Autoreferat

dr inż. Teresa Mroczek

27 września 2023

Spis treści

1 Wykształcenie

2009 Politechnika Wrocławska, Wydział Informatyki i Zarządzania. Uzyskanie stopnia doktora nauk technicznych

- obrona rozprawy doktorskiej: 6 stycznia 2009 r.
- nadanie stopnia doktora nauk technicznych przez Radę Wydziału Informatyki i Zarządzania Politechniki Wrocławskiej: 29 września 2009 r.

2001 Politechnika Rzeszowska im. Ignacego Łukasiewicza, Wydział Elektrotechniki i Informatyki. Uzyskanie tytułu zawodowego magistra inżyniera

- obrona pracy magisterskiej: 18 czerwca 2001 r.
- kierunek studiów: Informatyka, specjalność: Systemy i sieci komputerowe.

2 Przebieg zatrudnienia

Okres	Miejsce zatrudnienia
Od VII 2010	Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie, Kolegium Informatyki Stosowanej (dawniej Wydział Informatyki Stosowanej), Katedra Sztucznej Inteligencji (dawniej Katedra Systemów Ekspertowych i Sztucznej Inteligencji), stanowisko: adiunkt, pracownik naukowo-dydaktyczny
2009–2010	Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie, Wydział Informatyki Stosowanej, Katedra Systemów Rozproszonych, stanowisko: adiunkt, pracownik naukowo-dydaktyczny
2020	Wyższa Szkoła Europejska im. ks. Józefa Tischnera w Krakowie, prowadzenie zajęć laboratoryjnych i wykładów (umowa o dzieło)
2009–2010	Wyższa Szkoła Administracji i Zarządzania w Zamościu, prowadzenie zajęć laboratoryjnych i wykładów (umowa o dzieło)
2001–2009	Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie, Wydział Informatyki Stosowanej, Katedra Elektroniki i Telekomunikacji (dawniej Zakład Sieci Komputerowych i Telekomunikacyjnych), stanowisko: asystent
2006	Policealne Studium Zawodowe w Jarosławiu, prowadzenie zajęć laboratoryjnych i wykładów (umowa o dzieło)

2001–2006	Policealne Studium Zawodowe Stowarzyszenia Promocji Przedsiębiorczości w Rzeszowie, prowadzenie zajęć laboratoryjnych i wykładów (umowa o dzieło)
2001	Wyższa Szkoła Informatyki i Zarządzania w Rzeszowie, Wydział Administracyjno - Informatyczny, stanowisko: stażysta

3 Omówienie osiągnięć, o których mowa w art. 219 ust. 1 pkt. 2 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478 z późn. zm.)

3.1 Tytuł i zakres osiągnięcia I

Podstawę wniosku o przeprowadzenie postępowania habilitacyjnego stanowi osiągnięcie I pt. **Rozwój metod eksploracji danych niekompletnych**.
Na przedmiotowe osiągnięcie I składa się 20 prac, z czego cztery pozycje z listy Journal Citation Reports (JCR), jedenaście w materiałach konferencyjnych międzynarodowych konferencji i pięć to rozdziały w książkach. Sumaryczny wskaźnik IF wymienionych, zgodnie z rokiem publikacji, prac wynosi 10.16, a liczba punktów MNiSW/MEiN wynosi 770. W ramach osiągnięcia I:

1. Uogólniono definicję maksymalnych bloków spójnych;
2. Zbadano właściwości maksymalnych bloków obliczonych z danych z brakującymi wartościami atrybutów interpretowanymi jako wartości utracone (lost values);
3. Zdefiniowano trzy rodzaje przybliżeń probabilistycznych: singleton, podzbiór oraz koncept, oparte na maksymalnych blokach spójnych. Zbadano właściwości zdefiniowanych przybliżeń oraz zweryfikowano skuteczność klasyfikacji i złożoność zbiorów reguł indukowanych z przybliżeń;
4. Wykazano, że niektóre opublikowane metody mogą generować bloki spójne, które nie są maksymalne, szczególnie w przypadku danych dla których relacja charakterystyczna jest nieprzechodnia;
5. Oszacowano liczbę maksymalnych bloków spójnych oraz złożoność czasową obliczania maksymalnych bloków spójnych;
6. Opracowano nową metodę obliczania maksymalnych bloków spójnych umożliwiającą ich praktyczne zastosowanie;
7. Opracowano dwa nowe typy przybliżeń probabilistycznych: globalne przybliżenie probabilistyczne oraz nasycone przybliżenie probabilistyczne;

8. Zweryfikowano zastosowanie globalnych i nasyconych przybliżeń probabilistycznych w kontekście skuteczności klasyfikacji oraz złożoności zbiorów reguł indukowanych z przybliżeń.

3.2 Lista prac wchodzących w zakres osiągnięcia I

[I.1] T.Mroczek (2023) *Handling the Complexity of Computing Maximal Consistent Blocks*. Electronics 12(10):2295. IF(2022):2.9. doi: 10.3390/electronics12102295

[I.2] P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek, R.Niemiec (2023) *Global and Saturated Probabilistic Approximations Based on Generalized Maximal Consistent Blocks*. Logic Journal of the IGPL 31(2):223–239. IF(2022): 0.868. doi: 10.1093/jigpal/jzac015

[I.3] T.Mroczek, R.Zheng (2022) *A New Approach to Constructing Maximal Consistent Blocks for Mining Incomplete Data*. Procedia Computer Science 207:1047–1056. doi: 10.1016/j.procs.2022.09.160

[I.4] P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek, R.Niemiec (2021) *Complexity of Rule Sets Induced from Data with Many Lost Values and "Do Not Care" Conditions*. In: Abraham, A., Siarry, P., Ma, K., Kaklauskas, A. (eds) Intelligent Systems Design and Applications. ISDA 2019. Advances in Intelligent Systems and Computing 1181:376–385. Springer, Cham. doi: 10.1007/978-3-030-49342-4_36

[I.5] P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek (2021) *Mining Incomplete Data Using Global and Saturated Probabilistic Approximations Based on Characteristic Sets and Maximal Consistent Blocks*. W: Ramanna, S., Cornelis, C., Ciucci, D. (eds) Rough Sets. IJCRS 2021. Lecture Notes in Computer Science 12872:3-17. Springer, Cham. doi: 10.1007/978-3-030-87334-9_1

[I.6] P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek (2021) *Complexity of Rule Sets in Mining Incomplete Data Using Characteristic Sets and Generalized Maximal Consistent Blocks*. Logic Journal of the IGPL 29(2):124–137. IF(2021): 0.868. doi: 10.1093/jigpal/jzaa041

[I.7] P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek, R.Niemiec (2020) *Complexity of Rule Sets Mined from Incomplete Data Using Probabilistic Approximations Based on Generalized Maximal Consistent Blocks*. Procedia Computer Science 176:1803–1812. doi: 10.1016/j.procs.2020.09.219

[I.8] P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek, R.Niemiec (2020) *Global and Saturated Probabilistic Approximations Based on Generalized Maximal Consistent Blocks*. In: de la Cal, E.A., Villar Flecha, J.R., Quintián, H., Corchado, E. (eds) Hybrid Artificial Intelligent Systems. HAIS 2020. Lecture Notes in Computer Science 12344:387–396. Springer, Cham. doi: 10.1007/978-3-030-61705-9_32

[I.9] P.Clark, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2020) *Mining Data with Many Missing Attribute Values Using Global and Saturated Probabilistic Approximations Based on Characteristic Sets*. W: Lopata, A., Butkienė, R., Gudonienė, D., Sukackė, V. (eds) Information and Software Technologies. ICIST 2020. Communications in Computer and Information Science 1283:72–83. Springer, Cham. doi: 10.1007/978-3-030-59506-7_7

[I.10] P.Clark, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2020) *Mining Incomplete Data—A Comparison of Concept and New Global Probabilistic Approximations*. In: Czarnowski, I., Howlett, R., Jain, L. (eds) Intelligent Decision Technologies 2019. Smart Innovation, Systems and Technologies 142:167–178. Springer, Singapore. doi: 10.1007/978-981-13-8311-3_15

[I.11] P.Clark, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2019) *Rule Set Complexity in Mining Incomplete Data Using Global and Saturated Probabilistic Approximations*. In: Damaševičius, R., Vasiljevienė, G. (eds) Information and Software Technologies. ICIST 2019. Communications in Computer and Information Science 1078:451–462. Springer, Cham. doi: 10.1007/978-3-030-30275-7_35

[I.12] P.Clark, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2019) *A Comparison of Global and Saturated Probabilistic Approximations Using Characteristic Sets in Mining Incomplete Data*. INTELLI 2019: The Eighth International Conference on Intelligent Systems and Applications. IARIA, pp. 10–15

[I.13] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek (2018). *A Comparison of Characteristic Sets and Generalized Maximal Consistent Blocks in Mining Incomplete Data*. In: Medina, J., et al. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations. IPMU 2018. Communications in Computer and Information Science 854:480–489. Springer, Cham. doi: 10.1007/978-3-319-91476-3_40

[I.14] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2018). *A Comparison of Concept and Global Probabilistic Approximations Based on Mining Incomplete Data*. In: Damaševičius, R., Vasiljevienė, G. (eds) Information and Software Technologies. ICIST 2018. Communications in Computer and Information Science 920:324–335. Springer, Cham. doi: 10.1007/978-3-319-99972-2_26

[I.15] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek (2018) *Characteristic sets and generalized maximal consistent blocks in mining incomplete data*. Information Sciences 453:66–79. IF(2018): 5.524. doi: 10.1016/j.ins.2018.04.025

[I.16] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2018) *Complexity of Rule Sets in Mining Incomplete Data Using Characteristic Sets and Generalized Maximal Consistent Blocks*. In: de Cos Juez, F., et al. Hybrid Artificial Intelligent Systems. HAIS 2018. Lecture Notes in Computer Science 10870:84–94. Springer, Cham. 10.1007/978-3-319-92639-1_8

[I.17] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek, R.Niemiec (2018) *Complexity of Rule Sets Induced by Characteristic Sets and Generalized Maximal Consistent Blocks*. In: Rutkowski, L., et al. (eds) Artificial Intelligence and Soft Computing. ICAISC 2018. Lecture Notes in Computer Science 10842:301-310. Springer, Cham. doi: 10.1007/978-3-319-91262-2_27

[I.18] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek (2018) *On the Number of Conditions in Mining Incomplete Data Using Characteristic Sets and Maximal Consistent Blocks*. In: Weckman, G., Grzymała-Busse, J.W. (eds) The Fourth International Conference on Big Data, Small Data, Linked Data and Open Data. ALLDATA 2018. IARIA pp. 84–89

[I.19] P.Clark, C.Gao, J.W.Grzymała-Busse, T.Mroczek (2017) *Characteristic Sets and Generalized Maximal Consistent Blocks in Mining Incomplete Data*. In: Polkowski, L., et al. Rough Sets. IJCRS 2017. Lecture Notes in Computer Science 10313:477–486. Springer, Cham. doi: 10.1007/978-3-319-60837-2_39

[I.20] J.W.Grzymała-Busse, T.Mroczek (2016) *Definability in Mining Incomplete Data*. Procedia Computer Science 96:179–186. doi: 10.1016/j.procs.2016.08.125

3.3 Omówienie osiągnięcia I

3.3.1 Wprowadzenie

Głównym obszarem badań wnioskodawcy, wchodzącym w skład przedmiotowego osiągnięcia I, jest rozwój metod eksploracji danych niekompletnych, w szczególności rozwój podejścia opartego na maksymalnych blokach spójnych oraz przybliżeniach probabilistycznych. Rzeczywiste dane są bardzo często niekompletne, występują w nich – z różnych powodów – brakujące wartości atrybutów. Brakująca, nieznana wartość nie oznacza wartości niemożliwej do analizy, dla takiej przewidziane jest specjalne oznaczenie w zbiorze. Może być natomiast rozważana jako wartość z dziedziny atrybutu [1]. Jednak z uwagi na fakt, że głównymi przyczynami niekompletności są wartości utracone albo wartości uznane za nieistotne, badania wnioskodawcy koncentrują się w głównej mierze wokół tych dwóch przyczyn. Wartość utracona – *lost value* – jest interpretowana jako wartość, do której nie mamy dostępu np. z powodu jej usunięcia albo niewprowadzenia do zbioru. Z kolei wartość nieistotna – *"do not care" condition* – interpretowana jest jako dowolna wartość z dziedziny atrybutu.
Pierwsze, oparte na zbiorach przybliżonych, podejście do analizy brakujących wartości atrybutów interpretowanych jako lost values, zostało opisane w [2]. Następnie w [3] przedstawiono modyfikację pierwotnego zamysłu w postaci podejścia opartego na wartościowanej relacji tolerancji oraz na zbiorach rozmytych. W [4] zaproponowano metodę indukcji reguł dla brakujących wartości atrybutów interpretowanych jako *"do not care" conditions*, w której każdą brakującą wartość atrybutu zastępowano wszystkimi możliwymi wartościami z dziedziny analizowanego atrybutu. W [5] rozwinięto podejście analizy danych z brakującymi wartościami interpretowanymi jako *"do not care" conditions* i uzupełniono o właściwości teoretyczne.
Na potrzeby omówienia osiągnięcia poniżej przedstawiono wprowadzenie precyzujące nomenklaturę związaną z niekompletnymi zbiorami danych, zbiorami charakterystycznymi oraz maksymalnymi blokami spójnymi.

Niekompletne zbiory danych

Idee teorii zbiorów przybliżonych są badane i rozwijane na podstawie danych zgromadzonych w postaci tablicy decyzji [6, 7]. Przykład niekompletnego zbioru danych, w formie tablicy decyzji, został przedstawiony w Tabeli 3. Wiersze tablicy decyzji opisują *przypadki*, zwane także *obiektami*. Zbiór wszystkich przypadków nazywany jest *uniwersum* i jest oznaczany jako U. W Tabeli 3 U = {1, 2, 3, 4, 5, 6, 7, 8}. Niezależne zmienne nazywane są *atrybutami* i są oznaczane jako A. W Tabeli 3 *Przebieg*, *Stłuczka* oraz *Wyposażenie* są atrybutami. Zbiór wszystkich wartości atrybutu a jest nazywany domeną a i oznaczany jest jako V_a. W Tabeli 3 $V_{Przebieg} = \{średni, długi\}$. Zmienna zależna – *Zakup* – jest *decyzją*. Zbiór wszystkich przypadków z taką samą wartością zmiennej decyzyjnej jest nazywany *konceptem*. W Tabeli 3 są dwa koncepty: zbiór {1, 2, 3, 4} wszystkich przypadków, dla których *Zakup* ma wartość *tak* i zbiór {5, 6, 7, 8}, gdzie wartość decyzji *Zakup* jest *nie*.

Tabela 3: Niekompletny zbiór danych

	Atrybuty			Decyzja
Przypadek	Przebieg	Stłuczka	Wyposażenie	Zakup
1	średni	tak	podstawowe	tak
2	?	nie	*	tak
3	długi	nie	?	tak
4	średni	*	podstawowe	tak
5	długi	tak	?	nie
6	*	*	luksusowe	nie
7	?	*	podstawowe	nie
8	średni	*	luksusowe	nie

Większość zastosowań opartych na teorii zbiorów przybliżonych, takich jak klasyfikacja, uczenie maszynowe, wspomaganie decyzji, czy odkrywanie wiedzy, bazuje na reprezentacji atrybut-wartość, jako podstawowej granule informacji [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Wartość *v* atrybutu *a* dla przypadku *x* jest oznaczana jako $a(x) = v$. *Blok par atrybut-wartość*, oznaczany jako $[(a, v)]$, jest zbiorem wszystkich przypadków z U, dla których atrybut *a* przyjmuje wartość *v*, $\{x \in U | a(x) = v\}$.
Jak wspomniano wcześniej, rozważane są dwie interpretacje brakujących wartości: *lost values* oznaczone w Tabeli 3 jako ? i *"do not care" conditions* oznaczone jako $*$. Zbiór wszystkich przypadków z U, dla których atrybut a jest *lost value* $\{x \in U | a(x) = ?\}$ jest oznaczany jako $[(a, ?)]$, natomiast zbiór wszystkich przypadków z U dla których atrybut a jest *"do not care" condition* $\{x \in U | a(x) = *\}$ jest oznaczany jako $[(a, *)]$.
Dla niekompletnych zbiorów danych definicja *bloku par atrybut-wartość* została zmodyfikowana [21] jak następuje:

- jeśli dla atrybutu a i przypadku x, $a(x) = ?$, wówczas przypadek x nie powinien być uwzględniany w blokach $[(a, v)]$ dla wszystkich znanych wartości v atrybutu a,
- jeśli dla atrybutu a i przypadku x, $a(x) = *$, wówczas przypadek x powinien być uwzględniany w blokach $[(a, v)]$ dla wszystkich znanych wartości v atrybutu a.

Dla danych z Tabeli 3 wszystkie pary atrybut-wartość są następujące:
$[(Przebieg, średni)] = \{1, 4, 6, 8\}$,
$[(Przebieg, długi)] = \{3, 5, 6\}$,
$[(Stłuczka, tak)] = \{1, 4, 5, 6, 7, 8\}$,

$[(Stłuczka, nie)] = \{2, 3, 4, 6, 7, 8\}$,
$[(Wyposażenie, podstawowe)] = \{1, 2, 4, 7\}$,
$[(Wyposażenie, luksusowe)] = \{2, 6, 8\}$.

Zbiory charakterystyczne

Zbiory charakterystyczne dla niekompletnych zbiorów danych z dowolną interpretacją brakujących wartości atrybutów zostały wprowadzone w [21] jako zbiory przypadków podobnych $x \in U$ w kontekście analizowanego zestawu atrybutów $B \subseteq A$:

- jeśli wartość $a(x)$ jest znana, wówczas $K(x, a)$ jest blokiem $[(a, a(x))]$ atrybutu a i jego wartości $a(x)$,
- jeśli $a(x) = ?$ lub $a(x) = *$, wówczas $K(x, a) = U$.

Ponadto dla niekompletnych zbiorów danych została zdefiniowana na U *B-relacja charakterystyczna* R(B) dla $(x, y) \in U$ [17]:

$$(x, y) \in R(B) \; wtedy \; i \; tylko \; wtedy, \; gdy \; y \in K_B(x),$$

.
Dla danych z Tabeli 3, dla $B = A$ można wskazać następujące zbiory charakterystyczne:
$K_A(1) = \{1, 4\}$,
$K_A(2) = \{2, 3, 4, 6, 7, 8\}$,
$K_A(3) = \{3, 6\}$,
$K_A(4) = \{1, 4\}$,
$K_A(5) = \{5, 6\}$,
$K_A(6) = \{2, 6, 8\}$,
$K_A(7) = \{1, 2, 4, 7\}$,
$K_A(8) = \{6, 8\}$

i relacje charakterystyczną $R(A)$ = {(1, 1), (1, 4), (2, 2), (2, 3), (2, 4), (2, 6), (2, 7), (2, 8), (3, 3), (3, 6), (4, 1), (4, 4), (5, 5), (5, 6), (6, 2), (6, 6), (6, 8), (7, 1), (7, 2), (7, 4), (7, 7), (8, 6), (8, 8)}.

W [I.20] zbadano siedem modyfikacji relacji charakterystycznej zdefiniowanej dla niekompletnych danych. Spośród tych siedmiu modyfikacji dwie ograniczają się do zbiorów danych zawierających wyłącznie brakujące wartości interpretowane jako "do not care" conditions. Tylko jedna jest globalnie definiowalna, jedna lokalnie, a pozostałych pięć nie jest nawet lokalnie definiowalnych. Tych pięć typów zmodyfikowanych relacji charakterystycznych nie należy używać do indukcji reguł decyzyjnych.

Maksymalne bloki spójne

Koncepcja maksymalnych bloków spójnych, została zaadaptowana z matematyki dyskretnej do niekompletnych zbiorów danych z brakującymi wartościami atrybutów interpretowanymi tylko jako *"do not care" conditions* [22]. Opisuje ona

maksymalną kolekcję obiektów, w której wszystkie obiekty są nierozróżnialne w kontekście rozważanego zestawu atrybutów.
Niech X będzie podzbiorem U, $B \subseteq A$. Zbiór X jest *B-spójny*, jeśli $(x, y) \in SIM(B)$ dla dowolnego $x, y \in X$, gdzie $SIM(B)$ jest relacją podobieństwa $\{(x, y) \in U \times U | \forall_b \in B,\ f_b(x) = f_b(y)\ lub\ f_b(x) = *\ lub\ f_b(y) = *\}$ [23, 24]. Jeśli nie istnieje podzbiór $Y \subseteq U$, taki że $Y \subset X$ i Y jest *B-spójny* to zbiór X jest *maksymalnym blokiem B-spójnym*.
Dla $B \subseteq A$ zbiór wszystkich maksymalnych bloków spójnych jest oznaczany jako $\mathscr{C}(B)$, natomiast zbiór wszystkich maksymalnych bloków spójnych ze względu na obiekty $x \in U$ jest oznaczany jako $\mathscr{C}(B)(x)$.

Tabela 4: Niekompletny zbiór danych z brakującymi wartościami atrybutów interpretowanymi jako "do not care" conditions

	Atrybuty			Decyzja
Przypadek	Przebieg	Stłuczka	Wyposażenie	Zakup
1	średni	tak	podstawowe	tak
2	*	nie	*	tak
3	długi	nie	*	tak
4	średni	*	podstawowe	tak
5	długi	tak	*	nie
6	*	*	luksusowe	nie
7	*	*	podstawowe	nie
8	średni	*	luksusowe	nie

Dla zbioru danych z Tabeli 4 i $B = A$, zbiór $\mathscr{C}(A)$ wszystkich maksymalnych bloków spójnych A to $\{\{1, 4, 7\}, \{2, 3, 6\}, \{2, 3, 7\}, \{2, 4, 7\}, \{2, 6, 8\}, \{5, 6\}, \{5, 7\}\}$. Natomiast zbiór maksymalnych bloków spójnych ze względu na obiekty zdeterminowane przez A, to:
$\mathscr{C}(A)(1) = \{\{1, 4, 7\}\}$,
$\mathscr{C}(A)(2) = \{\{2, 3, 6\}, \{2, 3, 7\}, \{2, 4, 7\}, \{2, 6, 8\}\}$,
$\mathscr{C}(A)(3) = \{\{2, 3, 6\}, \{2, 3, 7\}\}$,
$\mathscr{C}(A)(4) = \{\{1, 4, 7\}, \{2, 4, 7\}\}$,
$\mathscr{C}(A)(5) = \{\{5, 6\}, \{5, 7\}\}$,
$\mathscr{C}(A)(6) = \{\{2, 3, 6\}, \{2, 6, 8\}, \{5, 6\}\}$,
$\mathscr{C}(A)(7) = \{\{2, 4, 7\}, \{5, 7\}\}$,
$\mathscr{C}(A)(8) = \{\{2, 6, 8\}\}$.

3.3.2 Omówienie celów naukowych ww. prac oraz osiągniętych wyników

Celem naukowym osiągnięcia I jest rozwój metod eksploracji danych niekompletnych, w szczególności rozwój koncepcji maksymalnych bloków spójnych umożliwiającej ich praktyczne zastosowanie, rozwój przybliżeń probabilistycznych oraz ocena różnych wariantów eksploracji danych niekompletnych.
Koncepcja maksymalnych bloków spójnych została wprowadzona dla niekompletnych zbiorów danych wyłącznie z brakującymi wartościami atrybutów interpretowanymi jako *"do not care" conditions*, z zastosowaniem tylko dolnych i górnych przybliżeń [22]. Stąd, w pierwszej kolejności, rozszerzono definicję maksymalnych bloków spójnych, umożliwiając stosowanie bloków dla dowolnej interpretacji brakujących wartości atrybutów oraz z przybliżeniami probabilistycznymi stanowiącymi rozwinięcie klasycznych przybliżeń znanych z teorii zbiorów przybliżonych. Zbadano właściwości maksymalnych bloków spójnych obliczanych z danych z brakującymi wartościami atrybutów interpretowanymi jako lost values oraz przybliżeń probabilistycznych opartych na blokach. Właściwości bloków dla danych z "do not care" conditons były już przedmiotem badań w [22]. W celu wskazania najlepszego podejścia do eksploracji danych z użyciem maksymalnych bloków spójnych oraz przybliżeń probabilistycznych przeprowadzono szereg eksperymentów na rzeczywistych zbiorach danych, w których oceniano jakość oraz złożoność indukowanych zbiorów reguł. Podjęto również badania, których celem było ustalenie co jest lepszym wyborem w eksploracji danych: zbiory charakterystyczne czy maksymalne bloki spójne.
Oceniając podejścia oparte na zbiorach charakterystycznych oraz maksymalnych blokach spójnych zauważono, że zbiory charakterystyczne można obliczać w czasie wielomianowym, natomiast obliczanie maksymalnych bloków spójnych, w szczególności w przypadku niekompletnych danych z brakującymi wartościami "do not care" conditons, jest procesem wymagającym większych zasobów obliczeniowych. Stąd, kolejnym celem było oszacowanie maksymalnej liczby bloków w niekompletnym zbiorze danych oraz ustalenie złożoności czasowej ich obliczania. Zdefiniowano zbiór *k-galaktyka*, który ilustruje najgorszy przypadek (ang. worst-case), czyli zbiór, z którego można wygenerować największą możliwą liczbę maksymalnych bloków spójnych i wykazano, że obliczanie maksymalnych bloków spójnych jest związane z wykładniczą złożonością czasową.
Jednocześnie, biorąc pod uwagę złożoność obliczeniową, dla praktycznego wykorzystania takich bloków należało zaproponować poprawę wydajności ich obliczania. W pierwszej kolejności zweryfikowano istniejące, opublikowane metody obliczania maksymalnych bloków spójnych. Wykazano, że niektóre z nich mogą generować bloki, które nie są maksymalne oraz, że złożoność czasowa niektórych znanych algorytmów obliczania maksymalnych bloków spójnych była niedoszacowana. Następnie opracowano całkowicie nowe podejście do obliczania maksymalnych bloków spójnych, przeznaczone dla środowisk wieloprocesorowych.
Niezależnie, dla niekompletnych zbiorów danych, poszukiwano przybliżeń najbardziej zbliżonych do aproksymowanego konceptu. Probabilistyczne przybliżenie konceptu jest powiązane z prawdopodobieństwem α, wartością pomiędzy 0 a 1. Jeśli $\alpha = 1$, przybliżenie probabilistyczne staje się dolnym przybliżeniem, natomiast jeśli

α jest małą liczbą dodatnią, przybliżenie probabilistyczne sprowadza się do górnego przybliżenia. Przybliżenie dolne powinno być jak największe, a przybliżenie górne możliwie najmniejsze. Opracowano zatem dwa nowe typy przybliżeń probabilistycznych: globalne przybliżenia probabilistyczne oraz nasycone przybliżenia probabilistyczne bliskie konceptowi, który ma być aproksymowany. Przybliżenia zdefiniowano z użyciem zbiorów charakterystycznych oraz maksymalnych bloków spójnych. W celu wskazania najlepszego podejścia do eksploracji niekompletnych danych z zastosowaniem globalnych i nasyconych przybliżeń probabilistycznych przeprowadzono szereg eksperymentów na rzeczywistych zbiorach danych, w których oceniano jakość oraz złożoność zbiorów reguł indukowanych z nowych przybliżeń.

1. Uogólnienie definicji maksymalnych bloków spójnych

Maksymalne bloki spójne zostały zdefiniowane wyłącznie dla danych z brakującymi wartościami atrybutów interpretowanymi jako "do not care" conditions [12]. W [I.15] definicja maksymalnych bloków spójnych została uogólniona dla dowolnej interpretacji brakujących wartości.
Niech X będzie podzbiorem U, $B \subseteq A$. Zbiór X jest *B-spójny*, jeśli $(x, y) \in R(B)$ dla dowolnego $x, y \in X$. Jeśli nie istnieje B-spójny podzbiór $Y \subseteq U$, taki że X jest podzbiorem właściwym zbioru Y to zbiór X jest *uogólnionym maksymalnym blokiem B-spójnym*.
Ponadto w [I.15], dla niekompletnych zbiorów danych została zdefiniowana relacja *S(B)*, nazywana implikowaną przez rodzinę $\mathscr{C}(B)$ maksymalnych bloków spójnych. Relacja tworzona jest ze wszystkich możliwych par $(x, y) \in U$ takich, że x i y są elementami tego samego B-maksymalnego bloku spójnego.
W ten sposób dla danych z Tabeli 3, dla $B = A$, zbiór $\mathscr{C}(A)$ wszystkich uogólnionych maksymalnych bloków spójnych to $\{\{1, 4\}, \{2, 6\}, \{2, 7\}, \{3\}, \{5\}, \{6, 8\}\}$ oraz relacja implikowana przez rodzinę $\mathscr{C}(A)$ maksymalnych bloków spójnych to $S(A)$ = {(1, 1), (1, 4), (2, 2), (2, 6), (2, 7), (3, 3), (4, 1), (4, 4), (5, 5), (6, 2), (6, 6), (6, 8), (7, 2), (7, 7), (8, 6), (8, 8)}.

Dalsze badania przeprowadzano z użyciem uogólnionych maksymalnych bloków spójnych.

2. Właściwości maksymalnych bloków spójnych obliczonych z danych z brakującymi wartościami atrybutów interpretowanymi jako wartości utracone (lost values)

Właściwości maksymalnych bloków spójnych dla danych ze wszystkimi brakującymi wartościami atrybutów interpretowanymi jako "do not care" conditions, zostały przedstawione w [22] oraz [I.20]. W pracy [I.15] dla zbiorów danych ze wszystkimi brakującymi wartościami interpretowanymi jako lost values wykazano, że:

- zdefiniowana relacja S(B) implikowana przez maksymalne bloki spójne jest relacją równoważności;
- rodzina wszystkich maksymalnych bloków spójnych $\mathscr{C}(B)$ jest podziałem zbioru U;

- zbiór X jest maksymalnym blokiem B-spójnym jeśli $K_B(X) = X$, $x \in X$

i $B \subseteq A$.

3. Przybliżenia probabilistyczne bazujące na maksymalnych blokach spójnych. Właściwości zdefiniowanych przybliżeń. Skuteczność klasyfikacji reguł i złożoność zbiorów reguł indukowanych z przybliżeń

Maksymalne bloki spójne zostały wprowadzone dla niekompletnych zbiorów danych ze wszystkimi brakującymi interpretowanymi jako "do not care" conditions z użyciem tylko dolnego i górnego przybliżenia [12]. W [I.15] zdefiniowano trzy rodzaje przybliżeń probabilistycznych bazując na maksymalnych blokach spójnych:

- przybliżenie probabilistyczne B-singleton zbioru X

  $$appr_{\alpha,B}^{MCB,Singleton}(X) = \{x \mid Y \in \mathscr{C}(B)(x), x \in U,\ Pr(X|Y)\ \geq \alpha\}$$

  gdzie $0 < \alpha \leq 1$, $Pr(X|Y) = \frac{|X \cap Y|}{|Y|}$ jest prawdopodobieństwem warunkowym X dla danego Y;

- przybliżenie probabilistyczne B-podzbiór zbioru X

  $$appr_{\alpha,B}^{MCB,Podzbiór}(X) = \cup\{Y \mid Y \in \mathscr{C}(B),\ Pr(X|Y)\ \geq \alpha\};$$

- przybliżenie probabilistyczne B-koncept zbioru X

  $$appr_{\alpha,B}^{MCB,Koncept}(X) = \cup\{Y \mid Y \in \mathscr{C}(B)(x), x \in X,\ Pr(X|Y)\ \geq \alpha\}.$$

Wykazano, że trzy rodzaje przybliżeń probabilistycznych oparte na maksymalnych blokach spójnych są sobie równe. Można je zredukować do jednego typu przybliżenia probabilistycznego:

$$appr_{\alpha,B}^{MCB}(X) = \cup\{Y \mid Y \in \mathscr{C}(B),\ Pr(X|Y)\ \geq \alpha\}.$$

Dla Tabeli 3 wszystkie prawdopodobieństwa warunkowe $Pr([(Zakup, tak)]|Y)$ oraz $Pr([(Zakup, nie)]|Y)$, gdzie $Y \in \mathscr{C}(A)$ przedstawiono w Tabeli 5.

Tabela 5: Prawdopodobieństwo warunkowe $Pr([(Zakup, tak)]|Y)$ oraz $Pr([(Zakup, nie)]|Y)$

Y	{1, 4}	{2, 6}	{2, 7}	{3}	{5}	{6, 8}
$Pr(\{1,2,3,4\}|Y)$	1	0.5	0.5	1	0	0
$Pr(\{5,6,7,8\}|Y)$	0	0.5	0.5	0	1	1

Wszystkie odrębne przybliżenia probabilistyczne, oparte na maksymalnych blokach spójnych, to:
$appr^{MCB}_{\alpha,0.5}(\{1,2,3,4\}) = \{1,2,4,3,6,7\}$,
$appr^{MCB}_{\alpha,1}(\{1,2,3,4\}) = \{1,3,4\}$,
$appr^{MCB}_{\alpha,0.5}(\{5,6,7,8\}) = \{2,5,6,7,8\}$,
$appr^{MCB}_{\alpha,1}(\{5,6,7,8\}) = \{5,6,8\}$.

Bazując na dwóch rodzajach granul – zbiorach charakterystycznych oraz uogólnionych maksymalnych blokach spójnych – przeprowadzono eksperymenty, w których porównano jakość reguł indukowanych z przybliżeń. W analizie, jako kryterium jakości, przyjęto poziom błędu obliczony jako rezultat zastosowania dziesięciokrotnej walidacji krzyżowej [I.15, I.19]. Eksperymenty przeprowadzono z użyciem dziewięciu zbiorów danych, pozyskanych z Repozytorium Uczenia Maszynowego Uniwersytetu Kalifornijskiego w Irvine, stosując dwie interpretacje brakujących wartości atrybutów: lost values oraz "do not care" conditions oraz trzy rodzaje przybliżeń probabilistycznych: dolne, środkowe ($\alpha = 0.5$) oraz górne.
Dla każdego zbioru danych został stworzony szablon. Szablon taki powstał poprzez zastąpienie losowo jak największej liczby określonych wartości atrybutów, wartościami interpretowanymi jako lost values. Maksymalny odsetek brakujących wartości atrybutów został ograniczony wymogiem, zgodnie z którym żaden wiersz zbioru danych nie może zawierać wyłącznie lost values. Te same szablony zostały użyte do skonstruowania zbiorów danych z brakującymi wartościami interpretowanymi jako "do not care" conditions, poprzez zastąpienie ? przez *.
W celu porównania wspomnianych metod użytych do indukcji reguł opartych na: dolnych, środkowych i górnych przybliżeniach, zbiorach charakterystycznych i uogólnionych maksymalnych blokach spójnych, zastosowano test sumy rang Friedmana w połączeniu z testem wielokrotnych porównań post hoc, z 5% poziomem istotności, dla każdej interpretacji brakujących wartości osobno.
Wykazano, że dla większości analizowanych zbiorów danych jakość reguł opartych na zbiorach charakterystycznych bądź uogólnionych maksymalnych blokach spójnych nie różni się istotnie, natomiast w przypadku gdy różnica jest istotna zastosowanie środkowego przybliżenia opartego na maksymalnych blokach spójnych jest najlepszym podejściem.
Zachowując parametry eksperymentu oceniono złożoności zbiorów reguł, w kontekście liczby warunków [I.18] oraz liczby reguł [I.17] w zbiorze reguł, indukowanych z użyciem zbiorów charakterystycznych i uogólnionych maksymalnych bloków spójnych oraz trzech rodzajów przybliżeń probabilistycznych: dolnym, środkowym ($\alpha = 0.5$) oraz górnym. Wykazano, że dla zbiorów danych z brakującymi wartościami atrybutów interpretowanymi jako lost values, nie ma znaczącej różnicy w złożoności modeli uczenia dla zastosowanych sześciu podejść do eksploracji danych niekompletnych. Natomiast dla zbiorów danych z brakującymi wartościami atrybutów interpretowanymi jako "do not care" conditions, zbiory reguł indukowane z górnych przybliżeń, w oparciu o zbiory charakterystyczne lub uogólnione maksymalne bloki spójne, są prostsze pod względem całkowitej liczby warunków. Biorąc

pod uwagę liczbę reguł, wybór między zbiorami charakterystycznymi a uogólnionymi maksymalnymi blokami spójnymi oraz między rodzajami przybliżeń probabilistycznych jest ważny, ponieważ istnieją statystycznie istotne różnice w złożoności indukowanych zbiorów reguł.
W [I.16, I.13] porównano cztery podejścia do eksploracji niekompletnych zbiorów danych, łącząc zbiory charakterystyczne i uogólnione maksymalne bloki spójne z dwiema interpretacjami brakujących wartości atrybutów, lost values i "do not care conditions". Eksperymenty przeprowadzono tym razem z użyciem ośmiu zbiorów z Repozytorium Uczenia Maszynowego, stosując probabilistyczne przybliżenie konceptu. Dla każdego zbioru danych tworzono szablon, poprzez zastąpienie losowo 35% znanych wartości atrybutów, wartościami interpretowanymi jako lost values. Te same szablony użyto do skonstruowania zbiorów danych z brakującymi wartościami atrybutów interpretowanymi jako "do not care" conditions, zamieniając lost values na "do not care" conditions. Analizowano liczbę reguł i całkowitą liczbę warunków w zbiorach reguł. Wykazano, że wybór między zbiorami charakterystycznymi a uogólnionymi maksymalnymi blokami spójnymi nie jest tak istotny, jak wybór pomiędzy interpretacjami brakujących wartości atrybutów. Najprostsze zestawy reguł były indukowane z niekompletnych zestawów danych z brakującymi wartościami interpretowanymi jako „do not care" conditions, co potwierdziły również wspomniane wcześniej badania.
Analizę złożoności zbioru reguł rozszerzono w [I.6], stosując trzy interpretacje brakujących wartości atrybutów: lost values, "do not care" conditions oraz atrybut-koncept. Wykazano, że najmniejsze zbiory reguł, w kontekście liczebności reguł w zbiorze, są indukowane na podstawie niekompletnych zbiorów danych z wartościami atrybut-koncept, podczas gdy najbardziej liczne i zawierające najwięcej warunków są zbiory reguł indukowane dla zbiorów danych z lost values.

4. Analiza istniejących metod obliczania maksymalnych bloków spójnych

Opublikowano kilka podejść do obliczania maksymalnych bloków spójnych: metodę brute force [25], metodę rekurencyjną [25] oraz metodę hierarchiczną [26] dla niekompletnych zbiorów danych z brakującymi wartościami atrybutów interpretowanymi wyłącznie jako "do not care" conditions. Dla niekompletnych zbiorów danych zawierających brakujące wartości typu lost values uproszczona metoda rekurencyjna, oparta na własności przechodniości relacji charakterystycznej R(B) dla takich zbiorów, została wprowadzona w [27]. Z kolei w [28] została zastosowana metoda obliczania maksymalnych bloków spójnych, dla dowolnych interpretacji brakujących wartości atrybutów, oparta na relacji charakterystycznej.
W [I.3] wykazano, że niektóre z powyższych metod mogą generować bloki spójne, które nie są maksymalne, szczególnie w przypadku zbiorów dla których relacja charakterystyczna jest nieprzechodnia. Wynika to z faktu, że zastosowane w algorytmach brute force, rekurencyjnym oraz hierarchicznym metody scalania są niewystarczające. W tym przypadku poszukiwanie maksymalnych bloków spójnych wymaga zweryfikowania czy każdy nowo dodany blok nie jest podzbiorem któregoś z istniejących oraz czy żaden z istniejących bloków nie jest podzbiorem stworzonego bloku. Taka metoda scalania lokalnych bloków spójnych jest NP-trudna.

W [I.3] zaproponowano nowe podejście do obliczania maksymalnych bloków spójnych dla niekompletnych danych, z dwiema interpretacjami brakujących wartości atrybutów: lost values i "do not care" conditions. Wykazano, że otrzymane bloki spójne są maksymalne oraz podejście jest szybsze niż powszechnie stosowana metoda generowania maksymalnych bloków spójnych bazująca na relacji charakterystycznej. Jednocześnie zoptymalizowano procedurę scalania bloków, poprzez wprowadzenie dodatkowego porównywania ich długości oraz przerywanie procesu porównania już przy pierwszym niezgodnym elemencie. Sprawdzono także inne potencjalne ulepszenia, takie jak sortowanie zbioru maksymalnych bloków spójnych według długości, jak sugerowano w [29] lub sortowanie zawartości zbiorów, a następnie używanie wyszukiwania binarnego zamiast wyszukiwania liniowego w celu eliminacji podzbiorów. Niestety żadna z wymienionych metod nie okazała się skuteczna w praktyce i zamiast przyspieszyć, znacznie spowolniła obliczenia, głównie ze względu na czas poświęcony na sortowanie, pomimo zastosowania algorytmu QuickSort.
Wyniki eksperymentów, w szczególności dla zbiorów danych z brakującymi wartościami interpretowanymi jako "do not care" conditions, sugerowały, że obliczanie maksymalnych bloków spójnych może być związane z wykładniczą złożonością czasową. Aby jednak to potwierdzić, należało w pierwszej kolejności oszacować całkowitą możliwą liczbę maksymalnych bloków spójnych.

5. Oszacowanie liczby maksymalnych bloków spójnych oraz złożoności czasowej obliczania maksymalnych bloków spójnych

Problem złożoności czasowej obliczania maksymalnych bloków spójnych nie był przedmiotem badań poprzedzających publikację [I.1]. Pojawiły się sugestie, że obliczanie maksymalnych bloków spójnych może być związane z wielomianową złożonością czasową [25].
W [I.1] oszacowano całkowitą możliwą liczbę maksymalnych bloków spójnych i udowodniono, że liczba bloków może rosnąć wykładniczo zależnie od liczby atrybutów dla niekompletnych danych z brakującymi wartościami atrybutów traktowanymi jako "do not care" conditions. W tym celu definiowano specjalny zbiór, który nazwano k-galaktyką.

Niech $|A| = k$, $V_a = \{1, 2, \ldots, k\}$ dla każdego $a \in A$. Zbiór jest *k-galaktyką* jeśli $a_j(i) = j$ gdy $k \cdot (j-1) < i \leq k \cdot j$; w przeciwnym razie, $a_j(i) = *$ dla każdego $i \in \{1, 2, ..., k^2\}$ i $j \in \{1, ..., k\}$. '$\cdot$' oznacza standardową operację mnożenia arytmetycznego.

K-galaktyka ilustruje najgorszy przypadek tzw.*worst-case*, dla którego liczba maksymalnych bloków spójnych zależy wykładniczo od k i wynosi k^k.

Przykład *k-galaktyki* (k = 3) przedstawiono w Tabeli 6. Dla *k-galaktyki* (k=3) z Tabeli 6, zbiór wszystkich maksymalnych bloków spójnych $\mathscr{C}(A)$ to {{1, 4, 7}, {1, 4, 8}, {1, 4, 9}, {1, 5, 7}, {1, 5, 8}, {1, 5, 9}, {1, 6, 7}, {1, 6, 8}, {1, 6, 9}, {2, 4, 7}, {2, 4, 8}, {2, 4, 9}, {2, 5, 7}, {2, 5, 8}, {2, 5, 9}, {2, 6, 7}, {2,

6, 8}, {2, 6, 9}, {3, 4, 7}, {3, 4, 8}, {3, 4, 9}, {3, 5, 7}, {3, 5, 8}, {3, 5, 9}, {3, 6, 7}, {3, 6, 8}, {3, 6, 9}}.

Tabela 6: Przykład k-galaktyki dla k=3

Przypadek	Atrybut 1	Atrybut 2	Atrybut 3	Decyzja
1	1	*	*	1
2	2	*	*	1
3	3	*	*	1
4	*	1	*	2
5	*	2	*	2
6	*	3	*	2
7	*	*	1	3
8	*	*	2	3
9	*	*	3	3

Szacując liczbę maksymalnych bloków spójnych wykazano, że w najgorszym przypadku złożoność czasowa generowania takich bloków wynosi $O(n^n)$, gdzie n to liczba obiektów w zbiorze danych. Jednocześnie wynik ten wskazuje, że złożoność czasowa niektórych znanych algorytmów obliczania maksymalnych bloków spójnych była niedoszacowana [25, 26].
Wskazanie najgorszego przypadku obliczania maksymalnych bloków spójnych pozwoliło określić ich wpływ na wydajność systemu i wyznaczyć górną granicę zasobów wymaganych w praktycznych zastosowaniach.

6. Metoda równoległego obliczania maksymalnych bloków spójnych

Biorąc pod uwagę złożoność czasową obliczania maksymalnych bloków spójnych zaproponowano metodę ich współbieżnego tworzenia.
W [I.3] wykazano, że koncepcja sekwencyjnej aktualizacji zbioru maksymalnych bloków spójnych w oparciu o Właściwość 5 z [22] jest bardziej efektywna niż powszechnie stosowana metoda konstruowania bloków oparta na relacji charakterystycznej. Jednak scalanie bloków, którego integralną częścią jest eliminacja podzbiorów występujących w dwóch łączonych zestawach maksymalnych bloków spójnych – dotychczasowym, uzyskanym na danym etapie analizy i nowo zbudowanym dla kolejnego atrybutu – ma znaczący wpływ na ogólną wydajność. Uwzględniając fakt, że inkluzja jest relacją przechodnią, w procedurze scalania można zastosować eliminację podzbiorów w dowolnej kolejności; w szczególności połączone zestawy bloków spójnych mogą być przetwarzane parami w tym samym czasie. Bazując na wspomnianych własnościach, opracowano zupełnie nowe podejście do obliczania maksymalnych bloków spójnych, przeznaczone dla środowisk wieloprocesorowych

[I.1]. W podejściu tym, obliczenia maksymalnych bloków spójnych dla lokalnie rozważanego zestawu atrybutów oraz procedura scalania bloków, rozdzielane są na odrębne zadania, wykonywane równolegle przez pulę dostępnych procesorów. Z uwagi na fakt, że wynikowe bloki spójne muszą być globalnie maksymalne, konieczne jest wielokrotne synchronizowanie wszystkich równolegle wykonywanych zadań, tworzenie nowego zestawu zadań i ponawianie procedury scalania, aż do uzyskania jednego, wynikowego zbioru bloków.
Żadna z dostępnych w literaturze metod obliczania maksymalnych bloków spójnych [22, 25, 26], nie umożliwiała ich współbieżnego tworzenia. Efektywność proponowanego rozwiązania przyczyniła się do jego zastosowania w module eksploracji niepełnych danych systemu detekcji upadków. System FRSystem [30], tworzony we współpracy z Domem Opieki nad Osobami Starszymi w Rzeszowie, przeznaczony będzie do monitoringu osób starszych. W systemach opartych na czujnikach mogą wystąpić różnego rodzaju zakłócenia, spowodowane np. awarią zasilania lub wyczerpaniem się akumulatorów. W pracy [31] zaproponowano dwie nowe metody imputacji niekompletnych danych w oparciu o teorię zbiorów rozmytych o wartościach przedziałowych i teorię zbiorów przybliżonych, w szczególności maksymalnych bloków spójnych, oraz ich zastosowanie w systemie detekcji postawy. Opracowane hybrydowe podejście do eksploracji niekompletnych danych zapewnia stabilną wydajność pomimo wzrostu liczby brakujących danych.

7. Globalne oraz nasycone przybliżenia probabilistyczne

Opracowano dwa nowe typy przybliżeń probabilistycznych, globalne przybliżenie probabilistyczne (ang. global) [I.14] oraz nasycone przybliżenie probabilistyczne (ang. saturated) [I.12]. Przybliżenia zostały zdefiniowane dla dwóch granul, zbiorów charakterystycznych oraz uogólnionych maksymalnych bloków spójnych. Z uwagi na wykładniczą złożoność czasową obliczania przybliżeń, w praktycznych zastosowaniach posługiwano się heurystycznymi wersjami przybliżeń.
Heurystyczna wersja globalnych przybliżeń probabilistycznych bazuje na algorytmie indukcji reguł MLEM2 [4] i jest związana z parametrem prawdopodobieństwa α, $0 < \alpha \leq 1$. W zależności od parametru α przybliżenie probabilistyczne reprezentuje całe spektrum przybliżeń, może się okazać potencjalnie bardziej przydatne w eksploracji danych niż standardowe przybliżenia. Eksperymenty wykazały, że zmiana parametru α ma wpływ na poziom błędu, mierzony z zastosowaniem dziesięciokrotnej walidacji krzyżowej.
MLEM2 globalne przybliżenia probabilistyczne konceptu są tworzone z granul, które są najbardziej pasujące do konceptu oraz mają prawdopodobieństwo warunkowe większe bądź równe od zadanego parametru α. Najbardziej pasujące do konceptu X granule to zbiory charakterystyczne lub uogólnione maksymalne bloki spójne, których $|X \cap K(x)|$ lub $|X \cap Y|$, $Y \in \mathscr{C}(A)$ jest największe. $Pr(X|K(x)) = \frac{|X \cap K(x)|}{|K(x)|}$ jest prawdopodobieństwem warunkowym X dla danego zbioru charakterystycznego $K(x)$, a $Pr(X|Y) = \frac{|X \cap Y|}{|Y|}$ jest prawdopodobieństwem warunkowym X dla danego uogólnionego maksymalnego bloku spójnego.
Z kolei heurystyczna wersja nasyconego przybliżenia probabilistycznego opiera się

na selekcji granul, przy jednoczesnym nadaniu wyższego priorytetu granulom o większym prawdopodobieństwie warunkowym. Dodatkowo, jeśli przybliżenie obejmuje wszystkie przypadki konceptu, dodawanie granul jest przerywane. Koncept został nasycony.
Dla konceptu *tak*, $X = \{1, 2, 3, 4\}$ z Tabeli 3 wszystkie odrębne globalne i nasycone przybliżenia probabilistyczne bazujące na uogólnionych maksymalnych blokach spójnych są następujące:

$appr_1^{gmcb,global}(\{1, 2, 3, 4\}) = \{1, 3, 4\}$,

$appr_{0.5}^{gmcb,global}(\{1, 2, 3, 4\}) = \{1, 2, 3, 4, 6, 7\}$,

$appr_1^{gmcb,saturated}(\{1, 2, 3, 4\}) = \{1, 3, 4\}$,

$appr_{0.5}^{gmcb,saturated}(\{1, 2, 3, 4\}) = \{1, 2, 3, 4, 6\}$.

Należy zauważyć, że $appr_{0.5}^{gmcb,global}(\{1, 2, 3, 4\})$ pokrywa przypadki 6 i 7 pomimo, iż przypadki 6 i 7 nie należą do konceptu {1, 2, 3, 4}. Przypadek 7 nie znajduje się wśród nasyconych przybliżeń probabilistycznych tego konceptu.
Opracowane przybliżenia probabilistyczne były przedmiotem licznych analiz, w kontekście jakości oraz złożoności zbiorów reguł indukowanych z ich użyciem oraz istotnych różnic w proponowanych kombinacjach podejść do eksploracji niekompletnych zbiorów danych.

8. Zastosowanie globalnych oraz nasyconych przybliżeń probabilistycznych w eksploracji danych niekompletnych

Eksperymenty przeprowadzono na ośmiu zbiorach danych, pozyskanych z Repozytorium Uczenia Maszynowego Uniwersytetu Kalifornijskiego w Irvine. Dla każdego zbioru danych stworzono szablon. W szablonie 35% istniejących wartości atrybutów losowo zamieniono na lost values. Te same szablony użyto następnie do skonstruowania zbiorów danych z brakującymi wartościami interpretowanymi jako "do not care" conditions lub atrybut-wartość, zastępując znaki ? znakami * lub –. Reguły decyzji indukowano z zastosowaniem algorytmu MLEM2 [32]. W analizie jako kryterium jakości przyjęto poziom błędu obliczony jako rezultat zastosowania dziesięciokrotnej walidacji krzyżowej. Złożoność reguł oceniano biorąc pod uwagę liczbę reguł oraz całkowitą liczbę warunków w zbiorze reguł. Podejścia porównano, przy użyciu nieparametrycznego testu Friedmana, a następnie testu post-hoc (wielokrotne porównania bez założeń o rozkładzie danych oparte na sumach rang Friedmana), przy poziomie istotności wynoszącym 5%.
Przeprowadzono serię eksperymentów w których analizowano różne podejścia do eksploracji niekompletnych zbiorów danych, łącząc:

- dwie interpretacje brakujących wartości atrybutów (lost values i "do not care" conditions) z dwoma rodzajami przybliżeń probabilistycznych (globalnym i konceptowym) opartymi na zbiorach charakterystycznych, w kontekście

oceny jakości klasyfikacji reguł indukowanych dla proponowanej kombinacji [I.10, I.14];

- dwie interpretacje brakujących wartości atrybutów (lost values i "do not care" conditions) z dwoma rodzajami przybliżeń probabilistycznych (globalnym i nasyconym) bazującymi na zbiorach charakterystycznych, w kontekście oceny jakości klasyfikacji [I.12] oraz złożoności zbioru reguł [I.11];

- dwie interpretacje brakujących wartości atrybutów (lost values i "do not care" conditions) z dwoma rodzajami przybliżeń probabilistycznych (globalnym i nasyconym) bazującymi na uogólnionych maksymalnych blokach spójnych, w kontekście oceny jakości klasyfikacji [I.8] oraz złożoności zbioru reguł [I.7];

- dwie interpretacje brakujących wartości atrybutów (lost values i "do not care" conditions) z dwoma rodzajami przybliżeń probabilistycznych (globalnym i nasyconym) oraz z dwoma rodzajami granul (zbiory charakterystyczne i uogólnione maksymalne bloki spójne), w kontekście oceny jakości klasyfikacji [I.5];

- trzy interpretacje brakujących wartości atrybutów (lost values, "do not care" conditions i atrybut-koncept) z dwoma rodzajami przybliżeń probabilistycznych (globalnym i nasyconym) oraz z dwoma rodzajami granul (zbiory charakterystyczne i uogólnione maksymalne bloki spójne) w kontekście oceny jakości klasyfikacji [I.2].

Głównym celem badań było ustalenie, które z proponowanych podejść do eksploracji danych jest najlepsze. Najważniejsze wnioski z przeprowadzonych badań to:

- istnieją istotne różnice między analizowanymi podejściami, jednak różnica pomiędzy zastosowanymi przybliżeniami probabilistycznymi nie jest znacząca. Statystycznie znacząca jest różnica pomiędzy użytymi interpretacjami brakujących wartości atrybutów;

- jakość klasyfikacji reguł indukowanych z globalnych przybliżeń probabilistycznych jest wyższa niż reguł indukowanych z konceptowych przybliżeń probabilistycznych, ale wyłącznie dla danych z brakującymi wartościami atrybutów interpretowanymi jako "do not care" conditions;

- reguły indukowane ze zbiorów danych z brakującymi wartościami interpretowanymi jako atrybut-wartość są najprostsze. Liczba reguł oraz całkowita liczba warunków w zbiorze reguł jest mniejsza niż w zbiorach reguł indukowanych z danych z pozostałymi interpretacjami brakujących wartości atrybutów;

- najbardziej złożone, pod względem liczby reguł oraz liczby warunków w zbiorze reguł są reguły indukowane ze zbiorów danych z brakującymi wartościami atrybutów interpretowanymi jako lost values;

- dla danych z brakującymi wartościami interpretowanymi jako lost values, poziom błędu związany z przybliżeniami globalnymi jest taki sam, jak poziom błędu związany z przybliżeniami nasyconymi. Wynika to z faktu, że dla większości analizowanych zbiorów danych maksymalne bloki spójne są singletonami (zestawami zawierającymi tylko jeden element).

Uzupełnieniem powyższych badań jest artykuł P.Clark, J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek, *Mining Incomplete Data Using Global and Saturated Probabilistic Approximations Based on Characteristic Sets and Maximal Consistent Blocks* zgłoszony w sierpniu br. do czasopisma Information Sciences. Na dzień dzisiejszy artykuł posiada trzy pozytywne recenzje.

Omówione powyżej wyniki dotyczą eksperymentów przeprowadzonych na zbiorach danych z 35% brakujących wartości atrybutów. Niezależnie, podjęto również analizę danych z wieloma brakującymi wartościami atrybutów, w kontekście oceny jakości zastosowanych podejść do eksploracji danych [I.9] oraz złożoności zbioru reguł [I.4]. W toku analizy użyto czterech różnych sposobów generowania reguł decyzji, łącząc dwie interpretacje brakujących wartości atrybutów (lost values, "do not care" conditions) z dwoma rodzajami przybliżeń probabilistycznych (globalnym i nasyconym) opartymi na zbiorach charakterystycznych. Analogicznie jak wcześniej, jako kryterium jakości przyjęto poziom błędu obliczony jako rezultat zastosowania dziesięciokrotnej walidacji krzyżowej. Złożoność reguł oceniano biorąc pod uwagę liczbę reguł oraz całkowitą liczbę warunków w zbiorze reguł. Wyniki eksperymentów dla czterech metod indukcji reguł porównano za pomocą testu sumy rang Friedmana oraz testu wielokrotnych porównań post hoc (poziom istotności 5%). Wykazano, że istnieją istotne różnice pomiędzy proponowanymi podejściami. Zastosowanie interpretacji brakujących wartości atrybutów jako „do not care" conditions zapewnia prostsze zestawy reguł niż użycie interpretacji lost values. Różnica pomiędzy użyciem obu typów przybliżeń probabilistycznych, globalnego i nasyconego, nie jest statystycznie znacząca. Nie można jednak wskazać najlepszego podejścia do eksploracji danych. W przypadku niekompletnego zbioru danych z wieloma brakującymi wartościami atrybutów najlepsze podejście do eksploracji danych należy wybrać przeprowadzając eksperymenty z uwzględnieniem wszystkich czterech możliwości.

3.4 Tytuł i zakres osiągnięcia II

Podstawę wniosku o przeprowadzenie postępowania habilitacyjnego stanowi osiągnięcie II pt. **Rozwój metod dyskretyzacji danych numerycznych**.
Na przedmiotowe osiągnięcie II składa się 5 prac, z czego trzy pozycje z listy Journal Citation Reports (JCR), jedna w materiałach konferencyjnych międzynarodowej konferencji i jedna to rozdział w książce. Sumaryczny wskaźnik IF wymienionych, zgodnie z rokiem publikacji, prac wynosi 6.73, a liczba punktów MNiSW/MEiN wynosi 195. W ramach osiągnięcia II:

1. Wykazano, że redukcja atrybutów w połączeniu z dyskretyzacją atrybutów numerycznych negatywnie wpływa na jakość klasyfikacji;

2. Zaproponowano selekcję atrybutów bazującą na redukcji atrybutów numerycznych podczas dyskretyzacji;

3. Oceniono wpływ łączenia przedziałów dyskretyzacji na jakość klasyfikacji;

4. Oceniono wpływu technik dyskretyzacji bazujących na entropii na jakość klasyfikacji oraz złożoność modelu uczenia.

3.5 Lista prac wchodzących w zakres osiągnięcia II

[II.1] J.W.Grzymała-Busse, Z.S.Hippe, T.Mroczek (2019) *Reduced Data Sets and Entropy-Based Discretization*. Entropy 21(11):1051. IF(2019): 2.494. doi: 10.3390/e21111051

[II.2] J.W.Grzymała-Busse, T.Mroczek (2018) *Attribute Selection Based on Reduction of Numerical Attributes During Discretization*. In: Stańczyk, U., Zielosko, B., Jain, L. (eds) Advances in Feature Selection for Data and Pattern Recognition. Intelligent Systems Reference Library 138:13–24. Springer, Cham. doi: 978-3-319-67588-6_2

[II.3] J.W.Grzymała-Busse, T.Mroczek (2018) *Merging of Numerical Intervals in Entropy-Based Discretization*. Entropy 20(11):880. IF(2018): 2.419. doi: 10.3390/e20110880

[II.4] J.W.Grzymała-Busse, T.Mroczek (2016) *A Comparison of Four Approaches to Discretization Based on Entropy*. Entropy 18(3):69. IF(2016): 1.821. doi: 10.3390/e18030069

[II.5] J.W.Grzymała-Busse, T.Mroczek (2015) *A Comparison of Two Approaches to Discretization: Multiple Scanning and C4.5*. In: Kryszkiewicz, M., Bandyopadhyay, S., Rybinski, H., Pal, S. (eds) Pattern Recognition and Machine Intelligence. PReMI 2015. Lecture Notes in Computer Science 9124:301–310. Springer, Cham. doi: 10.1007/978-3-319-19941-2_5

3.6 Omówienie osiągnięcia II

3.6.1 Wprowadzenie

Eksploracja numerycznych zbiorów danych wymaga, w niektórych przypadkach, dodatkowego etapu zwanego *dyskretyzacją*. Dyskretyzacja to proces przekształcania wartości liczbowych na przedziały. Ciągły zakres wartości atrybutów jest dzielony na kilka punktów odcięcia (ang. cut-points), które determinują granice przedziałów. Niniejsze badania koncentrują się wokół rozwoju metod dyskretyzacji opartych na entropii. Dyskretyzacja oparta na entropii warunkowej konceptu danego atrybutu (cechy) uważana jest za jedną z najskuteczniejszych technik dyskretyzacji [33, 34, 35, 36, 37, 38, 39, 40, 41].

Nowa technika dyskretyzacji, zwana skanowaniem wielokrotnym (ang. Multiple Scanning), wprowadzona w [37], okazała się bardzo skuteczna w połączeniu z indukcją reguł i systemem klasyfikacji LERS [42]. W metodzie skanowania wielokrotnego:

- cały zbiór atrybutów jest skanowany, dla każdego atrybutu wyznaczany jest najlepszy punkt odcięcia w oparciu o minimalną entropię warunkową. Oryginalny zbiór danych jest dzielony na podtabele wyznaczone przez najlepsze punkty odcięcia,
- jeżeli ustalona liczba skanów nie została osiągnięta, częściowo zdyskretyzowane atrybuty skanowane są ponownie i wybierane są najlepsze odpowiadające im punkty odcięcia. Obliczany jest najlepszy punkt odcięcia dla każdej podtabeli. Wybierany jest najlepszy punkt odcięcia spośród wszystkich podtabel,
- jeżeli ustalona liczba skanów została osiągnięta, a zbiór danych wymaga dalszej dyskretyzacji, dla pozostałych podtabel stosuje się technikę atrybutu dominującego (ang. Dominant Attribute) [35, 36],
- proces jest zatrzymywany, gdy kryterium spójności [33], bazujące na teorii zbiorów przybliżonych [6, 7], osiągnie wartość 1. Oznacza to, że zdyskretyzowany zbiór jest spójny.

Ostatnim etapem dyskretyzacji jest łączenie przedziałów w celu redukcji ich ilości przy jednoczesnym zachowaniu spójności zbioru. W badaniach stosowano łączenie właściwe (ang. proper merging) polegające na łączeniu dowolnych dwóch sąsiednich przedziałów zdyskretyzowanego atrybutu jeśli nowy przedział, będący wynikiem połączenia, nie zmniejsza poziomu spójności zbioru.

3.6.2 Omówienie celów naukowych ww. prac oraz osiągniętych wyników

Celem naukowym osiągnięcia II jest rozwój metod dyskretyzacji danych numerycznych opartych na entropii, w szczególności ocena wpływu różnych technik dyskretyzacji, redukcji przedziałów oraz atrybutów na jakość procesu uczenia oraz złożoność modeli uczenia.
Proces dyskretyzacji może znacząco wpłynąć na efektywność i jakość procesu uczenia. Zastąpienie wielu wartości atrybutu numerycznego niewielką liczbą przedziałów, wynikających z dyskretyzacji, może poprawić wydajność oraz dokładność klasyfikacji, zwłaszcza w przypadku nie w pełni poprawnych danych uczących. Ważny jest wybór odpowiedniej techniki dyskretyzacji. Z tego powodu podjęto badania, których celem jest ocena różnych podejść do procesu dyskretyzacji i wskazanie najlepszych wariantów.
Niektóre podejścia omawiane w ramach osiągnięcia II, jak połączenie redukcji zbioru atrybutów i dyskretyzacji atrybutów numerycznych, były już przedmiotem badań [43, 44, 45, 46, 47]. W [46] przeprowadzono eksperymenty na dziesięciu bazach z atrybutami numerycznymi, w których jako klasyfikatorów użyto SVM [48] i C4.5

[49]. Jednak przedstawione wyniki testu Friedmana nie są jednoznaczne, więc korzyści wynikające z redukcji nie są jasne. W [47] algorytmy genetyczne i sztuczne sieci neuronowe użyto do: dyskretyzacji, redukcji cech i przewidywania indeksu cen akcji. Trudno ocenić, w jaki sposób redukcja cech przyczyniła się do wyników końcowych. Prace [50, 44, 45] omawiają redukty połączone z dyskretyzacją. Brakowało jednak jednoznacznych wyników potwierdzających wpływ redukcji w połączeniu z dyskretyzacją na jakość klasyfikacji. Z tego powodu podjęto badania w tym kierunku. Jednocześnie zaproponowano metodę selekcji zdyskretyzowanych atrybutów, która nie była przedmiotem wcześniejszych badań, oceniając także jej wpływ na proces klasyfikacji.

1. Wpływ redukcji atrybutów w połączeniu z dyskretyzacją na jakość klasyfikacji

Idea reduktu pozwala wybrać minimalne podzbiory atrybutów zachowujących charakterystykę całego zbioru atrybutów. Należy zauważyć, że każdy algorytm znajdowania wszystkich reduktów ma wykładniczą złożoność czasową. W zastosowaniach praktycznych konieczne jest podejście heurystyczne. W ocenie wpływu redukcji w połączeniu z dyskretyzacją na jakość klasyfikacji przeprowadzono eksperymenty z użyciem prawych i lewych reduktów, które tworzono bazując na teorii zbiorów przybliżonych [II.1]. Użyto trzynastu zbiorów danych z atrybutami numerycznymi. Zbiory te, poza zbiorem *bankructwa*, zostały zapożyczone z repozytorium uczenia maszynowego przechowywanego na Uniwersytecie Kalifornijskim w Irvine. Zbiór danych *bankructwa* jest znanym zbiorem danych, używanym przez Altmana do przewidywania upadłości przedsiębiorstw [51]. W procesie dyskretyzacji zastosowano globalne wersje metod: równa szerokość przedziałów i równa częstość przedziałów, opartych na entropii [33]. Dla każdego zbioru danych rozważono trzy jego wersje:

- oryginalny (niezredukowany) zdyskretyzowany zbiór danych,
- zbiór danych oparty na lewym redukcie uprzednio zdyskretyzowanego zbioru danych,
- zbiór danych oparty na prawym redukcie uprzednio zdyskretyzowanego zbioru danych.

Uzyskane zbiory wprowadzono do systemu generującego drzewa decyzyjne C4.5 [49]. Poziom błędu obliczono przy użyciu wewnętrznego mechanizmu dziesięciokrotnej walidacji krzyżowej systemu C4.5. W eksperymentach, dla reduktów lewych i prawych, wykluczono wewnętrzny mechanizm dyskretyzacji C4.5, ponieważ w tym przypadku zbiory danych zostały zdyskretyzowane metodami globalnymi. Wyniki analizowano za pomocą testu sumy rang Friedmana oraz testu wielokrotnych porównań, z poziomem istotności wynoszącym 5%. Wykazano, że redukcja zdyskretyzowanych zbiorów zwiększa poziom błędu. Dodatkowo porównano złożoność wygenerowanych drzew decyzyjnych. Drzewa decyzyjne wygenerowane ze zredukowanych zbiorów nie okazały się prostsze niż drzewa decyzyjne wygenerowane z niezredukowanych zbiorów danych.

2. Selekcja atrybutów bazująca na redukcji atrybutów numerycznych podczas dyskretyzacji

Podczas dyskretyzacji zbiorów danych z atrybutami numerycznymi niektóre atrybuty mogą zostać zredukowane, gdy cała dziedzina atrybutu jest odwzorowywana w jednym przedziale. Nowy zbiór danych jest wówczas tworzony poprzez usunięcie atrybutów z pierwotnego zbioru danych. Takie zbiory danych nazwane zostały zbiorami zredukowanymi. Głównym celem badań było porównanie jakości klasyfikacji zbiorów danych z atrybutami typu numerycznego, oryginalnych i zredukowanych, przy użyciu systemu generowania drzew decyzji C4.5 [II.2]. W momencie publikacji wyników, w dostępnej literaturze, nie wzmiankowano o przeprowadzeniu podobnych badań.
Badania zostały przeprowadzone na piętnastu zbiorach danych z atrybutami numerycznymi, zapożyczonymi z repozytorium uczenia maszynowego przechowywanego na Uniwersytecie Kalifornijskim w Irvine. Zastosowano metodę dyskretyzacji atrybutu dominującego [35, 36]. Poziom błędu klasyfikacji obliczono stosując metodę dziesięciokrotnej walidacji krzyżowej. Wykazano, że poziom błędu jest istotnie większy (5% poziom istotności, test dwustronny) dla zredukowanych zbiorów danych. Jednak drzewa decyzyjne generowane na podstawie zredukowanych zbiorów danych są znacznie prostsze niż drzewa decyzyjne generowane na podstawie oryginalnych zbiorów danych. Złożoność drzew decyzyjnych mierzono parametrami: głębokość i rozmiar.

3. Wpływ łączenia przedziałów dyskretyzacji na jakość klasyfikacji

Wstępne wyniki dotyczące wpływu łączenia przedziałów na jakość klasyfikacji, w których wnioskodawca brał udział, wykazały, że nie ma uniwersalnej metody – różnica w jakości klasyfikacji sześciu analizowanych metod dyskretyzacji, oceniana dla dziewięciu zestawów danych, nie była znacząca statystycznie [52].
W [II.4, II.5] wykazano, że technika dyskretyzacji skanowania wielokrotnego jest istotnie lepsza niż pozostałe znane techniki: atrybutu dominującego oraz zglobalizowane wersje techniki równej szerokości przedziałów oraz równej częstości przedziałów. Z tego powodu w ocenie wpływu łączenia przedziałów na jakość klasyfikacji do dyskretyzacji zastosowano metodę skanowania wielokrotnego [II.3]. Zastosowano trzy podejścia do łączenia przedziałów w ostatnim etapie dyskretyzacji:

- brak łączenia,
- właściwe łączenie w oparciu o minimalną wartość entropii warunkowej, oraz
- właściwe łączenie w oparciu o maksymalną wartość entropii warunkowej.

Zdyskretyzowane zbiory danych zostały użyte do indukcji drzew decyzji metodą C4.5 [49]. Eksperymenty przeprowadzono z użyciem siedemnastu zbiorów danych z atrybutami numerycznymi, zapożyczonymi z repozytorium uczenia maszynowego przechowywanego na Uniwersytecie Kalifornijskim w Irvine oraz z [51]. Poziomy błędów obliczono za pomocą dziesięciokrotnej walidacji krzyżowej i porównano za

pomocą testu sumy rang Friedmana w połączeniu z testem wielokrotnych porównań, przy poziomie istotności wynoszącym 5%.
Wykazano, że różnice pomiędzy wszystkimi trzema podejściami do łączenia przedziałów są nieistotne statystycznie. Nie można zatem wskazać najlepszego podejścia do ostatniego etapu dyskretyzacji – łączenia przedziałów. Stąd kolejnym celem było sprawdzenie różnicy między wszystkimi trzema podejściami dla konkretnego zbioru danych. Przeprowadzono szeroko zakrojone eksperymenty, powtarzając 30 dziesięciokrotnych walidacji krzyżowych dla każdego zbioru danych rejestrując średnie i standardowe odchylenia, aby następnie zastosować standardowy test różnic między wartościami średnimi wielu serii pomiarowych. Wykazano, że istnieją statystycznie wysoce istotne różnice (z poziomem istotności 1%) pomiędzy tymi trzema podejściami do łączenia, w zależności od zbioru danych. Należy zatem, podczas dyskretyzacji zbiorów danych z atrybutami numerycznymi za pomocą metody skanowania wielokrotnego, zweryfikować różne podejścia łączenia przedziałów i zastosować najlepszy wybór.

4. Ocena wpływu technik dyskretyzacji bazujących na entropii na jakość klasyfikacji oraz złożoność modelu uczenia

W ocenie jakości technik dyskretyzacji przeprowadzono porównanie:

- dwóch metod dyskretyzacji: techniki dyskretyzacji skanowania wielokrotnego połączonej z systemem klasyfikacji C4.5 oraz techniki dyskretyzacji wewnętrznej występującej w systemie C4.5 [II.5],
- czterech metod dyskretyzacji: techniki dyskretyzacji wewnętrznej występującej w systemie C4.5 i trzech technik globalnych: techniki dyskretyzacji skanowania wielokrotnego, techniki równej szerokości przedziałów oraz równej częstości przedziałów [II.4]

przy użyciu dwóch kryteriów: poziomu błędu ocenianego metodą dziesięciokrotnej walidacji krzyżowej oraz rozmiaru drzewa decyzyjnego wygenerowanego przez system C4.5.
Eksperymenty przeprowadzono z użyciem siedemnastu zbiorów danych w [II.4] oraz dwunastu zbiorów danych w [II.5] z atrybutami numerycznymi. Zbiory te, poza zbiorem *bankructwa* [51], zostały pobrane z repozytorium uczenia maszynowego przechowywanego na Uniwersytecie Kalifornijskim w Irvine. Do analizy wyników eksperymentów użyto testu sumy rang Friedmana w połączeniu z testem porównań wielokrotnych, z poziomem istotności 5% w [II.4] oraz testu kolejności par Wilcoxona, z poziomem istotności 5% w [II.5].
Wykazano, że technika dyskretyzacji skanowania wielokrotnego jest istotnie lepsza niż inne metody dyskretyzacji użyte w analizie: dyskretyzacja wewnętrzna zastosowana w systemie C4.5, dyskretyzacja atrybutem dominującym, zglobalizowane metody: równa szerokość przedziału i równa częstość w przedziałach, pod względem poziomu błędów obliczonego przez dziesięciokrotną walidację krzyżową (test dwustronny, poziom istotności 5%). Dodatkowo technika dyskretyzacji skanowania wielokrotnego jest istotnie lepsza niż wprowadzony przez Fayyada i Irani wariant

dyskretyzacji oparty na entropii warunkowej, zwany atrybutem dominującym. Jednocześnie drzewa decyzyjne generowane na podstawie danych zdyskretyzowanych metodą skanowania wielokrotnego są prostsze, biorąc pod uwagę ich głębokość i rozmiar, w porównaniu z drzewami decyzyjnymi generowanymi bezpośrednio przez C4.5 dla tych samych zbiorów danych. Różnice pomiędzy wydajnością C4.5, zglobalizowanymi wersjami metod dyskretyzacji o jednakowej szerokości przedziałów i równej częstości przedziałów oraz skanowaniem wielokrotnym są nieistotne statystycznie.

4 Istotna aktywność naukowa realizowana w więcej niż jednej uczelni, instytucji naukowej lub instytucji kultury, w szczególności zagranicznej

Znacząca część prac związanych z rozwojem metod eksploracji danych niekompletnych oraz rozwojem metod dyskretyzacji danych numerycznych realizowana była we współpracy z Katedrą Elektrotechniki i Informatyki Uniwersytetu w Kansas, USA. Wynikiem wspólnych prac są publikacje wchodzące w zakres osiągnięcia I oraz osiągnięcia II, będących przedmiotem niniejszego wniosku.
Efektem aktywności naukowej wnioskodawcy, realizowanej we współpracy z inną uczelnią – Uniwersytetem Rzeszowskim oraz Domem Opieki nad Osobami Starszymi w Rzeszowie, jest także hybrydowy system o nazwie FRSystem [30]. System stosuje metody rozmyte i przybliżone w analizie detekcji upadków. Rozwinięciem systemu są nowe metody imputacji niekompletnych danych, zapewniające stabilne działanie w przypadku wystąpienia zakłóceń spowodowanych np. awarią zasilania lub wyczerpaniem się akumulatorów w urządzeniach monitorujących [31].
Innym udokumentowanym rezultatem współpracy międzynarodowej z ośrodkami naukowymi we Francji (Fastlite, Amplitude, Centre National de Recherche Scientifique) oraz Niemczech (JenLab GMBH, Friedrich-Alexander-Universität Erlangen-Nürnberg) jest grant naukowy *Versatile infrared light source for advanced illumination*, który uzyskał finansowanie w ramach programu Horyzont Europa. Celem projektu jest opracowanie nowych, opłacalnych i uniwersalnych źródeł światła. Natomiast rolą wnioskodawcy jest opracowanie skutecznej metodyki automatycznego rozpoznawania obrazów zmian melanocytowych wykonanych z użyciem nowego źródła światła.
Ponadto, w zakresie zwiększania wpływu badań i rozwoju DeepTech na dobrobyt ludzi i planety, wnioskodawca uczestniczy w projekcie *DeepTech in Higher Education Institutions and Ecosystems through Entrepreneurial Education+ (SFF.DeepT+)* finansowanym z środków EIT HEI Initiative, będącej organem UE i integralną częścią programu ramowego Horyzont Europa. Projekt realizowany jest w ramach współpracy międzynarodowej. Konsorcjum tworzą przedstawiciele licznych i uznanych w świecie ośrodków: Universidade de Aveiro, Portugalia (lider), VIA University College (Dania), Martin Luther University Halle-Wittenberg (Niemcy), The Queen's University of Belfast (Wlk. Brytania), Edinburgh Napier University (Wlk. Brytania), Dundalk Institute of Technology (Irlandia), Strascheg

Center for Entrepreneurship (Niemcy), Ikonomicheski Universitet – Varna (Bułgaria), Fundació Tecnocampus Mataró-Maresme (Hiszpania), Josip Juraj Strossmayer University of Osijek (Chorwacja). Wnioskodawca jest członkiem zespołu realizującego projekt, specjalistą w zakresie sztucznej inteligencji.
Niezależnie, w ramach badań o charakterze interdyscyplinarnym, wnioskodawca nawiązał współpracę naukową z Katedrą Przedsiębiorczości i Innowacji Społecznych Uniwersytetu Pedagogicznego w Krakowie oraz Katedrą Finansów, Bankowości i Rachunkowości Politechniki Rzeszowskiej. Rezultaty współpracy zostały omówione poniżej.

Współpraca interdyscyplinarna

Od 2014 r. wnioskodawca współpracuje z zespołem ekonomistów nad zastosowaniem metod uczenia maszynowego w eksploracji danych w kontekście różnych problemów ekonomicznych m.in wskazania zestawu najważniejszych czynników wpływających na przedsiębiorczość w prywatnej służbie zdrowia w Polsce, zdefiniowania optymalnego rozmiaru sektora finansów publicznych z punktu widzenia gospodarek państw UE, czy zdiagnozowania poziomu kompetencji przedsiębiorczych studentów. Niekonwencjonalne podejście do analizy kompetencji zostało dostrzeżone przez naukowców z University of Ljubljana i zaowocowało zaproszeniem do międzynarodowej współpracy. Efektem wspólnie prowadzonych badań jest artykuł zgłoszony do czasopisma Education Sciences.
Opublikowanymi rezultatami współpracy interdyscyplinarnej są:

- T.Mroczek, A.Kaszuba-Perz, M.Czyżewska (2022) *Assessment of Young People's Attitudes Towards Sustainable Entrepreneurship with Machine Learning Techniques Use*. Horizons of Politics 13(42):187–202. doi: 10.35765/hp.2214

- T.Mroczek, A.Kaszuba-Perz, M.Czyżewska (2021) *Pro-social Attitudes of the Young Generation Planning their own Business in the Time of COVID-19 Pandemic – A Decision Tree Approach; Innovation management and information technology impact on global economy in the era of pandemic*. In: Soliman, K.S. (ed) Proceedings of the 37th International Business Information Management Association Conference pp. 11164

- M.Czyżewska, T.Mroczek (2020) *Data Mining in Entrepreneurial Competencies Diagnosis*. Education Sciences 10(8):196. doi: 10.3390/educsci10080196

- T.Skica, T.Mroczek, M. Leśniowska-Gontarz (2019) *The impact of selected factors on new business formation in the private healthcare sector*. International Entrepreneurship and Management Journal 15(1):307–320

- T.Skica, T.Mroczek, J.Rodzinka (2019) *Impact of regional diversification in the size of the public finance sector on the EU countries economies*. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu 63(2)2:65–80

- T.Skica, J.Rodzinka, T.Mroczek (2018) *Application of probabilistic inference in defining impact of general government sector's size on economy and determining size of sector by economy*. Finansowy Kwartalnik Internetowy e-Finanse 14(1):1–11
- T.Skica, J.Rodzinka, T.Mroczek (2016) *Selection of relevant variables identifying the relationship between the general government sector size and the economy*. Modern Management Review 23(3):131–167
- T.Skica, J.Rodzinka, T.Mroczek (2015) *Data mining approach to determine the relationships between the economy and the general government sector size*. e-Finanse 11(3):1–21
- M.Czyżewska, T.Mroczek (2014) *Bayesian Approach to the Process of Identification of the Determinants of Innovativeness*. e-Finanse 10(2):44–56

5 Osiągnięcia dydaktyczne, organizacyjne oraz popularyzujące naukę

5.1 Działalność dydaktyczna

Kształcenie kadry

Promotor 111 prac inżynierskich oraz 23 prac magisterskich. W 2019 r. pod kierunkiem wnioskodawcy powstała praca magisterska Marka Jędryki nt. *Predykcja notowań akcji spółek indeksu WIG20 na podstawie informacji prasowych*, która została wyróżniona w siódmej edycji konkursu na najlepsze prace licencjackie, magisterskie i doktorskie o nagrodę Prezesa Zarządu Giełdy Papierów Wartościowych w Warszawie.

Prace magisterskie:

1. Michał Gudyka – Analiza aplikacji odzyskiwania danych w systemach Windows i Linux
2. Iwona Lalak – Optymalizacja zarządzania projektami w firmie na przykładzie rozwoju aplikacji Top Projekty
3. Marcin Stelmach – Porównanie i analiza dostępnych na rynku narzędzi do automatyzacji kompilacji, testowania i wdrażania oprogramowania
4. Dominika Marchlewska – Optymalizacja funkcjonowania przedsiębiorstwa poprzez wdrożenie systemu zintegrowanego klasy ERP
5. Konrad Krężel – Optymalizacja procesu aktualizacji danych osobowych
6. Krzysztof Borowy – Wpływ dyskretyzacji atrybutów ciągłych na jakość klasyfikacji

7. Bartłomiej Kida – Analiza porównawcza systemów bazodanowych MySQL i PostgreSQL w praktycznym zastosowaniu
8. Marek Jędryka – Predykcja notowań akcji spółek indeksu WIG20 na podstawie informacji prasowych
9. Mateusz Pałka – Optymalizacja bezpieczeństwa infrastruktury informatycznej firmy
10. Michał Nabożny – Optymalizacja procesu wersjonowania aplikacji na przykładzie firmy Sagitum
11. Jakub Kuśnierz – Centralne systemy uwierzytelniania - przykłady rozszerzeń
12. Kinga Pachla – Systemy rekomendacyjne
13. Piotr Stachaczyński – Dobór optymalnego algorytmu uczenia ze wzmocnieniem do nauki poruszania się postaci wyposażonej w fizykę
14. Mateusz Mazurek – Wpływ selekcji cech na skuteczność klasyfikacji
15. Aleksander Ligęza – Badanie wpływu zastosowania mikroserwisów na wydajność i odporność aplikacji internetowej
16. Damian Olchawa – Optymalizacja zasad bezpieczeństwa IT w firmie o profilu handlowym
17. Daniel Belcik – Palo Alto Network. Bezpieczeństwo w sieci
18. Damian Lubera – Wpływ redukcji atrybutów na jakość klasyfikacji
19. Dominik Rogosz – Badanie struktur ekranowych sklepów internetowych
20. Marcin Podolak – Ewaluacja rozwiązań do tworzenia kopii zapasowych i odzyskiwania danych dedykowanych dla firmy o profilu produkcyjnym
21. Paweł Sroczyk – Inteligentna analiza danych cechująca się brakującymi wartościami decyzji
22. Radosław Kamiński – Analiza wpływu redukcji wymiarowości na jakość klasyfikacji
23. Eryk Trojanowski – Badanie skuteczności narzędzi do tworzenia aplikacji internetowych

Prace inżynierskie:

1. Rafał Solarski – Serwis internetowy Sfera Dźwięków
2. Wojciech Marek – Zdecentralizowany Blockchain dla własnej kryptowaluty
3. Łukasz Kuznecki – Aplikacja e-commerce artykułów ozdobnych

4. Nazarii Kurash – Projekt i implementacja aplikacji Lego Store
5. Jakub Hajkowicz – System zarządzania nieobecnościami i urlopami pracowników
6. Kamil Piękoś – Projekt i implementacja aplikacji Unitravel do zamawiania transportu
7. Michał Kycia – Aplikacja internetowa - VegaFruit
8. Aleksander Witko – Yogani - aplikacja pomagająca w nauce jogi
9. Maciej Weber – Aplikacja do organizacji grupowych wydatków
10. Maciej Pelc – Projekt i implementacja aplikacji Unreal Home do sterowania urządzeniami codziennego użytku
11. Krzysztof Socha – Aplikacja webowa umożliwiająca integrację osób o wspólnych zainteresowaniach
12. Weronika Górska – Eyetrackingowa aplikacja rejestracji i analizy atencji do zastosowań w systemach HMI
13. Hubert Kobos – Dedykowany system sprzedaży kruszyw
14. Dominik Bednarz – Aplikacja internetowa - Study.me
15. Dawid Mielniczek – Aplikacja internetowa przeznaczona do integracji społeczności aktywnej fizycznie
16. Mateusz Hajder – System elektronicznej weryfikacji dyplomów - eDyplom
17. Tomasz Ożóg – Narzędzie webowe KZK Kreator zestawów komputerowych
18. Konrad Marciszewski – Projekt lokalnej sieci komputerowej dla średniej wielkości firmy.
19. Dorota Styś – Aplikacja do szacowania wartości nieruchomości z zastosowaniem metod sztucznej inteligencji
20. Aleksandra Krusiec – Aplikacja mobilna W bunkry! Edycja: Helskie fortyfikacje
21. Jakub Bednarski – Inteligentny system kontroli dostępu dla budynków wielorodzinnych
22. Krzysztof Babiuch – Zintegrowane środowisko do zarządzania przepływem realizowanych dokumentacji projektów
23. Ewa Basara – Serwis internetowy przedsiębiorstwa stolarskiego BASMEB
24. Krzysztof Bigos – Aplikacja wspomagająca proces zamawiania karnetu na codzienne posiłki

25. Marek Kluk – Projekt elektronicznej książeczki zdrowia
26. Daniel Młynarski – Projekt i implementacja systemu internetowego rozwijającego umiejętność szybkiego czytania
27. Kateryna Antoniuk – Stworzenie uniwersalnego internet serwisu do sprzedaży towarów z możliwością dynamicznego rozszerzania treści
28. Przemysław Szmyd – Identyfikacja zagrożeń oraz wdrożenie zabezpieczeń danych w sieciach komputerowych
29. Daniel Woś – Zabezpieczenia i stabilność w sieciach lokalnych
30. Kamil Kotulak – Responsywny serwis salonu kosmetyczno-podologicznego Odpicowane w Jaśle
31. Piotr Płonka – Projekt i implementacja gry komputerowej typu RPG Jura Forest Lost Tales
32. Adrian Rosnowski – Projekt i implementacja serwisu internetowego firmy handlowo-usługowej DAKK Krzysztof Duliński
33. Kamil Sagan – Projekt i implementacja szablonu nowoczesnego serwisu internetowego
34. Wojciech Siuzdak – Projekt i implementacja aplikacji webowej password manager
35. Mateusz Zając – Projekt i Implementacja Aplikacji Piłkarskiej
36. Nazar Prokudin – Aplikacja mobilna Ping serwis do wynajęcia mieszkań
37. Andrii Yakovyna – Aplikacja webowa Fitness
38. Michał Latos – Serwis internetowy pozwalający na uruchomienie bloga internetowego
39. Szymon Wiktor – Serwis internetowy do zarządzania służbami dla Świadków Jehowy
40. Konrad Przytuła – Aplikacja internetowa do umawiania i odbywania wizyt lekarskich za pośrednictwem wideoczatu
41. Piotr Skwara – Gra wyścigowa w środowisku Unity 3d z wykorzystaniem technologii VR
42. Adam Arciszewski – Aplikacja internetowa do zarządzania bazą produktów spożywczych oraz przepisów kucharskich
43. Beata Dudek – Aplikacja webowa służąca do analizy składu kosmetyków.
44. Hubert Kogut – Aplikacja webowa BackOffice służąca do zarządzania towarami wraz z aplikacją webową sklepu internetowego

45. Bartłomiej Rzeszut – Gra ekonomiczna z elementami gry typu clicker
46. Filip Mika – Audyt bezpieczeństwa sieci w średnim przedsiębiorstwie
47. Patryk Skoczylas – Projekt i implementacja aplikacji internetowej służącej do rezerwacji transportu
48. Bartosz Ochędowski – YourScale aplikacja służąca do tworzenia dynamicznego bilansu kalorycznego
49. Adrian Chrobak – System zarządzania zasobami ludzkimi
50. Damian Oliwa – Aplikacja internetowa jako system doradzający wybór komponentów do komputera PC
51. Tomasz Kalicki – Projekt i implementacja zręcznościowej gry komputerowej
52. Monika Drapała – Aplikacja mobilna "Być w formie"
53. Karol Pitra – Inteligentny magazyn
54. Mykhailo Luzhetskyi – Strona internetowa dla branży nieruchomości
55. Michał Ciężczak – System zarządzania licencjami oprogramowania klientów
56. Katarzyna Matejkowska – Aplikacja mobilna Mam się w co ubrać
57. Dominik Mądro – Aplikacja internetowa pomagająca w zarządzaniu domowymi finansami
58. Joanna Kuta – Aplikacja mobilna Fit Storage na system Android
59. Piotr Gajdek – System do fakturowania
60. Jakub Nowak – Automatyzacja procesu obsługi klienta
61. Karol Szeliga – Zarządzanie inteligentnym domem za pomocą technologii mobilnych
62. Kinga Wywrot – System internetowy do nauki języków programowania
63. Miłosz Winnicki – Aplikacja wspomagająca zarządzanie projektami
64. Maksymilian Matuła – Inteligentne Lustro
65. Agnieszka Zaguła – Aplikacja internetowa do zarządzania transportem przedsiębiorstwa dla produkcyjnego
66. Rafał Wywrot – Inteligentny serwis dla warsztatu samochodowego
67. Grzegorz Wójcik – System obsługi klienta firmy Arkonsoft
68. Dariusz Piechota – System zarządzania salonem fryzjerskim

69. Kamil Szczerba – Projekt sieci lokalnej dla firmy średniej wielkości z uwzględnieniem podstawowych zabezpieczeń.
70. Sławomir Lech – Intranet - System monitorujący czas pracy pracowników
71. Marcin Makara – Wizualizacja oświetlenia Wieży Katedralnej w Przemyślu
72. Marcin Urban – Aplikacja do zarządzania lokalną bazą filmów na komputery osobiste z systemem Windows
73. Przemysław Czachor – Aplikacja internetowa do zarządzania projektami
74. Michał Borowiec – HelpDesk - system do zarządzania problemami
75. Sebastian Białek – System CRM do zarządzania małą firmą
76. Wacław Golas – Serwis ogłoszeniowy dla budownictwa
77. Michał Jachyra – System zarządzania czasem pracy
78. Damian Szelest – System monitoringu wizyjnego w firmie Meritum
79. Mateusz Wróbel – Projekt i implementacja systemu do zarządzania firmą transportową na przykładzie firmy Trans Way Sp. z o.o.
80. Karol Buraczewski – Platforma e-learning dla e-sportowców
81. Artur Pelczar – System zarządzania hodowlą matek pszczelich
82. Artur Merecik – Realizacja rozszerzonej rzeczywistości na platformie mobilnej
83. Michał Krzyżak – Moduł zarządzania sklepem internetowym
84. Mateusz Klich – Sklep internetowy oparty o platformę WooCommerce
85. Piotr Golec – Projekt i implementacja systemu wspomagania napraw serwisowych urządzeń mobilnych
86. Mikołaj Kozak – Strona internetowa szkoły oparta na CMS
87. Jakub Szybisty – Zdalna aplikacja do ewaluacji kompetencji pracownika
88. Sylwester Nowosiad – Uruchomienie serwera pracującego pod kontrolą systemu operacyjnego Linux Ubuntu wraz z konfiguracją usług serwerowych
89. Mykola Avram – Portal ogłoszeniowy
90. Iryna Pietraszek – Blog internetowy
91. Rafał Pietraszek – Optymalizacja i pozycjonowanie strony internetowej z antykami w wyszukiwarce Google
92. Iryna Kovalchuk – Modelowanie domu jednorodzinnego w 3D max

93. Krzysztof Stec – System wspomagający pracę mechanika samochodowego
94. Yuliia Stechyshyn – Wizualizacja i animacja układu pokarmowego człowieka w Adobe After Effects
95. Adam Tarka – Graficzny serwis społecznościowy przeznaczony do publikacji wpisów w oparciu o system CMS Wordpress
96. Alan Uzar – Serwis internetowy z ogłoszeniami motoryzacyjnymi
97. Vasylyna Gulyn – Techniki animacji komputerowej w ćwiczeniach poprawiających wzrok
98. Vladyslava Kashychenko – Trójwymiarowy model Zamku Lubomirskich w Rzeszowie
99. Dmytro Chykalo – Internetowy sklep motoryzacyjny
100. Marek Łaska – Uruchomienie ukrytej usługi w sieci TOR oraz węzła pośredniczącego
101. Piotr Kocot – Wizualizacja i animacja samochodu Bentley 4.5l w systemie 3DS-MAX
102. Nataliia Strukalo – Migracja adresów wersji IPv4 na adresy wersji Ipv6
103. Karol Kostrubiec – Fotorealistyczna wizualizacja sprzętu muzycznego
104. Liudmyla Burtnyk – Serwis internetowy firmy Pa-Dent
105. Tetiana Povar – Drzewa decyzji w analizie danych
106. Mateusz Chudzik – Zastosowanie technologii Adobe Flash w projektowaniu stron WWW
107. Szymon Pociask – Odzyskiwanie danych z nośników pamięci masowych
108. Karol Krępa – Analiza skuteczności metod pozycjonowania oraz promocji stron
109. Paweł Skocz – Monitoring sieciowych systemów i usług informatycznych
110. Piotr Kozioł – Wirtualizacja systemów operacyjnych
111. Piotr Ruła – Projekt i implementacja bazodanowego systemu wspomagającego działalność Przychodni Weterynaryjnej wykonanej w technologii Java i MySQL

Prowadzone przedmioty

Wykłady:

- Wstęp do informatyki (I stopień I rok KIS WSIiZ)
- Bazy danych (I stopień II rok KIS WSIiZ; III stopień I rok CSP WSIiZ)
- Sztuczna inteligencja (I stopień II rok KIS WSIiZ)
- Technologie internetowe (III stopień I rok CSP WSIiZ)
- Eksploracja danych (I stopień II rok ESW im. ks. Tisznera)

Laboratoria i projekty:

- Eksploracja danych (I stopień III rok KIS WSIiZ)
- Komunikacja człowiek komputer (I stopień III rok KIS WSIiZ)
- Administrowanie systemami baz danych (I stopień III rok KIS WSIiZ)
- Eksploracja danych medycznych (I stopień III rok KIS WSIiZ)
- Wykład monograficzny (Uczenie maszynowe) (II stopień II rok KIS WSIiZ)
- Projektowanie interfejsów multimedialnych (II stopień II rok KIS WSIiZ)
- Systemy wspomagania decyzji (II stopień I rok KIS WSIiZ)
- Zastosowanie informatyki w eksploracji danych medycznych (II stopień II rok KIS WSIiZ)
- Zastosowanie informatyki w projektowaniu interaktywnych systemów komputerowych (II stopień II rok KIS WSIiZ)

Dodatkowa działalność dydaktyczna

- Opracowanie kursu e-learning do przedmiotu Sztuczna inteligencja
- Opracowanie kursu e-learning do przedmiotu Bazy danych
- Wielokrotny udział w dniach otwartych WSIiZ i dniach otwartych Kolegium
- Współautorstwo skryptu *Wybrane metody eksploracji danych 2. Analiza danych sprzecznych*
- Wykłady otwarte dla uczniów Akademickiego Liceum Ogólnokształcącego w Rzeszowie

5.2 Działalność organizacyjna

- Opiekun Kierunku Informatyka od 2015 r.
- z-ca Kierownika Katedry Sztucznej Inteligencji WSIiZ od 2019 r.
- Pełnomocnik Prorektora ds. Nauki na Wydziale Informatyki Stosowanej w latach 2015 – 2018
- z-ca Kierownika Katedry Systemów Rozproszonych WSIiZ w 2010 r.
- członek Wydziałowego Zespołu do Spraw Jakości Kształcenia, od 2013 r.
- koordynator projektu PITAGORAS - program pomocy osobom z uszkodzeniem słuchu w latach 2003-2005

5.3 Działalność popularyzująca naukę

- Organizacja trzech międzynarodowych konferencji naukowych Human System Interactions
- Artykuły popularno-naukowe na blogu naukowym Uczelni
- Uczestnictwo w cyklu *Ciekawa lekcja* oraz *Przybij piątkę nauce*

5.4 Nagrody i wyróżnienia

- wyróżnienie zespołowe w konkursie Contest for the Most Influential Article on Rough Sets co-authored by Polish Researchers in 2020-2021 w 2022 r.
- nagroda indywidualna II Prezydenta WSIiZ w Rzeszowie za wybitne osiągnięcia publikacyjne w 2019 r.
- Best Paper Awards on the Fourth International Conference on Big Data, Small Data, Linked Data and Open Data w 2018 r.
- nagroda zespołowa Rektora i Kanclerza WSIiZ w Rzeszowie za sukces w organizacji międzynarodowej konferencji Human System Interactions w 2010 r.
- nagroda indywidualna Rektora i Kanclerza WSIiZ w Rzeszowie za wysoki poziom doktoratu i wyniki w pracy naukowo-badawczej w 2009 r.
- nagroda zespołowa Rektora i Kanclerza WSIiZ w Rzeszowie za dorobek naukowy w 2003 r.

Literatura

[1] W. Lipski, On semantic issues connected with incomplete information database, ACM Transactions on Databases Systems 4 (1979) 262–296.

[2] J. W. Grzymala-Busse, On the unknown attribute values in learning from example, in: Proceedings of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent System, 1991, p. 368–377.

[3] J. Stefanowski, A. Tsoukias, On the extension of rough sets under incomplete information, in: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'1999, 1999, pp. 73–81.

[4] J. W. Grzymala-Busse, A. Wang, Modified algorithms lem1 and lem2 for rule induction from data with missing attribute values, in: Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), 1997, p. 69–72.

[5] M. Kryszkiewicz, ough set approach to incomplete information systems, in: Proceedings of the Second Annual Joint Conference on Information Sciences, 1995, p. 194–197.

[6] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.

[7] Z. Pawlak, Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.

[8] Z. Pawlak, J. W. Grzymala-Busse, R. Slowinski, W. Ziarko, Rough sets, Communications of the ACM 38 (1995) 89–95.

[9] Z. Pawlak, S. A., Rough sets: some extensions, Information Sciences 177 (2007) 28–40.

[10] J. Stefanowski, A. Tsoukias, Incomplete information tables and rough classification, Computational Intelligence 17 (3) (2001) 545–566.

[11] J. W. Grzymala-Busse, Data with missing attribute values: Generalization of indiscernibility relation and rule induction, Transactions on Rough Sets 1 (2004) 78–95.

[12] Y. Leung, W. Wu, W. Zhang, Knowledge acquisition in incomplete information systems: A rough set approach, European Journal of Operational Research 168 (2006) 164–180.

[13] J. W. Grzymala-Busse, W. Rzasa, Local and global approximations for incomplete data, in: Proceedings of the Fifth International Conference on Rough Sets and Current Trends in Computing, 2006, pp. 244–253.

[14] H. A. Nabwey, A probabilistic rough set approach to rule discovery, International Journal of Advanced Science and Technology 30 (2011) 25–34.

[15] P. G. Clark, J. W. Grzymala-Busse, Experiments on probabilistic approximations, in: Proceedings of the 2011 IEEE International Conference on Granular Computing, 2011, pp. 144–149.

[16] J. W. Grzymala-Busse, Characteristic relations for incomplete data: A generalization of the indiscernibility relation, in: Proceedings of the Fourth International Conference on Rough Sets and Current Trends in Computing, 2004, pp. 244–253.

[17] J. Grzymala-Busse, Data Mining: Foundations and Practice. Studies in Computational Intelligence, Springer, Berlin, Heidelberg, 2008, Ch. Three Approaches to Missing Attribute Values: A Rough Set Perspective.

[18] Y. Y. Yao, Probabilistic rough set approximations, International Journal of Approximate Reasoning 49 (2008) 255–271.

[19] Y. S. Qi, H. Sun, X. B. Yang, Y. Song, Q. Sun, Approach to approximate distribution reduct in incomplete ordered decision system, Journal of Information and Computing Science 3 (2008) 189–198.

[20] M. Chen, X. Xia, An extended rough set model based on a new characteristic relation, in: Proceedings of the IEEE Conference on Granular Computing, 2011, pp. 100–105.

[21] J. W. Grzymala-Busse, Rough set strategies to data with missing attribute values, in: Notes of the Workshop on Foundations and New Directions of Data Mining, in conjunction with the Third International Conference on Data Mining, 2003, pp. 56–63.

[22] Y. Leung, D. Li, Maximal consistent block technique for rule acquisition in incomplete information systems, Information Sciences 153 (2003) 85–106.

[23] M. Kryszkiewicz, Rough set approach to incomplete information systems, Information Sciences 112 (1998) 39–49.

[24] M. Kryszkiewicz, Rules in incomplete information systems, Information Sciences 113 (3-4) (1999) 271–292.

[25] X. Liu, M. Shao, Approaches to computing consistent blocks, in: Proceedings of the International Conference on Machine Learning and Cybernetics, Springer, 2014, pp. 264–274.

[26] J. Y. Liang, B. L. Wang, Y. H. Qian, D. Y. Li, An algorithm of constructing maximal consistent blocks in incomplete information systems, International Journal of Computer Science and Knowledge Engineering 2 (1) (2008) 11–18.

[27] R. Zheng, Algorithms for computing maximal consistent blocks, Master's thesis, University of Kansas (2019).

[28] P. G. Clark, C. Gao, J. W. Grzymala-Busse, T. Mroczek, R. Niemiec, Characteristic sets and generalized maximal consistent blocks in mining incomplete data, in: Rough Sets. IJCRS 2017. Lecture Notes in Computer Science, Vol. 10313, 2017, pp. 477–486.

[29] D. Yellin, Algorithms for subset testing and finding maximal sets, Computational Intelligence (1992) 386–392.

[30] B. Pękala, T. Mroczek, D. Gil, M. Kepski, Application of fuzzy and rough logic to posture recognition in fall detection system, Sensors 22 (4) (2022). URL `https://www.mdpi.com/1424-8220/22/4/1602`

[31] T. Mroczek, B. Pękala, D. Gil, Fuzzy and rough approach to the problem of missing data in fall detection system, Fuzzy Sets and Systems, (pozytywne recenzje).

[32] J. W. Grzymala-Busse, P. G. Clark, M. Kuehnhausen, Generalized probabilistic approximations of incomplete data, International Journal of Approximate Reasoning 132 (2014) 180–196.

[33] M. R. Chmielewski, J. W. Grzymala-Busse, Global discretization of continuous attributes as preprocessing for machine learning, International Journal of Approximate Reasoning 15 (4) (1996) 319–331.

[34] T. Elomaa, J. Rousu, General and efficient multisplitting of numerical attributes, Machine Learning 36 (1999) 201–244.

[35] U. M. Fayyad, K. B. Irani, On the handling of continuous-valued attributes in decision tree generation, Machine Learning 8 (1992) 87–102.

[36] U. M. Fayyad, K. B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence, 1993, pp. 1022–1027.

[37] J. W. Grzymala-Busse, A multiple scanning strategy for entropy based discretization, in: Proceedings of the 18th International Symposium on Methodologies for Intelligent Systems, 2009, pp. 25–34.

[38] R. Kohavi, M. Sahami, Error-based and entropy-based discretization of continuous features, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 114–119.

[39] H. S. Nguyen, S. H. Nguyen, Discretization methods in data mining, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery 1: Methodology and Applications, Physica-Verlag, Heidelberg, 1998, pp. 451–482.

[40] J. Stefanowski, Handling continuous attributes in discovery of strong decision rules, in: Proceedings of the First Conference on Rough Sets and Current Trends in Computing, 1998, pp. 394–401.

[41] J. Stefanowski, Algorithms of Decision Rule Induction in Data Mining, Poznan University of Technology Press, Poznan, Poland, 2001.

[42] J. W. Grzymala-Busse, A new version of the rule induction system LERS, Fundamenta Informaticae 31 (1997) 27–39.

[43] D. Tian, X.-J. Zeng, J. Keane, Core-generating approximated minimum entropy discretization for rough set feature selection in pattern classification, International Journal of Approximate Reasoning 52 (2011) 863–880.

[44] H. S. Nguyen, Discretization problem for rough sets methods, in: Proceedings of the 1-st International Conference RSCTC 1998 on Rough Sets and Current Trends in Computing, Springer-Verlag, Berlin, Heidelberg, 1998, pp. 545–552.

[45] R. W. Swiniarski, Rough set methods in feature reduction and classification, International Journal of Applied Mathematics and Computer Science 11 (2001) 656–582.

[46] Q. Hu, D. Yu, Z. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (2006) 414–423.

[47] K.-j. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Systems with Applications 19 (2000) 125–132.

[48] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.

[49] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[50] R. Jensen, Q. Shen, Fuzzy-rough sets for descriptive dimensionality reduction, in: Proceedings of the International Conference on Fuzzy Systems FUZZ-IEEE 2002, 2002, pp. 29–34.

[51] E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, Journal of Finance 23 (4) (1968) 589–609.

[52] P. Blajdo, J. W. Grzymala-Busse, Z. S. Hippe, M. Knap, T. Mroczek, L. Piatek, A comparison of six approaches to discretization—a rough set perspective, in: Proceedings of the Rough Sets and Knowledge Technology Conference, 2008, pp. 31–38.