POLITECHNIKA WROCŁAWSKA
WYDZIAŁ ELEKTRONIKI

KIERUNEK: INFORMATYKA
SPECJALNOŚĆ: SYSTEMY INFORMATYKI W MEDYCYNIE

PRACA DYPLOMOWA MAGISTERSKA

Analiza porównawcza wybranych metod selekcji cech w zadaniu klasyfikacji danych niezbalansowanych.

Comparative analysis of feature selection techniques in inbalanced data classification problems.

AUTOR:
Maciej Hajduk

OPIEKUN PRACY:
dr inż. Paweł Trajdos, KSSK

WROCŁAW 2021

Spis treści

1 Wstęp **3**
1.1 Wprowadzenie . . . 3

2 Analiza problemu **5**
2.1 Przegląd literaturowy . . . 5
2.2 Cel selekcji cech . . . 6
2.3 Podstawowy podział . . . 6
2.3.1 Metody rankingowe . . . 7
2.3.2 Metody opakowane . . . 7
2.3.3 Metody wbudowane . . . 7
2.4 Selekcja cech a ekstrakcja . . . 7

3 Szczegółowa charakterystyka zagadnienia **10**
3.1 Metody klasyfikacji danych niezbalansowanych . . . 11
3.1.1 Metody na poziomie danych . . . 11
3.1.2 Metody na poziomie algorytmów . . . 11
3.1.3 Podejścia hybrydowe . . . 12

4 Założenia i plan eksperymentu **14**
4.1 Ocena działania algorytmów . . . 14

5 Wyniki **14**

6 Wnioski **16**

7 Bibliografia **18**

8 Zawartość płyty CD **20**

1 Wstęp

1.1 Wprowadzenie

Celem pracy jest porównanie różnych metod selekcji cech w problemie trenowania algorytmów uczenia maszynowego na danych niezbalansowanych. W jej ramach, przedstawione i opisane zostaną popularne obecnie metody selekcji oraz przeprowadzone zostaną eksperymenty dla przykadowych zbiorów danych, zarówno naturalnych jak i sztucznie wygenerowanych, celem których będzie stworzenie rankingu algorytmów. Autor sprawdzi, jak właściwie przeprowadzona selekcja wpływa na jakość wyników dostarczanych przez klasyfikator i jak przytoczone przez niego metody radzą sobie z danymi, w których występuje znaczna przewaga liczebności jednej bądź kilku klas. Aspekt inżynieryjny opierać się będzie na implementacji zaproponowanych w pracy eksperymentów, co pozwoli na kompleksowe porównanie algorytmów.

Praca swoim zakresem obejmie porównanie popularnych metod selekcji cech w ramach kilku wybranych zbiorów danych. Napisany w jej ramach program pozwoli …

Praca składa się z pięciu rozdziałów:

Rozdział pierwszy: Na rozdział pierwszy skłądają się omówienie analizy wybranego problemu, przedstawienie motywacji podjęcia tego tematu oraz przegląd literatury. Zostały opisane w nim również podstawowe metody selekcji cech i wyjaśnienie różnicy pomiędzy selekcją oraz ekstrakcją.

Rozdział drugi: Szczegółowa charakterystyka zagadnienia opisana w rozdziale zawiera opis problemu jakim jest niezrównoważnony rozkład klas w algorytmie uczenia maszynowego. Rozdział zawiera też szczegółowy opis poszczególnych, wykorzystanych później metod…

Rozdział trzeci: Założenia i plan eksperymentu. W rodziale trzecim zawarte zostaną informacje związane z inżynieryjnym aspektem pracy, czyli projekt systemu, plan poszczególnych ekpserymentów i opis danych, jakie użyte zostaną podczas doświadczeń. Znajdują się tutaj również instrunkcje instalacji i wdrożenia systemu dla potencjalnych środowisk docelowych.

Rozdział czwarty: Rozdział zawiera podsumawnie uzyskanyh wyników oraz przedstawienie ich w czytelny i zrozumiały sposób.

Rozdział piąty: W rozdziale piątym zawarta zostanie interpretacja oraz konfrontacja wyniki z hipotezą postawioną na początku pracy. Przedstawione zostaną ewentualne możliwości rozwoju projektu.

Udało się zrealizować wszystkie postawione cele.

2 Analiza problemu

Uczenie maszynowe to bardzo dynamicznie rozwijająca się gałąź informatyki. Niezwykła ekspansja wynika z zapotrzebowania na wykrywanie prawidłowości, uogólnianie oraz precyzowanie danych. Takie możliwości pozwoliły znaleźć zastosowanie dla algorytmów sztucznej inteligencji w bardzo wielu różnych branżach - począwszy od medycyny, poprzez finanse, produkcję i branżę rozrywkową. Tak duży przekrój różnych zastosowań wymaga ciągłego ulepszania istniejących już wzorców oraz wymyślania nowych, lepszych i bardziej efektywnych algorytmów. W większości praktycznych problemów do klasyfikacji obiektów, autor programu operuje na dużej ilości cech. Należy jednak pamiętać, że w tym przypadku wiele, nie oznacza wczale lepszych rezultatów. Należy przytoczyć pojęcie „przekleństwa wielowymiarowości". Oznacza ono, że większy wymiar wymaga od programisty znacznie większej ilości danych, oraz wraz ze wzrostem ilości cech wykładniczo rośnie liczba możliwych wariantów, co znacznie zwiększa złożoność obliczeniową naszych algorytmów.

Aby uniknąć problemów generowanych przez zbyt dużą ilość cech, a jednocześnie wykorzystać cech, które zapewniają jak najlepszą separowalność klas, zazwyczaj pierwszym krokiem w zadaniu klasyfikacji jest selekcja lub ekstrakcja najodpowiedniejszych cech.

2.1 Przegląd literaturowy

W rozdziale zostaną pokrótce opisane wybrane artykuły naukowe zajmujące się tematyką selekcji cech. Jest to, szczególnie ostatnio, często poruszany 1problem, co skutkuje dużym przekrojem prac, również w ujęciu czysto dziedzinowym - jak wykorzystanie konkretnych algorytmów dla bardzo konkretnych zastosowań.

Wstępną analizę problemu przedstawił Jakub Piątkowski w pracy *Analiza i rozwój metod doboru cech dla dużych problemów klasyfikacyjnych*. Autor wymiania istniejące metody doboru cech i szczegółowo je opisuje. Jest to dobre wprowadzenie do problematyki klasyfikacji i selekcji cech, a każda przytoczona metoda ma również podłoże matematyczne opisane odpowienimi wzorami. W podobnej pracy *Redukcja wymiarowości i selekcja cech w zadaniach klasyfikacji i regresji z wykorzystaniem uczenia maszynowego*, twórca zacytował wyniki swoich eksperymentów, które pozwoliły mu na tworzenie rankingów algorytmów. Tezy postawione przez autora zostaną sprawdzone w tej pracy. Szczegółowo do problemu podszedł mgr Wiesław Chmielnicki, w swojej rozprawie: *Efektywne metody selekcji cech i rozwiązywania problemu wieloklasowego w nadzorowanej klasyfikacji danych*. Oprócz opisu znanych metod, artykuł zawiera również sugestie dotyczące nowych algorytmów hybrydowych, które w niektórych przypadkach dają lepsze wyniki niż metody tradycyjne. Mark A. Hall porusza interesujący temat w swojej książce *Correlation-based Feature Selection for Machine Learning*, badając wartość zestawu cech na podstawie korelacji pomiędzy nimi. Autor przeprowadził szereg eksperymentów, porównał swoją metodę z metodami powszechnie stosowanymi, starając się między innymi wyodrębnić

problemy, dla których jego algorytm jest najbardziej skuteczny. Twórcy pracy *A Survey on Evolutionary Computation Approaches to Feature Selection* zajęli się przeglądem znanych metod tworząc dokument podsumowujący każdą z nich, z jej wadami oraz zaletami. Artykuł jest oparty na przeglądzie najnowszych prac w zadanej dziedzinie i pozwala na dobranie odpowiedniej metody do zadanego zadania. Ważna dla tematu tej pracy jest również rozprawa *Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji* mgr inż. Piotra Płońskiego, podsumowująca niejako cały proces uczenia maszynowego i roli, jaką pełni selekcja cech w kontekscie dobrze działającego systemu analizy danych.

2.2 Cel selekcji cech

Selekcja cech polega na identyfikacja tych elementów puli cech, które uznawane są za najlepsze deskryptory rozważanych kategorii. Zaletą selekcji jest możliwość zbadania tych deskryptorów, które są istotne z punktu widzenia danego zadania klasyfikacji, czyli jednocześnie zrozumienia różnic między analizowanymi kategoriami. Poprzez proces selekcji cech tracimy niestety bezpowrotnie część początkowych cech, a wiedza w cechach wybranych jest często dublowana. Z tego powodu nie selekcja, a ekstrakcja cech jest obecnie najpowszechniejszą strategią służącą przygotowaniu reprezentacji analizowanych danych. \ Selekcja cech jest odpowiedzialna za wybór najbardziej istotnych atrybutów badanych obiektów, co przekłada się bezpośrednio na poprawne działanie klasyfikatora. Dyspozycja coraz większymi bazami danych zmusza do optymalizacji tego procesu. Gwałtownie rosnąca liczba cech stanowi poważny problem - powoduje nie tylko wydłużenie procesu uczenia oraz wzrost złożoności klasyfikatora, ale niesie ze sobą także ryzyko spadku poprawnej klasyfikacji. Związane jest to z tak zwanym ”przekleństwem wymiarowości”. Zjawisko to zachodzi, gdy ilość cech znacznie przewyższa liczebność samego zbioru danych. Zadaniem selekcji cech jest również lepsze zrozumienie problemu oraz zmniejszenie kosztów archiwizacji przyszłych danych. W kolejnych rozdziałach opisane zostaną trzy główne metody tworzenia algorytmów selekcji: metody rankingowe - zwane filtrami, metody opakowane oraz metody wbudowane. Dla każdej z wymienionych metod zostanie określona idea, oraz przedstawione zostaną algorytmy reprezentujące daną metodologie.

2.3 Podstawowy podział

W kolejnych rozdziałach opisane zostaną trzy główne metody tworzenia algorytmów selekcji: metody rankingowe - zwane filtrami, metody opakowane oraz metody wbudowane. Dla każdej z wymienionych metod zostanie określona idea, oraz przedstawione zostaną algorytmy reprezentujące daną metodologie.

2.3.1 Metody rankingowe

Najprostsze podejście do problemu selekcji cech reprezentowane jest właśnie poprzez metody rankingowe, nazywane też filtrami. Jak sama nazwa wskazuje do zadania selekcji przy pomocy metod rankingowych podchodzimy wyróżniając w zbiorze cech następujące grupy: cechy istotne, nieistotne i redundantne. Istotne to takie cechy, które odróżniają od siebie klasy, nieistotne to takie których wartości dla problemu klasyfikacji są przypadkowe, a cechy redundantne to takie których role z powodzeniem mogą przyjąć inne cechy. Metody rankingowe polegają więc na znalezieniu pewnej miary pozwalającej stworzyć taki ranking cech, a potem wybrać najlepsze cechy, a odrzucić najgorsze. Metody rankingowe zazwyczaj są najszybsze i - co istotne - nie zależą one od używanej metody analizy danych. Ich istotną wadą stanowi brak możliwości uwzględnienia zależności pomiędzy cechami. Kolejne opisane typy metod selekcji cech tej wady nie posiadają.

2.3.2 Metody opakowane

Podstawowymi metodami selekcji cech są metody opakowane, tak zwane wrappery. W przeciwieństwie do metod rankingowych, w których selekcja cech i klasyfikator pozostają niezależne, w algorytmach opakowanych selekcji, ocena atrybutów dokonuje się przy użyciu konkretnego modelu. To właśnie efektywność samego klasyfikatora służy za miarę skuteczności metody. Zaletą tej metody jest jej uniwersalność i dokładność, natomiast wadą - wysoka złożoność obliczeniowa. Dla efektywności tych algorytmów istotny jest sposób ustalania podzbioru cech. Wśród wielu metod wyszukiwania tegoż, wyróżnić można najprostszą - przeszukanie całego zbioru podzbiorów. Jest to jednak rozwiązanie bardzo kosztowne. Wobec tego typowymi strategiami są: przeszukiwanie w przód, przeszukiwanie wstecz oraz tworzenie indywidualnego rankingu.

2.3.3 Metody wbudowane

Metody wbudowane zawierają się w algorytmie klasyfikacji i to na etapie tworzenia modelu przypisuje się poszczególnym cechom wagi lub przeprowadza się ich eliminację. Do algorytmów klasyfikacji z wbudowaną metodą selekcji zaliczyć można popularne LASSO i RIDGE. W literaturze natknąć się też można na przypasowanie do tej kategorii metody wektorów nośnych (SVM) czy też analizy składowych głównych (PCA). Zaletą tych metod jest ich szybkość, ponieważ użycie ich nie wiąże się z dodatkowymi operacjami na zbiorze.

2.4 Selekcja cech a ekstrakcja

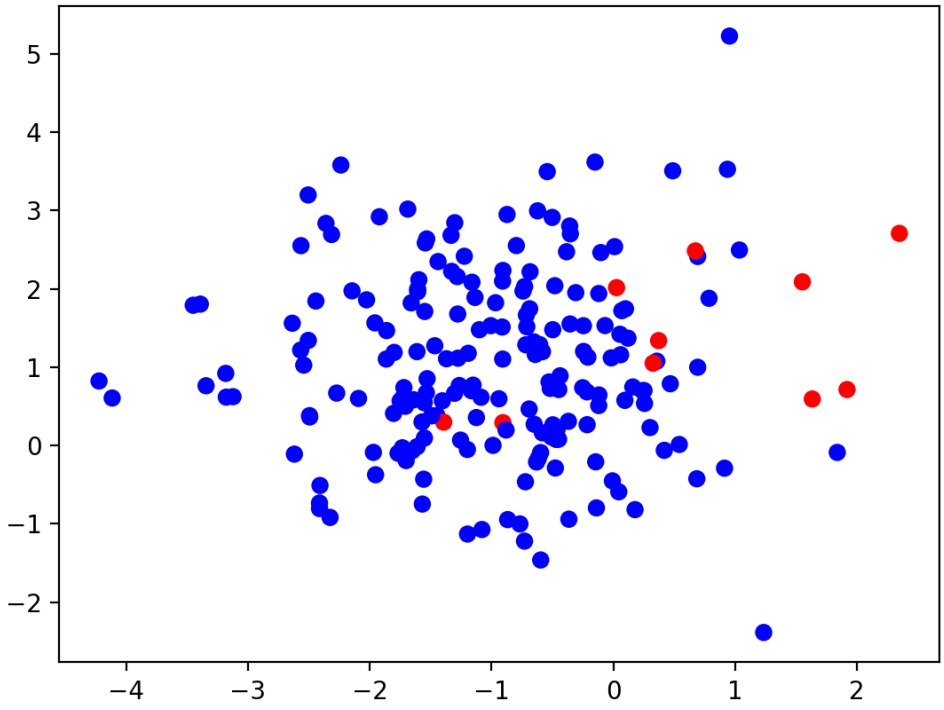
Selekcja cech ma na celu wybranie pewnych atrybutów opisujących dane pod kątem tego, czy nadają się one do dalszego wykorzystania w klasyfikacji, przy jednoczesnym odrzuceniu innych danych. Zawsze rozważana jest ona w kontekście kolejnych zadań i nie można oceniać jej skuteczności w oderwaniu od wyników metody klasyfikacji wykorzystującej wybrane zmienne. W

większości przypadków budowany jest złożony model, który może zawierać jeden lub więcej algorytmów selekcji i co najmniej jeden klasyfikator. Ekstrakcja cech natomiast polega na utworzeniu nowego zestawu zmiennycg poprzez liniową lub nieliniową kombinację oryginalnych danych. W przeciwieństwie do selekcji, gdzie celem jest zawsze uzyskanie podzbioru wszystkich atrybutów, wykorzystanie ekstrakcji wiąże się z wymiarem przestrzennym mniejszym, równy lub nawet większy od wymiaru przestrzeni startowej.

3 Szczegółowa charakterystyka zagadnienia

Wsród wielu dobrze zbadanych i szeroko wykorzystywanych rozwiązań bazujących na uczeniu maszynowym, najbardziej obiecującymi są te, mające ratować ludzkie życie. Złożone choroby, takie jak rak mózgu, stanowią poważne dla niego zagrożenie. Postęp w dziedzinie sztucznej inteligencji i metodach statystycznych stworzyły nowe możliwości klasyfikacji i diagnozy najbardziej śmiertelnych chorób, takich jak rak, choroba Alzheimera, cukrzyca itp. Z przypadkami takimi wiąże się jednak problem niezrównoważonej dystrybucji klas.

Niezrównoważony rozkład klas ma miejsce, gdy co najmniej jedna klasa jest niewystarczająco reprezentowana i przytłoczona przez inne klasy. Model klasyfikacji dla niezrównoważonych danych stwarza wiele przeszkód w uczeniu się algorytmów i przedstawia liczne konsekwencje dla rzeczywistych zastosowań. Ten problem powoduje niedocenianie przykładów klas mniejszościowych i powoduje niedokładne wyniki klasyfikacji w stosunku do przykładów klas większościowych. Klasyfikacja niezrównoważonego zbioru danych staje się trudniejsza przy ograniczonej liczbie próbek i ogromnej liczbie cech. Przykład takiego problemu zaobserwować można na poniższej grafice. Zawiera ona 200 elementów z których tylko 5% należy do klasy mniejszościowej - czerwonej.



Rysunek 1: Przykład niezrównoważonego rozkładu klas

Taka sytuacja jest problemem, ponieważ większość tradycyjnych algorytmów uczenia maszynowego trenowana na podobnym zbiorze, obciążona jest biasem w stosunku do klasy bardziej licznej. Jednocześnie, zazwyczaj lepsze zrozumienie klas mniej licznych jest istotniejsze z punktu widzenia problemu w ujęciu biznesowym. Problemem jest również określenie jakości algorytmu. Jakość klasyfikacji używana jako metryka ewaluacji może być w takim przypadku niewystarczająca, gdyż nawet model o skuteczności 95% - co jest na ogół wartością bardzo dobrą - mógłby nie rozpoznawać żadnego elementu klasy mniejszościowej.

3.1 Metody klasyfikacji danych niezbalansowanych

Problem nierównoważnego rokładu przyciąga w ostatnim czasie zainteresowanie dużej części społeczności zajmującej się uczeniem maszynowym i eksploracją danych, zarówno ze środkowik akademickich jak i w przemyśle co znajduje odbicie w dużej liczbie statupów opierających swoje produkty i usługi na rozwiązaniach *machine-learningowych*. W ciągu kilkunastu ostatnich lat wyklarowały się trzy główne podejścia do uczenia modeli na danych niezbalansowanych.

3.1.1 Metody na poziomie danych

Metody na poziomie danych (Data-level methods), modyfikują dostępne instancje problemu w celu jego zbalansowania. Można je dalej podzielić na podgrupy: metody próbkowania danych (data-sampling) i metody wyboru cech (feature selection methods). Metody nadpróbkowania i podpróbkowania stanowią dwie podgrupy metod próbkowania danych, w których próbkowanie danych z danego zbioru danych odbywa się losowo lub z wykorzystaniem określonego wzoru / algorytmu. W procesie oversamplingu (nadpróbkowania) do danego zbioru danych dodawane są instancje klasy mniejszościowej (poprzez replikację), gdzie replikacja odbywa się losowo lub z wykorzystaniem inteligentnych algorytmów. W procesie undersamplingu natomiast, większość wystąpień klasy zostanie usuniętych z danego zbioru danych, a usuwanie odbywa się w dużej mierze losowo. Chociaż metody selekcji cech są powszechnie stosowane w celu poprawy wyników klasyfikacji, mogą one również pomóc w wyborze najbardziej wpływowych cech w celu wygenerowania unikalnej wiedzy do klasyfikacji. Zmniejsza to niekorzystny wpływ nierównowagi klas na wyniki klasyfikacji.

3.1.2 Metody na poziomie algorytmów

Metody na poziomie algorytmów (Algorithm-level methods), modyfikują istniejące algorytmy uczenia maszynowego. Można dalej podzielić na metody wrażliwe na koszty (cost-sensitive methods) i metody zintegrowane. Pierwsza z nich opiera się na zasadzie przypisywaniu większej wagi instancjom w przypadku błędnej klasyfikacji, na przykład fałszywie negatywnym przewidywaniom można przypisać wyższy koszt niż fałszywie dodatnim przewidywaniom. Metody zintegrowane mogą być również stosowane jako metody wrażliwe na koszty, w przypadku których wynikiem klasyfikacji jest pewna kombinacja wielu klasyfikatorów zbudowanych na zbiorze

danych. Bagging i Boosting to dwa powszechne typy metod uczenia zintegrowanego. Bagging minimalizuje wariancję, generując kilka zestawów uczących z danego zestawu danych i generując klasyfikator dla każdego zestawu uczącego, a następnie łącząc ich odpowiednie modele w celu ostatecznej klasyfikacji. Boosting wykorzystuje również wiele zestawów treningowych z danego zestawu danych, a po iteracyjnym przypisaniu różnych wag do każdego klasyfikatora w oparciu o ich błędne klasyfikacje, łączy je metodą ważenia wyników każdego klasyfikatora, aby uzyskać ostateczną klasyfikację.

3.1.3 Podejścia hybrydowe

Metody hybrydowe mają na celu rozwiązanie znanych problemów spowodowanych metodami próbkowania danych, metodami wyboru cech, metodami wrażliwymi na koszty i podstawowymi algorytmami uczenia się (takimi jak Naive Bayes). W niektórych przypadkach podgrupy metod na poziomie danych lub podgrupy metod na poziomie algorytmu można łączyć jako ogólną metodę rozwiązywania problemu niezbalansowania klas. Na przykład popularny klasyfikator losowego lasu (Random Forest) jest wersją oryginalnego algorytmu losowego lasu decyzyjnego (Random Decision Forest) i jest zintegrowanym algorytmem uczenia się, który dodatko implementuje Bagging.

4 Założenia i plan eksperymentu

Temat projeku zakłada przeprowadzenie szeregu eksperymentów porównujących skuteczność popularnych metod selekcji cech. Praca swoim zakresem obejmie eksperymenty przeprowadzone na kilku, wybranych zbiorach danych. Zakłada się użycie zbiorów naturalnych - to znaczy zebranych w ramach rzeczywistych pomiarów. Bazy danych, użyte w ramach badań implikują skupienie się w większości na problemach wieloklasowych, których elementy są opisywane przez dużą liczbę cech. Hipoteza, z którą twórca będzie konfrontować wyniki eksperymentów, to założeniem że wszystkie, badane metody selekcji poradzą sobie podobnie z postawionym zadaniem, a poza względami wydajnościowymi, nie ma znaczenia funkcja, która zostanie użyta. Technologia, w jakiej zostaną przeprowadzone doświadczenie to jezyk Python w wersji `3.8` oraz biblioteki `sckit-learn` i `pandas`.

4.1 Ocena działania algorytmów

Ważnym elementem badań jest sposób, w jaki weryfikowane są otrzymane wyniki. Sposoby te, różnić się będą w zależności od zastosowanego algorytmu, gdyż dla przykładu, metody opakowane wykonują selekcje cech równolegle z klasyfikacją danych. Dla metod rankingowych i opakowanych klasyfikatorami badającymi skuteczność powziętych działań będzie KNN - Algorytm K Najbliższych Sąsiadów. Pod uwagę brane będą funkcja straty (loss), która informuje o dopasowaniu modelu do danych, oraz dokładności (accuracy), która wylicza skuteczność klasyfikacji. Porównany zostanie również ranking cech uzyskany przez każdą z metod. Bardzo ważnym kryterium będzie czas potrzebny na wykonanie danej operacji. Eksperymenty zostaną uruchomione kilkukrotnie w izolowanym środowisku a czasy potrzebne na pełną klasyfikację uśrednione dla wykluczenia czynników zewnętrznych.

W celu określenia, która z testowanych metod daje najlepsze wyniki klasyfikacji wykorzystany zostanie test statystyczny - test Wilcoxona. Do jego wykonania użyte zostaną wartości dokładności (accuracy) uzyskane dla każdej z badanych metod.

5 Wyniki

6 Wnioski

7 Bibliografia

[1] Feature Selection for High-Dimensional and Imbalanced Biomedical Data Based on Robust Correlation Based Redundancy and Binary Grasshopper Optimization Algorithm, Garba Abdulrauf Sharifai and Zurinahni Zaino; 2020.

[2] A Survey on Evolutionary Computation Approaches to Feature Selection, Bing Xue; Mengjie Zhang; Will N. Browne; Xin Yao, 2015.

[3] Correlation-based Feature Selection for Machine Learning, Mark A Hall, University of Waikato, 1999.

[4] Efektywne metody selekcji cech i rozwiazywania problemu wieloklasowego w nadzorowanej klasyfikacji danych, Wieszław Chmielnicki, IPPT PAN, 2012

[6] Zastosowanie wybranych metod przekształcania i selekcji danych oraz konstrukcji cech w zadaniach klasyfikacji i klasteryzacji, Piotr Płoński, Politechnika Warszawka, 2016.

[7] Analiza i rozwój metod selekcji cech dla dużych problemów klasyfikacyjnych, Jakub Piątkowski, Uniwersytet Mikołaja Kopernika, 2006.

[8] Feature Selection in Imbalance data sets, International Journal of Computer Science Issues, 2013.

8 Zawartość płyty CD

Do pracy dołączono płytę CD o następującej zawartości:

- kod źródłowy programu znajdujący się w folderze `/src`
- gotową, zbudowaną w katalogu `/dist` aplikację
- katalog `/docs` zawierający kod źródłowy tej pracy
- plik w formacie `pdf` zawierający pracę