Politechnika Warszawska

Wydział Matematyki
i Nauk Informacyjnych

Praca dyplomowa licencjacka

na kierunku Matematyka i Analiza Danych

Modyfikacja submodularnego algorytmu selekcji zmiennych dla danych nieprecyzyjnie etykietowanych przy użyciu różnych miar zależności i funkcji wagowych

Piotr Bartosiewicz

promotor

dr Krzysztof Rudaś

WARSZAWA 2025

Streszczenie

Modyfikacja submodularnego algorytmu selekcji zmiennych dla danych nieprecyzyjnie etykietowanych przy użyciu różnych miar zależności i funkcji wagowych

Problem uczenia na danych niejednoznacznie etykietowanych (ang. *Partial Label Learning*, PLL) stanowi uogólnienie klasyfikacji wieloklasowej. Jest to istotny i dynamicznie rozwijający się obszar uczenia maszynowego, znajdujący liczne zastosowania praktyczne [14, 34, 36]. Pomimo rosnącego zainteresowania wciąż istnieje wiele otwartych zagadnień badawczych. Jednym z nich jest selekcja zmiennych. Celem niniejszej pracy jest omówienie submodularnego algorytmu w zagadnieniu Partial Label Learning służącemu rozwiązaniu problemu selekcji oraz zaproponowanie autorskich modyfikacji poprawiających jego działanie. Na początku pracy wprowadzono podstawowe pojęcia związane z niejednoznacznym oznaczeniem danych, w tym definicję zbiorów niejednoznacznie etykietowanych, oraz dokonano przeglądu literatury dotyczącej tego obszaru. Następnie omówiono zagadnienia związane z selekcją zmiennych, miarami zależności i klasyfikacją, a także opisano wybrane algorytmy wykorzystywane w PLL. Część zasadnicza pracy została poświęcona algorytmowi SAUTE oraz analizie jego problemów implementacyjnych. Przedstawiono również autorskie modyfikacje tego algorytmu, oparte na wcześniejszych rozważaniach. Zaproponowane rozwiązania zostały przetestowane na danych rzeczywistych i porównane z oryginalną wersją SAUTE. Wyniki eksperymentów zostały poddane szczegółowej analizie, pozwalającej na ocenę zdolności klasyfikacyjnych modeli PLL, w których do selekcji użyto metody SAUTE lub jej modyfikacji.

Słowa kluczowe: Uczenie Maszynowe, Klasyfikacja, Selekcja Zmiennych, Dane Niejednoznacznie Oznaczone

Abstract

Modification of submodular feature selection algorithm for partial labelling data using different dependence measures and weight functions

The problem of learning from partially labeled data (Partial Label Learning, PLL) is a generalization of multi-class classification. It constitutes an important and dynamically developing area of machine learning, with numerous practical applications [14, 34, 36]. Despite the growing interest, many research questions remain open. One of them is variable selection. The aim of this work is to discuss a submodular algorithm in the context of Partial Label Learning used to address the problem of selection, and to propose original modifications to improve its performance. The thesis begins with the introduction of fundamental concepts related to partial labeling, including the definition of partially labeled sets, followed by a review of the existing literature in this area. Subsequently, issues related to feature selection, dependency measures, and classification are discussed, along with an overview of the selected algorithms used in PLL. The core part of the thesis is devoted to the SAUTE algorithm and the analysis of its implementation challenges. Furthermore, author's modifications of this algorithm, motivated by previous considerations, are presented. The proposed solutions were evaluated on real-world datasets, compared with the original version of SAUTE. The results of the experiments were subjected to a detailed analysis, allowing for the assessment of the classification capabilities of the PLL models to which SAUTE method or its modifications were applied.

Keywords: Machine Learning, Classification, Feature Selection, Partial Label Data

Spis treści

1. Wstęp

1.1. Wprowadzenie

Problem uczenia na danych niejednoznacznie etykietowanych (ang. *Partial Label Learning*) jest uogólnieniem problemu klasyfikacji wieloklasowej. Polega ono na tym, że dla każdej obserwacji naszą zmienną odpowiedzi jest zbiór etykiet, wśród których znajduje się jedna prawdziwa. Technikę takiego oznaczania danych nazywamy Częściowym Etykietowaniem (ang. *Partial Labeling*). Niejednoznaczność w oznaczeniu etykiet może być wyrazem niepewności osoby klasyfikującej co do prawdziwej etykiety obserwacji [35]. Przykładowo podczas sondy ulicznej ankieter pokazuje uczestnikom kolejne zdjęcia zwierząt prosząc o wskazanie jakie zwierzę znajduje się na zdjęciu. Uczestnik nie będąc pewien, czy na danym zdjęciu widzi psa czy wilka z powodu ich podobieństwa udziela niejednoznaczną odpowiedź (przykład 1.1). Niejednoznaczne oznaczanie może być również wynikiem decyzji mającej na celu zmniejszenie kosztów procesu oznaczania obserwacji [16, 33, 34]. Przykładowo pewna firma ma potrzebę wytrenowania modelu uczenia maszynowego rozpoznającego osobę publiczną, która znajduje się na zdjęciu. Firma posiada zbiór zdjęć wraz z oznaczeniem osób widocznych na zdjęciu. W przypadku zdjęć, na których znajduje się więcej niż jedna osoba, firma staje przed wyzwaniem, jak oznaczyć te fotografie. Jednym ze sposobów jest użycie algorytmów, które wyciągną twarze z tych zdjęć i z każdej osoby obecnej na zdjęciu stworzą nową fotografię. Wówczas te nowe zdjęcia mogą być jednoznacznie oznaczone przez fizyczną osobę. Wykonanie tego zlecenia polegałoby na czasochłonnym oznaczaniu każdej obserwacji oddzielnie. Czas pracy tej osoby, a zarazem koszty poniesione przez firmę można zredukować używając częściowego oznaczania przypisując każdemu nowemu zdjęciu oryginalne oznaczenie z pierwotnego zdjęcia (patrz przykład 1.2). Niestety brak precyzji w oznakowaniu obserwacji zazwyczaj negatywnie wpływa na zdolności predykcyjne ze względu na ograniczone zdolności nadzorowania procesu treningu. Stawia to nowe wyzwania, nieobecne w klasycznym uczeniu maszynowym. Większość dostępnych algorytmów polega na próbie oszacowania prawdziwych ukrytych etykiet poprzez odpowiednie modyfikacje klasycznych modeli uczenia maszynowego [5, 20]. Innym istotnym zagadnieniem w tym obszarze jest problem selekcji zmiennych. Temat ten jest szczególnie ważny z powodu korzyści takich jak większa interpretowalność

modelu, poprawienie skuteczności klasyfikacji oraz zmniejszenie złożoności obliczeniowej procesu uczenia. Jeden z ciekawszych algorytmów w tym obszarze - *SubmodulAr featUre selecTion for partial labEl learning* (SAUTE) został opisany w pracy [1]. Podstawową zasadą działania tego algorytmu jest zachłanna selekcja zmiennych objaśniających maksymalizująca *informację wzajemną* między nimi a częściowo obserwowalną zmienną odpowiedzi z uwzględnieniem interakcji pomiędzy zmiennymi objaśniającymi. Celem niniejszej pracy jest przeanalizowanie algorytmu SAUTE, wprowadzenie modyfikacji mających na celu poprawę działania algorytmu oraz ocenę zdolności klasyfikacyjnych na danych rzeczywistych podstawowego modelu PLL: PL-KNN, w którym jako metody selekcji użyto algorytmu SAUTE lub jego modyfikacji. Jedną z najważniejszych proponowanych zmian jest wprowadzenie nowego sposobu szacowania wpływu zmiennych niezależnych na zmienną odpowiedzi z uwzględnieniem zależności między zmiennymi objaśniającymi opisaną w pracy [2] oraz modyfikacje aktualizacji macierzy pewności poprzez zastosowanie metodologii zaprezentowanej w algorytmie IPAL [18].

Przykład 1.1. Przykładem danych częściowo etykietowanych są wyniki sondy ulicznej, w której osoby proszone są o określenie obiektu znajdującego się na zdjęciu. Osoby te mogą nie mieć pewności czy zdjęcie przedstawia wilka czy psa przyznając obie te etykiety na raz.

{ Wilk, Pies }

Rysunek 1.1: Samiec psa rasy Wilczak Czechosłowacki

Przykład 1.2. Kolejnym zastosowaniem Partial Labelingu jest znaczące zredukowanie kosztów oznaczania danych. W tym podejściu twarze osób wyciągniętych z fotografii oznaczone są poprzez opis pierwotnego zdjęcia, na którym znajduje się więcej niż jedna osoba.

Rysunek 1.2: Podejście niejednoznacznego oznaczania obrazów

1.2. Dane niejednoznacznie oznaczone w uczeniu maszynowym

Uczenie maszynowe stanowi obszar sztucznej inteligencji zajmujący się konstruowaniem modeli zdolnych do identyfikowania wzorców w danych oraz formułowania prognoz lub decyzji bez konieczności definiowania przez użytkownika wszystkich reguł postępowania. W ramach uczenia nadzorowanego wyróżnia się dwa podstawowe typy zadań: regresję oraz klasyfikację. Regresja obejmuje zagadnienia związane z przewidywaniem wartości liczbowych zmiennej objaśnianej na podstawie zestawu zmiennych objaśniających. Modele regresyjne zwracają wyniki w postaci wartości ciągłych. Klasyfikacja natomiast odnosi się do przypisywania obserwacji do jednej z dyskretnych kategorii. Modele klasyfikacyjne zwracają wyniki w postaci etykiet klas. W klasycznym uczeniu maszynowym w problemie klasyfikacji zakłada się, że każda instancja posiada dokładnie jedną prawidłową etykietę klasową.

Zadanie klasyfikacji formalnie polega na znalezieniu funkcji $c^* : \mathbb{R}^d \to S$ nazywanej klasyfikatorem. Uzyskuje się go, szukając rozwiązania bądź przybliżenia następującego problemu:

$$c^*(x) = \arg\max_c \ \mathcal{P}(C = c \mid X = x). \tag{1.1}$$

Formalnie, dla klasycznej klasyfikacji rozwiązanie problemu (1.1) uzyskujemy używając następujących danych:

$$\mathcal{D} = \{(x_i, c_i)\}_{i=1}^n,$$

gdzie:

- $x_i \in \mathbb{R}^d$ — wektor zmiennych objaśniających i-tej obserwacji,
- $c_i \in S$ — etykieta klasy dla i-tej obserwacji
- $S = \{1, 2, \ldots, q\}$ — zbiór wartości etykiet.

Problem niejednoznacznego etykietowania jest naturalnym uogólnieniem tego podejścia. W tym przypadku dla każdej obserwacji posiadamy częściowe etykiety, czyli zbiory potencjalnych klas, z których tylko jedna jest poprawna, ale nie wiadomo która. Dla tego problemu klasyfikator uzyskiwany jest przy użyciu następujących danych:

$$\mathcal{D} = \{(x_i, s_i)\}_{i=1}^{n},$$

gdzie:

- $x_i \in \mathbb{R}^d$ — wektor zmiennych objaśniających i-tej instancji,
- $s_i \subseteq S$ — zbiór kandydatów na etykiety (częściowe etykiety), taki że $c_i \in s_i$, gdzie c_i to częściowo obserwowalna, prawdziwa etykieta dla i-tej obserwacji,
- $S = \{1, 2, \ldots, q\}$ — zbiór wartości etykiet.

1.3. Przegląd literatury

Uczenie z częściowo oznaczonymi etykietami (Partial Label Learning, PLL) zalicza się do kategorii uczenia słabo nadzorowanego. PLL charakteryzuje się tym, że w przeciwieństwie do klasycznego uczenia maszynowego nie posiadamy jawnej klasy obserwacji, a dysponujemy pewnym podzbiorem etykiet wśród, których znajduje się prawdziwa klasa. Problem ten występuje w rzeczywistych zastosowaniach, takich jak eksploracja danych internetowych [4, 14] na przykład obrazów, gdzie podejście PL pozwala rozwiązać problem etykietowania niejednoznacznych instancji. Dotychczasowe podejścia do PLL dzielą się na dwie główne strategie: *uśredniania*, gdzie wszystkie elementy z podzbioru etykiet są traktowane równorzędnie [4], oraz *identyfikacji*, gdzie ukryta prawdziwa etykieta szacowana jest iteracyjnie [6, 7]. Nowsze metody, jak SURE [3], bazują na rozszerzeniu funkcji celu do problemu danych niejednoznacznie etykietowanych poprzez wprowadzenie mechanizmu pseudoadnotacji. Inne metody jak IPAL [18] stanowią przykład nowoczesnego podejścia do PLL, gdzie łączy się lokalną analizę sąsiedztwa obserwacji z iteracyjnym wzmacnianiem przekonań klasyfikacyjnych, co pozwala skutecznie ograniczać wpływ fałszywych etykiet.

Chociaż poprawa generalizacji modeli poprzez redukcję wymiarowości jest powszechnie stosowana w innych dziedzinach uczenia słabo nadzorowanego [11] dotychczas w problemie PLL skupiano się głównie na oszacowaniu prawdziwych ukrytych etykiet, nie poruszając tematu redukcji wymiarowości. Nieliczne prace, takie jak [8, 9, 10] wykorzystujące redukcję wymiarowości, opierają się na transformacji zmiennych [12] odpowiednio dostosowując standardowe metody jak na przykład LDA. Jednak problem selekcji zmiennych mogący nieść takie korzyści jak poprawa interpretowalności i eliminacja redundancji [13] nie został odpowiednio dobrze przebadany. Algorytm SAUTE, aby zapełnić tę lukę, wykorzystuje submodularną selekcję zmiennych w zagadnieniu PLL [1]. Metoda ta maksymalizuje informację wzajemną między zmienną objaśniającą a częściowo obserwowalną zmienną odpowiedzi uwzględniając interakcje pomiędzy zmiennymi objaśniającymi. Prowadzi to do znaczącej poprawy jakości klasyfikacyjnych istniejących modeli PLL.

Zagadnienie selekcji zmiennych w kontekście nieparametrycznym w klasycznym problemie klasyfikacji zostało pogłębione w pracy [2], gdzie przeanalizowano kryteria oparte na informacji wzajemnej, takie jak Joint Mutual Information (JMI) [2, 16], będącym przybliżeniem warunkowej informacji wzajemnej. Pokazano również, że JMI przy precyzyjnej analizie entropii i informacji wzajemnej w mieszankach rozkładów normalnych może wspomagać wybór zmiennych istotnych informacyjnie.

W dostępnej literaturze nie istnieją metody łączące zagadnienie selekcji zmiennych w problemie PLL z większością istniejących kryteriów opartych na informacji wzajemnej, takich jak JMI. Jedynie metoda SAUTE wykorzystuje kryterium Minimum Redundancy Maximum Relevance (mRMR) [22]. W celu wypełnienia tej luki opiszemy autorską modyfikację metody SAUTE uwzględniającą miarę JMI do szacowania warunkowej informacji wzajemnej.

W tej pracy przedstawimy kluczowe zagadnienia powiązane z selekcją zmiennych, w tym najpopularniejsze algorytmy oraz podstawowe miary stosowane w tym obszarze. Następnie omówimy algorytmy klasyfikacyjne PLL oraz algorytm SAUTE będący najważniejszym algorytmem selekcji zmiennych w PLL. Na koniec przedstawimy zaproponowane przez nas modyfikacje algorytmu SAUTE oraz omówimy wyniki eksperymentów na danych rzeczywistych porównujących zaproponowane modyfikacje z bazową metodą SAUTE.

2. Problem selekcji zmiennych w uczeniu maszynowym

2.1. Wybór zmiennych

Selekcja zmiennych jest jednym z najważniejszych zagadnień w *Uczeniu Maszynowym*. Zadanie to polega na wybraniu podzbioru zmiennych objaśniających $\hat{\mathcal{F}}_T = \{f_{t_1}, f_{t_2}, ..., f_{t_{d'}}\} \subset \mathcal{F} = \{f_1, f_2, \ldots, f_d\}$, gdzie $d^{'} < d$. Rozwiązanie tego problemu może nieść za sobą wiele korzyści jak np. zwiększenie interpretowalności modelu, zmniejszenie złożoności obliczeniowej modelu, zmniejszenie wymiarowości danych. Główną motywacją stojącą za selekcją zmiennych oprócz wyżej wymienionych korzyści jest chęć uniknięcia przeuczenia modelu. Jest to sytuacja, w której model jest zbytnio dopasowany do danych treningowych co skutkuje zmniejszeniem jego zdolności predykcyjnych. Może to być spowodowane istnieniem predyktorów w modelu, które mogą być same w sobie, bądź w obecności innych zmiennych nieistotne bądź zbędne. Usunięcie ich może nie wpłynąć znacznie na jakość przewidywań na nowych danych, bądź nawet znacząco ją poprawić. Problem selekcji zmiennych możemy zdefiniować następująco:

$$\mathcal{F}^* = \arg\max_{\hat{\mathcal{F}} \subset F} g(\hat{\mathcal{F}}), \tag{2.1}$$

gdzie g jest pewnym kryterium według którego oceniamy podzbiory zmiennych. Najpopularniejsze typy kryteriów zostaną opisane w kolejnym podrozdziale. Zauważmy, że problem (2.1) posiada wykładniczą złożoność z względu na liczbę zmiennych objaśniających w oryginalnym zbiorze $\mathcal{F}$ [23]. Z tego powodu do problemu selekcji zmiennych stosuje się heurystyczne podejścia jak np. algorytmy zachłanne polegające na dodaniu/usunięciu w każdym kroku predyktora maksymalizującego/minimalizującego wybrane kryterium.

2.2. Typy algorytmów selekcji zmiennych

W literaturze wyróżnia się trzy główne kategorie algorytmów selekcji zmiennych: metody filtrujące, metody wrapper oraz metody wbudowane. Metody filtrujące dokonują selekcji zmiennych niezależnie od używanego modelu predykcyjnego. Ocena istotności zmiennych objaśniających opiera się na właściwościach statystycznych zbioru treningowego. Ocena ta jest mierzona za

pomocą różnych zależności między zmienną odpowiedzi, a zmienną objaśniającą takich jak korelacja Pearsona, test niezależności Chi-kwadrat czy informacja wzajemna. Na podstawie wartości tych miar dla poszczególnych zmiennych objaśniających tworzy się ranking, za pomocą którego używając ustalonego punktu odcięcia wybiera się zmienne znajdujące się najwyższej w rankingu. Algorytmy te cechuje wysoka wydajność obliczeniowa i odporność na przeuczenie, jednak mogą nie uwzględniać złożonych interakcji między zmiennymi objaśniającymi [13]. W odpowiedzi na ten problem wprowadzono nowe kryteria bazujące na informacji wzajemnej takie jak Joint Mutual Information (JMI) [16], Conditional Infomax Feature Extraction (CIFE) [15] i Minimum Redundancy Maximum Relevance (mRMR) [22]. Kolejna grupa algorytmów nazywana metodami wrapper traktuje selekcję zmiennych jako proces optymalizacji, w którym podzbiór zmiennych jest oceniany na podstawie wydajności danego modelu predykcyjnego. Algorytm iteracyjnie testuje różne kombinacje zmiennych, budując modele i porównując ich jakość za pomocą wybranej miary (np. Accuracy, AUC, MSE), na każdej iteracji dobierając bądź usuwając zmienne. Choć metody te często prowadzą do wysokiej skuteczności predykcyjnej, ich główną wadą jest kosztowność obliczeniowa, zwłaszcza przy dużej liczbie zmiennych [24]. Do najczęściej stosowanych strategii wrapper zalicza się selekcja sekwencyjna (np. forward selection, backward elimination), algorytmy genetyczne, metody oparte na przeszukiwaniu siatki podzbiorów zmiennych. Kolejną grupą algorytmów są metody wbudowane. Integrują one proces selekcji zmiennych z etapem uczenia modelu. W przeciwieństwie do podejścia filtrującego i wrapper, które rozdzielają selekcję i modelowanie, metody wbudowane automatycznie identyfikują istotne zmienne w trakcie konstruowania modelu, wykorzystując wbudowane mechanizmy regularyzacji bądź oceny ważności zmiennych. Są one z reguły bardziej wydajne niż metody wrapper i dokładniejsze niż metody filtrujące [25]. Najpopularniejsze metody wbudowane obejmują między innymi Regresję Lasso (Regresja z Karą l_1), Adaptacyjną Regresję Lasso, Sieci Elastyczne oraz Drzewa Decyzyjne.

Każda z opisanych metod posiada zalety i ograniczenia, a ich efektywność zależy od specyfiki zadania, rodzaju danych oraz dostępnych zasobów obliczeniowych. W praktycznych problemach często stosuje się kombinacje metod lub ich hybrydyzację, aby osiągnąć kompromis pomiędzy dokładnością a złożonością obliczeniową.

2.3. Miary zależności między zmiennymi zależnymi, a zmienną odpowiedzi

W metodach filtrujących kryterium oceny przydatności zmiennej objaśniającej zazwyczaj jest pewną miarą zależności między zmienną objaśniającą, a zmienną odpowiedzi. Przykładem takiej miary jest informacja wzajemna, wykorzystywana w algorytmie SAUTE, będącym głównym

obiektem badań tej pracy. Stąd w dalszej części skupimy się wyłącznie na kryteriach selekcji opartych na informacji wzajemnej. W tym celu niech $(\Omega, \mathcal{F}_\omega, \mathcal{P})$ oznacza przestrzeń probabilistyczną. W dalszej części rozdziału wprowadzając zmienne losowe, będziemy zakładać, że są one określone na tej przestrzeni probabilistycznej. Wprowadźmy następujące definicje:

Definicja 2.1. (Entropia zmiennej losowej)
Entropią zmiennej losowej X o nośniku $\text{supp}(X) = \mathcal{X}$, i o rozkładzie p_X nazywamy:

$$H(X) = -\int_{\mathcal{X}} p_X(x) \log_2 p_X(x)\, dx. \tag{2.2}$$

Pojęcie entropii wywodzi się z teorii informacji C. Shannona [17], możemy ją interpretować jako średnią liczbę informacji mierzoną w bitach, którą dostarcza nam zaobserwowanie próbki.

Uwaga 2.2. (Własności entropii)
Niech X będzie dyskretną zmienną losową z rozkładem prawdopodobieństwa $\{p_1, p_2, ..., p_n\}$, wtedy zachodzi:

- $H(X) \geqslant 0$, ponadto równość zachodzi wtedy i tylko wtedy, gdy X ma rozkład jednopunktowy.
- $H(X) \leqslant \log_2 n$, ponadto równość zachodzi wtedy i tylko wtedy, gdy $X \sim \mathcal{U}(\{1, 2, \ldots, n\})$, gdzie $\mathcal{U}(\{1, 2, \ldots, n\})$ oznacza dyskretny rozkład jednostajny.

Przykład 2.3. Zmienna losowa Z opisuje wybór środka transportu przez studenta dojeżdżającego na uczelnię. Z prawdopodobieństwem 0.6 student wybierze autobus , z prawdopodobieństwem 0.25 rower, a z prawdopodobieństwem 0.15 pójdzie pieszo.

Wtedy entropia zmiennej Z wynosi:

$$H(Z) = -0.6 \log_2 0.6 - 0.25 \log_2 0.25 - 0.15 \log_2 0.15 \approx 1.36$$

Oznacza to, że obserwacja sposobu dojazdu studenta dostarcza nam średnio około 1.36 bita informacji na temat środka transportu wybranego przez niego danego dnia.

Przykład 2.4. Niech zmienna losowa Y opisuje pogodę w Warszawie w sezonie jesienno-zimowym. Prawdopodobieństwo wystąpienia deszczu wynosi 0.05, prawdopodobieństwo zachmurzenia wynosi 0.94, a prawdopodobieństwo wystąpienia słońca wynosi 0.01. Wtedy entropia zmiennej Y wynosi:

$$H(Y) = -0.05 \log_2 0.05 - 0.94 \log_2 0.94 - 0.01 \log_2 0.01 \approx 0.37$$

Oznacza to, że sprawdzenie pogody przynosi nam średnio około 0.37 bita informacji.

Entropię możemy również interpretować jako miarę niepewności co do wartości zmiennej losowej. Ilustrują to Przykłady 2.3, 2.4 oraz Uwaga 2.2. W przypadku rozkładu, w którym prawdopodobieństwa poszczególnych wartości są rozłożone w przybliżeniu równomiernie (jak w Przykładzie 2.3), szanse na uzyskanie każdej wartości są podobne. Oznacza to, że przed obserwacją zmiennej losowej mamy wysoki poziom niepewności, co implikuje wysoką entropię, czyli bliską wartości $\log_2 n$, gdzie n to liczba wartości przyjmowanych przez zmienną losową. Z kolei gdy masa prawdopodobieństwa jest silnie skoncentrowana w jednym punkcie (jak w Przykładzie 2.4), szansa na wystąpienie tej konkretnej wartości przewyższa pozostałe. Przekłada się to na większą pewność co do realizacji zmiennej losowej, co skutkuje niską wartości entropii w okolicach zera. Możemy tą myśl podsumować następującym zdaniem: Im bardziej równomierny rozkład, tym wyższa entropia, a więc większa niepewność; im bardziej skupiony, tym niższa entropia i większa przewidywalność.

W ramach omawianego zagadnienia interesować nas będą zależności między zmiennymi objaśniającymi a zmienną odpowiedzi, w tym celu wprowadzimy pojęcia entropii łącznej i entropii warunkowej.

Definicja 2.5. (Entropia łączna)
Entropią łączną między dwiema zmiennymi losowymi X, Y o rozkładach prawdopodobieństwa kolejno p_X, p_Y i nośnikach $\text{supp}(X) = \mathcal{X}$, $\text{supp}(Y) = \mathcal{Y}$ oraz łącznym rozkładzie prawdopodobieństwa $p_{X,Y}$ zmiennej losowej (X, Y) definiujemy jako:

$$H(X,Y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} p_{X,Y}(x,y)\log_2\left(p_{X,Y}(x,y)\right)dydx.$$

Definicja 2.6. (Entropia warunkowa)
Entropią między dwiema zmiennymi losowymi X, Y o rozkładach prawdopodobieństwa kolejno p_X, p_Y i nośnikach $\text{supp}(X) = \mathcal{X}$, $\text{supp}(Y) = \mathcal{Y}$ oraz łącznym rozkładzie prawdopodobieństwa $p_{X,Y}$ zmiennej losowej (X, Y) definiujemy:

$$H(X|Y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} p_{X,Y}(x,y)\log_2\left(\frac{p_{X,Y}(x,y)}{p_Y(y)}\right)dydx.$$

Z entropią ściśle powiązane jest pojęcie informacji wzajemnej dwóch zmiennych losowych, które będzie kluczowe w kontekście algorytmów selekcji zmiennych w problemie PLL:

Definicja 2.7. (Informacja wzajemna zmiennych losowych)
Informacją wzajemną między dwiema zmiennymi losowymi X, Y o rozkładach kolejno p_X, p_Y i nośnikach $\text{supp}(X) = \mathcal{X}$, $\text{supp}(Y) = \mathcal{Y}$ oraz gęstości łącznej $p_{X,Y}$ zmiennej losowej (X, Y) definiujemy:

$$I(X,Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{X,Y}(x,y) \log_2 \left(\frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right) dydx.$$

Informację wzajemną zmiennych losowych X i Y można interpretować jako ilość informacji mierzoną w bitach, którą dzielą z sobą X i Y. Związek Informacji wzajemnej z entropią przedstawiony został w pierwszym punkcie Uwagi 2.8.

Uwaga 2.8. (Własności informacji wzajemnej)
Niech X i Y będą zmiennymi losowymi, wtedy;

- $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$,
- $I(X,Y) \geqslant 0$,
- $I(X,Y) = I(Y,X)$.

Definicja 2.9. (Warunkowa informacja wzajemna) Niech X, Y, Z będą zmiennymi losowymi o rozkładach prawdopodobieństwa kolejno p_X, p_Y, p_Z i nośnikach $\text{supp}(X) = \mathcal{X}$, $\text{supp}(Y) = \mathcal{Y}$, $\text{supp}(Z) = \mathcal{Z}$ oraz rozkładach warunkowych $p_{X|Z}$, $p_{Y|Z}$, $p_{X,Y|Z}$ zmiennych losowych kolejno $X|Z$, $Y|Z$, $(X,Y)|Z$. Warunkową informacją wzajemną (CMI) nazywamy:

$$I(X,Y \mid Z) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \int_{\mathcal{Z}} p_Z(z)\, p_{X,Y|Z}(x,y|z) \log_2 \left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z) p_{Y|Z}(y|z)} \right) dzdydx.$$

W ramach niniejszej pracy jednym z kluczowych zagadnień było zastosowanie powyższych miar do problemu selekcji zmiennych dla danych PL. W szczególności miara CMI będzie dla nas przydatna. Rozważmy sytuację, w której mamy wybrany podzbiór predyktorów $\hat{\mathcal{F}}_T \subset \mathcal{F}$ i chcemy dobrać do niego kolejną zmienną objaśniającą tak, aby maksymalizowała ona informację wzajemną między nią a zmienną odpowiedzi pod warunkiem wybranego już podzbioru $\hat{\mathcal{F}}_T$, problem ten możemy zapisać jako

$$\underset{f \in \mathcal{F} \setminus \hat{\mathcal{F}}_T}{\arg\max}\ I(f, C \mid \hat{\mathcal{F}}_T) \tag{2.3}$$

Niestety powyższy problem w realnych zastosowaniach jest często trudny bądź niemożliwy do bezpośredniego rozwiązania. W celu oszacowania tej wartości wprowadzimy miary JMI [16] oraz mRMR [22] będące aproksymacją problemu przedstawionego w równaniu (2.3).

Definicja 2.10. (Minimum Redundancy Maximum Relevance)
Niech X_S będzie wektorem losowym o wymiarze S oraz X_j będzie zmienną losową taką, że $X_j \notin X_S$, miarę mRMR definiujemy następująco:

$$mRMR(X_j, C \mid X_S) = \frac{1}{|S|} I(X_j, C) - \frac{1}{|S|^2} \sum_{i \in S} I(X_i, X_j). \tag{2.4}$$

Definicja 2.11. (Joint Mutual Information Criterion)
Niech X_S będzie wektorem losowym o wymiarze S oraz X_j będzie zmienną losową taką, że $X_j \notin X_S$, miarę JMI definiujemy następująco:

$$JMI(X_j, C \mid X_S) = I(X_j, C) + \frac{1}{|S|} \sum_{i \in S} [I(X_i, X_j \mid C) - I(X_i, X_j)]. \tag{2.5}$$

Zauważmy, że do miary JMI służącej do aproksymacji kryterium opisanego równaniem (2.3), używamy warunkowej informacji wzajemnej, jednakże warunkowanie odbywa się po innej zmiennej, co przy pewnych założeniach umożliwia policzenie jej bezpośrednio. Dzieje się tak, ponieważ unikamy warunkowania po wielowymiarowej zmiennej, którego rozkładu łącznego zazwyczaj nie znamy, sprowadzając nasz problem do operacji na zmiennych losowych dwu bądź trzywymiarowych. W dalszej części pracy będziemy posługiwać się pojęciami kryteriów JMI oraz mRMR, przez które będziemy rozumieć zastosowanie wyżej wymienionych miar w metodach filtrujących do selekcji zmiennych.

3. Wybrane algorytmy klasyfikacyjne Partial Labelingu

Klasyfikacja stanowi jedno z podstawowych zagadnień w dziedzinie uczenia maszynowego, znajdując zastosowanie w szerokim spektrum problemów, od rozpoznawania obrazów po analizę tekstu i diagnostykę medyczną [26, 27, 28]. Celem klasyfikacji jest przypisanie obiektowi jednej z uprzednio zdefiniowanych klas na podstawie zestawu jego zmiennych objaśniających. Do najważniejszych modeli w tym obszarze możemy zaliczyć m.in. regresję logistyczną, metodę k-najbliższych sąsiadów i sieci neuronowe. Niestety z powodu braku jednoznaczności w oznaczaniu zmiennej odpowiedzi w problemie partial labelingu niemożliwe jest bezpośrednie zastosowanie klasycznych algorytmów uczenia maszynowego. Algorytmy klasyfikacyjne dostępne w literaturze dla tego problemu możemy podzielić na modyfikacje klasycznych modeli poprzez zastosowanie funkcji agregujących dla zmiennych odpowiedzi oraz oryginalne metody stworzone wyłącznie dla problemu PLL. Do najważniejszych algorytmów opisanych w literaturze możemy zaliczyć PL-KNN [5] będący modyfikacją klasycznego algorytmu k-najbliższych sąsiadów, należący do grupy algorytmów agregujących i IPAL [18] ujednoznaczniający zmienną odpowiedzi poprzez rozwiązanie zadania programowania kwadratowego, należący do grupy oryginalnych algorytmów dla problemu PLL.

Znaczna część metod w problemie PLL opiera się na wykorzystaniu macierzy $Y \in \mathbb{R}^{m \times q}$ nazywanej macierzą pewności etykiet bądź macierzą zaufania. Interpretujemy ją następująco: $Y(i, l)$ oznacza oszacowanie prawdopodobieństwa, że etykieta l jest prawdziwą klasą obserwacji x_i.

3.1. PL-KNN

W tej sekcji przybliżymy metodę PL-KNN uznawaną za bazowy algorytm klasyfikacyjny w problemie PLL przede wszystkim z powodu bycia naturalnym rozszerzeniem klasycznego algorytmu KNN. Rozważmy zadanie klasyfikacji w problemie PL. Niech $\mathcal{N}_k(x_i)$ oznacza zbiór k-najbliższych sąsiadów obserwacji x_i, czyli zbiór k obserwacji z ramki danych, które mają najmniejszą odległość od obserwacji względem wybranej metryki - najczęściej euklidesowej. Klasy-

fikacji dokonujemy według wzoru:

$$c_i = \arg\max_{l \in S} \sum_{x_j \in \mathcal{N}_k(\mathbf{x}_i)} \omega_j \, Y(j,l), \tag{3.1}$$

gdzie ω_j to waga odpowiadającą j-temu najbliższemu sąsiadowi obserwacji x_i. W ramach tej pracy będziemy przyjmować $\omega_j = (k - j + 1)$. Ponadto dla algorytmu PL-KNN zazwyczaj przyjmuje się następującą postać macierzy pewności:

$$Y(i,l) = \begin{cases} \frac{1}{|s_i|}, & \text{jeżeli } l \in s_i \\ 0, & \text{w.p.p.} \end{cases} \quad \text{dla wszystkich } 1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant q \tag{3.2}$$

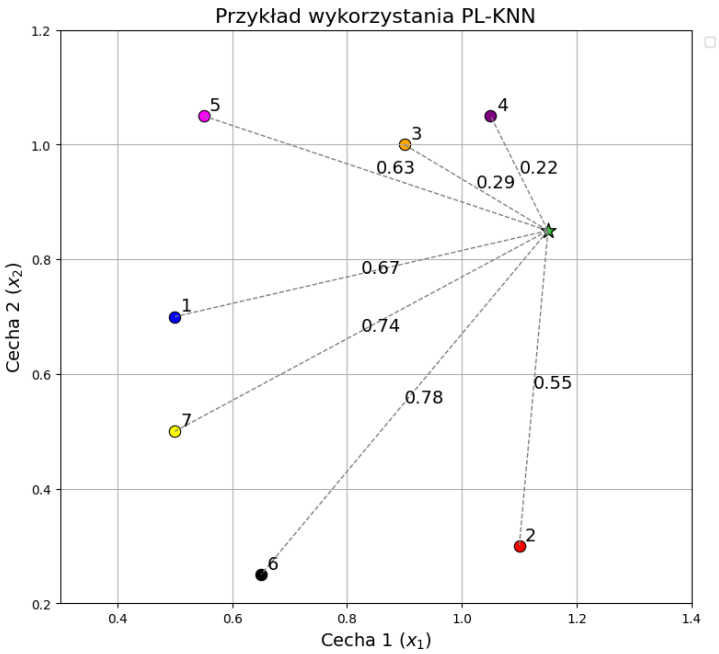
Działanie tego algorytmu możemy podsumować następującym zdaniem. Algorytm uznaje jako bardziej prawdopodobne etykiety, które pojawiają się często wśród najbliższych sąsiadów, zwłaszcza wśród tych, którzy są najbliżej danej obserwacji oraz jednocześnie ich prawdopodobieństwo bycia prawdziwą etykietą dla danego sąsiada jest wysokie.

Przykład 3.1. (Wykorzystanie PL-KNN)

Rozważmy następujący problem klasyfikacji czteroklasowej w problemie PLL. Do predykcji użyjemy algorytmu PL-KNN z liczbą najbliższych sąsiadów $k = 4$. Nasze dane treningowe prezentują się następująco:

ID	**Etykieta PL** (s)	**Cecha 1** (x_1)	**Cecha 2** (x_2)
1	{0,2}	0.50	0.70
2	{1,2}	1.10	0.30
3	{1,2,3}	0.90	1.00
4	{0,2,3}	1.05	1.05
5	{0,1,2}	0.55	1.05
6	{1,2}	0.65	0.25
7	{1,3}	0.50	0.50

Chcemy dokonać klasyfikacji dla obserwacji o parametrach $x_1 = 1.15$ oraz $x_2 = 0.85$, wtedy nasza sytuacja wygląda następująco:



Najbliższymi sąsiadami obserwacji, którą mamy zaklasyfikować są kolejno obserwacje numer: 4, 3, 2, 5. Wówczas używając wzoru (3.1) uzyskujemy następujące wyniki etykiet dla obserwacji klasyfikowanej:

Etykieta:	**0**	**1**	**2**	**3**
Wynik:	$\frac{5}{3}$	$\frac{7}{3}$	$\frac{11}{3}$	$\frac{7}{3}$

Najwyższy wynik uzyskała etykieta numer 2, stąd nowa obserwacja zostaje zaklasyfikowana jako należąca do klasy numer 2.

3.2. IPAL

Metoda Instance based Partial Label learning (IPAL) używa podejścia opartego na relacjach między obserwacjami i na bezpośrednim rozstrzygnięciu niejednoznaczności etykiet. W przeciwieństwie do tradycyjnych metod, które traktują wszystkie potencjalne etykiety równorzędnie na poziomie pojedynczej obserwacji i agregują prawdopodobieństwa poszczególnych etykiet, w algorytmie IPAL można wyróżnić dwie fazy: aktualizacja macierzy pewności, w której zmniejszona zostaje niejednoznaczność etykiet oraz klasyfikacji bazującej na zaktualizowanej macierzy pewności. Precyzując algorytm IPAL możemy opisać następująco: metodę rozpoczynamy od konstrukcji ważonego grafu skierowanego $G = (V, E)$, którego wierzchołkami są wszystkie obserwacje: $V = \{x_1, x_2, \ldots, x_m\}$. Dla każdego x_j wybieranych jest k najbliższych sąsiadów $\mathcal{N}_k(x_j)$ (według metryki euklidesowej), a następnie tworzone są krawędzie skierowane:

$E = \{(x_i, x_j) \mid i \in \mathcal{N}_k(x_j),\ i \neq j\}$. Krawędź (x_i, x_j) istnieje wtedy i tylko wtedy, gdy x_i należy do zbioru najbliższych sąsiadów $\mathcal{N}_k(x_j)$. Następnie dla każdej krawędzi z grafu (x_{i_a}, x_j) wyznaczamy wagę $w_{i_a,j}$ rozwiązując następujące zadanie programowania kwadratowego:

$$\min_{w_j=(w_{i_1},\ldots,w_{i_k})} \left\| x_j - \sum_{a=1}^{k} w_{i_a,j} \cdot x_{i_a} \right\|^2 \quad \text{p.o.} \quad w_{i_a,j} \geqslant 0. \tag{3.3}$$

W przypadku, kiedy krawędź (x_i, x_j) nie należy do grafu to wadze $w_{i,j}$ przypisujemy wartość 0. Następnie konstruujemy macierz wag $W = [w_{i,j}] \in \mathbb{R}^{m \times m}$, która w j-tej kolumnie posiada niezerowe wartości uzyskane przez rozwiązanie (3.3) tylko i wyłącznie w wierszach indeksowanych jej najbliższymi sąsiadami. Warto zauważyć, że macierz W nie musi być symetryczna. W kolejnym kroku przechodzimy do kolejnej fazy algorytmu, czyli iteracyjnej propagacji etykiet. Na początek normalizujemy macierz wag kolumnowo:

$$H = WD^{-1}, \quad \text{gdzie} \quad D = \operatorname{diag}(d_1, \ldots, d_m),\ d_j = \sum_{i=1}^{m} w_{i,j}. \tag{3.4}$$

Następnie inicjalizujemy macierz pewności Y jako:

$$Y(i,l) = \begin{cases} \frac{1}{|s_i|}, & \text{jeżeli } l \in s_i, \\ 0, & \text{w przeciwnym razie.} \end{cases} \tag{3.5}$$

Przechodzimy do kluczowej fazy algorytmu, czyli iteracyjnej aktualizacji macierzy $P^{(t)}$, gdzie:

$$\tilde{P}^{(t)} = \alpha H^\top P^{(t-1)} + (1-\alpha)Y, \tag{3.6}$$

gdzie $P^{(0)} = Y$ oraz $\alpha \in [0,1]$ jest hiperparametrem metody, będącym pewnym zabezpieczeniem przed zbyt dynamiczną aktualizacją macierzy mogącą skutkować potencjalną utratą informacji na temat częściowo obserwowanej etykiety.

Ponadto na każdej iteracji t normalizujemy macierz $\tilde{P}^{(t)}$ oraz eliminujemy wpływ fałszywych etykiet, uzyskując macierz $P^{(t)}$ w następujący sposób:

$$P^{(t)}(i,l) = \begin{cases} \dfrac{\tilde{P}^{(t)}(i,l)}{\sum\limits_{\tilde{l} \in s_i} \tilde{P}^{(t)}(i,\tilde{l})}, & \text{jeżeli } l \in S_i, \\ 0, & \text{w przeciwnym razie,} \end{cases} \tag{3.7}$$

gdzie $P^{(t)}(i,l), \tilde{P}^{(t)}(i,l)$ są (i,l)-tym elementem macierzy kolejno $P^{(t)}, \tilde{P}^{(t)}$. Po T iteracjach dostajemy finalną macierz pewności $\hat{Y}$ ($P^{(T)} \to \hat{Y}$). Korzystając z $\hat{Y}$ dokonujemy wyboru najbardziej prawdopodobnej etykiety dla każdej próbki:

$$c_i = \arg\max_{l \in S} \frac{n_l}{\hat{n}_l} \cdot \hat{Y}(i,l), \ \text{gdzie: } n_l = \sum_{i=1}^{m} Y(i,l), \quad \hat{n}_l = \sum_{i=1}^{m} \hat{Y}(i,l), \tag{3.8}$$

gdzie $Y(i,l), \hat{Y}(i,l)$ są (i,l)-tym elementem macierzy kolejno $Y, \hat{Y}$. Na koniec przechodzimy do ostatniej fazy algorytmu, w której dokonujemy klasyfikacji nowych obserwacji. Rozpoczynamy od znalezienia k-najbliższych sąsiadów dla nowej obserwacji x^* wśród przykładów treningowych. Następnie dobieramy odpowiednie wagi dla najbliższych sąsiadów obserwacji x^*, poprzez rozwiązanie zadania programowania kwadratowego:

$$\min_{w^*=(w_{i_1},\ldots,w_{i_k})} \left\| x^* - \sum_{a=1}^{k} w^*_{i_a} \cdot x_{i_a} \right\|^2 \quad \text{p.o.} \quad w^*_{i_a} \geqslant 0. \tag{3.9}$$

Ostatecznie przypisujemy etykietę według wzoru:

$$c^* = \arg\min_{l \in S} \left\| x^* - \sum_{a=1}^{k} \mathbb{I}(c_{i_a} = l) \cdot w^*_{i_a} \cdot x_{i_a} \right\|^2. \tag{3.10}$$

Ostateczne przypisanie etykiety możemy interpretować następująco: algorytm za najbardziej prawdopodobną uznaje etykietę, która często występuj u k-najbliższych sąsiadów, zwłaszcza wśród tych, którzy otrzymali wysokie wagi w wyniku rozwiązania zagadnienia programowania kwadratowego (3.9).

4. Algorytm SAUTE

SAUTE (SubmodulAr featUre selecTion for partial labEl learning) to algorytm selekcji zmiennych zaprojektowany dla problemu Partial Label Learning. Celem algorytmu jest wybór podzbioru o zadanej liczności $d^{'}$ najbardziej informatywnych zmiennych objaśniających względem częściowo oznaczonej zmiennej odpowiedzi. Ograniczenie liczby zmiennych może przyczynić się do poprawy zdolności predykcyjnych klasyfikatora uczonego na niejednoznacznie oznaczonych danych, ponadto zwiększyć interpretowalność modelu oraz zmniejszyć złożoność obliczeniową. Algorytm rozpoczyna się od zainicjowania macierzy pewności etykiet $Y \in \mathbb{R}^{m \times q}$:

$$Y(i,l) = \begin{cases} \frac{1}{|s_i|}, & \text{jeżeli } l \in s_i \\ 0, & \text{w.p.p.} \end{cases} \quad \text{dla wszystkich } 1 \leqslant i \leqslant m,\ 1 \leqslant l \leqslant q \tag{4.1}$$

Macierz tą będziemy interpretować w następujący sposób, wartość $Y(i,l)$ oznacza prawdopodobieństwo, że etykieta l jest prawdziwą klasą dla obserwacji x_i, czyli w ten sam sposób jak w algorytmie IPAL. Naszym celem jest rozwiązanie następującego problemu optymalizacji:

$$\mathcal{F}_T{}^* = \underset{\mathcal{F}_T \subset \mathcal{F}, |\mathcal{F}_T| = d'}{\arg\max} I(\mathcal{F}_T, C), \tag{4.2}$$

to znaczy chcemy wybrać optymalny zbiór $\mathcal{F}_T{}^*$ maksymalizujący informację wzajemną między zmiennymi objaśniającymi należącymi do tego zbioru a zmienną losową C opisującą częściowo obserwowalną, prawdziwą klasę zmiennej odpowiedzi. Niepewną informację na temat rozkładu zmiennej C obserwujemy przy pomocy macierzy Y. Niestety problem (4.2) jest NP-trudny [21]. Aby znaleźć przybliżone rozwiązanie tego problemu można zauważyć, że funkcja $I(\mathcal{F}_T, C)$ jest submodularna, nieujemna, niemalejąca, pod warunkiem słabej warunkowej niezależności [19]. Jedną z konsekwencji powyższych własności jest to, że rozwiązanie problemu (4.2) może być przybliżone przez odpowiedni algorytm zachłanny, dokładający iteracyjnie kolejne zmienne do zbioru $\mathcal{F}_T$. Na początku załóżmy, że $I(\emptyset, C) = 0$ oraz ustalmy $\mathcal{F}_0 = \emptyset$. W p-tym kroku będziemy dobierać zmienną objaśniającą w następujący sposób:

$$f_p^* = \underset{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}}{\arg\max}\ I(\mathcal{F}_{p-1} \cup f, C). \tag{4.3}$$

W ten sposób na p-tej iteracji dostajemy zbiór $\mathcal{F}_p$ będący sumą zbioru $\mathcal{F}_{p-1}$ składającego się z $p-1$ wybranych w wcześniejszych krokach pętli iteracyjnej oraz zmiennej f_p^* będącej rozwiązaniem problemu (4.3). Po d' iteracjach uzyskujemy zbiór $\mathcal{F}_{greedy}$, który dzięki wcześniej wspomnianym własnościom funkcji $I(\mathcal{F}_T, C)$ spełnia następującą zależność [21]:

$$\max_{\mathcal{F}_T \subset \mathcal{F}, |\mathcal{F}_T| = d'} I(\mathcal{F}_T, C) \geqslant I(\mathcal{F}_{greedy}, C) \geqslant (1 - \frac{1}{e}) \max_{\mathcal{F}_T \subset \mathcal{F}, |\mathcal{F}_T| = d'} I(\mathcal{F}_T, C). \tag{4.4}$$

Zależność ta mówi nam o stopniu bliskości rozwiązania naszego algorytmu zachłannego względem rozwiązania problemu (4.2). Gwarantuje ona, że wartość funkcji $I(\mathcal{F}_{greedy}, C)$ będzie nie gorsza niż około 63% optymalnej wartości tej funkcji.

Problem opisany w równaniu (4.3) wymaga liczenia wielowymiarowych gęstości rozkładów prawdopodobieństwa, co zazwyczaj jest bardzo kosztowne obliczeniowo. Jednakże przy założeniu niezależności zmiennych objaśniających, możemy uniknąć tego problemu w następujący sposób:

$$\begin{aligned}
f_p^* &= \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} I(\mathcal{F}_{p-1} \cup \{f\}; C) \\
&= \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} H(\mathcal{F}_{p-1} \cup \{f\}) - H(\mathcal{F}_{p-1} \cup \{f\} \mid C) && (4.5) \\
&= \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} \big[H(\mathcal{F}_{p-1}) + H(f)\big] - \big[H(\mathcal{F}_{p-1} \mid C) + H(f \mid C)\big] && (4.6) \\
&= \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} \big[H(f) - H(f \mid C)\big] && (4.7) \\
&= \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} I(f; C). && (4.8)
\end{aligned}$$

Przejście (4.5) wynika z własności informacji wzajemnej, (4.6) wynika z założenia niezależności zmiennych oraz (4.7) jest konsekwencją tego, że $H(\mathcal{F}_{p-1})$ oraz $H(\mathcal{F}_{p-1}|C)$ na p-tym kroku są stałymi z względu na $f \in \mathcal{F} \setminus \mathcal{F}_{p-1}$. Jednakże w większości problemów klasyfikacji, zmienne są między sobą zależne. Aby to uwzględnić modyfikujemy problem (4.8) następująco:

$$f_p^* = \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} \left[I(f, C) - \frac{1}{|\mathcal{F}_{p-1}|} \sum_{f \in \mathcal{F}_{p-1}} I(f, f_i) \right]. \tag{4.9}$$

Nietrudno zauważyć, że mimo powyższej argumentacji na temat doboru następnej zmiennej f_p^*, wzór (4.9) jest zastosowaniem kryterium mRMR opisanego równaniem (2.4) do problemu selekcji zmiennej.

Następnie, aby ułatwić obliczenia, korzystając z własności informacji wzajemnej z Uwagi 2.8 oraz faktu, że $H(C)$ jest stała z względu na $f \in \mathcal{F} \setminus A_{p-1}$. przekształcamy równanie (4.9) do:

$$f_p^* = \arg\max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} \left[-H(C \mid f) - \frac{1}{|\mathcal{F}_{p-1}|} \sum_{f \in \mathcal{F}_{p-1}} I(f, f_i) \right]. \tag{4.10}$$

Zanim przejdziemy do kolejnych etapów algorytmu omówimy problemy implementacyjne równania (4.10). Szczególne trudności pojawiają się przy obliczaniu entropii warunkowej $H(C \mid f)$, która odgrywa kluczową rolę w ocenie istotności zmiennych. Głównym problemem jest brak możliwości zaobserwowania prawdziwego rozkładu zmiennej C dla danych typu PL. W celu przezwyciężenia tego ograniczenia, w SAUTE zaproponowano estymację $H(C \mid f)$ w ramach uczenia z niejednoznacznie oznaczonymi etykietami. W tym celu, tworzymy dla każdej etykiety ($l \in S$) zbiór D_l składający się z obserwacji (x_i, s_i), dla których zachodzi $Y(i,l) \geqslant \frac{1}{|s_i|}$. W ten sposób dostajemy macierz D zdefiniowaną następująco:

$$D(i,l) = \begin{cases} 1, & \text{jeżeli } Y(i,l) \geqslant \frac{1}{|s_i|}, \\ 0, & \text{w.p.p.} \end{cases}. \tag{4.11}$$

Dla każdej etykiety l zakładamy, że rozkład warunkowy zmiennej f pod warunkiem etykiety l jest opisywany rozkładem normalnym $f \mid l \sim \mathcal{N}(\mu_f^l, \sigma_f^l)$, gdzie μ_f^l oraz σ_f^l oznaczają odpowiednio średnią i odchylenie standardowe. Są one oszacowane na zbiorze D_l w następujący sposób:

$$\hat{\mu}_l^f = \frac{\sum_{i \in D_l} x_i^f}{|D_l|}, \tag{4.12}$$

$$\hat{\sigma}_l^f = \sqrt{\frac{\sum_{i \in D_l} (x_i^f - \hat{\mu}_l^f)^2}{|D_l| - 1}}. \tag{4.13}$$

Poprzez oznaczenie $|D_l|$ rozumiemy liczbę obserwacji należących do zbioru D_l. Następnie, w celu oszacowania prawdopodobieństwo etykiety l pod warunkiem zmiennej f: $p(l \mid f)$, stosuje się regułę Bayesa:

$$\hat{p}(l \mid f) = \frac{\hat{p}(f \mid l) \cdot \hat{p}(l)}{\sum_{u \in \mathcal{L}} \hat{p}(f \mid u) \cdot \hat{p}(u)}, \tag{4.14}$$

przy czym $\hat{p}(u)$ obliczane jest na podstawie macierzy zaufania:

$$\hat{p}(u) = \frac{1}{m} \sum_{i=1}^{m} Y(i,u), \tag{4.15}$$

a $\hat{p}(f \mid l)$ jest oszacowaniem rozkładu $p(f \mid l)$ uzyskanym z rozkładu normalnego z parametrami $\hat{\mu}_l^f$ i $\hat{\sigma}_l^f$. Ze względu na to, że zmienna klasowa C ma charakter dyskretny, entropia warunkowa definiowana jest poprzez:

$$H(C \mid f) = -\int_{X_f} p(f) \sum_{l=1}^{q} p(l \mid f) \log p(l \mid f)\, df. \tag{4.16}$$

Celem oszacowania tej wartości używamy sumy po wszystkich m obserwacjach, zakładając równomierne prawdopodobieństwo dla każdej obserwacji:

$$\hat{H}(C \mid f) = -\sum_{j=1}^{m} \frac{1}{m} \sum_{l=1}^{q} \hat{p}(l \mid x_j^f) \log \hat{p}(l \mid x_j^f), \tag{4.17}$$

gdzie x_j^f oznacza wartość zmiennej objaśniającej f dla j-tej obserwacji.

Kolejny problem stanowi wyznaczenie informacji wzajemnej $I(f, f_i)$ między zmiennymi objaśniającymi. Aby uniknąć kosztownych obliczeniowo całek, przyjęliśmy podejście oparte na dyskretyzacji. Każda zmienna f zostaje przekształcona na zmienną dyskretną przy użyciu przedziałów wyznaczanych względem średniej μ_f i odchylenia standardowego σ_f estymowanych z wartości obserwacji dla cechy f, zgodnie z:

$$\hat{x}_i^f = \begin{cases} -2, & x_i^f \leqslant \mu_f - 2\sigma_f \\ -1, & \mu_f - 2\sigma_f < x_i^f \leqslant \mu_f - \sigma_f \\ 0, & \mu_f - \sigma_f < x_i^f \leqslant \mu_f + \sigma_f \\ 1, & \mu_f + \sigma_f < x_i^f \leqslant \mu_f + 2\sigma_f \\ 2, & x_i^f > \mu_f + 2\sigma_f \end{cases} . \tag{4.18}$$

Następnie w każdej iteracji algorytmu dobieramy kolejną zmienna według wzoru (4.10). Po ukończeniu pracy pętli otrzymujemy ramkę danych $\mathcal{D}' = \{(x_i', s_i) \mid 1 \leqslant i \leqslant m\}$, ograniczoną do zmiennych wybranych przez algorytm, czyli x_i' jest obserwacją x_i uwzględniająca tylko wybrane zmienne objaśniające. Przy użyciu zbioru $\mathcal{D}'$ będziemy aktualizować macierz pewności Y. W tym celu wprowadzimy macierz uczącą L wymiaru $m \times q$:

$$L(i, j) = \sum_{\mathbf{x}'_{i_a} \in \mathcal{N}_k(\mathbf{x}'_i)} Y(i_a, j) \cdot \omega_a, \tag{4.19}$$

gdzie $\mathcal{N}_k(\mathbf{x}'_i)$ oznacza zbiór k-najbliższych sąsiadów $\mathbf{x}'_i$, $\omega_a = k - a + 1$ oznacza wagę przypisaną a-temu najbliższemu sąsiadowi ($1 \leqslant a \leqslant k$).

Następnie dla wybranego $\alpha \in [0, 1]$ będącego hiperparametrem algorytmu, aktualizujemy macierz Y:

$$Y' = (1 - \alpha) \cdot Y + \alpha \cdot L \tag{4.20}$$

Na koniec, aby zachować interpretację macierzy pewności etykiety, dokonujemy normalizacji:

$$Y_{\text{nowy}}(i, l) = \begin{cases} \dfrac{Y'(i, l)}{\sum_{b \in s_i} Y'(i, b)}, & \text{jeżeli } l \in s_i, \\ 0, & \text{w.p.p.} \end{cases} . \tag{4.21}$$

Całą procedurę powtarzamy ustaloną liczbę razy. Pseudokod algorytmu SAUTE został przedstawiony w tabeli (4.1).

Po zakończeniu działania algorytm SAUTE zwraca zbiór danych ograniczony do wybranych zmiennych ze zbioru $\mathcal{F}_{greedy}$ oraz zaktualizowaną macierz pewności Y. Dzięki temu, integrując

SAUTE jako algorytm selekcji zmiennych z metodami klasyfikacyjnymi wykorzystującymi macierz pewności, można zamiast procedury opisanej w równaniu (4.1) użyć bezpośrednio macierzy pewności zwróconej przez SAUTE.

4.1. Modyfikacje algorytmu SAUTE

Algorytm SAUTE został zaprojektowany jako jedno z pierwszych podejść do selekcji zmiennych w problemie PLL. Naturalnymi miejscami, w których można dokonać modyfikacji algorytmu i potencjalnie uzyskać poprawę jego działania jest zmiana kryterium doboru kolejnej zmiennej oraz sposób aktualizacji macierzy pewności etykiet.

Pierwszą zaproponowaną przez nas modyfikacją jest zastąpienie oryginalnej metody aktualizacji macierzy pewności Y poprzez metodę zaproponowaną w algorytmie IPAL. Eksperymenty przeprowadzone na danych rzeczywistych wykazały wyższą skuteczność klasyfikacyjną metody IPAL niż PL-KNN [18]. Dlatego aktualizacja macierzy pewności jest naturalnym miejscem, w którym można zbadać czy przewaga w skuteczności predykcji IPAL przełoży się na poprawę działania SAUTE. Modyfikacja ta polega na zmianie sposobu tworzenia macierzy uczącej L opisanej w równaniu (4.19). Zamiast wykorzystywać metodę k-najbliższych sąsiadów użyjemy podejścia przedstawionego w algorytmie IPAL. W tym celu tworzymy graf skierowany $G = (V, E)$, gdzie $V = \{x_1, x_2, \dots x_m\}$ i $E = \{(x_i, x_j) \mid i \in \mathcal{N}_k(x_j),\ i \neq j\}$. Następnie wykonujemy kroki opisane w algorytmie IPAL oznaczone równaniami (3.3) - (3.7). Dostajemy w ten sposób macierz $\hat{Y}$, której używamy do stworzenia macierzy uczącej L według wzoru:

$$L(i,l) = \begin{cases} \dfrac{n_l}{\hat{n}_l} \cdot \hat{Y}(i,l) & \text{jeżeli } l \in s_i, \\ 0, & \text{w.p.p.} \end{cases}, \tag{4.22}$$

gdzie $n_l = \sum_{i=1}^{m} Y(i,l), \hat{n}_l = \sum_{i=1}^{m} \hat{Y}(i,l)$. Tą modyfikacje bazowego algorytmu nazwaliśmy SAUTE-IPALFRAC. Kolejna zaproponowana przez nas modyfikacja jest szczególnym przypadkiem SAUTE-IPALFRAC. W tej sytuacji macierz ucząca L wykorzystywana jest w szczególny sposób - w równaniu (4.20) parametr α przyjmuje wartość 0. W ten sposób w algorytmie SAUTE aktualizacja macierzy pewności Y w pełni zachodzi w ten sam sposób jak w algorytmie IPAL. Modyfikację tę nazwano SAUTE-IPALALL. Motywacją stojącą za wyszczególnieniem tego przypadku jest to, że algorytm IPAL sam sobie posiada parametr będący odpowiednikiem parametru α w metodzie SAUTE.

Kolejna modyfikacja dotyczy kryterium wyboru zmiennej. W oryginalnej pracy [1] ten dobór w zachłannym algorytmie odbywa się przy użyciu kryterium mRMR uwzględniającym istnienie

zależności między zmiennymi objaśniającymi. Naszym celem jest zastąpienie sposobu doboru nowej zmiennej poprzez kryterium JMI, którego teoretyczne właściwości zostały zbadane w pracy [2], będącym naturalną aproksymacją warunkowej informacji wzajemnej opisanej równaniem (2.3). Stąd kolejna modyfikacja algorytmu SAUTE nazwana SAUTE-JMI polega na zastąpieniu równania (4.10) poprzez:

$$\begin{aligned} f_p^* &= \arg \max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} JMI(f, C \mid \mathcal{F}_{p-1}) \\ &= \arg \max_{f \in \mathcal{F} \setminus \mathcal{F}_{p-1}} \left[I(f, C) + \frac{1}{|\mathcal{F}_{p-1}|} \sum_{i=1}^{p-1} [I(f, f_i \mid C) - I(f, f_i)] \right]. \end{aligned} \tag{4.23}$$

Wprowadzając wzór (4.23) pojawia się problem implementacji $I(f, C)$ oraz $I(f, f_i \mid C)$, który możemy rozwiązać w następujący sposób: podobnie jak poprzednio pierwszy wyraz korzystając z własności z Uwagi 2.8 zastępujemy przez $-H(C \mid f)$, który obliczamy według wzoru (4.17). Drugi wyraz będziemy liczyć w następujący sposób, korzystając z Definicji 2.9:

$$I(f_j, f_i \mid C) = \frac{1}{m^2} \sum_{l \in S} \sum_{x_i \in f_i} \sum_{x_j \in f_j} p(l)\ p(x_j, x_i \mid l)\ \log_2 \left(\frac{p(x_j, x_i \mid l)}{p(x_j \mid l)\ p(x_i \mid l)} \right) \tag{4.24}$$

Obliczenie rozkładu $p(x_j, x_i \mid l)$ wymaga przyjęcia dodatkowego założenia dotyczącego współzależności zmiennych objaśniających w obrębie etykiety l, w tym celu zakładamy, że rozkład warunkowy $f_i, f_j \mid l$ jest rozkładem normalnym:

$$f_i, f_j \mid l \sim \mathcal{N} \left(\begin{pmatrix} \mu_l^{f_i} \\ \mu_l^{f_j} \end{pmatrix}, \begin{bmatrix} \sigma_l^{f_i 2} & \operatorname{cov}(f_i^l, f_j^l) \\ \operatorname{cov}(f_j^l, f_i^l) & \sigma_l^{f_j 2} \end{bmatrix} \right),$$

gdzie $\mu_l^{f_i}$ oraz $\sigma_l^{f_i}$ są szacowane za pomocą wzorów kolejno (4.12) i (4.13). Natomiast kowariancje szacujemy przy użyciu następującego wzoru:

$$\hat{\operatorname{cov}}(f_i^l, f_j^l) = \frac{1}{|D_l| - 1} \sum_{k \in D_l} \left(x_k^{f_i} - \hat{\mu}_l^{f_i} \right) \left(x_k^{f_j} - \hat{\mu}_l^{f_j} \right). \tag{4.25}$$

Wyniki eksperymentów badających działanie zaproponowanych przez nas modyfikacji omówimy szczegółowo w następnym rozdziale.

Tabela 4.1: Pseudokod algorytmu SAUTE.

dane wejściowe:	$\mathcal{D} = \{(x_i, s_i) \mid 1 \leqslant i \leqslant m\}$: zbiór treningowy PL, gdzie $x_i \in \mathbb{R}^d$, $s_i \subseteq S = \{1, 2, \ldots, q\}$, d': liczba zmiennych do wybrania, α: współczynnik uczenia, k: liczba najbliższych sąsiadów.
dane wyjściowe:	$\mathcal{D}' = \{(x'_i, s_i) \mid 1 \leqslant i \leqslant m\}$: zbiór treningowy ograniczony do d' wybranych zmiennych. Y: zaktualizowana macierz pewności etykiet
Algorytm:	
1.	Stwórz macierz pewności etykiet Y $m \times q$ używając równania (4.1);
2.	**powtarzaj:**
3.	Stwórz $\mathcal{F}_0 = \emptyset$;
4.	**dla** $p = 1$ **do** d' **rób:**
5.	Oblicz $\hat{H}(C \mid f)$ dla każdego $f \in F \setminus \mathcal{F}_{p-1}$ używając (4.17);
6.	Oblicz $\sum_{f_i \in \mathcal{F}_{p-1}} I(f; f_i)$ dla wszystkich $f \in F \setminus \mathcal{F}_{p-1}$ dyskretyzując obserwacje według wzoru (4.18);
7.	Wybierz f_p^* zgodnie z (4.10);
8.	Zaktualizuj $\mathcal{F}_p = \mathcal{F}_{p-1} \cup \{f_p^*\}$;
9.	**skończ pętle dla.**
10.	Skonstruuj $\mathcal{D}' = \{(x'_i, s_i) \mid 1 \leqslant i \leqslant m\}$ używając wybranych zmiennych;
11.	Dla każdego $\mathbf{x}'_i$, znajdź k-najbliższych sąsiadów $\mathcal{N}_k(\mathbf{x}'_i)$;
12.	Oblicz macierz uczącą L według wzoru (4.19);
13.	Oblicz macierz pośrednią Y' według wzoru (4.20);
14.	Zaktualizuj macierz pewności etykiet Y_{nowy} według wzoru (4.21);
15.	Ustaw $Y = Y_{\text{nowy}}$;
16.	**dopóki** algorytm nie zbiegnie.
17.	Stwórz d' wymiarowy zbiór danych $\mathcal{D}'$ na podstawie $\mathcal{F}_{greedy}$;
18.	**zwróć** $\mathcal{D}'$ i Y.

5. Eksperymenty

5.1. Opis danych

Eksperymenty zostały przeprowadzone na częściowo etykietowanych zbiorach rzeczywistych **Lost** [29] i **FG-NET** [30]. Dane te są dostępne na stronie palm.seu.edu.cn/zhangml. Zbiór danych **Lost** został przygotowany na podstawie 16 odcinków popularnego serialu telewizyjnego "Lost" i zawiera 1122 obserwacje, z których każda odpowiada jednej scenie filmowej, zidentyfikowanej jako osobny kadr. Każda obserwacja zawiera wybrane zmienne opisujące wizualne i semantyczne aspekty sceny. Do każdej obserwacji przypisany jest zbiór etykiet będący możliwymi numerami odcinków, w których mogła pojawić się ta scena. Głównym zadaniem klasyfikacyjnym w ramach tego zbioru jest przypisanie konkretnej sceny do odpowiedniego odcinka serialu, z którego ona pochodzi. Zbiór danych **FG-NET** Aging Database został opracowany w ramach europejskiego projektu w celu wspierania badań nad procesem starzenia się twarzy. Zawiera 1002 zdjęcia 82 osób w wieku od 0 do 69 lat, przy czym największe zagęszczenie przypadków przypada na przedział 0–40 lat. Zbiór zawiera szczegółowy opis każdego zdjęcia obejmujący charakterystyczne punkty twarzy oraz metadane takie jak płeć, wyraz twarzy, obecność okularów, zarostu czy nakrycia głowy. Każdej obserwacji przypisany jest zbiór etykiet składający się z prawdziwego wieku danej osoby oraz losowo dobranych innych lat życia. Celem klasyfikacji jest określenie wieku osoby będącej na zdjęciu. Dla obydwu zbiorów dysponujemy również wartościami prawdziwych etykiet, których nie uwzględniamy w procesie uczenia, ale posłużą nam do ewaluacji poszczególnych metod.

5.2. Opis eksperymentów

Celem naszych eksperymentów było zbadanie w jaki sposób użycie poszczególnych modyfikacji algorytmu SAUTE wpływa na porawę zdolności klasyfikacyjnej algorytmu PL-KNN. Jako miarę jakości klasyfikacji przyjęliśmy Accuracy, zdefiniowaną następująco:

Definicja 5.1. (miara Accuracy)
Niech (x_i, s_i), gdzie $i = 1, 2, \ldots, m$ będzie zbiorem zbiorem częściowo oznaczonych danych, c_i - prawdziwą etykietą, a c^* będzie klasyfikatorem, wówczas miarę Accuracy definiujemy następująco:

$$Accuracy(c^*) = \frac{\sum_{i=1}^{m} \mathbb{I}\left(c^*(x_i) = c_i\right)}{m}.$$

Miara Accuracy określa, jaki odsetek przykładów na danym zbiorze został zaklasyfikowany prawidłowo przez klasyfikator. Wartość maksymalną równą 1 miara przyjmuje, kiedy model dokona wszystkich klasyfikacji prawidłowo. Wartość minimalną równą 0 przyjmuje, kiedy wszystkie obserwacje zostały zaklasyfikowane błędnie. Zaletą tej miary jest prostota zarówno implementacji i interpretacji. Natomiast w przypadku niezbalansowanych klas (np. gdy jedna klasa dominuje w danych), wartość miary Accuracy może być myląca – model może mieć wysoką dokładność, mimo że praktycznie ignoruje klasę mniejszościową.

Wydajność modeli uczenia maszynowego w dużej mierze zależy od doboru hiperparametrów. Jedną z technik wyboru optymalnych hiperparametrów jest przeszukiwanie siatki. Polega ona na systematycznym testowaniu modelu używając wszystkich kombinacji wartości z iloczynu kartezjańskiego wybranych hiperparametrów. Zastosowanie tej techniki może przynieść korzyści w postaci modelu uzyskującego znacznie lepsze wyniki niż model z hiperparametrami dobranymi losowo. Nasze eksperymenty z użyciem przeszukania siatki hiperparametrów odbyły się na metodach SAUTE, SAUTE-IPALFRAC i SAUTE-IPALALL zintegrowanych z algorytmem PL-KNN o k=10 najbliższych sąsiadach. Wybrane dla tych metod siatki hiperparametrów przedstawiano w Tabelach 5.1 - 5.3.

Hiperparametr	**Wartości**
$\text{knn}_{\text{SAUTE}}$	{4, 8, 20}
α_{SAUTE}	{0.3, 0.6, 0.9}

Tabela 5.1: Zestaw wartości hiperparametrów użytych w przeszukiwaniu siatki dla metody SAUTE

Eksperyment przeprowadzony został dla trzech różnych proporcji wybranych zmiennych do modelu, kolejno: $d' = 0.3, 0.5, 0.7$. Eksperyment polegał na $10-$krotnym podziale zbioru danych na trzy rozłączne podzbiory: treningowy, walidacyjny i testowy w proporcji 49:21:30. Ta nietypowa proporcja wzięła się z podziału zbioru najpierw na dwa zbiory w proporcji 7:3, a następnie ponownym podziale większego podzbioru w proporcji 7:3. Następnie dla każdego podziału zbioru, używając techniki przeszukania siatki najpierw dopasowano na zbiorze treningowym mo-

Hiperparametr	**Wartości**
$\text{knn}_{\text{SAUTE}}$	{4, 8, 20}
knn_{IPAL}	{4, 8, 20}
α_{SAUTE}	{0.3, 0.6, 0.9}
α_{IPAL}	{0.3, 0.6, 0.9}

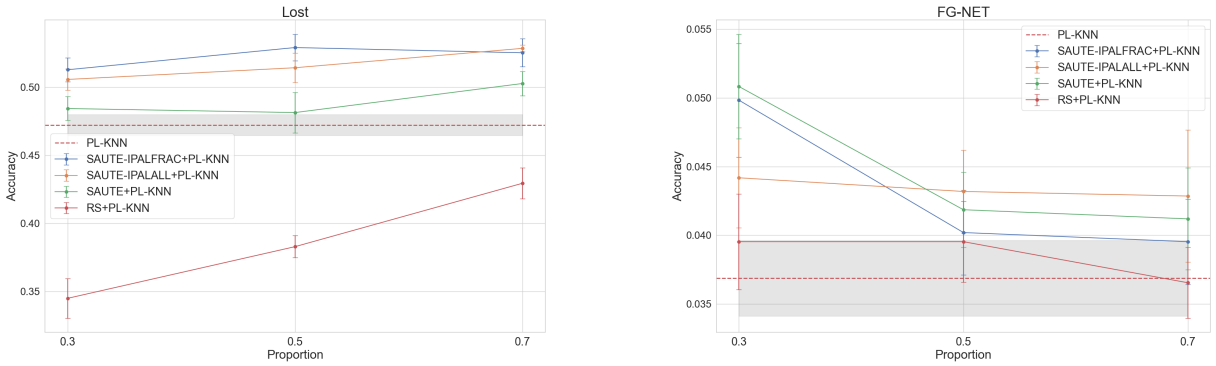
Tabela 5.2: Zestaw wartości hiperparametrów użytych w przeszukiwaniu siatki dla metody SAUTE-IPALFRAC

Hiperparametr	**Wartości**
$\text{knn}_{\text{SAUTE}}$	{4, 8, 20}
knn_{IPAL}	{4, 8, 20}
α_{IPAL}	{0.3, 0.6, 0.9}

Tabela 5.3: Zestaw wartości hiperparametrów użytych w przeszukiwaniu siatki dla metody SAUTE-IPALALL

dele integrujące daną metodę selekcji zmiennych z algorytmem PL-KNN dla każdej kombinacji parametrów. Następnie dla każdego z tych modeli policzono Accuracy na zbiorze walidacyjnym. Uporządkowano modele malejąco względem uzyskanych wyników tworząc w ten sposób ranking kombinacji hiperparametrów. Korzystając z niego wybieramy tą kombinacje, która miała największe Accuracy. Wybrana kombinacja hiperparametrów następnie zostawała użyta do dopasowania modelu na zbiorze będącym sumą teoriomnogościową zbioru treningowego i walidacyjnego. Na koniec, w celu obliczania zdolności predykcyjnych tego modelu ponownie była używana miara Accuracy na zbiorze testowym. Grafika ilustrująca ten proces przedstawiona jest na rysunku 5.1. Niestety z względu na złożoność obliczeniową zastosowanie techniki przeszukania siatki parametrów dla metody SAUTE-JMI było niemożliwe do uzyskania w sensownym czasie. W celu oceny tej metody został zastosowany pięciokrotny sprawdzian krzyżowy na następujących hiperparametrach: $\alpha_{SAUTE} = 0.6, knn_{SAUTE} = 8$ zarówno dla metody SAUTE-JMI jak i SAUTE będącej punktem odniesienia dla nowo wprowadzonej modyfikacji. W tej sytuacji metody SAUTE-JMI i SAUTE również zostały zintegrowane z modelem PL-KNN z liczbą najbliższych sąsiadów równą 10.

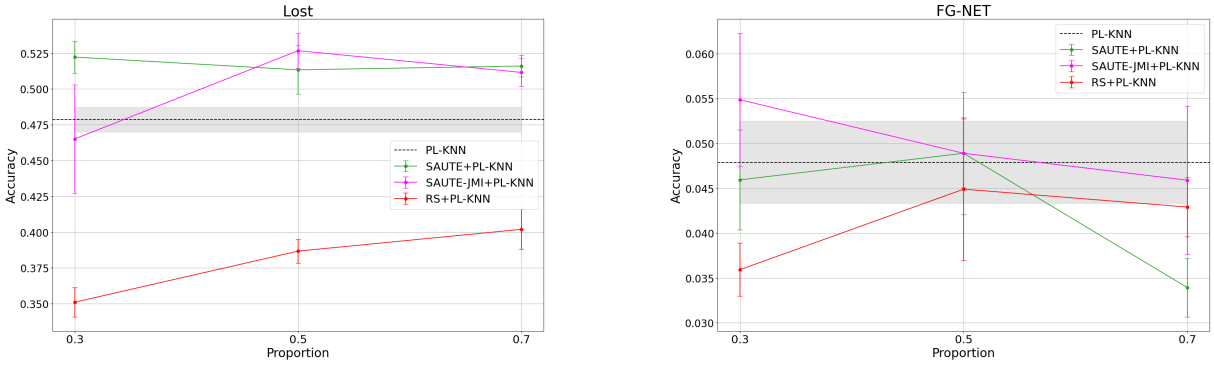
5.3. Opis wyników



(a) Lost (b) FG-NET

Rysunek 5.2: Wyniki eksperymentów dla modyfikacji bazujących na metodzie IPAL.

Na rysunku 5.2 przedstawiono wyniki porównania miary Accuracy dla metod: PL-KNN na wszystkich zmiennych, PL-KNN z selekcją metodą SAUTE, PL-KNN z selekcją metodą SAUTE-IPALALL, PL-KNN z selekcją metodą SAUTEIPAL-FRAC, PL-KNN z selekcją metodą losowego wyboru (RS) uzyskanych na zbiorach kolejno Lost i FG-NET w zależności od proporcji zmiennych d' wybranych w procedurze selekcji. Wartości uzupełniono o przedziały ufności. Analiza wskazuje, iż metody selekcji zmiennych oparte na algorytmie SAUTE uzyskują wyraźnie wyższe wyniki w stosunku do metody bazowej PL-KNN zastosowanej na pełnym zbiorze zmiennych. Na zbiorze Lost szczególnie istotne jest to, że dwie modyfikacje – SAUTE-IPALFRAC oraz SAUTE-IPALALL – osiągają znacząco lepsze wyniki od klasycznej wersji SAUTE. Na zbiorze FG-NET ciężko jednoznacznie wskazać modyfikację SAUTE, która uzyskała najlepsze wyniki, gdyż różnicę między nimi nie są istotne statystycznie. Z kolei metoda losowego doboru zmiennych objaśniających - RS osiąga zauważalnie niższą skuteczność niż pozostałe podejścia, co dowodzi poprawności metody SAUTE i jej modyfikacji. Podsumowując, uzyskane rezultaty wskazują na przewagę metod SAUTE nad podejściem bazowym, a dodatkowo pokazują potencjał modyfikacji SAUTE-IPALFRAC oraz SAUTE-IPALALL, które mimo niewielkiej skali różnic mogą przewyższać klasyczną wersję algorytmu. Wyniki te stanowią przesłankę do dalszych badań nad rozwijaniem i optymalizacją proponowanych modyfikacji.



(a) Lost

(b) FG-NET

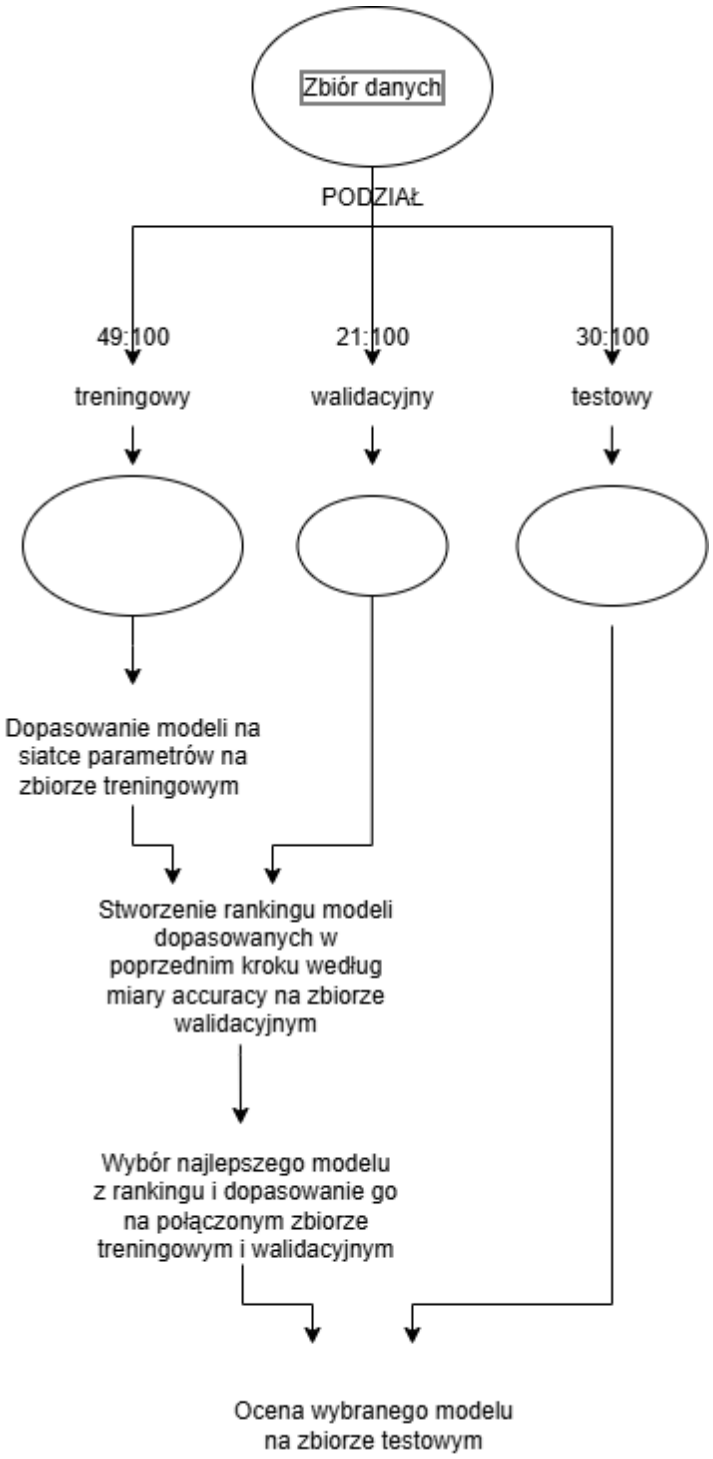
Rysunek 5.3: Wyniki eksperymentów dla modyfikacji bazującej na mierze JMI.

Przejdźmy teraz do wyników zastosowania kryterium JMI w metodzie SAUTE. Rysunek 5.3 przedstawia wyniki porównania trzech metod PL-KNN zintegrowanych z algorytmami selekcji zmiennych kolejno: z SAUTE, z SAUTE-JMI oraz z losowym wyborem zmiennych (RS). Dla zbioru Lost zarówno SAUTE, jak i SAUTE-JMI w początkowej proporcji osiągały wyniki zbliżone lub wyższe od PL-KNN na wszystkich zmiennych. Obie metody z grupy SAUTE znacząco poprawiły zdolności klasyfikacyjne algorytmu PL-KNN. Dla zbioru FG-NET wyniki były mniej jednoznaczne, ale również w tym przypadku metody SAUTE i SAUTE-JMI w większości proporcji zmiennych osiągały nie gorsze wyniki niż PL-KNN bez algorytmu selekcji zmiennych. Najsłabsze rezultaty odnotowała losowa selekcja zmiennych, co również potwierdza, że zarówno SAUTE jak i SAUTE-JMI poprawnie identyfikują istotne zmienne. Mimo braku istotnej statystycznie przewagi SAUTE-JMI nad bazowym SAUTE, modyfikacja osiągnęła zbliżone rezultaty potwierdzające jej działanie jako algorytmu identyfikującego istotne zmienne objaśniające.

5.4. Podsumowanie eksperymentów

Przeprowadzone eksperymenty potwierdzają przewagę metod SAUTE nad podejściem bazowym PL-KNN oraz losową selekcją zmiennych osiągając zarówno na zbiorze Lost, jak i FG-NET wyższe wartości miary Accuracy. Szczególnie widoczny efekt uzyskano dla zaproponowanych przez nas modyfikacji SAUTE-IPALFRAC oraz SAUTE-IPALALL, które zazwyczaj przewyższały lub były porównywalne z klasyczną wersją SAUTE. Wyniki były szczególnie korzystne w przypadku zbioru Lost. Na zbiorze FG-NET różnice między wariantami SAUTE były mniej wyraźne. W przypadku zastosowania kryterium JMI metoda SAUTE-JMI uzyskała wyniki lepsze od losowej selekcji, lecz jednoznacznie nie przewyższała klasycznej wersji SAUTE. Tym samym modyfikacja ta nie poprawiła jakości podstawowego algorytmu. Podsumowując, wyniki

badań wskazują, że metody z rodziny SAUTE stanowią skuteczną metodę selekcji zmiennych w problemie PLL znacznie poprawiając zdolności klasyfikujące. Modyfikacje SAUTE-IPALFRAC, SAUTE-IPALALL i SAUTE-JMI wykazują potencjał i mogą stanowić ciekawy temat przyszłych badań.



Rysunek 5.1: schemat eksperymentów

6. Podsumowanie

W niniejszej pracy omówiono submodularny algorytm selekcji zmiennych dla danych częściowo oznaczonych oraz zaproponowano jego modyfikacje. Praca obejmowała część teoretyczną i eksperymentalną. W pierwszej kolejności przeanalizowano literaturę dotyczącą algorytmu SAUTE i jego dotychczasowych zastosowań. Ponadto przedstawiono najważniejsze pojęcia związane z tym obszarem takie jak miary zależności między zmiennymi bądź algorytmy selekcji zmiennych. Punktem wyjścia dla modyfikacji była analiza możliwych wrażliwych punktów algorytmu. W celu ich poprawy przeprowadzono analizę istniejących rozwiązań z obszaru PLL oraz miar zależności między zmiennymi objaśniającymi a zmienną odpowiedzi. Na tej podstawie zaproponowano autorskie modyfikacje polegające na zmianie mechanizmu aktualizacji macierzy pewności oraz udoskonaleniu sposobu doboru kolejnej zmiennej, których celem była poprawa efektywności procesu optymalizacji. Następnie opracowano implementację modyfikacji oraz przygotowano środowisko testowe obejmujące dwa rzeczywiste zbiory danych pozwalające na obiektywną ocenę rezultatów. W eksperymentach porównano oryginalną wersję algorytmu ze zmodyfikowanymi wariantami, a wyniki poddano analizie. Rezultaty potwierdziły skuteczność bazowej wersji algorytmu oraz wykazały, że zaproponowane zmiany w określonych warunkach pozwalają osiągnąć lepsze rezultaty. Jednocześnie skuteczność modyfikacji okazała się silnie uzależniona od charakterystyki środowiska testowego, co wskazuje na potrzebę dalszych badań w kierunku zwiększenia uniwersalności podejść. Podsumowując, praca dowiodła, że możliwe jest ulepszenie algorytmu SAUTE poprzez wprowadzenie modyfikacji w jego strukturze. Otrzymane wyniki z jednej strony dostarczają dowodów na zasadność dalszych badań w tym obszarze, z drugiej mogą stanowić punkt odniesienia dla inżynierów i badaczy poszukujących metod poprawy efektywności algorytmów klasyfikujących w problemie PLL. W świetle przeprowadzonych badań można wskazać kilka potencjalnych kierunków dalszych prac. Warto rozważyć rozszerzenie analizy na bardziej złożone i dynamiczne środowiska testowe w celu oceny uniwersalności zaproponowanych rozwiązań. Kolejnym tematem może być wprowadzenie dodatkowych modyfikacji, np. zastosowanie miary CIFE [15] jako kryterium oceny istotności zmiennych objaśniających lub zmiana strategii selekcji na backward bądź bidirectional. Ciekawym kierunkiem badań może być również metoda łącząca modyfikacje zaproponowane w niniejszej pracy, jednocześnie wykorzystująca miarę JMI oraz aktualizację ma-

cierzy pewności metodą IPAL. Ostatecznie, praca zrealizowała założone cele badawcze i wykazała zasadność dalszej eksploracji omawianej tematyki. Wyniki dowodzą, że nawet niewielkie zmiany w istniejących algorytmach mogą prowadzić do zauważalnych usprawnień, otwierając drogę do pogłębionych badań i potencjalnych innowacji w obszarze PLL.

Bibliografia

[1] Wei-Xuan Bao, Jun-Yi Hang, and Min-Ling Zhang. Submodular Feature Selection for Partial Label Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, pages 1534–1544, Washington D.C., August 2022. ACM.

[2] Łazęcka Małgorzata, Mielniczuk Jan: Analysis of Information-Based Nonparametric Variable Selection Criteria, Entropy, MDPIAG, vol. 22, no. 9, 2020, pp. 974-992, DOI:10.3390/e22090974

[3] Lei Feng and Bo An. Partial Label Learning with Self-Guided Retraining. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 3542–3549, Honolulu, Hawaii, 2019. AAAI Press.

[4] Timothee Cour, Benjamin Sapp, and Ben Taskar. Learning from Partial Labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

[5] Eyke Hüllermeier and Johannes Beringer. Learning from Ambiguous Data. *Intelligent Data Analysis*, 10(5):419–439, 2006.

[6] Bing Liu and Thomas G. Dietterich. A Conditional Multinomial Mixture Model for Superset Label Learning. In *Advances in Neural Information Processing Systems*, pages 553–561, 2012.

[7] Feng Yu and Min-Ling Zhang. Maximum Margin Partial Label Learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 2645–2651, 2016.

[8] W.-X. Bao, J.-Y. Hang, and M.-L. Zhang. 2021. Partial label dimensionality reduction via confidence-based dependence maximization.

[9] J.-H. Wu and M.-L. Zhang. 2019. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction.

[10] M.-L. Zhang, J.-H. Wu, and W.-X. Bao. 2022. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction.

[11] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[12] Min-Ling Zhang and Zhi-Hua Zhou. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[13] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] L. Jie and F. Orabona. 2010. Learning from candidate labeling sets. In Advances in Neural Information Processing Systems 23. Cambridge, MA, 1504–1512.

[15] Lin, D.; Tang, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In European Conference on Computer Vision 2006 May 7; Springer: Berlin/Heidelberg, Germany, 2006; pp. 68–82.

[16] Yang, H.H.; Moody, J. Data visualization and feature selection: New algorithms for non-gaussian data. Adv. Neural. Inf. Process Syst. 1999, 12, 687–693.

[17] David MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003.

[18] M.-L. Zhang and F. Yu. 2015. Solving the partial label learning problem: An instance-based approach. In Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 4048–4054

[19] A. Krause and C. Guestrin. 2005. Near-optimal nonmyopic value of information in graphical models. In Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland. 324–331

[20] N. Nguyen and R. Caruana. 2008. Classification with partial labels. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, NV, 381–389.

[21] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. Mathematical programming 14, 1 (1978), 265–294

[22] Hanchuan Peng, Fuhui Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,"in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[23] C. Qian, Y. Yu, K. Tang, X. Yao, and Z.-H. Zhou. 2019. Maximizing submodular or monotone approximately submodular functions by multi-objective evolutionary algorithms. Artificial Intelligence 275 (2019), 279–294.

[24] Kohavi, R., John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1–2), 273–324.

[25] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society

[26] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[27] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249–268.

[28] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

[29] T. Cour, B. Sapp, and B. Taskar. 2011. Learning from partial labels. Journal of Machine Learning Research 12 (2011), 1501–1536.

[30] G. Panis and A. Lanitis. 2014. An overview of research activities in facial age estimation using the FG-NET aging database. In Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerlan, 737–750.

[31] German, B. (1987). Glass Identification [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5WW2P.

[32] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

[33] C.-H. Chen, V. M. Patel, and R. Chellappa. 2018. Learning from ambiguously labeled face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 7 (2018), 1653–1667.

[34] F. Briggs, X. Z. Fern, and R. Raich. 2012. Rank-loss support instance machines for MIML instance annotation. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 534–542.

[35] C. Zhou, M. Prabhushankar, and G. AlRegib, "Perceptual Quality-based Model Training under Annotator Label Uncertainty," in International Meeting for Applied Geoscience & Energy (IMAGE) 2023, Houston, TX, Aug. 28-Sept. 1, 2023.

[36] D. Zhou, Z. Zhang, M.-L. Zhang, and Y. He. 2018. Weakly supervised POS tagging without disambiguation

Spis rysunków

Spis tabel