Wybrane aspekty interaktywnej ekstrakcji cech
(ang. Selected aspects of interactive feature extraction)

(Autoreferat rozprawy doktorskiej)

Marek Grzegorowski

1 Wprowadzenie

W większości zastosowań algorytmów uczenia maszynowego, głównym zadaniem analityka jest odpowiednie zamodelowanie badanego problemu, w tym opracowanie odpowiedniej reprezentacji obiektów opisanych w danych. Ekstrakcja cech to proces przetwarzania otrzymanych danych, który prowadzi do uzyskania reprezentacji odpowiednio sprofilowanej do analizowanego problemu. Przyczynia się do poprawy wydajności przetwarzania informacji i optymalizacji procesu modelowania oraz umożliwia pozyskiwanie atrybutów, które mogą być wykorzystywane zarówno przez ekspertów dziedzinowych, jak i algorytmy uczenia maszynowego. Wyróżnia się dwie zasadnicze fazy tego procesu: pierwsza to konstrukcja nowych atrybutów (ang. feature engineering), natomiast druga to wybór najistotniejszych spośród uzyskanych w ten sposób atrybutów (ang. feature selection). Mimo tego, że istnieje wiele podejść do automatyzacji procesu tworzenia i selekcji atrybutów [18, 21, 28, 33, 39], potrzeba rozwoju metod wspierających interakcję z użytkownikami, które uwzględniałyby wiedzę dziedzinową pozyskiwaną od ekspertów, ich doświadczenie i preferencje, jest konsekwentnie wskazywana przez wielu naukowców jako jeden z najtrudniejszych i najważniejszych obszarów badawczych [5, 41, 42].

Interaktywna eksploracja danych pozwala na płynne przechodzenie od formułowania do weryfikacji hipotez dotyczących badanych zjawisk, poprzez odpowiednią wizualizację wyników cząstkowych analiz oraz wykorzystanie tak zdobytej wiedzy w kolejnych etapach eksploracji. Automatyzacja tego procesu wydaje się szczególnie ambitnym zadaniem. Już na poziomie postrzegania informacji napotykamy istotne różnice między człowiekiem a maszyną. *Lotfi Zadeh* zwraca uwagę, że ludzie postrzegają i opisują rzeczywistość w sposób nieostry, podczas gdy wszelkie pomiary są bardzo precyzyjne [62]. Rzeczywiście, w codziennych dyskusjach o pogodzie rzadko podajemy precyzyjne wartości temperatury, ciśnienia czy opadów atmosferycznych, korzystamy raczej z ogólnych określeń, takich jak gorąco, ciepło, zimno. W przypadku danych generowanych maszynowo pojedyncze wartości zamieniają się w całe łańcuchy odczytów numerycznych i stają się jeszcze bardziej niezrozumiałe i trudne do interpretacji [27,36]. W tym kontekście, jednym z wyzwań, na które zwraca uwagę *Leslie Valiant*, jest potrzeba określenia elementarnych bloków obliczeniowych niezbędnych w procesie poznawczym [49].

Opisanie danych w możliwie intuicyjny sposób – na wyższym poziomie abstrakcji – jest jednym z kluczowych założeń paradygmatu granulacji informacji (ang. inforamtion granulation) [11, 61]. W odniesieniu do inżynierii atrybutów[1] opisujących dane, możemy odwołać się do tzw. techniki *przesuwnego okna czasowego* pozwalającej ograniczyć ilość atrybutów poprzez zastąpienie ich wysokopoziomowymi statystykami wyznaczonymi z fragmentów badanych szeregów czasowych, które są znacznie bardziej zrozumiałe dla człowieka [16]. Warto zwrócić uwagę, że reprezentowanie badanych zjawisk za pomocą łatwych w interpretacji, zrozumiałych cech jest ważne w wielu dziedzinach, takich jak medycyna czy wymiar sprawiedliwości [6, 45]. Innym ważnym problemem jest możliwość wzbogacania danych generowanych maszynowo dodatkowymi atrybutami odpowiadającymi wiedzy dziedzinowej pozyskiwanej od ekspertów. W ten sposób możemy dostarczać do procesu uczenia maszynowego

[1] W naszych rozważaniach *atrybut*, *cecha* oraz *wymiar* danych są stosowane wymiennie.

dodatkowych informacji na temat zachodzących zjawisk, które nie są zawarte w danych historycznych. W jednym z analizowanych zastosowań metod uczenia maszynowego, związanym z przewidywaniem wstrząsów sejsmicznych w kopalniach węgla kamiennego, połączenie danych maszynowych pozyskiwanych z sieci czujników z atrybutami odpowiadającymi ocenom ekspertów pozytywnie wpłynęło na jakość i niezawodność predykcji [23].

Przeciwstawny kierunek badań stanowią metody uczenia reprezentacji (ang. representation learning), które dążą do uzyskania całkowitej autonomiczności algorytmów uczenia maszynowego i uniezależnienia procesów odkrywania wiedzy od interakcji z użytkownikiem [2]. W tym wypadku wejściem do algorytmu są nieprzetworzone dane, a odkrywanie w nich najważniejszych czynników przyczyniających się do skutecznego wnioskowania oraz efektywne zamodelowanie badanego problemu odbywają się całkowicie automatycznie. Koncepcje głębokich sieci neuronowych, uczenia się zanurzeń dla danych tekstowych czy nienadzorowanego wstępnego trenowania modeli stanowią znaczące przykłady sukcesów w tej dziedzinie [7, 44]. Pokrewnym kierunkiem badań są również metody redukcji wymiaru (ang. dimensionality reduciton), takie jak analiza głównych składowych czy skalowanie wielowymiarowe. Umożliwiają one przedstawienie otrzymanych obserwacji w nowej przestrzeni o zadanym wymiarze, niższym niż w przypadku oryginalnych danych [19]. Pozwala to na rzutowanie danych na przestrzeń dwu- lub trzy-wymiarową, które łatwo przedstawić graficznie, dlatego też metody te stosunkowo często wykorzystuje się do wizualizacji [32]. Jakkolwiek, należy zwrócić uwagę, że takie przekształcenie danych prowadzi do powstania zupełnie nowej przestrzeni, w której każdy z rzutowanych wymiarów stanowi kombinację wszystkich (lub prawie wszystkich) oryginalnie otrzymanych atrybutów, co znacząco utrudnia, a często uniemożliwia ich interpretację [20, 45]. W związku z powyższym te podejścia nie będą dalej rozpatrywane w tej rozprawie.

Bardzo duża ilość wymiarów stanowi oczywiście istotny problem w kontekście interaktywności. Co prawda w takim przypadku, użytkownicy mogą nadal interpretować znaczenie pojedynczych atrybutów, ale nawigacja po ich podzbiorach staje się istotnie trudniejsza i w przypadku wielowymiarowych szeregów czasowych może wykraczać poza ramy ludzkiej percepcji. Problem ten dotyczy także metod uczenia maszynowego i określany jest mianem przekleństwa wymiarowości. Taka sytuacja wprowadza duży szum informacyjny, utrudnia odkrycie istotnych czynników wpływających na badany problem decyzyjny, często uniemożliwiając jego uogólnienie. Jednym ze sposobów radzenia sobie z taką sytuacją może być operowanie całymi klastrami atrybutów indukującymi podobne informacje [25]. Alternatywnie, można zastosować techniki selekcji atrybutów, które zastąpiłyby duże zbiory atrybutów znacznie mniejszymi podzbiorami dostarczającymi porównywalnych informacji o danych. Wyszukiwanie takich podzbiorów jest dobrze znanym zadaniem w teorii zbiorów przybliżonych (ang. rough set theory) [40]. Dysponując początkowym zestawem atrybutów można wyznaczyć tzw. redukty, które indukują (prawie) taką samą informację jak wszystkie rozważane atrybuty. Mimo że odnajdywanie minimalnych reduktów należy do klasy problemów NP-trudnych, to w literaturze odnajdujemy szereg metod heurystycznych pozwalających wyznaczyć dostatecznie małe podzbiory atrybutów spełniające wybrane kryterium redukcji [4, 57, 58]. W ramach rozprawy zagadnienie selekcji cech jest rozważane na wielu płaszczyznach. Z jednej strony może dotyczyć wyboru ważnych źródeł danych, np. czujników generujących istotne dla modelu odczyty. W takim przypadku kluczowa jest interakcja z ekspertami dziedzinowymi i dostarczenie odpowiednich raportów czy wizualizacji, aby umożliwić im dokonanie właściwej oceny. Z drugiej strony możemy je traktować jako problem wyznaczenia najistotniejszych spośród atrybutów, np. spośród statystyk wyliczanych w ramach procesu ekstrakcji cech opartej o technikę przesuwnego okna czasowego.

W wybranych zastosowaniach, związanych np. z monitorowaniem zagrożeń lub diagno-

styką medyczną, oprócz dokładności, szybkości i niezawodności modeli predykcyjnych, niemniej ważna jest ich odporność na braki lub niedoskonałości danych. W szczególności, awaria pojedynczego czujnika, która zazwyczaj powoduje brak całego wymiaru danych, nie może skutkować niemożnością oceny sytuacji. Jedną z możliwości radzenia sobie z takimi sytuacjami jest imputacja brakujących wartości [3, 35]. Techniki imputacji są szczególnie skuteczne gdy braki w danych stanowią chwilową anomalię pracy sytemu i są stosunkowo nieduże. Jakkolwiek, w sytuacji długotrwałej niedostępności danego wymiaru danych, skutkującej dużymi lukami w szeregach czasowych, imputacja odczytów obarczona by była dużym błędem i znacząco podnosiłaby poziom niepewności w odniesieniu do późniejszej predykcji [47, 56]. Jedna z metod, która nie wymaga uzupełniania brakujących wartości, polega na uczeniu modeli predykcyjnych na zagregowanych cechach wyższego poziomu, zamiast na pojedynczych atrybutach. Operowanie na granulach cech jest mniej wrażliwe na braki wartości pojedynczych atrybutów [15, 60]. Inny podejściem jest wykorzystanie zespołów klasyfikatorów [23, 59] lub modeli predykcyjnych wzbogaconych nadmiarowymi informacjami na potrzeby weryfikacji ich działania, takimi jak dodatkowe cięcia weryfikacyjne (ang. verifying cuts) w drzewach decyzyjnych [1].

Powyższe podejścia opierają się na założeniu, że cechy wybrane do procesu uczenia maszynowego zawierają dodatkową (nadmiarową) wiedzę, pozwalającą na wykonanie prawidłowej predykcji nawet w sytuacji braku niektórych atrybutów. Tymczasem wszystkie nowoczesne techniki selekcji cech dążą do usunięcia tej nadmiarowości. Naszym zadaniem jest sformułowanie takich zależności, dzięki którym algorytmy selekcji cech zagwarantują, że wybrane podzbiory atrybutów będą zawierały wystarczającą ilość informacji o rozważanych zmiennych docelowych, nawet jeśli niektóre z tych atrybutów będą niedostępne. Jednym z możliwych podejść do tego problemu jest zbudowanie swoistego zespołu podzbiorów atrybutów, np. poprzez teoriomnogościowe zsumowanie wyników istotnie różnych algorytmów selekcji. W takim wypadku model bazowałby na zbiorze różnych podzbiorów cech wraz z odpowiadającymi im modelami predykcyjnymi traktowanymi jako zespół. Innym podejściem jest poszukiwanie zestawów cech, które gwarantowałyby zapewnienie wystarczającej mocy predykcyjnej, nawet w razie braku niektórych ich elementów [9]. W rozprawie wprowadzamy nową koncepcję resilientnej selekcji cech (resilient feature selection). Jako przykład resilientnego podejścia do selekcji cech wprowadzamy r-$\mathbb{C}$-redukty – tj. nieredukowalne podzbiory atrybutów, które zapewniają satysfakcjonujący poziom informacji o rozważanej zmiennej docelowej, nawet w wypadku niedostępności ich dowolnych r atrybutów [14] .

W rozprawie poruszono problem interaktywnej ekstrakcji cech oraz zaproponowano szereg innowacyjnych podejść do automatyzacji procesu ich tworzenia i selekcji, rozważając możliwość angażowania w ten proces użytkowników. Przedstawiono aktualny stan wiedzy w dziedzinie ekstrakcji atrybutów oraz zaprezentowano zastosowania komercyjne tego procesu. Omówiono wyzwania związane z przetwarzaniem dużych zbiorów danych, ze szczególnym naciskiem na przetwarzanie wielowymiarowych szeregów czasowych. Poddano dyskusji problem opracowania takiej reprezentacji danych, która byłaby zrozumiała dla ekspertów dziedzinowych. W tym celu, przedyskutowano możliwość wykorzystania atrybutów uzyskiwanych metodą przesuwnego okna czasowego oraz granulacji atrybutów. Opracowane metody i algorytmy ekstrakcji cech poddano weryfikacji eksperymentalnej oraz przedstawiono ich zastosowania w wybranych projektach naukowych.

Wybrane wyniki rozprawy były już prezentowane na międzynarodowych konferencjach i warsztatach oraz zostały opublikowane w artykułach, które ukazały się w recenzowanych materiałach konferencyjnych oraz w renomowanych czasopismach. W szczególności publikacje związane z granulacją przestrzeni atrybutów oraz z resilientnymi metodami selekcji cech zaprezentowano w ramach [9, 11, 14], przy czym praca [9] otrzymała tytuł najlepszej pracy

studenckiej podczas konferencji IJCRS'2016. Publikacje autora rozprawy dotyczące problemu wyznaczania różnorodnych podzbiorów atrybutów oraz budowania w oparciu o nie zespołów modeli predykcyjnych to m.in. [10, 53]. Ponadto przeprowadzone badania nad zastosowaniami ekstrakcji cech w zakresie analizy danych z sieci czujników w górnictwie oraz w służbie pożarniczej zostały przedstawione w ramach następujących publikacji [15,16,23,24]. Badania nad inteligentnymi systemami, modelami danych i mechanizmami przetwarzania informacji wspierającymi interaktywną ekstrakcję cech zostały ujęte w [8,13,17,30,54,55,63]. Wybrane wyniki rozprawy zostały również opublikowane w artykułach technicznych i monografiach w języku polskim [12, 22].

2 Główne wyniki rozprawy

Tablica 1: Przykładowa tabela decyzyjna $\mathbb{S}$.

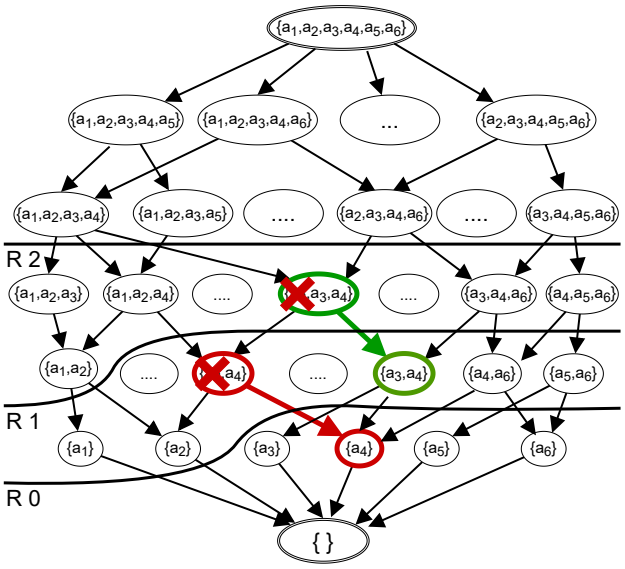
$U \backslash A$	a_1	a_2	a_3	a_4	a_5	a_6	d
u_1	0	'a'	$\triangledown$	$\circ$	$\square$	'x'	*tak*
u_2	0	'b'	$\triangledown$	$\odot$	$\bullet$	'x'	*tak*
u_3	0	'c'	$\triangle$	$\oslash$	$\diamond$	'x'	*tak*
u_4	0	'd'	$\triangledown$	$\otimes$	$\triangleleft$	'x'	*tak*
u_5	0	'e'	$\triangledown$	$\ominus$	$\star$	'y'	*tak*
u_6	0	'f'	$\triangledown$	$\oslash$	$\triangleright$	'y'	*tak*
u_7	1	'g'	$\triangle$	$\otimes$	$\square$	'y'	*nie*
u_8	1	'h'	$\triangle$	$\ominus$	$\bullet$	'z'	*nie*
u_9	1	'i'	$\triangledown$	$\oplus$	$\diamond$	'z'	*nie*

W badaniach nad interaktywnością procesu ekstrakcji cech poruszono problemy związane z uzyskiwaniem użytecznych i zrozumiałych dla ekspertów atrybutów z wielowymiarowych danych, a także możliwość ograniczenia ilości tych atrybutów w celu uzyskania możliwie najprostszych, ale dokładnych modeli. Zaproponowane w rozprawie nowe metody interaktywnej ekstrakcji cech wykraczają poza obecnie znane standardy, umożliwiając skuteczniejszy sposób wyrażania wiedzy dziedzinowej związanej z najważniejszymi podzbiorami atrybutów. Zaproponowane algorytmy konstrukcji i doboru cech wykorzystują różne formy granulacji przestrzeni atrybutów, a także pozwalają na wydajne przetwarzanie dużych danych (ang. Big Data) poprzez zrównoleglenie obliczeń. Na szczególną uwagę zasługuje zaproponowana metoda uodpornienia algorytmów selekcji atrybutów na ewentualne braki w danych, która pozwala znacząco zmniejszyć wymiarowość danych, gwarantując jednocześnie zachowanie niezbędnego poziomu informacji (wg zadanego kryterium) do predykcji zmiennej celu, nawet po usunięciu określonej liczby atrybutów.

2.1 Resilientna selekcja cech w oparciu o r-$\mathbb{C}$-redukty

Selekcja cech pozwala wyznaczyć minimalny podzbiór atrybutów, który zachowuje najistotniejsze charakterystyki całego zbioru atrybutów. Odgrywa ona istotną rolę w procesie konstruowania modeli predykcyjnych, ponieważ uczenie modeli na uproszczonej reprezentacji pozwala je lepiej uogólnić, zapewniając większą stabilność predykcji oraz ułatwia interpretację wyników. Jakkolwiek, myśląc o wdrożeniu komercyjnym wytrenowanego modelu, należy rozważyć ciągłe zmiany zachodzące w środowisku i ich wpływ na jego działanie. Jednym z większych zagrożeń dla stabilności i jakości działania są braki danych. Wybrane wymiary danych mogą być po pewnym czasie niedostępne, np. w wyniku wyłączenia wybranych źródeł danych. Mając na uwadze tak postawiony problem, sformalizujemy pojęcie resilientnej selekcji cech poprzez wprowadzenie r-$\mathbb{C}$-reduktów - nieredukowalnych podzbiorów atrybutów zapewniających satysfakcjonujący poziom informacji o zmiennej docelowej według zadanego kryterium $\mathbb{C}$, nawet po usunięciu z nich $r \geqslant 0$ elementów. Proponowane podejście opiera się na uogólnieniu (przybliżonych) reduktów decyzyjnych, znanych z teorii zbiorów przybliżonych (RST) [57]. Jednakże, poprzez odwołanie się do wybranych właściwości funkcji $\mathbb{C}$ zamiast do jej konkretnych implementacji, zaproponowane metody obejmują

bardzo szeroką rodzinę kryteriów oceny podzbiorów atrybutów. W naturalny sposób pozwalają zdefiniować resilientne wersje dla wielu znanych metod selekcji cech.



Rysunek 1: Krata atrybutów dla danych w Tabeli 1. Linie oznaczone jako R0, R1 oraz R2 odpowiadają poziomom resiliencji zapewnianymi przez podzbiory atrybutów (tj. r-$\mathbb{C}$-redukty) powyżej każdej z nich.

Jedną z metodologii eksploracji danych, w której duży nacisk kładzie się na granulację przestrzeni atrybutów i selekcję cech, jest teoria zbiorów przybliżonych (RST) [34, 48]. Zakłada ona, że cała dostępna wiedza o obiekcie $u \in U$ jest reprezentowana w strukturze zwanej tablicą decyzyjną $\mathbb{S} = (U, A \cup \{d\})$, gdzie U jest skończonym, niepustym zbiorem obiektów, A jest skończonym, niepustym zbiorem atrybutów, a d odpowiada decyzji (por. Tabela 1). Podstawowym pojęciem związanym z selekcją cech jest redukt. Omówmy jego znaczeie odwołując się do kraty atrybutów na Rysunku 1, gdzie owale odpowiadają podzbiorom atrybutów, a strzałki odpowiadają relacji zawierania się zbiorów ($\subset$). Linia oznaczona jako $R0$ wizualizuje granicę, powyżej której podzbiory atrybutów pozwalają, dla każdego obiektu (tj. wiersza w Tabeli 1), wyznaczyć jego decyzję (tj. kolumna d) tak samo dokładnie jak przy wykorzystaniu całego zbioru atrybutów A. Wszystkie zbiory powyżej linii $R0$ nazywamy superreduktami, natomiast te znajdujące się bezpośrednio nad nią nazywamy reduktami. Naszym celem jest uogólnienie i rozszerzenie tego pojęcia poprzez sformułowanie nowych więzów, dzięki którym wybrane podzbiory cech zapewnią rozróżnialność obiektów względem decyzji, nawet jeśli niektóre z wybranych cech będą niedostępne [9].

W pierwszej kolejności uogólnijmy sposób wnioskowania o jakości podzbiorów atrybutów określając oczekiwane własności funkcji oceny $\mathbb{C}$, która dla tabeli decyzyjnej $\mathbb{S} = (U, A \cup \{d\})$ i podzbioru atrybutów R zwraca binarną ocenę badanego podzbioru cech:

Definicja 1 (Kryterium jakości atrybutów $\mathbb{C}$)**.**
Niech dana będzie tablica decyzyjna $\mathbb{S} = (U, A \cup \{d\})$. Funkcję $\mathbb{C} : \mathcal{P}(A) \to \{0, 1\}$ określoną na podzbiorach atrybutów tablicy $\mathbb{S}$ nazwiemy kryterium jakości atrybutów, jeżeli $\forall_{X,Y \subseteq A}$ zachodzi $X \Rrightarrow Y$ to $\mathbb{C}(X) \geqslant \mathbb{C}(Y)$

Wiele znanych kryteriów redukcji atrybutów, opartych na rozróżnialności, entropii, czy też funkcji γ posiada własności określone w Definicji 1. Możemy je uogólnić poniższą definicją $\mathbb{C}$-reduktu:

Definicja 2 ($\mathbb{C}$-redukt)**.**
Niech dane będą tablica decyzyjna $\mathbb{S} = (U, A \cup \{d\})$ oraz kryterium redukcji $\mathbb{C}$. Podzbiór atrybutów $R \subseteq A$ nazwiemy $\mathbb{C}$-superreduktem, jeżeli spełnia $\mathbb{C}(R) = 1$. Jeżeli dodatkowo zachodzi $\forall_{R' \subsetneq R}\ \mathbb{C}(R') = 0$, wówczas podzbiór atrybutów R nazwiemy $\mathbb{C}$-reduktem.

Przykładowo, odwołując się do konstrukcji $\mathbb{C}$, sformułujemy powszechnie znane kryterium redukcji w oparciu o relację rozróżnialności [40], a w szczególności o odpowiadającą jej miarę $Disc$[2]:

[2]Niech dane będą $\mathbb{S} = (U, A \cup \{d\})$ i $R \subseteq A$. $Disc(R) = |\{(u, u') : \exists a \in R, a(u) \neq a(u') \wedge d(u) \neq d(u')\}|$

$$\mathbb{C}^{Disc}(R) = \begin{cases} 1 & \text{jeśli } Disc(R) = Disc(A) \\ 0 & \text{w p.p.} \end{cases} \tag{1}$$

Bazując na własnościach funkcji $\mathbb{C}$ wynikających z Definicji 1 możemy łatwo skonstruować resilientną wersję zadanego kryterium r-$\mathbb{C}$ jako:

$$r\text{-}\mathbb{C}(R) = \begin{cases} 1 & \text{jeśli } \min_{R' \subseteq R : |R'| \geqslant max(|R|-r,0)} \mathbb{C}(R') = 1 \\ 0 & \text{w p.p.} \end{cases} \tag{2}$$

Poniżej przedstawiamy definicję resilientnego redutku (r-$\mathbb{C}$-reduktu):

Definicja 3 (r-$\mathbb{C}$-redukt)**.**
Niech dane będą tablica decyzyjna $\mathbb{S} = (U, A \cup \{d\})$, kryterium $\mathbb{C}$ i oczekiwany poziom resiliencji r. Podzbiór atrybutów $\breve{R} \subseteq A$ nazwiemy r-$\mathbb{C}$-superreduktem, jeśli spełnia r-$\mathbb{C}(\breve{R}) = 1$. Jeśli dodatkowo zachodzi $\forall_{R \subsetneq \breve{R}}$ r-$\mathbb{C}(R) = 0$, wówczas $\breve{R}$ nazwiemy r-$\mathbb{C}$-reduktem.

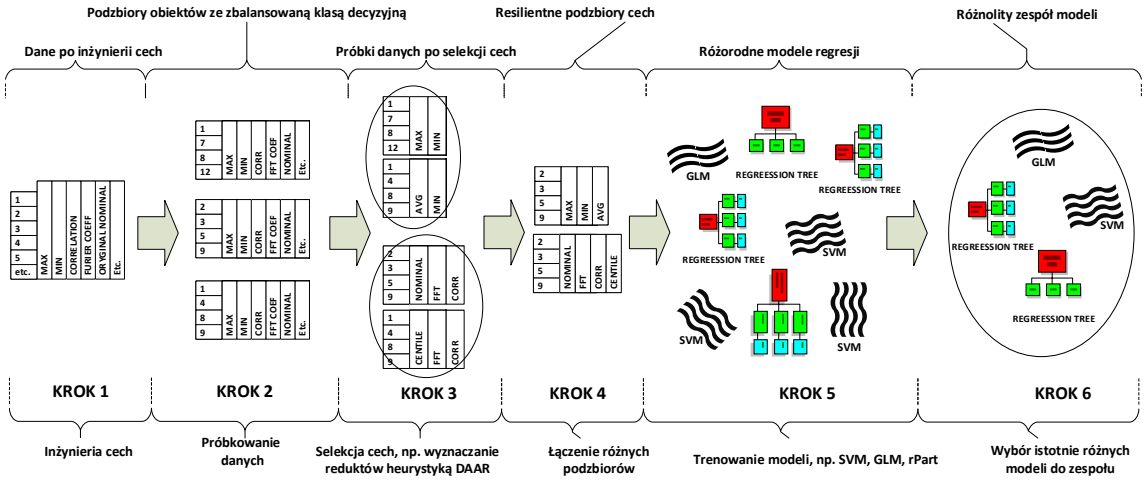
Zaprezentowana koncepcja r-$\mathbb{C}$-reduktu istotnie rozszerza znane z teorii zbiorów przybliżonych podejście do zagadnienia redukcji cech. Definiując współczynnik resiliencji r w sposób umożliwiający łączenie go z dowolnym kryterium redukcji spełniającym własności Definicji 1, uogólniamy oba te sformułowania (por. Definicja 3). Chcąc podkreślić znaczenie resiliencji, odwołamy się jeszcze raz do kraty atrybutów na Rysunku 1. Przypomnijmy, że linia oznaczona jako $R0$ wizualizuje granicę, powyżej której podzbiory spełniają $\mathbb{C}$, $R1$ odpowiada granicy, powyżej której każdy podzbiór jest r-$\mathbb{C}$-superreduktem, a $R2$ wyznacza podzbiory, które są r-$\mathbb{C}$-superreduktami. Omówmy jaki wpływ miałoby usunięcie z danych atrybutu a_2, odnosząc się do dwóch wyróżnionych przykładów na Rysunku 1. Możemy zauważyć, że usunięcie atrybutu a_2 z podzbioru $\{a_2, a_4\}$ powoduje znaczną utratę informacji, uniemożliwiając rozróżnienie decyzji, ponieważ po usunięciu $\mathbb{C}(\{a_4\}) = 0$. Natomiast wpływ usunięcia a_2 z $\{a_3, a_4\}$ jest niezauważalny, ponieważ $\mathbb{C}(\{a_3, a_4\}) = 1$.

Znaczącym teoretycznym wynikiem rozprawy jest sformułowanie i dowód Twierdzenia 1, w którym wykazujemy, że każde NP-trudne zadanie selekcji atrybutów polegające na znalezieniu — dla wejściowej tablicy decyzyjnej $\mathbb{S} = (U, A \cup \{d\})$ — minimalnego $\mathbb{C}$-reduktu pozostaje NP-trudne również w resilientnej wersji znalezienia minimalnego r-$\mathbb{C}$-reduktu.

Twierdzenie 1. *(Problem r-$\mathbb{C}^P$ jest NP-trudny dla każdego NP-trudnego P)*
Niech P będzie problemem znalezienia minimalnego podzbioru atrybutów $R \subseteq A$ spełniającego warunek wyrażony przez kryterium $\mathbb{C}^P : \mathcal{P}(A) \to \{0, 1\}$ dla dowolnej danej na wejściu tablicy decyzyjnej $\mathbb{S} = (U, A \cup \{d\})$. Jeśli P jest problemem NP-trudnym, to znalezienie minimalnego zbioru $\breve{R}$ spełniającego r-$\mathbb{C}^P$ (por. równanie (2)) również jest NP-trudne.

Bazując na Twierdzeniu 1 dowodzimy, że problemy wyznaczenia minimalnego (H, ε)-reduktu aproksymacyjnego [51], (γ, ε)-reduktu aproksymacyjnego [57], a także szeroka rodzina problemów optymalizacyjnych dla redukcji przybliżonej opartej na rozróżnialności [37, 38], pozostają problemami NP-trudnymi także w ich resilientnej wersji. Ten sam mechanizm można zastosować w wielu przypadkach, w szczególności dla różnych sformułowań $\mathbb{C}^{(Q,\varepsilon)}$-reduktów [50, 52].

W rozprawie omawiamy również możliwości wyznaczania resilientnych podzbiorów cech z tablic decyzyjnych przy wykorzystaniu zarówno algorytmów zachłannych, jak i wybranych heurystyk oraz oceniamy złożoność obliczeniową proponowanych algorytmów. Z jednej strony rozważamy możliwość wyznaczania minimalnych r-$\mathbb{C}$-reduktów stosując przeszukiwanie kraty atrybutów wszerz (ang. breadth first search). W tym celu, odpowiednio dostosowujemy dobrze znany algorytm Apriori [46]. Rozważamy również podejścia algorytmiczne



Rysunek 2: Przebieg procesu resilientnej selekcji cech oraz budowania różnolitego zespołu modeli.

oparte o przeszukiwanie kraty atrybutów w głąb (ang. depth first search), inspirowane znanymi z RST algorytmami permutacyjnymi [52, 57]. Proponujemy także algorytmy aproksymacyjne, które bazując na zaproponowanych heurystykach, pozwalają znacznie szybciej wyznaczać podzbiory atrybutów zapewniające wymagany poziom resiliencji r, rezygnując z minimalności wyznaczonych podzbiorów. Warto też podkreślić, że omawiana idea resilientnej selekcji cech jest zdecydowanie bardziej ogólna i można ją uznać za rozszerzenie wielu innych metod, niekoniecznie wywodzących się z teorii zbiorów przybliżonym.

2.2 Ekstrakcja cech i budowanie różnolitych zespołów modeli

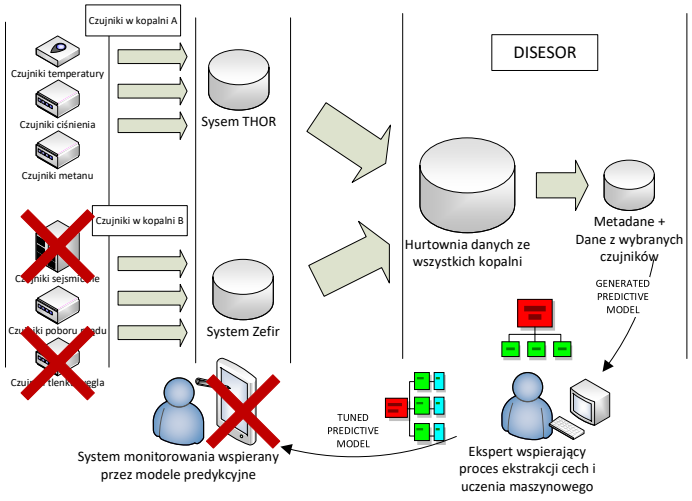
W rozprawie przedstawiamy kompleksowe podejście do interaktywnej ekstrakcji cech, przy czym szczególny nacisk kładziemy na przetwarzanie wielowymiarowych szeregów czasowych [16]. Główną ideą naszego podejścia jest inteligentne wykorzystanie paradygmatu granulacji informacji w kontekście agregacji, selekcji i inżynierii cech [11]. W zakresie inżynierii cech, staramy się reprezentować dane w możliwie intuicyjny sposób, wykorzystując statystyki charakteryzujące przesuwne okna czasowe. Koncepcja granulacji informacji jest przydatna również przy selekcji cech, szczególnie w przypadku dużych i złożonych zbiorów danych. Nasza metodologia inżynierii i selekcji cech została opracowana tak, aby odpowiadała różnym rodzajom granulek cech, które mogą być indukowane np. przez hierarchiczne grupowanie atrybutów lub analizę ich wymienności. Z drugiej strony pokazujemy, że znaczące granulacje można uzyskać zgodnie z różnymi interpretacjami relacji podobieństwa, wyrażanymi przez odpowiednio dobrane miary odległości albo bazując na funkcji lub znaczeniu rozważanych cech [23].

Istotą jest operowanie na granulkach atrybutów utworzonych przy użyciu różnych algorytmów odkrywania wiedzy [10, 53]. W wielu przypadkach, czego przykładem są omówione zastosowania, granulki zbudowane w przestrzeni atrybutów mogą reprezentować relacje semantyczne ważne dla ekspertów dziedzinowych. W przypadku wielowymiarowych szeregów czasowych, np. pozyskanych z dużej sieci sensorów, gromadzone elementy są bardzo złożone i niejednokrotnie ich interpretacja jest możliwa tylko w kontekście dodatkowej wiedzy dziedzinowej uzyskanej od ekspertów [54, 55]. Jedno z opracowanych rozwiązań zakłada rozszerzenie procesu ekstrakcji cech z okien przesuwnych o możliwość dodania atrybutów odzwierciedlających oceny ekspertów dziedzinowych [23]. Reprezentowanie problemów eksploracji danych za pomocą zrozumiałych, wysokopoziomowych cech, wzbogaconych ocenami eksperckimi, jest jednym z istotnych aspektów z punktu widzenia interaktywności.

Przyjrzyjmy się kolejnym krokom opracowanej metodologii na przykładzie jednego z

analizowanych zastosowań, dotyczącego przewidywania niebezpiecznych stężeń metanu. Na Rysunku 2 przedstawiono przebieg opracowanego procesu selekcji cech, uczenia maszynowego i budowania zespołu modeli. Przedstawiony proces rozpoczyna działanie w KROKU 1, na bogatej reprezentacji danych, uzyskanej poprzez opracowane mechanizmy inżynierii cech, wzbogaconej atrybutami odpowiadającymi ocenom ekspertów (inżynieria cech ma miejsce we wcześniejszej fazie procesu, nie ujętej na Rysunku 2). W KROKU 2 ma miejsce próbkowanie danych, którego celem jest uzyskanie podzbiorów obserwacji o możliwie zrównoważonym rozkładzie atrybutu decyzyjnego – jest to szczególnie istotne, gdy ważniejsza klasa decyzyjna (np. dotycząca ryzyka wybuchu metanu [15]) jest niedostatecznie reprezentowana w zebranych obserwacjach. W KROKU 3 odbywa się selekcja cech na próbkach danych np. przy użyciu heurystyki DAAR [26]. W KROKU 4 scalamy istotnie różne spośród wyznaczonych w poprzednim kroku podzbiorów atrybutów, aby uzyskać odpowiedni poziom resiliencji w procesie uczenia modeli. W KROKU 5 otrzymane wcześniej podzbiory obiektów i atrybutów są wykorzystywane do trenowania modeli przy użyciu wcześniej wybranych algorytmów. Wreszcie w KROKU 6 najważniejsze modele tworzą zespół. W celu zapewnienia wysokiej jakości predykcji, kryteria doboru modeli do zespołu uwzględniają zarówno jakość poszczególnych regresorów oraz stopień zróżnicowania powstałego zespołu. Przebieg kroków 5 i 6 jest szczegółowo wyjaśniony w Algorytmie 1. W wyniku połączenia różnych modeli ostateczny zespół minimalizuje wpływ dryfowania koncepcji i zapewnia lepszą jakość predykcji. Wyniki przeprowadzonych eksperymentów potwierdziły, że pomysł jest bardzo obiecujący, a odwołanie się do resiliencji w procesie uczenia znacząco ogranicza wpływ utraty danych na wyniki predykcji.

2.3 Ewaluacja i wybrane zastosowania zaproponowanych rozwiązań



Rysunek 3: Architektura systemu DISESOR.

Omawiając znaczenie i potencjał komercyjnego wykorzystania opracowanych mechanizmów możemy odwołać się do projektu DISESOR [31, 43], którego celem było stworzenie systemu wspomagania decyzji (por. Rysunek 3). Wybrane spośród omawianych zastosowań dotyczyły podniesienia poziomu bezpieczeństwa w górnictwie węgla kamiennego, gdzie występuje wiele zagrożeń naturalnych, spośród których dwa najistotniejsze to zagrożenia metanowe i sejsmiczne [22]. Obecnie dominującą rolę w ocenie ryzyka w kopalniach odgrywają metody eksperckie. Naszym celem było wytrenowanie modeli prognostycznych wspierających to zadanie. Dane zebrane do analizy tego problemu pochodziły z wybranych systemów monitorowania zagrożeń, które zostały zintegrowane z hurtownią danych systemu DISESOR [12]. Pierwotnym źródłem informacji są sieci czujników w kopalniach, których odczyty mogą przybierać różne formy i wymagają specyficznego przetwarzania [53]. Ich wartości mogą wyrażać zjawiska ciągłe, takie jak ciśnienie, temperatura lub stężenie metanu czy wartości binarne takie jak włączenie/wyłączenie danego urządzenia (por. Rysunek 3).

Analiza i prawidłowa ocena potencjalnie niebezpiecznego stężenia metanu [53] i zda-

Algorithm 1: Algorytm konstrukcji różnolitego zespołu modeli.

Data:
- *attSubsets* - wyznaczone resilientne podzbiory atrybutów
- *objectSamples* - próbki danych
- *testSet* - zbiór testowy
- *regressionAlgorithms*, wybrane algorytmy ML, np. { rPart, SVM, glm }
- *allowedAttempts* oraz *minQuality* - parametry jakości i zróżnicowania zespołu modeli

Result: *zespół modeli*

```
ensemble ← ∅; weakAttempts ← 0
alg ← regressionAlgorithms.removeFirst
while TRUE do
    a1, a2 ← attSubsets.drawAndRemoveTwo
    b1, b2 ← objectSamples.drawAndRemoveTwo
    model ← alg.trainAndEvaluate(a1, b1, a2, b2)
    score ← model.score(testSet)
    if score > minQuality ∧ ¬ensemble.containsSimilar(model, score) then
        ensemble ← ensemble ∪ {model ⊕ score}
    else
        weakAttempts ← weakAttempts + 1
        if weakAttempts < allowedAttempts then
            continue;
        end
        if regressionAlgorithms ≠ ∅ then
            alg ← regressionAlgorithms.removeFirst
            weakAttempts ← 0
        else
            break;
        end
    end
end
return ∑_{s∈ensemble.scores} s;
```

rzeń sejsmicznych [23] znacząco poprawia bezpieczeństwo i obniża koszty podziemnego wydobycia węgla. Jakkolwiek, w związku z ekstremalnymi warunkami panującymi w kopalniach, zapewnienie ciągłości działania systemów monitorowania i predykcji zagrożeń stanowi istotne wyzwanie. Uzależnienie pracy tych systemów od dostępności danych pochodzących z wielu strumieni odczytów z dużej sieci czujników zlokalizowanych pod ziemią, podnosi ryzyko związane z niedostępnością informacji. Z powodu niebezpiecznych zdarzeń czujniki lub przewody przesyłowe mogą ulec uszkodzeniu (oznaczone czerwonym kolorem na rysunku 3). Może to powodować luki w zebranych odczytach, a tym samym wpływać na niedostępność całych wymiarów danych. W ten sposób, wszystkie modele predykcyjne wykorzystujące cechy uzyskane z odczytów uszkodzonych czujników stałyby się bezużyteczne, dlatego zapewnienie kontrolowanego poziomu redundancji informacji – na tyle małego, żeby nie wprowadzać niepotrzebnego szumu informacyjnego, ale na tyle dużego, aby zapewnić ciągłość działania w sytuacji awarii wybranych czujników – jest bardzo ważnym zagadnieniem i stanowi trzon naszych rozważań.

Dążąc do gruntownej oceny wszechstronności opracowanego rozwiązania przeprowadziliśmy eksperymenty na danych dotyczących istotnie różnych problemów w wielu dziedzinach zastosowań. Poza analizą danych dotyczących wydobycia węgla kamiennego, zajmujemy się predykcją cen zasobów w chmurach obliczeniowych oraz problemem monitorowania bezpieczeństwa strażaków podczas akcji pożarniczych. W pierwszym przypadku problem dotyczy

zmienności cen zasobów kupowanych w tzw. modelu spot. Dostawcy rozwiązań starają się zoptymalizować wykorzystanie serwerów, aby uniknąć ich bezczynności i pików obciążeń [29]. Doprowadziło to do powstania swoistej giełdy zasobów komputerowych, w ramach której operator chmury obliczeniowej i podmioty z niej korzystające mogą licytować obecnie niewykorzystywaną moc obliczeniową w czasie zbliżonym do rzeczywistego. Istotą jest to, że cena i dostępność takich zasobów może się zmieniać w czasie, a jeśli wcześniej zalicytowana oferta zostanie przebita, usługa może zostać niezwłocznie wyłączona, co może zakłócić trwające obliczenia. Pokazujemy, że możliwe jest skuteczne prognozowanie przyszłych cen na giełdzie zasobów spot na podstawie analizy ich historycznych ofert. W tym celu operujemy na bardzo prostych i zrozumiałych technikach, jak modele ARIMA i reprezentacja danych przedstawionych jako świece japońskie, ponieważ obie te techniki są stosunkowo łatwe do interpretacji i często stosowane przez ekspertów [17]. Uniwersalność naszego rozwiązania podkreślamy poprzez jego adaptację do zadania związanego z monitorowaniem aktywności strażaków w trakcie akcji gaszenia pożaru. W tym wypadku, dane pochodzą z sieci czujników umieszczonych na ciele strażaków, które monitorują aktywność (akcelerometry i żyroskopy umieszczone na różnych częściach ciała) oraz funkcje życiowe strażaków (takie jak puls czy oddech) [16]. Wyniki przeprowadzonych eksperymentów potwierdzają jakość rozwiązania, wydajność obliczeniową i uniwersalność opracowanej metodologii, co skutkuje bardzo krótkim czasem potrzebnym na adaptację do istotnie różnych problemów.

W rozprawie przedstawiamy w jaki sposób metody interaktywnej ekstrakcji cech i budowy różnolitego zespołu modeli predykcyjnych można zastosować do analizy wielowymiarowych strumieni czasowych w różnych dziedzinach. Ponadto, opisujemy szereg międzynarodowych konkursów analizy danych, które z jednej strony umożliwiły obiektywną weryfikację zaproponowanego podejścia na danych dotyczących rzeczywistych problemów komercyjnych. Z drugiej zaś strony, pozwoliły porównać opracowane metody z rozwiązaniami przygotowanymi przez wiele grup badawczych z różnych krajów [24, 27, 36]. W tym miejscu warto podkreślić, że w przypadku analizy danych górniczych, podejście oparte na budowie różnolitego zespołu modeli w oparciu o resilientną selekcję cech osiągnęło doskonałą jakość predykcji, przewyższając inne rozwiązania [10, 15, 53]. W ramach przeprowadzonych badań, wykorzystujemy dane pokonkursowe do badania tzw. problemu zimnego startu, który jest szczególnie istotny przy wdrażaniu systemów wspomagania decyzji na nowych danych, np. pochodzących z nowych wyrobisk węgla [23]. Zaproponowane podejście do ekstrakcji cech daje doskonałe wyniki, nawet w połączeniu z bardzo prostymi technikami predykcji, takimi jak regresja logistyczna. Ponadto, koncepcja ta sprawdza się w sytuacji, gdy niektóre źródła danych (tj. czujniki) są uszkodzone lub nieaktywne.

3 Podsumowanie

W badaniu nad interaktywnością metodologii ekstrakcji cech dążymy do uzyskania istotnych i zrozumiałych atrybutów z surowych odczytów czujników oraz do zmniejszenia liczby tych atrybutów w celu uzyskania możliwie najprostszych, ale dokładnych modeli. Zaproponowane algorytmy konstrukcji i selekcji cech wykorzystują różne formy granulacji przestrzeni atrybutów [11], odpowiadając na potrzebę intuicyjnego wyrażania złożonych pojęć. W pracy formułujemy także zagadnienie resilientnej selekcji cech, przedstawiając je w oparciu o r-$\mathbb{C}$-redukty. W ramach badań dowodzimy NP-trudności tego zagadnienia oraz proponujemy kilka algorytmów (zachłannych i aproksymacyjnych) pozwalających wyznaczać r-$\mathbb{C}$-redukty z wielowymiarowych danych i przedstawiamy analizę ich złożoności [9, 14]. Proponowane podejścia do interaktywnej ekstrakcji cech zostały opracowane w oparciu o doświadczenia zdobyte w trakcie kilku projektów badawczych w dziedzinie przetwarzania

strumieni z wielu czujników w różnych dziedzinach [10, 16], a także analizy danych tekstowych [13] oraz analizy cen zasobów obliczeniowych w kilkunastu regionach operatora chmury obliczeniowej AWS [17]. Przedstawione metody zostały zweryfikowane pod względem jakości uzyskanych cech, jak również przepustowości, skalowalności i stabilności ich działania, a omawiana metodologia została z powodzeniem zastosowana w kilku rzeczywistych problemach związanych z analizą szeregów czasowych [8, 13, 23, 63].

W rozprawie omawiamy zadanie interaktywnej ekstrakcji cech i przedstawiamy kilka innowacyjnych podejść do automatyzacji procesu ich tworzenia i selekcji. Odpowiadamy na szereg wyzwań związanych, między innymi, ze zrozumiałą i zwięzłą reprezentacją analizowanych danych czy możliwością osadzenia w danych wiedzy dziedzinowej. Opracowane rozwiązania zostały dogłębnie rozważone zarówno ze strony teoretycznej jak i praktycznej oraz gruntownie ocenione pod kątem różnych aspektów jakościowych, uwzględniając różnorodność rozwiązania czy jego odporność na braki danych. Rozprawa dostarcza kompleksowego uzasadnienia dla tego kierunku badań, budując solidne podstawy teoretyczne do dalszych rozważań związanych z interaktywnością procesu ekstrakcji cech i uczenia maszynowego.

Literatura

[1] J. G. Bazan, S. Bazan-Socha, S. Buregwa-Czuma, Ł. Dydo, W. Rząsa, and A. Skowron. A Classifier Based on a Decision Tree with Verifying Cuts. *Fundam. Informaticae*, 143(1-2):1–18, 2016.

[2] Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[3] R. Bruni, C. Daraio, and D. Aureli. Imputation techniques for the reconstruction of missing interconnected data from higher educational institutions. *Knowledge-Based Systems*, 212:106512, 2021.

[4] C. Cornelis, R. Jensen, G. H. Martín, and D. Ślęzak. Attribute Selection with Fuzzy Decision Reducts. *Information Sciences*, 180(2):209–224, 2010.

[5] S. Dutta, A. Jankowski, G. Rozenberg, and A. Skowron. Linking reaction systems with rough sets. *Fundam. Informaticae*, 165(3-4):283–302, 2019.

[6] M. Garbulowski, K. Smolińska, K. Diamanti, G. Pan, K. Maqbool, L. Feuk, and J. Komorowski. Interpretable machine learning reveals dissimilarities between subtypes of autism spectrum disorder. *Frontiers in Genetics*, 12:73, 2021.

[7] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. *CoRR*, abs/1802.06893, 2018.

[8] M. Grzegorowski. Scaling of Complex Calculations over Big Data-Sets. In *AMT*, volume 8610 of *Lecture Notes in Computer Science*, pages 73–84. Springer, 2014.

[9] M. Grzegorowski. Governance of the Redundancy in the Feature Selection Based on Rough Sets' Reducts. In V. Flores, F. Gomide, A. Janusz, C. Meneses, D. Miao, G. Peters, D. Ślęzak, G. Wang, R. Weber, and Y. Yao, editors, *Rough Sets - International Joint Conference, IJCRS 2016, Santiago de Chile, Chile, October 7-11, 2016, Proceedings*, volume 9920 of *Lecture Notes in Computer Science*, pages 548–557, 2016.

[10] M. Grzegorowski. Massively Parallel Feature Extraction Framework Application in Predicting Dangerous Seismic Events. In *Proceedings of FedCSIS 2016*, pages 225–229, 2016.

[11] M. Grzegorowski, A. Janusz, D. Ślęzak, and M. S. Szczuka. On the Role of Feature Space Granulation in Feature Selection Processes. In J. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, and M. Toyoda, editors, *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*, pages 1806–1815. IEEE Computer Society, 2017.

[12] M. Grzegorowski, M. Kalisch, M. Kozielski, and Ł. Wróbel. Hurtowania danych i procesy ETL. In P. Przystałka and M. Sikora, editors, *Zintegrowany, szkieletowy system wspmagania decyzji dla systemów monitorowania procesów, urządzeń i zagrożeń*, chapter 3, pages 31–40. Monograficzna Seria Wydawnicza Instyututu Technik Innowacyjnych EMAG, 2017.

[13] M. Grzegorowski, P. W. Pardel, S. Stawicki, and K. Stencel. SONCA: Scalable Semantic Processing of Rapidly Growing Document Stores. In M. Pechenizkiy and M. Wojciechowski, editors, *New Trends in Databases and Information Systems, Workshop Proceedings of the 16th East European Conference, ADBIS 2012, Poznań, Poland, September 17-21, 2012*, volume 185 of *Advances in Intelligent Systems and Computing*, pages 89–98. Springer, 2012.

[14] M. Grzegorowski and D. Ślęzak. On resilient feature selection: Computational foundations of r-C-reducts. *Inf. Sci.*, 499:25–44, 2019.

[15] M. Grzegorowski and S. Stawicki. Window-Based Feature Engineering for Prediction of Methane Threats in Coal Mines. In Y. Yao, Q. Hu, H. Yu, and J. W. G. . Busse, editors, *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing - 15th International Conference, RSFDGrC 2015, Tianjin, China, November 20-23, 2015, Proceedings*, volume 9437 of *Lecture Notes in Computer Science*, pages 452–463. Springer, 2015.

[16] M. Grzegorowski and S. Stawicki. Window-based feature extraction framework for multi-sensor data: A posture recognition case study. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015*, pages 397–405. IEEE, 2015.

[17] M. Grzegorowski, E. Zdravevski, A. Janusz, P. Lameski, C. Apanowicz, and D. Ślęzak. Cost Optimization for Big Data Workloads Based on Dynamic Scheduling and Cluster-Size Tuning. *Big Data Research*, 25:100203, 2021.

[18] I. Guyon, M. Nikravesh, S. R. Gunn, and L. A. Zadeh, editors. *Feature Extraction - Foundations and Applications*, volume 207 of *Studies in Fuzziness and Soft Computing*. Springer, 2006.

[19] Y. h. Taguchi. *Unsupervised Feature Extraction Applied to Bioinformatics*. Springer, 2020.

[20] B. Hosseini and B. Hammer. Interpretable discriminative dimensionality reduction and feature selection on the manifold. In U. Brefeld, É. Fromont, A. Hotho, A. J. Knobbe, M. H. Maathuis, and C. Robardet, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part I*, volume 11906 of *Lecture Notes in Computer Science*, pages 310–326. Springer, 2019.

[21] L. Hu and Z. Zhang, editors. *EEG Signal Processing and Feature Extraction*. Springer, Singapore, 2019.

[22] A. Janusz, M. Grzegorowski, M. Kozielski, Z. Krzystanek, M. Sikora, D. Ślęzak, and Ł. Wróbel. Przykłady zastosowania systemu DISESOR w analizie i predykcji zagrożeń. In P. Przystałka and M. Sikora, editors, *Zintegrowany, szkieletowy system wspmagania decyzji dla systemów monitorowania procesów, urządzeń i zagrożeń*, chapter 11, pages 31–40. Monograficzna Seria Wydawnicza Instyututu Technik Innowacyjnych EMAG, 2017.

[23] A. Janusz, M. Grzegorowski, M. Michalak, Ł. Wróbel, M. Sikora, and D. Ślęzak. Predicting Seismic Events in Coal Mines Based on Underground Sensor Measurements. *Engineering Applications of Artificial Intelligence*, 64:83–94, 2017.

[24] A. Janusz, M. Sikora, Ł. Wróbel, S. Stawicki, M. Grzegorowski, P. Wojtas, and D. Ślęzak. Mining Data from Coal Mines: IJCRS'15 Data Challenge. In Y. Yao, Q. Hu, H. Yu, and J. W. Grzymała-Busse, editors, *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing - 15th International Conference, RSFDGrC 2015, Tianjin, China, November 20-23, 2015, Proceedings*, volume 9437 of *Lecture Notes in Computer Science*, pages 429–438. Springer, 2015.

[25] A. Janusz and D. Ślęzak. Rough Set Methods for Attribute Clustering and Selection. *Applied Artificial Intelligence*, 28(3):220–242, 2014.

[26] A. Janusz and D. Ślęzak. Computation of Approximate Reducts with Dynamically Adjusted Approximation Threshold. In *Proceedings of ISMIS 2015*, volume 9384 of *Lecture Notes in Computer Science*, pages 19–28. Springer, 2015.

[27] A. Janusz, D. Ślęzak, M. Sikora, and Ł. Wróbel. Predicting Dangerous Seismic Events: AAIA'16 Data Mining Challenge. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland, September 11-14, 2016*, volume 8 of *Annals of Computer Science and Information Systems*, pages 205–211. IEEE, 2016.

[28] M. Kang and J. Tian. *Machine Learning: Data Pre-processing*, pages 111–130. 2019.

[29] R. Keller, L. Häfner, T. Sachs, and G. Fridgen. Scheduling Flexible Demand in Cloud Computing Spot Markets. *Business & Information Systems Engineering*, 62(1):25–39, 2020.

[30] M. Kowalski, D. Ślęzak, K. Stencel, P. W. Pardel, M. Grzegorowski, and M. Kijowski. RDBMS Model for Scientific Articles Analytics, booktitle = Intelligent Tools for Building a Scientific Information Platform. volume 390 of *Studies in Computational Intelligence*, pages 49–60. Springer, 2012.

[31] M. Kozielski, M. Sikora, and Ł. Wróbel. DISESOR - decision support system for mining industry. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015*, volume 5 of *Annals of Computer Science and Information Systems*, pages 67–74. IEEE, 2015.

[32] A. L. Levada. Parametric PCA for unsupervised metric learning. *Pattern Recognition Letters*, 135:425 – 430, 2020.

[33] H. Liu and H. Motoda, editors. *Feature Extraction, Construction and Selection*. Springer, 1998.

[34] J. Luo, H. Fujita, Y. Yao, and K. Qin. On modeling similarity and three-way decision under incomplete information in rough set theory. *Knowledge-Based Systems*, 191:105251, 2020.

[35] A. J. Mason. *Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies*. PhD thesis, Imperial College London, 2009.

[36] M. Meina, A. Janusz, K. Rykaczewski, D. Ślęzak, B. Celmer, and A. Krasuski. Tagging Firefighter Activities at the Emergency Scene: Summary of AAIA'15 Data Mining Competition at Knowledge Pit. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015*, volume 5 of *Annals of Computer Science and Information Systems*, pages 367–373. IEEE, 2015.

[37] M. J. Moshkov, M. Piliszczuk, and B. Zielosko. On Construction of Partial Reducts and Irreducible Partial Decision Rules. *Fundamenta Informaticae*, 75(1-4):357–374, 2007.

[38] H. S. Nguyen and D. Ślęzak. Approximate Reducts and Association Rules – Correspondence and Complexity Results. In N. Zhong, A. Skowron, and S. Ohsuga, editors, *Proceedings of RSFDGrC 1999*, pages 137–145. Springer, 1999.

[39] M. S. Nixon and A. S. Aguado. *Feature Extraction and Image Processing for Computer Vision*. Academic Press, fourth edition edition, 2020.

[40] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data*, volume 9 of *System Theory, Knowledge Engineering and Problem Solving*. Kluwer, 1991.

[41] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[42] W. Pedrycz. Granular computing for data analytics: a manifesto of human-centric computing. *IEEE CAA J. Autom. Sinica*, 5(6):1025–1034, 2018.

[43] P. Przystałka and M. Sikora, editors. *Zintegrowany, szkieletowy system wspmagania decyzji dla systemów monitorowania procesów, urządzeń i zagrożeń*. Monograficzna Seria Wydawnicza Instyututu Technik Innowacyjnych EMAG, 2017.

[44] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[45] C. Rudin. Please stop explaining black box models for high stakes decisions. *CoRR*, abs/1811.10154, 2018.

[46] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. *Data Mining and Knowledge Discovery*, 4(2-3):89–125, 2000.

[47] J. S. Shah. *Novel statistical approaches for missing values in truncated high-dimensional metabolomics data with a detection threshold.* PhD thesis, University of Louisville, 2017.

[48] A. Skowron and S. Dutta. Rough sets: past, present, and future. *Nat. Comput.*, 17(4):855–876, 2018.

[49] A. Skowron and P. Wasilewski. Interactive information systems: Toward perception based computing. *Theor. Comput. Sci.*, 454:240–260, 2012.

[50] D. Ślęzak. Normalized Decision Functions and Measures for Inconsistent Decision Tables Analysis. *Fundamenta Informaticae*, 44(3):291–319, 2000.

[51] D. Ślęzak. Approximate Entropy Reducts. *Fundamenta Informaticae*, 53(3-4):365–390, 2002.

[52] D. Ślęzak. Rough Sets and Functional Dependencies in Data: Foundations of Association Reducts. *Transactions on Computational Science*, 5:182–205, 2009.

[53] D. Ślęzak, M. Grzegorowski, A. Janusz, M. Kozielski, S. H. Nguyen, M. Sikora, S. Stawicki, and Ł. Wróbel. A Framework for Learning and Embedding Multi-Sensor Forecasting Models into a Decision Support System: A Case Study of Methane Concentration in Coal Mines. *Information Sciences*, 451-452:112–133, 2018.

[54] D. Ślęzak, M. Grzegorowski, A. Janusz, and S. Stawicki. Interactive Data Exploration with Infolattices. Abstract Materials of BAFI 2015.

[55] D. Ślęzak, M. Grzegorowski, A. Janusz, and S. Stawicki. Toward Interactive Attribute Selection with Infolattices - A Position Paper. In *IJCRS (2)*, volume 10314 of *Lecture Notes in Computer Science*, pages 526–539. Springer, 2017.

[56] M. Smuk. *Missing Data Methodology: Sensitivity analysis after multiple imputation.* PhD thesis, University of London, 2015.

[57] S. Stawicki, D. Ślęzak, A. Janusz, and S. Widz. Decision Bireducts and Decision Reducts – A Comparison. *International Journal of Approximate Reasoning*, 84:75–109, 2017.

[58] R. W. Świniarski and A. Skowron. Rough Set Methods in Feature Selection and Recognition. *Pattern Recognition Letters*, 24(6):833–849, 2003.

[59] A. Wójtowicz. *Ensemble classification of incomplete data – a non-imputation approach with an application in ovarian tumour diagnosis support.* PhD thesis, University in Poznań, 2017.

[60] A. Wójtowicz, P. Żywica, A. Stachowiak, and K. Dyczkowski. Solving the problem of incomplete data in medical diagnosis via interval modeling. *Appl. Soft Comput.*, 47:424–437, 2016.

[61] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[62] L. A. Zadeh. From Computing with Numbers to Computing with Words - From Manipulation of Measurements to Manipulation of Perceptions. In B. Azvine, N. Azarmi, and D. D. Nauck, editors, *Intelligent Systems and Soft Computing: Prospects, Tools and Applications*, volume 1804 of *Lecture Notes in Computer Science*, pages 3–40. Springer, 2000.

[63] E. Zdravevski, P. Lameski, A. Dimitrievski, M. Grzegorowski, and C. Apanowicz. Cluster-size optimization within a cloud-based ETL framework for Big Data. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*, pages 3754–3763. IEEE, 2019.