Rozdział 4

WPŁYW TECHNIK WSTĘPNEGO PRZYGOTOWANIA DANYCH NA SKUTECZNOŚĆ KLASYFIKACJI OBIEKTÓW BAZY DERMATOLOGY ZA POMOCĄ ALGORYTMU LEM2

Dariusz Jankowski*

Streszczenie Choroby skóry mają wiele różnych odmian i są częstym obiektem badań w medycynie. Szczególnie niebezpieczne są wszelkie odmiany choroby prowadzące do powstania raka. Z dotychczasowych badań medycznych wynika, że wczesne wykrycie symptomów choroby nowotworowej pozwala na znaczne zmniejszenie prawdopodobieństwa powstania raka lub też jego rozwoju. Najnowsze doniesienia medyczne wskazują, że liczba pacjentów z chorobami skóry stale zwiększa się. Według prognoz do 2025 roku w Polsce liczba zachorowań na czerniaka skóry podwoi się. Trend jest taki sam również w innych krajach. Na całym świecie są prowadzone liczne badania nad poszukiwaniem najbardziej efektywnych metod budowy modeli danych opisujących choroby dermatologiczne skóry człowieka. Modele te budowane są na bazie niepełnych danych - na podstawie wybranej próby statystycznej, co wymusza zastosowanie metod, pozwalających na uogólnianie wyników uzyskanych z próby. Głównym problemem badawczym niniejszej pracy było poszukiwanie odpowiedzi na pytanie: jaki wpływ na jakość klasyfikacji chorób skóry modelu opartego na algorytmie LEM2 mają techniki wstępnego przygotowania danych (podziału tabeli, usuwania brakujących wartości, dyskretyzacji, selekcji atrybutów) w pozyskiwaniu wiedzy z bazy danych Dermatology, zawierającej historię osób cierpiących na choroby skóry. Podczas poszukiwania podobnych badań zauważono małą liczbę wyników z wykorzystaniem LEM2 i bazy Dermatology.

Słowa kluczowe: klasyfikacja chorób skóry, algorytm LEM2, zbiory przybliżone, selekcja atrybutów, dyskretyzacja

* Wydział Informatyki, Politechnika Białostocka, Wiejska 45A, 15-351 Białystok, d.jankowski@pb.edu.pl

DOI 10.24427/978-83-66391-58-1_4

Wprowadzenie

Choroby skóry mają wiele różnych odmian i są częstym obiektem badań w medycynie. Szczególnie niebezpieczne są wszelkie odmiany choroby prowadzące do powstania raka. Z dotychczasowych badań medycznych wynika, że wczesne wykrycie symptomów choroby nowotworowej pozwala na znaczne zmniejszenie prawdopodobieństwa powstania raka lub też jego rozwoju. Podkreśla to Amerykańska Fundacja Raka Skóry (Foundation, 2020), która definiuje raka skóry jako „niekontrolowany wzrost nieprawidłowych komórek w naskórku, najbardziej zewnętrznej warstwie skóry, spowodowany przez nienaprawialne, uszkodzone DNA, które wyzwala mutacje. Mutacje te prowadzą do szybkiego namnażania się komórek skóry i tworzenia złośliwych guzów. Główne typy raka skóry to rak podstawnokomórkowy (BCC), rak płaskonabłonkowy (SCC), czerniak i rak z komórek Merkla (MCC)". „Skin cancer is the out-of-control growth of abnormal cells in the epidermis, the outermost skin layer, caused by unrepaired DNA damage that triggers mutations. These mutations lead the skin cells to multiply rapidly and form malignant tumors. The main types of skin cancer are basal cell carcinoma (BCC), squamous cell carcinoma (SCC), melanoma and Merkel cell carcinoma (MCC)". Ich badania ustaliły, że przyczyną większości chorób raka skóry są: szkodliwe promienie ultrafioletowe (UV) słoneczne oraz korzystanie z solariów UV. „The two main causes of skin cancer are the sun's harmful ultraviolet (UV) rays and the use of UV tanning machines" (Foundation, 2020).

Najnowsze doniesienia medyczne wskazują, że liczba pacjentów z chorobami skóry stale zwiększa się (Didkowska, Wojciechowska, Czderny, Olasek i Ciuba, 2019). Według prognoz do 2025 roku w Polsce liczba zachorowań na czerniaka skóry podwoi się (Didkowska, Wojciechowska i Zatorski, 2009). Rosnący trend zachorowalności jest taki sam również w innych krajach. Zgodnie ze statystykami (Wojciechowska i Didkowska, 2020) w Polsce jest prawie o połowę mniejsza zachorowalność na czerniaka skóry niż w Unii Europejskiej, natomiast nieco większa niż przeciętna umieralność (o około 20%).

Na całym świecie prowadzone są liczne badania nad poszukiwaniem najbardziej efektywnych metod budowy modeli danych opisujących choroby dermatologiczne skóry człowieka. Modele te budowane są w oparciu o niepełne dane - na podstawie wybranej próby statystycznej, co wymusza zastosowanie metod, pozwalających na uogólnianie wyników.

Za budowę i przetwarzanie modeli danych medycznych obecnie odpowiedzialne są wyspecjalizowane programy komputerowe, wykorzystujące metody statystyczne i metody sztucznej inteligencji. Na ich podstawie proces wykrywania chorób oraz proces podejmowania decyzji o sposobie leczenia pacjentów staje się coraz bardziej efektywny.

Wśród baz danych ogólnodostępnych, które umożliwiają poszukiwanie nowych metod analitycznych, służących budowie modeli danych, dominują:

1. HAM1000.
2. Dermatology (Dua i Graff, 2017).
3. Melanoma Gene Database (MGDB).
4. Melanoma (Australia bioplatforms data portal).
5. MelanomaDB.

Ze względu na rodzaj informacji w nich zawartych o chorobach skóry, bazy te można podzielić na:

- Bazy obrazów skóry z etykietami (np. HAM10000).
- Bazy genów (np. MelanomaDB).
- Bazy danych opisowych (np. Dermatology).

Bazy danych opisowe reprezentowane są formalnie jako systemy informacyjne.

Definicja 4.1. Systemem informacyjnym S (Pawlak, 1980) (Pawlak, 1991) nazywamy układ:

$$SI =< U, A, V, f > \tag{4.1}$$

gdzie:

- U - niepusty, skończony zbiór obiektów zwany uniwersum,
- A - niepusty, skończony zbiór atrybutów opisujących obiekty uniwersum,
- $V = \cup_{a \in A} V_a$, gdzie V_a jest zbiorem wartości atrybutu a, zaś card(V_a)>1,
- $f : U \times A \rightarrow V$ - funkcja informacji, taka że: $\forall_{u \in U, a \in A}\ f(u,a) \in V_a$.

Baza danych Dermatology jest przykładem reprezentacji systemu informacyjnego.

Zgodnie z Pawlak (2005) oraz Stepaniuk (2008) podczas analizy danych w systemach informacyjnych podstawową kwestią jest poszukiwanie wzorców wśród danych, w celu odnalezienia zależności pomiędzy wybranymi zbiorami atrybutów.

Charakterystyczną cechą systemów informacyjnych, zawierających dane medyczne, jest problem występowania brakujących wartości oraz błędnie wprowadzonych danych do systemu informacyjnego lub błędnie zmierzonymi wartościami, o czym piszą: Little i in. (2012), Dziura, Post, Zhao, Fu i Peduzzi (2013), O'Neill i Temple (2012), Pezoulas i in. (2019), Cao, Stojkovic i Obradovic (2016), Khare i in. (2017), Tremblay, Hevner i Berndt (2012), Thabane i in. (2013), Kannan, Manoj i Arumugam (2015). Powyższe badania potwierdzają, że problem jest wciąż aktualny. Od jakości danych zależy jakość znajdowanych wzorców danych, które służą do budowy systemów decyzyjnych, a następnie trafności decyzji takich systemów.

Do innych, często spotykanych problemów w analizie danych medycznych są wykorzystywane systemy informacyjne zawierające historię o małej liczbie pacjentów, lecz wielu atrybutach (np. bazy z danymi genetycznymi) oraz systemy o bardzo dużej liczbie obiektów i wielu atrybutach. W zależności od metody analitycznej, przetworzenie wszystkich informacji może nie być możliwe i dlatego wymagany jest

dodatkowy etap preselekcji atrybutów, aby zmniejszyć wymiarowość przestrzeni poszukiwanych rozwiązań.

W przypadku wielu badań eksperymentalnych, można spotkać bazy zawierające małą liczbę obiektów i atrybutów, co związane może być np. z ograniczoną liczbą ochotników biorących udział w eksperymencie. W takim przypadku, wyniki obarczone są dodatkowym ryzykiem niedopasowania wzorców do całej populacji.

Należy zauważyć, że powyższe problemy z danymi dotyczą wszystkich systemów informacyjnych, a nie tylko medycznych.

Chcąc odpowiedzieć na bieżące problemy, autor niniejszej pracy przedstawił nowe możliwości wykorzystania algorytmu LEM2 (Grzymala-Busse, 1992) do pozyskiwania modeli danych z medycznych baz danych opisowych na podstawie ogólnodostępnej bazy Dermatology (Dua i Graff, 2017), zawierającej historię osób cierpiących na choroby skóry. Wyniki tych badań mają charakter uniwersalny i można je wykorzystać przy analizie innych systemów informacyjnych.

Algorytm LEM2 należy do zbioru metod poszukiwania minimalnego zbioru reguł w systemach informacyjnych za pomocą indukcji reguł. Umożliwia przetwarzanie tablic decyzyjnych zawierających sprzeczności. Różni się pod tym względem od pozostałych metod, jak np. drzewa decyzyjne, które wymagają usunięcia sprzeczności przed etapem budowy modelu danych. Algorytm ten nie zastępuje metod badania obrazów skóry chorób pacjentów czy też genów, a jedynie je uzupełnia.

Analiza porównawcza dostępnych badań naukowych nad bazą Dermatology oraz wykorzystania algorytmu LEM2 do badań medycznych nad nią, wykazała małe zainteresowanie wykorzystaniem tej metody do budowy klasyfikatorów wykrywających choroby skóry i symptomów raka. Większość opublikowanych badań wykorzystuje modele oparte o metody sieci neuronowych, drzewa decyzyjne i inne. Dotychczasowe badania z wykorzystaniem algorytmu LEM2 w stosunku do danych o chorobach skóry, skupiały się w głównej mierze na benchmarkingu metod. Opublikowane badania wykazują skuteczność predykcji klasyfikatorów opartych o algorytm LEM2 w zakresie 87-90%, a jednocześnie wysoką skuteczność takich metod jak: sieci neuronowe, drzewa decyzyjne, SVM - gdzie uzyskano skuteczność klasyfikacji na poziomie 95 - 100%. Zestawienie skuteczności klasyfikatorów dla różnych baz danych, w tym bazy Dermatology, przygotował Zhang, Liu, Zhang i Almpanidis (2017). Zgodnie z jego zestawieniem, najskuteczniejszą metodą klasyfikacji bazy Dermatology jest klasyfikator zbudowany na podstawie algorytmu SVM, dla którego współczynnik skuteczności predykcji wyniósł 100%.

Kusunoki i Inuiguchi (2006) na podstawie algorytmu LEM2 w stosunku do bazy Dermatology zbudowali klasyfikator o skuteczności predykcji 90.24%, a Borowik, Kraśniewski i Łuba (2015) uzyskali skuteczność 87,77% używając systemu RSES oraz 78% wykorzystując metodę autorską.

Srimani i Koti (2014) wykorzystali co prawda algorytm LEM2 do wygenerowania reguł, otrzymując współczynnik pokrycia równy 90%, jednak nie zbudowali klasyfikatora i testów jego skuteczności. Badania Koti (2014) również objęły ana-

lizę pokrycia reguł wygenerowanych z użyciem algorytmu LEM2 w systemie RSES (bez budowy klasyfikatora), a także badania zbioru PIMA (zawierającego przypadki pacjentów cierpiących na cukrzycę), dla którego skuteczność algorytmu LEM2 wyniosła 76%.

Metoda przyśpieszająca generowanie reduktów, zaprezentowana w Borowik (2019) pozwoliła na obliczenie wszystkich reduktów bazy danych Dermatology w 2 minuty. Autor zwrócił uwagę, że w systemie RSES obliczenie reduktów nie było możliwe z powodu dużego zużycia pamięci.

W dalszej części artykułu, autor pracy prezentuje wyniki i możliwości dalszego rozwoju prac nad wykorzystaniem algorytmu LEM2 w analizie danych medycznych.

4.1 Metodyka badań

4.1.1 Opis danych i stanowiska badawczego

Zgodnie z wprowadzeniem, do badania została wybrana baza danych Dermatology, opublikowana w Dua i Graff (2017). Od strony programistycznej zostały wykorzystane biblioteki języka R oraz Python, z wyszczególnieniem bibliotek: RoughSets, arules oraz scikit-learn. Poniżej przedstawiona jest charakterystyka bazy Dermatology.

- liczba przebadanych pacjentów: 366,
- liczba atrybutów opisujących: 34,
- liczba atrybutów decyzyjnych: 1 (6 klas decyzyjnych),
- liczba rekordów zawierających brakujące dane: 8 - wszystkie dotyczą atrybutu Age. Wartości brakujące zostały zastąpione znakiem '?'.

Oznaczenie klas decyzyjnych:

Kod klasy - Nazwa klasy - Liczba obiektów,
1 - psoriasis (łuszczyca) - 112,
2 - seboreic dermatitis (łojotokowe zapalenie skóry) - 61,
3 - lichen planus (liszaj płaski, liszaj czerwony, liszaj Wilsona) - 72,
4 - pityriasis rosea (łupież różowy Giberta) - 49,
5 - cronic dermatitis (przewlekłe zapalenie skóry) - 52,
6 - pityriasis rubra pilaris (łupież czerwony mieszkowy) - 20.

Podane choroby charakteryzuje jedna właściwość: trudno jest je rozpoznać za pomocą obserwacji i najczęściej potrzebne jest wykonanie biopsji, lecz niestety choroby te mają wiele cech histopatologicznych. Według autorów bazy, pacjenci byli w pierwszej kolejności badani klinicznie. Wyniki tych badań reprezentują cechy

o numerach: 1-11, 34. W drugiej kolejności badane były próbki skóry, a wyniki zapisane za pomocą cech o numerach: 12-33. Wszystkie nazwy cech dostępne są na stronie projektu Dua i Graff (2017). Należy zauważyć, że jedynie cecha Age zawiera rzeczywisty wiek pacjenta. Pozostałe cechy warunkowe są zakodowane. W przypadku cechy "family history"mamy dwie możliwe wartości: 1 - oznacza, że choroba wystąpiła wcześniej u innego członka rodziny pacjenta, a 0 - brak wystąpienia. Pozostałe cechy mogą przybierać wartości: 0, 1, 2, 3, gdzie: 0 - oznacza brak wystąpienia cechy, 3 - oznacza największą możliwą wartość (przedział), a 1 i 2 - odpowiednio wartości pośrednie.

4.1.2 Przygotowanie danych treningowych i testowych

Zgodnie z założeniami uczenia maszynowego, proces budowy rozwiązań dzieli się na etap trenowania (budowy modelu danych, klasyfikatora) oraz predykcji nowych obiektów (nieznanych na etapie trenowania) i zebraniu mierników jakościowych. Etap ten powtarza się wielokrotnie dla różnych kombinacji parametrów i różnych podziałów zbioru danych, a następnie wybiera się jeden z najlepszych.

Z uwagi na mały rozmiar danych (366 obiektów) oraz brak dostępu do nowych pacjentów, każdy nowy eksperyment rozpoczyna się podziałem bazy na 2 tabele: treningową i testową. Tabela treningowa używana jest wyłącznie na etapie trenowania, a tabela testowa wyłącznie w końcowej ocenie jakości klasyfikacji (zastępuje dane o nowych pacjentach). Przed dokonaniem podziału rekordy bazy Dermatology są losowo przestawiane.

Podział danych był dokonywany w 2 wariantach: 80:20 oraz 90:10.

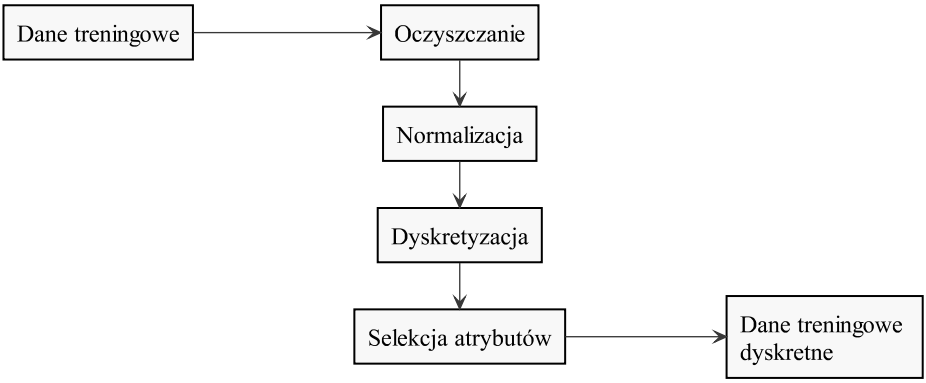
W badaniu starano się zachować równomierne proporcje klas w tabeli walidacyjnej, aby nie dopuścić do sytuacji, że większość przypadków należałaby do klasy 1 (najbardziej licznej).

Jak wspomniano wcześniej, baza Dermatology zawiera 8 rekordów z brakującymi wartościami cechy Age. W celu dostosowania bazy do dalszej analizy, obliczono dla każdej klasy decyzyjnej najczęściej występującą wartość i zgodnie z tym kryterium uzupełniono brakujące wartości w bazie danych.

Ponieważ baza danych została znormalizowana przez ich twórców, dlatego etap normalizacji został pominięty.

Z całej bazy, tylko 1 cecha (Age) wymagała przeprowadzania procesu dyskretyzacji. Do tego celu wykorzystywano zamiennie metody z pakietu RoughSets oraz arules.

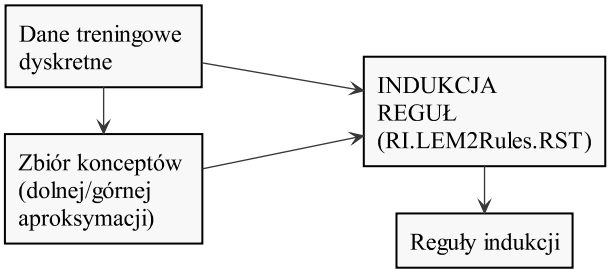
Do etapu selekcji atrybutów wybrano metody znajdowania reduktów z pakietu RoughSets.



Rysunek 4.1: Przygotowanie danych treningowych dla algorytmu LEM2

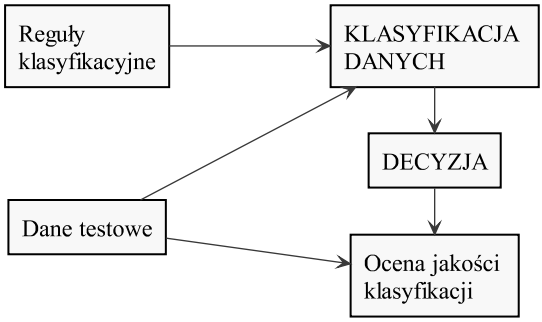
4.1.3 Indukcja reguł i ocena ich jakości

Indukcja reguł za pomocą algorytmu LEM2 wymaga określenia podzbioru danych treningowych, tworzących zbiór konceptów, a dokładnie ich dolnej lub górnej aproksymacji zgodnie z metodą zastosowaną w systemie LERS (Grzymala-Busse, 1997). W przypadku funkcji z biblioteki RoughSets (RI.LEM2Rules.RST), zbiorem konceptów jest zbiór obiektów należących do obszaru pozytywnego w rozumieniu teorii zbiorów przybliżonych.



Rysunek 4.2: Indukowanie reguł z użyciem algorytmu LEM2

Na podstawie reguł utworzonych na etapie indukcji przeprowadzono wstępną klasyfikację obiektów, korzystając ze zbioru testowego), utworzonego w I etapie przetwarzania. Podczas klasyfikacji, informacja o przyporządkowanej klasie była niedostępna dla klasyfikatora. Wyniki klasyfikacji zostały zebrane w formie tabelarycznej a następnie porównane z wartościami wcześniej zapisanymi w zbiorze testowym.



Rysunek 4.3: Wstępna weryfikacja reguł klasyfikacyjnych za pomocą danych testowych

Na podstawie wzorów dla wyznaczenia macierzy pomyłek i typowych mierników jakości przedstawionych przez Manliguez (2016) i Fawcett (2006) obliczono wskaźnik skuteczności klasyfikacji całego zbioru testowego (w celu porównania osiągniętych wyników z przedstawionymi we wprowadzeniu innymi badaniami) oraz dla każdej klasy obliczono takie wartości jak: czułość, specyficzność, PPV, NPV, skuteczność zrównoważoną oraz ich wartość ważoną wg wzoru:

$$WWM = \frac{\sum_{i=1}^{k} (WM_i) * (IL_i)}{\sum_{i=1}^{k} IL_i} \tag{4.2}$$

gdzie:

- WWM - wartość ważona miernika,
- WM_i - wartość miernika dla klasy i,
- IL_i - liczba obiektów klasy i,
- k - liczba klas.

Do obliczeń mierników wykorzystano funkcję confusionMatrix z biblioteki R - caret oraz funkcję F1_score z biblioteki MLmetrics.

4.1.4 Dziesięciokrotna walidacja krzyżowa z 20 powtórzeniami

W celu znalezienia najlepszego modelu danych, wykorzystano metodę dziesięciokrotnej walidacji krzyżowej z dwudziestoma powtórzeniami, zachowującą równomierne rozłożenie klas w każdej próbce danych. Do obliczeń użyto autorski estymator i funkcje RepeatedStratifiedKFold oraz GridSearchCV z biblioteki Python - scikit-learn, odpowiedzialne za wielokrotne budowanie modeli w języku R.

Podczas walidacji krzyżowej pominięto etap selekcji atrybutów i poszukiwano najlepsze modele danych, podstawiając na przemian następujące parametry dyskretyzacji z biblioteki arules:

- metoda dyskretyzacji: frequency, cluster, interval,
- liczba podziałów: 2, 3, 4, 5.

Do oceny jakości klasyfikacji użyta została funkcja classification_report z pakietu scikit-learn oraz statystyki, które zebrała metoda walidacji krzyżowej, tj. średnią wartość dokładności klasyfikacj zbioru testowego i odchylenie standardowe dla każdego zbioru parametrów wejściowych.

Podczas walidacji krzyżowej z powtórzeniami dla każdego zestawu parametrów, zostały wyznaczone najlepsze modele oraz zebrane statystyki, tj. średnia wartość dokładności klasyfikacj zbioru testowego i odchylenie standardowe.

Zestaw parametrów dla którego średnia wartość dokładności klasyfikacji zbioru testowego była największa, został wybrany jako najlepszy, a wyznaczony na jej podstawie najlepszy model, został wybrany jako końcowy model danych wyznaczony za pomocą metody poszukiwania hiperparametrów połączonej z walidacją krzyżową oraz poddano go ostatecznej ocenie jakości klasyfikacji na podstawie całego zbioru danych.

4.2 Rezultaty

W wyniku badań nad bazą Dermatology z użyciem algorytmu LEM2 i metod pomocniczych zauważono, że odpowiedni dobór metody dyskretyzacji oraz liczba punktów podziału poprawił znacząco skuteczność klasyfikacji obiektów w stosunku do przedstawionych podobnych badań nad bazą Dermatology z użyciem algorytmu LEM2, bez potrzeby używania etapu selekcji atrybutów.

Przy podziale wejściowego zbioru w proporcji 90:10, nie stosując walidacji krzyżowej, uzyskano skuteczność równie wysoką, jak najlepsze algorytmy z zestawienia, które przygotował Zhang i in., tj. 99% w przypadku najlepszego modelu danych. Ocenę jakości trzech najlepszych modeli prezentuje tabela 4.1.

Podział zbioru w proporcji 80:20 wykazał dokładność klasyfikacji na poziomie 95% za pomocą miernika jakości skuteczności zrównoważonej. Podział ten obarczony jest mniejszym błędem generalizacji niż w pierwszym przypadku. Różnica jakości klasyfikacji wynosi 4%. Wyniki prezentuje tabela 4.2.

Dla modelu danych o skuteczności zrównoważonej ważonej równej 99%, z badania o identyfikatorze W20, została wyznaczona macierz pomyłek (tabela 4.3) wraz ze szczegółowymi wskaźnikami klasowymi (tabela 4.4). Macierz pomyłek potwierdza, że tylko jeden przypadek testowy został błędnie zaklasyfikowany.

Tabela 4.1: Ocena klasyfikacji najlepszych modeli dla podziału 90:10

	Model 1	**Model 2**	**Model 3**
Identyfikator badania (ID)	W3	W16	W20
Metoda dyskretyzacji cechy Age	frequency	interval	cluster
Liczba punktów podziału cechy Age	3	3	3
Skuteczność	0,975	0,8919	0,973
F1-score	0,96	0,857	1
Czułość ważona	0,98	0,89	1
Specyficzność ważona	0,99	0,96	1
PPV ważona	0,98	0,91	0,97
NPV ważona	1	0,98	0,99
Skuteczność zrównoważona ważona	0,98	0,93	0,99

Tabela 4.2: Ocena klasyfikacji najlepszych modeli dla podziału 80:20

	Model 1	**Model 2**	**Model 3**
Identyfikator badania (ID)	W32	W40	W47
Metoda dyskretyzacji cechy Age	interval	frequency	cluster
Liczba punktów podziału cechy Age	3	3	3
Skuteczność	0,919	0,904	0,9189
F1-score	0,857	0,93	0,96
Czułość ważona	0,94	0,88	0,87
Specyficzność ważona	0,97	0,96	0,98
PPV ważona	0,92	0,9	0,93
NPV ważona	0,98	0,98	0,99
Skuteczność zrównoważona ważona	0,95	0,92	0,93

Tabela 4.3: Macierz pomyłek dla badania o identyfikatorze W20

	Reference					
Prediction	**1**	**2**	**3**	**4**	**5**	**6**
1	12	0	0	0	0	0
2	0	6	0	0	0	0
3	0	0	6	0	0	0
4	0	1	0	4	0	0
5	0	0	0	0	6	0
6	0	0	0	0	0	2

Tabela 4.4: Wskaźniki klasowe macierzy pomyłek dla badania o identyfikatorze W20

	Class					
	1	**2**	**3**	**4**	**5**	**6**
Sensitivity	1.0000	0.8571	1.0000	1.0000	1.0000	1.00000
Specificity	1.0000	1.0000	1.0000	0.9697	1.0000	1.00000
PPV	1.0000	1.0000	1.0000	0.8000	1.0000	1.00000
NPV	1.0000	0.9677	1.0000	1.0000	1.0000	1.00000
Prevalence	0.3243	0.1892	0.1622	0.1081	0.1622	0.05405
Detection Rate	0.3243	0.1622	0.1622	0.1081	0.1622	0.05405
Detection Prevalence	0.3243	0.1622	0.1622	0.1351	0.1622	0.05405
Balanced Accuracy	1.0000	0.9286	1.0000	0.9848	1.0000	1.00000

Końcowe poszukiwania najlepszego modelu danych zostały przeprowadzone przy pomocy metody poszukiwania najlepszych parametrów dyskretyzacji w połączeniu z metodą walidacji krzyżowej z powtórzeniami. Pozwoliły one wyznaczyć model danych o skuteczności klasyfikacji równej 100%, dla parametrów:

- metoda dyskretyzacji: interval,
- liczba podziałów: 4.

Najlepszy klasyfikator został wybrany spośród wygenerowanych 2 400 modeli danych.

Średnia skuteczność walidacji dla najlepszych parametrów wyniosła 87,5% z odchyleniem standardowym +/-0.100 (dokładność klasyfikacji najsłabszego modelu wyniosła 70,27% - 34 reguły, o maksymalnej długości równej 8).

W tabeli Tabela 4.5 zostały przedstawione wyniki oceny klasyfikacji wszystkich modeli opartych o zestawy parametrów podczas walidacji krzyżowej.

Zbiór testowy dla najlepszego modułu zawierał 37 przypadków, w którym klasy od 1 do 6 pokrywały odpowiednio: 11,7,7,5,5,2 przypadków. 100% skuteczność klasyfikacji modelu nie wyklucza mocnego dopasowania do zbioru danych, jednak obserwując wygenerowane reguły przy użyciu algorytmu LEM2 należy stwierdzić, że model został znacząco uogólniony w stosunku do zbioru wejściowego - zmniejszyła się liczba reguł z 366 do 32 przy równoczesnym ograniczeniu długości reguł. Przykładowo, długość reguły nr 11 o wsparciu 82 (22,4%) wyniosła 5. Poza tym, reguły nr 12 i 15, o długości równej 1, pokrywają dużą liczbę zbioru - 55 i 47 przypadków.

Najlepszy model danych został wyznaczony na podstawie następujących reguł, uzyskanych podczas indukcji algorytmem LEM2:

1 (disappearance of the granular layer,0) & (band-like infiltrate,0) & (koebner phenomenon,0) & (knee and elbow involvement,0) & (elongation of the rete ridges,0) & (hyperkeratosis,0) & (scaling,2) -> (class,2)

Tabela 4.5: Ocena jakości klasyfikacji hiperparametrów

Lp	Metoda dyskretyzacji	Liczba podziałów	Średnia skuteczność	Odchylenie standardowe
1	frequency	2	0.866	+/-0.099
2	frequency	3	0.870	+/-0.095
3	frequency	4	0.868	+/-0.097
4	frequency	5	0.869	+/-0.095
5	cluster	2	0.872	+/-0.104
6	cluster	3	0.870	+/-0.097
7	cluster	4	0.869	+/-0.101
8	cluster	5	0.871	+/-0.097
9	interval	2	0.872	+/-0.103
10	interval	3	0.862	+/-0.098
11	interval	4	0.875	+/-0.100
12	interval	5	0.869	+/-0.098

2 (fibrosis of the papillary dermis,0) & (disappearance of the granular layer,0) & (band-like infiltrate,0) & (koebner phenomenon,0) & (knee and elbow involvement,0) & (parakeratosis,0) -> (class,2)
3 (parakeratosis,2) & (acanthosis,2) & (erythema,2) & (spongiosis,3) -> (class,2)
4 (fibrosis of the papillary dermis,0) & (disappearance of the granular layer,0) & (band-like infiltrate,0) & (perifollicular parakeratosis,0) & (koebner phenomenon,0) & (scalp involvement,0) & (hyperkeratosis,0) & (PNL infiltrate,0) -> (class,2)
5 (koebner phenomenon,0) & (spongiosis,2) & (scaling,3) -> (class,2)
6 (disappearance of the granular layer,0) & (acanthosis,2) & (thinning of the suprapapillary epidermis,0) & (PNL infiltrate,1) -> (class,2)
7 (PNL infiltrate,2) & (knee and elbow involvement,0) & (age,(37.5,56.2]) & (parakeratosis,2) -> (class,2)
8 (fibrosis of the papillary dermis,0) & (koebner phenomenon,0) & (disappearance of the granular layer,0) & (band-like infiltrate,0) & (eosinophils in the infiltrate,1) -> (class,2)
9 (acanthosis,1) & (band-like infiltrate,2) -> (class,2)
10 (PNL infiltrate,1) & (knee and elbow involvement,1) & (focal hypergranulosis,0) & (elongation of the rete ridges,0) -> (class,2)
11 (spongiosis,0) & (fibrosis of the papillary dermis,0) & (eosinophils in the infiltrate,0) & (follicular papules,0) & (exocytosis,0) -> (class,1)
12 (thinning of the suprapapillary epidermis,2) -> (class,1)
13 (spongiosis,0) & (polygonal papules,0) & (oral mucosal involvement,0) & (fibrosis of the papillary dermis,0) & (perifollicular parakeratosis,0) & (definite borders,2) -> (class,1)

14 (definite borders,3) & (polygonal papules,0) & (fibrosis of the papillary dermis,0) -> (class,1)
15 (band-like infiltrate,3) -> (class,3)
16 (band-like infiltrate,2) & (scalp involvement,0) -> (class,3)
17 (band-like infiltrate,2) & (parakeratosis,2) -> (class,3)
18 (fibrosis of the papillary dermis,0) & (knee and elbow involvement,0) & (scalp involvement,0) & (PNL infiltrate,0) & (hyperkeratosis,0) & (itching,0) -> (class,4)
19 (fibrosis of the papillary dermis,0) & (family history,0) & (saw-tooth appearance of retes,0) & (PNL infiltrate,0) & (erythema,2) & (exocytosis,2) & (parakeratosis,1) -> (class,4)
20 (saw-tooth appearance of retes,0) & (PNL infiltrate,0) & (scaling,2) & (definite borders,2) & (exocytosis,3) -> (class,4)
21 (saw-tooth appearance of retes,0) & (eosinophils in the infiltrate,0) & (inflammatory monoluclear inflitrate,2) & (spongiosis,2) & (erythema,1) & (scaling,1) -> (class,4)
22 (inflammatory monoluclear inflitrate,2) & (age,(18.8,37.5]) & (disappearance of the granular layer,0) & (scaling,2) & (hyperkeratosis,2) -> (class,4)
23 (spongiform pustule,0) & (saw-tooth appearance of retes,0) & (hyperkeratosis,0) & (age,(18.8,37.5]) & (disappearance of the granular layer,0) & (koebner phenomenon,1) -> (class,4)
24 (eosinophils in the infiltrate,0) & (definite borders,0) & (spongiosis,3) & (parakeratosis,2) -> (class,4)
25 (thinning of the suprapapillary epidermis,0) & (saw-tooth appearance of retes,0) & (disappearance of the granular layer,1) -> (class,4)
26 (itching,0) & (erythema,1) & (koebner phenomenon,2) -> (class,4)
27 (PNL infiltrate,0) & (koebner phenomenon,0) & (band-like infiltrate,0) & (knee and elbow involvement,0) & (clubbing of the rete ridges,0) & (spongiosis,0) -> (class,5)
28 (PNL infiltrate,0) & (koebner phenomenon,0) & (disappearance of the granular layer,0) & (band-like infiltrate,0) & (follicular horn plug,0) & (scaling,1) -> (class,5)
29 (fibrosis of the papillary dermis,1) -> (class,5)
30 (fibrosis of the papillary dermis,0) & (koebner phenomenon,0) & (munro microabcess,0) & (age,(0,18.8]) & (PNL infiltrate,0) -> (class,6)
31 (perifollicular parakeratosis,2) -> (class,6)
32 (knee and elbow involvement,3) & (follicular papules,2) -> (class,6)

Tabela 4.6: Wskaźniki oceny reguł

Nr reguły	Dł. reguły	Pokrycie	Laplace	RI	Confidence
1	7	23	6.28	0.827586	1
2	6	19	5.19	0.8	1
3	4	3	0.82	0.444444	1
4	8	9	2.46	0.666667	1
5	3	7	1.91	0.615385	1
6	4	13	3.55	0.736842	1
7	4	3	0.82	0.444444	1
8	5	15	4.1	0.761905	1
9	2	1	0.27	0.285714	1
10	4	2	0.55	0.375	1
11	5	82	22.4	0.943182	1
12	1	55	15.03	0.918033	1
13	6	66	18.03	0.930556	1
14	3	24	6.56	0.833333	1
15	1	47	12.84	0.90566	1
16	2	16	4.37	0.772727	1
17	2	7	1.91	0.615385	1
18	6	20	5.46	0.807692	1
19	7	11	3.01	0.705882	1
20	5	4	1.09	0.5	1
21	6	3	0.82	0.444444	1
22	5	1	0.27	0.285714	1
23	6	6	1.64	0.583333	1
24	4	2	0.55	0.375	1
25	3	15	4.1	0.761905	1
26	3	2	0.55	0.375	1
27	6	33	9.02	0.871795	1
28	6	31	8.47	0.864865	1
29	1	7	1.91	0.615385	1
30	5	14	3.83	0.75	1
31	1	11	3.01	0.705882	1
32	2	1	0.27	0.285714	1

Podsumowanie

Celem niniejszej pracy było lepsze poznanie możliwości i wpływu metod, przetwarzających dane przed rozpoczęciem procesu indukcji reguł. Badania przeprowadzono na podstawie medycznej bazy danych Dermatology, przechowującej informacje o pacjentach chorych na 6 różnych chorób skóry. Przed rozpoczęciem badań zapoznano się również z innymi opracowaniami na ten temat stwierdzając, iż ich liczba jest niewielka w stosunku do badań nad innymi metodami uczenia maszynowego.
W niniejszej pracy poprawiono znacząco uzyskane do tej pory wyniki, uzyskując rozwiązanie o 100% skuteczności klasyfikacji. Wyniki potwierdziły, że algorytm

LEM2 może być bardzo skuteczny, ale trzeba zachować szczególną uwagę, w jaki sposób przygotowuje się dla niego dane. Z przedstawionych rezultatów wynika, że mierniki jakościowe ważone dokładniej opisują skuteczność klasyfikacji reguł opartych o algorytm LEM2, a zatem również wpływ metod odpowiedzialnych za przygotowanie danych dla algorytmu LEM2. Badania potwierdziły dokładność znajdowania modeli danych o dużej skuteczności klasyfikacji za pomocą przeszukiwania przestrzeni parametrów przy równoczesnym zastosowaniu walidacji krzyżowej z powtórzeniami.

Wprowadzenie do oceny jakości klasyfikacji mierników ważonych pozwoliło spojrzeć na bazę Dermatology i możliwości algorytmu LEM2 w nowy sposób, niespotykany dotąd w literaturze. Z uwagi na dysproporcje w liczbie obiektów w każdej z klas, mierniki ważone, a szczególnie ważony miernik skuteczności zrównoważonej (ang. balanced accuracy), wydają się bardziej adekwatne do oceny jakości tego typu modeli danych. W przypadku modelu o 100% skuteczności, nie miały jednak większego znaczenia.

Używając wyników tej pracy, autor zamierza kontynuować badania na większych zbiorach danych i doskonalić metodę oceny jakości klasyfikacji za pomocą wskaźników ważonych. Baza Dermatology jest znormalizowana przez jej autorów, więc dostęp do pierwotnego zbioru danych mógłby umożliwić wyznaczenie lepszych wartości dyskretyzacji.

Bibliografia

Borowik, G. (2019). *Methods and algorithms of logic synthesis in data analysis and data mining*. Unpublished doctoral dissertation. Źródło: `https://wcy.wat.edu.pl/sites/default/files/gb_autoreferat_en.pdf`, dostęp: 04.12.2020.

Borowik, G., Kraśniewski, A. i Łuba, T. (2015). Rule Induction Based on Logic Synthesis Methods. *Advances in Intelligent Systems and Computing*, *1089*, 813–816.

Cao, X. H., Stojkovic, I. i Obradovic, Z. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics*, *17*, 359.

Didkowska, J., Wojciechowska, U., Czderny, K., Olasek, P. i Ciuba, A. (2019). *Nowotwory złośliwe w Polsce w 2017 roku.* Źródło: `http://onkologia.org.pl/wp-content/uploads/nowotwory_2017.pdf`, dostęp: 04.12.2020.

Didkowska, J., Wojciechowska, U. i Zatorski, W. (2009). *Prognozy zachorowalności i umieralności na nowotwory złośliwe w Polsce do 2025 roku.*

Dua, D., i Graff, C. (2017). *UCI machine learning repository.* Źródło:

https://archive.ics.uci.edu/ml/datasets/Dermatology, dostęp: 04.12.2020.

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z. i Peduzzi, P. (2013). Strategies for dealing with missing data in clinical trials: from design to analysis. *The Yale journal of biology and medicine*, *86*, 343–58.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Foundation, S. C. (2020). *Skin cancer 101.* Źródło: https://www.skincancer.org/skin-cancer-information/, dostęp: 04.12.2020.

Grzymala-Busse, J. W. (1992). LERS-A System for Learning from Examples Based on Rough Sets. W: R. Slowinski (red.), *Intelligent decision support. handbook of applications and advances of the rough sets theory,* Springer, 3–18.

Grzymala-Busse, J. W. (1997). A New Version of the Rule Induction System LERS. *Fundamenta Informaticae*, *31*, 27–39.

Kannan, K. S., Manoj, K. i Arumugam, S. (2015). Labeling Methods for Identifying Outliers. *International Journal of Statistics and Systems(IJSS)*, *10*, 231–238.

Khare, R., Utidjian, L., Ruth, B. J., Kahn, M. G., Burrows, E., Marsolo, K., Patibandla, N., Razzaghi, H., Colvin, R., Ranade, D., Kitzmiller, M., Eckrich, D. i Bailey, L. C. (2017). A longitudinal analysis of data quality in a large pediatric data research network. *Journal of the American Medical Informatics Association*, *24*, 1072–1079.

Koti, M. S. (2014). *RST Approach for the Prediction of Rules and Cost Effective Feature Selection in Medical Data*. Unpublished doctoral dissertation, Bharathiar University. Źródło: http://hdl.handle.net/10603/97869, dostęp: 04.12.2020.

Kusunoki, Y., i Inuiguchi, M. (2006). Rule Induction Via Clustering Decision Classes. W: S. Greco i in. (red.), *Rough sets and current trends in computing,* Springer, 928–938.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P. i Stern, H. (2012). The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, *367*, 1355–1360.

Manliguez, C. (2016). Generalized Confusion Matrix for Multiple Classes. *Machine Learning*.

O'Neill, R. T., i Temple, R. (2012). The Prevention and Treatment of Missing Data in Clinical Trials: An FDA Perspective on the Importance of Dealing With It. *Clinical Pharmacology & Therapeutics*, *91*, 550–554.

Pawlak, Z. (1980). Toward the Theory of Information Systems. W: *CS PAS Reports 419/80,* 1–35.

Pawlak, Z. (1991). *Rough Sets Theoretical Aspects of Reasoning about Data*. Springer, Dordrecht. Źródło: https://bcpw.bg.pw.edu.pl/Content/

1845/download/, dostęp: 04.12.2020.

Pawlak, Z. (2005). A Treatise on Rough Sets. W: *Transactions on Rough Sets IV,* Springer, 1–17.

Pezoulas, V. C., Kourou, K. D., Kalatzis, F., Exarchos, T. P., Venetsanopoulou, A., Zampeli, E., Gandolfo, S., Skopouli, F., De Vita, S., Tzioufas, A. G. i Fotiadis, D. I. (2019). Medical data quality assessment: On the development of an automated framework for medical data curation. *Computers in Biology and Medicine*, *107*, 270–283.

Srimani, P. K., i Koti, M. S. (2014). Knowledge discovery in medical data by using rough set rule induction algorithms. *Indian Journal of Science and Technology*, *7*, 905–915.

Stepaniuk, J. (2008). *Rough – Granular Computing in Knowledge Discovery and Data Mining* (152). Berlin, Heidelberg: Springer Berlin Heidelberg.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M., Giangregorio, L., Dennis, B., Kosa, D., Debono, V. B., Dillenburg, R., Fruci, V., Bawor, M., Lee, J., Wells, G. i Goldsmith, C. H. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Medical Research Methodology*, *13*, 92.

Tremblay, M. C., Hevner, A. R. i Berndt, D. J. (2012). Design of an information volatility measure for health care decision making. *Decision Support Systems*, *52*, 331–341.

Wojciechowska, U., i Didkowska, J. (2020). Zachorowania i zgony na nowotwory złośliwe w Polsce - Czerniak skóry (C43). Lata 1965-2010. *Krajowy Rejestr Nowotworów, Narodowy Instytut Onkologii im. Marii Skłodowskiej-Curie – Państwowy Instytut Badawczy*. Źródło: http://onkologia.org.pl/czerniak-skory-c43/, dostęp: 04.12.2020.

Zhang, C., Liu, C., Zhang, X. i Almpanidis, G. (2017, oct). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, *82*, 128–150.