrybutów.

Niech (1) będzie systemem informacyjnym i $P, B \subseteq Q$. Zbiór atrybutów B zależy w stopniu k od zbioru atrybutów P , jeżeli $k = \gamma_P(B)$. Zależność tę będziemy zapisywali w postaci $P \xrightarrow{k} B$ dla $k < 1$ oraz w postaci $P \rightarrow B$ dla $k = 1$.

Jeżeli $k = 1$ to B całkowicie zależy od P .

Jeżeli $0 < k < 1$ to B częściowo zależy od P .

Jeżeli $k = 0$ to B nie zależy od P .

Pojęcie reduktu atrybutu wykorzystywane jest do eliminacji zbędnych danych zawartych w systemie informacyjnym.

Niech (1) będzie systemem informacyjnym i $P \subseteq Q$. Powiemy, że zbiór atrybutów P jest niezależny w SI, jeżeli dla każdego podzbioru właściwego $B \subseteq Q$ zachodzi $\tilde{P} \neq \tilde{B}$; w przeciwnym wypadku P jest zależny w SI.

Zbiór $P \subseteq B \subseteq Q$ jest reduktom B w SI, jeżeli P jest minimalnym niezależnym podzbiorem B .

Atrybut $p \in P$ jest nieusuwalny z P jeżeli $\tilde{P} - \{p\} \neq \tilde{P}$; w przeciwnym przypadku atrybut p jest zbędny w P .

Zbiór wszystkich nieusuwalnych atrybutów z P nazywa się rdzeniem P i jest zdefiniowany następująco:

$$(12) \quad CORE(P) = \{p \in P : \tilde{P} - \{p\} \neq \tilde{P}\}$$

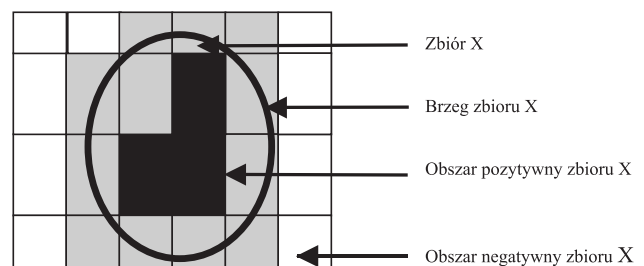
Tablicę decyzyjną można traktować jako reprezentację zbioru reguł decyzyjnych. Każdy wiersz tablicy decyzyjnej zawiera regułę, która specyfikuje decyzje, jakie muszą być podjęte, gdy odpowiednie warunki zostają spełnione. Zbiory warunków i decyzji są przy tym określone przez wartości atrybutów warunkowych i decyzyjnych.

Formalnie tablicą decyzyjną nazywamy uporządkowaną piątkę

$$(13) \quad DT = \langle U, C, D, v, f \rangle$$

gdzie U, C, D są skończonymi zbiorami, v jest odwzorowaniem przyporządkowującym każdemu elementowi $a \in C \cup D$ skończony zbiór V_a , a f jest funkcją z $U \times (C \cup D)$ w sumę V zbiorów V_a taką, że $f(x, a) \in V$ dla wszystkich x oraz a . Zbiór U nazywany jest uniwersum tablicy decyzyjnej, elementy C nazywane są atrybutami warunkowymi, elementy D nazywane są atrybutami decyzyjnymi, V_a jest dziedziną atrybutu a . Funkcja f nazywa się funkcją decyzyjną.

Informacje zawarte w tablicy mają swoje odwzorowanie na płaszczyźnie



Rys. 2 Graficzna interpretacja zawartości tablicy 1.

Rysunek 2 obrazuje graficzną interpretację zawartości tablicy 1. Prostokąty odpowiadają zbiorom elementarnym. Jeśli do poszczególnych zbiorów elementarnych przyporządkujemy specyficzne reguły decyzyjne (gdyż przedsiębiorstwa zaklasyfikowane do poszczególnych zbiorów elementarnych się nie różnią), to jeśli nowe przedsiębiorstwo zostanie zaklasyfikowane do określonego zbioru elementarnego to możemy prognozować jego stan według reguł które wcześniej zostały określone. Jeśli w interesującym nas zbiorze X znajdują się stany elementarne, które nie w całości tworzą zbioru elementarne, to mówimy wtedy o przybliżoności zbioru X . W przykładzie zaprezentowanym na rysunku 2 zbiór X jest zbiorom przybliżonym, ponieważ obszar pozytywny zbioru X nie pokrywa całego obszaru zbioru.

Do dolnej aproksymacji zbioru X zakwalifikujemy tylko te elementy zbioru X , które na pewno za pomocą wybranych atrybutów mogą być zakwalifikowane do zbioru X . Natomiast do górnej aproksymacji zbioru X zaliczamy wszystkie elementy, które należą do rozpa-trywanego zbioru X .

W oprogramowaniu RSES 2.2.1 zamiast relacji nierozróżnialności zaproponowanej przez Z. Pawlaka w roku 1982 zastosowano nowsze algorytmy tworzące zbiory elementarne takie jak: algorytm wyczerpujący - [Exhaustive algorithm], algorytm genetyczny - [Genetic algorithm], algorytm pokryciowy - [Covering algorithm] oraz algorytm LEM2 - [LEM2 algorithm].

3. Wykorzystanie oprogramowania RSES2.2.1 do budowy systemu przewidującego kondycję ekonomiczno-finansową

Do zaprezentowania procedury budowy modeli wykorzystano dane przedsiębiorstw wchodzących w skład sekcji I (601 - 642) - transport, gospodarka magazynowa i łączność z województwa lubelskiego².

Pierwszą czynnością, jaką wykonujemy po uruchomieniu oprogramowania RSES jest utworzenie w nowym projekcie obiektu "Tablica Danych". Do nowo powstałej tablicy importujemy dane z przygotowanego wcześniej pliku tekstowego zawierającego wszystkie przypadki w badanej grupie wraz z opisem poszczególnych atrybutów. Przed przystąpieniem do dalszej pracy z danymi należy dokonać ich dyskretyzacji, czyli podzielenia atrybutów ciągłych na rozłączne podzbiory.

Oprogramowanie RSES posiada dwie metody wykonywania dyskretyzacji: globalną i lokalną. Wykorzystując metodę globalną do dyskretyzacji sekcji I (601 - 642) do dalszej pracy system jako istotne zaklasyfikował dziewięć atrybutów. Każdy z atrybutów został podzielony na dwa rozłączne podzbiory, gdzie granicą jest wartość podana w nawiasach kwadratowych.

- X₂ [0,69637]
- X₆ [0,00485]
- X₁₃ [0,2046]
- X₁₅ [0,007065]
- X₁₇ [2,38986]
- X₂₂ [1,33492]
- X₂₄ [7,05E-4]

Z₁ [2,6]

Z₇ [4,08]

Lokalna metoda dyskretyzacji wykorzystuje dwanaście niżej zaprezentowanych zmiennych. Należy zaznaczyć, iż w przypadku wskaźników X₁, X₂, X₃, X₈, X₁₉ i X₂₂ zbiory zostały podzielone na większą liczbę przedziałów.

X₁ [50,06283 60,499205 70,374165 81,95134]

X₂ [0,15267 42,153335]

X₃ [46,74302 89,64382]

X₄ [8,6869249]

X₆ [0,42168]

X₈ [0,001785 0,0023]

X₁₃ [0,1684749]

X₁₇ [5,7804199]

X₁₉ [0,06185 0,19354]

X₂₂ [0,8783749 1,334919]

X₂₄ [7,05E-4]

Z₄ [-9,75]

Dobre rezultaty uzyskane dzięki wykorzystaniu lokalnej metody dyskretyzacji danych opisane w artykule [8] nie zostały w tym przypadku potwierdzone. W obecnym badaniu, metoda ta daje znacznie gorsze wyniki, powodując dopasowanie wygenerowanych reguł do danych uczących, a przez to złe rezultaty walidacji modelu na grupie testowej.

W dalszej części artykułu wykorzystywane będą dane dyskretyzowane metodą globalną.

Zbiór 185 występujących w sekcji I przedsiębiorstw podzielono losowo na dwa zbiory: uczący (138 elementów) - na podstawie którego zbudowano reguły decyzyjne oraz walidacyjny (47 elementów) - testujący wygenerowane reguły na nieznanym systemowi nowych przypadkach.

Kolejne dwa kroki analizy to wygenerowanie zbioru reguł decyzyjnych, a następnie sprawdzenie ich skuteczności na zbiorze walidacyjnym. Najlepszym algorytmem budującym zbiory reguł okazał się algorytm wyczerpujący. Jego skuteczność została wcześniej potwierdzona w trakcie badań na danych E. Altmanna [8]. Wygenerowano 380 reguł, które w 87% poprawie klasyfikują przypadki ze zbioru walidacyjnego. Niestety ze względu na

		Predicted				
		1	0	No. of obj.	Accuracy	Coverage
Actual	1	71	16	87	0.816	1
	0	5	20	25	0.8	1
True positive rate		0.93	0.56			
Total number of tested objects: 112						
Total accuracy: 0.812						
Total coverage: 1						

Rys. 3 Macierz wyników dla sekcji I - test na całej próbie, województwo podkarpackie.

² Szczegółowy opis wstępnego przygotowania danych przez urzędy statystyczne znajduje się w artykule M. Kowerskiego - Koncepcja badań sektora małych i średnich przedsiębiorstw w projekcie "System przeciwdziałania bezrobociu na obszarach słabo zurbanizowanych", opublikowanego w niniejszym opracowaniu.

bardzo małą reprezentację przypadków przedsiębiorstw w złej kondycji ekonomiczno-finansowej (23 przypadki) przeważająca część reguł dotyczy przedsiębiorstw w dobrej kondycji finansowej. Przedsiębiorstwa w dobrej kondycji finansowej rozpoznawane są w 90% poprawnie, zła kondycja finansowa rozpoznana jest poprawnie w 67%. W związku z dość dużą dysproporcją w poprawnym rozpoznawaniu sytuacji ekonomicznej przedsiębiorstwa, kolejne czynności zmierzały do polepszenia skuteczności reguł rozpoznających złą kondycję ekonomiczno-finansową. W rezultacie uzyskano zbiór 368 reguł rozpoznających przedsiębiorstwa w złej kondycji finansowej z dokładnością 100% na zbiorze testowym. Nieznacznemu pogorszeniu uległo poprawne klasyfikowanie przedsiębiorstw znajdujących się w dobrej kondycji finansowo-ekonomicznej (76%), jednak niesymetryczna funkcja straty badanego problemu, przemawia za pozosta-

wieniem reguł w wygenerowanej postaci. Niepokój budzi zbyt dobra (100%) klasyfikacja złej kondycji, jednakże w zbiorze walidacyjnym występowało tylko sześć przypadków przedsiębiorstw o złej kondycji ekonomicznej. Aby potwierdzić poprawność klasyfikacji przeprowadzono dwa eksperymenty. W pierwszym przetestowano posiadane reguły na całym zbiorze danych 185 przypadków. Uzyskane rezultaty poprawności klasyfikacji to: 82% poprawnych klasyfikacji przedsiębiorstw o dobrej kondycji finansowej i 86% o złej. Drugi eksperyment przeprowadzono na danych statystycznych sekcji I z województwa podkarpackiego. Jest to nowy nieznan systemowi zbiór 112 przypadków również z sekcji I - nowy zbiór walidacyjny, tym trudniejszy, iż pochodzący z innego województwa. Macierz wyników obrazująca wynik testu okazała się bardzo zadowolająca (Rys 3).

Results of experiments by train&test method: R601_642_A_GI_0.3_lo							
		Predicted					
Actual	0	0	1	No. of obj.	Accuracy	Coverage	
	1	5	0	0	5	1	1
	4	25		29	0.862	1	
	True positive rate	0.56	1				
Total number of tested objects: 34							
Total accuracy: 0.882							
Total coverage: 1							

Rys. 4 Macierz wyników dla sekcji I - test na próbie walidacyjnej, województwo podkarpackie.

Na rysunku trzecim widzimy, iż wskaźnik Total coverage wyrażający stosunek obiektów sklasyfikowanych do wszystkich obiektów testowanych przybiera wartość 1. Wartość taka oznacza, iż wszystkie obiekty ze zbioru testowego zostały sklasyfikowane z wykorzystaniem wcześniej opracowanych reguł.

Wskaźnik Total accuracy wskazuje stosunek poprawnie sklasyfikowanych obiektów do wszystkich przetestowanych obiektów. Na powyższym rysunku możemy zauważyć, iż system w 81% poprawnie dokonał klasyfikacji.

Jak widzimy na rysunku 6 wiersze odpowiadają klasom decyzyjnym, a kolumny odpowiedziom wygenerowanym przez algorytm decyzyjny. W opracowanym przykładzie wskaźnik Accuracy - oznacza stosunek poprawnie sklasyfikowanych obiektów do wszystkich obiektów sklasyfikowanych z danej klasy i wynosi on 81% dla 1 (dobra kondycja finansowa) i 80% dla 0 (zła kondycja finansowa).

Coverage for decision classes to stosunek obiektów sklasyfikowanych do wszystkich obiektów z danej klasy decyzyjnej. Wartość 1 oznacza prawidłowe zaklasyfikowanie wszystkich przypadków testowych do odpowiednich klas.

Wartym podkreślenia jest fakt zbudowania wspólnego zestawu reguł poprawnie klasyfikujących kondycję ekonomiczną przedsiębiorstw na poziomie 80% dla dwóch województw lubelskiego i podkarpackiego.

Zbudowany model dla województwa podkarpackiego dla sekcji I (601 - 642) charakteryzuje się jeszcze lepszymi parametrami.

Jak wspomniano we wstępie artykułu, ograniczony zbiór reguł mógłby posłużyć doradcom ekonomicznym do poprawy wybranych wskaźników przedsiębiorstwa, a dzięki temu uniknięcia bankructwa. Do takiego celu posiadany zbiór 368 reguł decyzyjnych, wydaje się zbyt rozbudowany aby mógł być wykorzystany do celów doradczych. Skracając i generalizując posiadane reguły uzyskano zbiór 11 reguł o nieznacznie gorszych wynikach predykcji, zaprezentowanych na rys. 5 i 6.

Nieco gorsze wyniki rozpoznawania przedsiębiorstw o dobrej kondycji finansowej spowodowane są losowym wyborem próby testowej. Test na grupie kontrolnej z województwa podkarpackiego sekcji I pokazuje trafność na poziomie 76%.

		Predicted			No. of obj.	Accuracy	Coverage
		1	0				
Actual	1	28	13	41	0.683	1	
	0	0	6	6	1	1	
True positive rate		1	0.32				
Total number of tested objects: 47							
Total accuracy: 0.723							
Total coverage: 1							

Rys. 5 Macierz wyników dla sekcji I, województwo lubelskie.

		Predicted			No. of obj.	Accuracy	Coverage
		1	0				
Actual	1	66	21	87	0.759	1	
	0	6	19	25	0.76	1	
True positive rate		0.92	0.48				
Total number of tested objects: 112							
Total accuracy: 0.759							
Total coverage: 1							

Rys. 6 Macierz wyników dla sekcji I - przy 11 regułach, województwo podkarpackie.

Poniżej zaprezentowano wygenerowane reguły. Trzy pierwsze dotyczą przedsiębiorstw o dobrej kondycji ekonomiczno-finansowej. Reguła pierwsza odnotowuje prawidłowość odnoszącą się do wskaźnika X_{13} . Jeśli wskaźnik X_{13} zawiera się w przedziale od 0,2046 do nieskończoności (Inf) to przedsiębiorstwo nie jest w złej kondycji ekonomiczno finansowej. Natomiast reguła numer cztery dotyczy przedsiębiorstwa znajdującego się w złej kondycji finansowo-ekonomicznej i przedstawia następującą prawidłowość: jeżeli wskaźnik X_{13} zawiera się w przedziałach od minus nieskończoności do 0,2046 lub od 0,2046 do nieskończoności i wskaźnik X_{15} zawiera się w przedziałach od 0,007065 do nieskończoności lub od minus nieskończoności do 0,007065 i wskaźnik X_{22} zawiera się w przedziałach od 1,33492 do nieskończoności lub od minus nieskończoności do 1,33492 i wskaźnik X_{24} zawiera się w przedziale od minus nieskończoności do 0,000705 i wskaźnik Z_1 zawiera się w przedziałach od 2,6 do nieskończoności lub od minus nieskończoności do 2,6 to przedsiębiorstwo jest w złej kondycji finansowej ($Y=0$) reguła ta jest poparta osiemnastoma przypadkami ze zbioru uczącego.

1. $(X_{13}="0,2046, Inf") \Rightarrow (Y=1[67])$
2. $(X_{22}="(-Inf, 1,33492)") \Rightarrow (Y=1[87])$
3. $(X_{24}="7,05E-4, Inf") \Rightarrow (Y=1[87])$
4. $(X_{13}="(-Inf, 0,2046)" | (0,2046, Inf)) \& (X_{15}="(0,007065, Inf)" | (-Inf, 0,007065)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (X_{24}="(-Inf, 7,05E-4)") \& (Z_1="(2,6, Inf)" | (-Inf, 2,6)) \Rightarrow (Y=0[18])$
5. $(X_6="(-Inf, 0,00485)" | (0,00485, Inf)) \& (X_{15}="(0,007065, Inf)" | (-Inf, 0,007065)) \&$

- $(X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (X_{24}="(-Inf, 7,05E-4)") \& (Z_7="(-Inf, 4,08)" | (4,08, Inf)) \Rightarrow (Y=0[18])$
6. $(X_{13}="(-Inf, 0,2046)" | (0,2046, Inf)) \& (X_{15}="(0,007065, Inf)" | (-Inf, 0,007065)) \& (X_{17}="(2,38986, Inf)" | (-Inf, 2,38986)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (Z_7="(-Inf, 4,08)" | (4,08, Inf)) \Rightarrow (Y=0[16])$
7. $(X_{13}="(-Inf, 0,2046)" | (0,2046, Inf)) \& (X_{15}="(0,007065, Inf)" | (-Inf, 0,007065)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (X_{24}="(-Inf, 7,05E-4)") \& (Z_7="(-Inf, 4,08)" | (4,08, Inf)) \Rightarrow (Y=0[18])$
8. $(X_2="(0,69637, Inf)" | (-Inf, 0,69637)) \& (X_{13}="(-Inf, 0,2046)" | (0,2046, Inf)) \& (X_{15}="(0,007065, Inf)" | (-Inf, 0,007065)) \& (X_{17}="(2,38986, Inf)" | (-Inf, 2,38986)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \Rightarrow (Y=0[16])$
9. $(X_2="(0,69637, Inf)" | (-Inf, 0,69637)) \& (X_{17}="(2,38986, Inf)" | (-Inf, 2,38986)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (Z_7="(4,08, Inf)" | (-Inf, 4,08)) \Rightarrow (Y=0[16])$
10. $(X_2="(0,69637, Inf)" | (-Inf, 0,69637)) \& (X_{13}="(-Inf, 0,2046)" | (0,2046, Inf)) \& (X_{15}="(0,007065, Inf)" | (-Inf, 0,007065)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (Z_1="(-Inf, 2,6)" | (2,6, Inf)) \Rightarrow (Y=0[16])$
11. $(X_{13}="(-Inf, 0,2046)" | (0,2046, Inf)) \& (X_{22}="(1,33492, Inf)" | (-Inf, 1,33492)) \& (Z_1="(-Inf, 2,6)" | (2,6, Inf)) \& (Z_7="(-Inf, 4,08)" | (4,08, Inf)) \Rightarrow (Y=0[16])$

4. Podsumowanie

Niniejszy artykuł opisuje wykorzystanie teorii zbiorów przybliżonych do predykcji kondycji finansowo-ekonomicznej przedsiębiorstw z województw podkarpackiego i lubelskiego dla sekcji I (601 - 642). Przedstawione wyniki potwierdzają słuszność włączenia teorii zbiorów przybliżonych do budowy syntetycznego modelu określającego kondycję ekonomiczno-finansową przedsiębiorstw.

Literatura:

1. Dimitras AI, Słowiński R, Susmara R, Zopounidis C. Business failure prediction using rough sets. *European Journal of Operational Research* 114, s. 263-280 1999
2. Gruszecki J.: Metody wspomaganie podejmowania decyzji, Wyd. WSZiA (w opracowaniu)
3. Knowledge Systems Group, Dept. of Computer and Information Science, NTNU, Norway, Logic Group, Inst. of Mathematics, Warsaw University, Poland: Rosetta <http://www.idi.ntnu.no/~aleks/rosetta/>
4. McKee TH.: Developing a Bankruptcy Prediction Model via Rough Sets Theory, *International Journal of Intelligent Systems in Accounting, Finance & Management*, John Wiley & Sons 2000
5. Mrózek A., Płonka L.: Analiza danych metodą zbiorów przybliżonych, Akademyka Oficyna Wydawnicza PLJ, Warszawa 1999
6. Pawlak Z.: *Rough Sets - Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, 1992
7. Rose2 Rough Set Data Explorer <http://www-idss.cs.put.poznan.pl/rose>
8. Sachajko J.: "Zastosowanie teorii zbiorów przybliżonych do analizy ryzyka bankructwa firmy" WSZiA 2006
9. Słowiński R.: "Wielokryterialne wspomaganie decyzji na podstawie reguł wyindukowanych ze zbiorów przybliżonych"
10. Słowiński R., Zopounidis C.: Application of the rough set approach to evaluation of bankruptcy risk. *International Journal of Intelligent Systems in Accounting, Finance, and Management* 4: No. 1, 1995
11. Witkowska D.: Sztuczne sieci neuronowe I metody statystyczne. Wybrane zagadnienia finansowe, Wydawnictwo C.H. Beck, 2002
12. Zopounidis C., Dimitras A.: *Multicriteria Decision Aid Methods for the Prediction of Business Failure*, Kluwer: Dordrecht.