KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Instrukcja współfinansowana przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego
w projekcie

*„Innowacyjna dydaktyka bez ograniczeń
– zintegrowany rozwój Politechniki Łódzkiej – zarządzanie Uczelnią,
nowoczesna oferta edukacyjna i wzmacniania zdolności
do zatrudniania osób niepełnosprawnych”*

Instrukcja jest dystrybuowana bezpłatnie

Instrukcja do laboratorium, część 5

dr inż. Artur Klepaczko

Eksploracja danych

Zadanie nr 13 – Studia podyplomowe „Przetwarzanie i analiza obrazów biomedycznych”

Politechnika Łódzka
Instytut Elektroniki

90-924 Łódź, ul. Żeromskiego 116,
tel. 042 631 28 83
www.kapitalludzki.p.lodz.pl

Spis treści

1 Wstęp

Ćwiczenia zawarte w niniejszej instrukcji dotyczą zagadnienia selekcji cech. Badane będą algorytmy selekcji typu filtr oraz typu powłoka (ang. wrapper) zarówno w odniesieniu do zbiorów danych pobranych z repozytorium UCI, jak i zbiorów tekstur. Wykonanie selekcji umożliwia przede wszystkim zakładka *Select attributes* modułu *Explorer* (zob. pkt 2.2). Ponadto dostęp do algorytmów selekcji możliwy jest za pośrednictwem filtru wstępnego przetwarzania danych o nazwie *AttributeSelection*, którego konfigurację oraz uruchomienie umożliwia zakładka *Preprocess* (także moduł *Explorer* zob. pkt 2.1).

Zasadniczym elementem dowolnej procedury selekcji cech jest miara, według której dokonuje się oceny przydatności poszczególnych atrybutów (bądź ich podzbiorów) do klasyfikacji wektorów danych. Wydaje się, iż naturalną funkcją oceny istotności cech jest błąd klasyfikacji. Istotnie, metoda ta jest bardzo popularna. Jednakże podprzestrzeń cech wyznaczona w ten sposób niekoniecznie zapewnia dobrą jakość klasyfikacji dla różnych algorytmów uczenia. Dlatego stosuje się również inne miary i będą one przedmiotem analizy obecnych ćwiczeń laboratoryjnych.

Oczekiwane efekty kształcenia:

→ Znajomość podstawowych strategii selekcji cech typu filtr oraz typu powłoka.

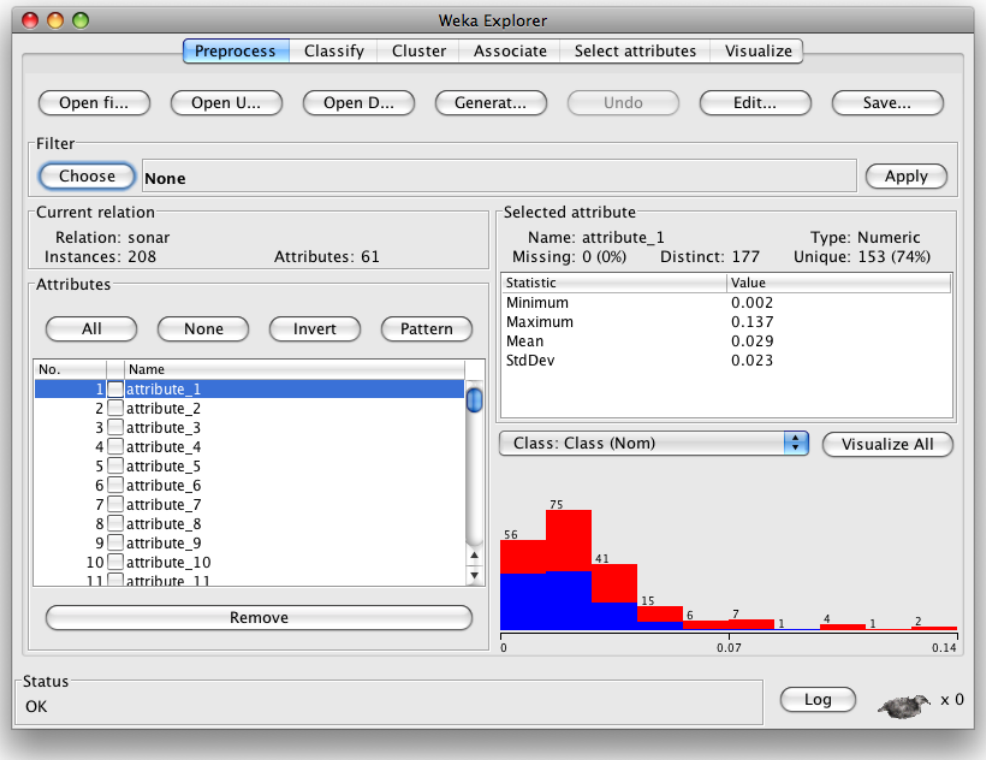
→ Umiejętność praktycznego zastosowania metod selekcji cech zaimplementowanych w programie Weka.

2 Selekcja cech w programie Weka

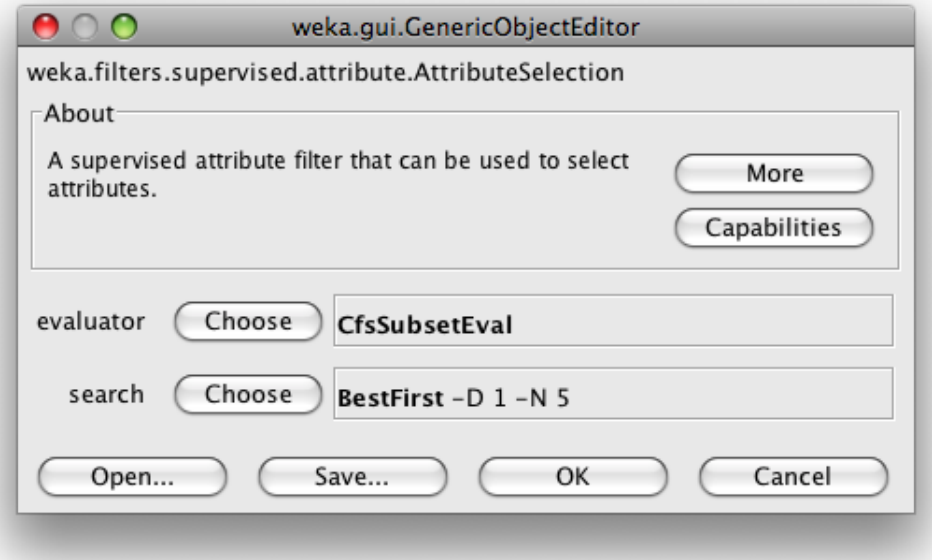
Zadanie selekcji cech w programie Weka można zrealizować na dwa sposoby. W przypadku, gdy selekcja cech ma stranowić pierwszy etap przetwarzania wektorów danych, po którym nastąpi np. ich klasyfikacja, należy skorzystać z filtru *Attribute selection* wywoływanego w zakładce *Preprocess*. Jeśli interesuje nas sam wynik selekcji (podzbiór cech znaczących) bez jego konkretnych zastosowań w innych zadaniach eksploracji danych, wówczas procedurę selekcji można uruchomić za pośrednictwem interfejsu dostępnego w zakładce *Select attributes*.

2.1 Selekcja cech jako etap wstępnego przetwarzania danych

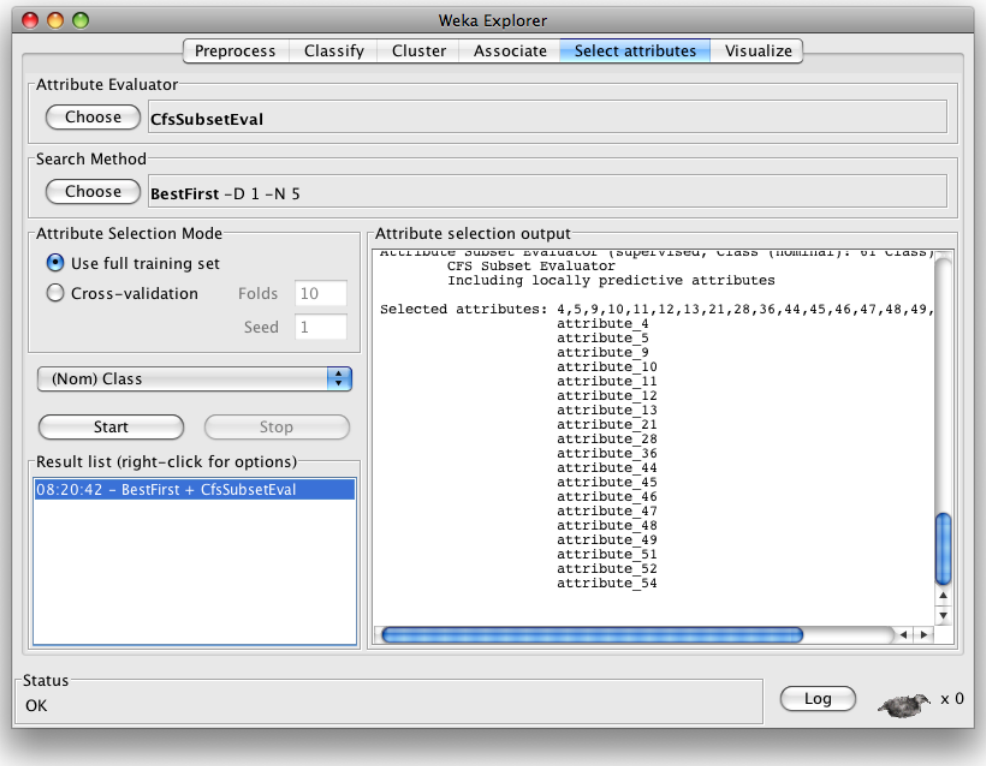
Wstępne przetwarzanie wektorów danych odbywa się za pośrednictwem sekcji *Filter*, znajdującej się w górnej części zakładki *Preprocess* (zob. rys. 1. Przycisk *Choose* służy do wyboru określonej metody przetwarzania, natomiast naciśnięcie przycisku *Apply* powoduje jej wykonanie. Filtr *Attribute selection* znajduje się w grupie *supervised/attribute*. Podobnie jak algorytmy klasyfikacji czy grupowania, filtry udostępniają szereg parametrów pozwalających szczegółowo skonfigurować ich zachowanie. W przypadku filtru selekcji cech, istotne są dwa parametry — *evaluator* oraz *search* (rys. 2). Pierwszy z nich służy do wskazania miary oceny istotności cech, drugi natomiast determinuje sposób przeszukiwania przestrzeni cech.



Rysunek 1: Widok zakładki *Preprocess*



Rysunek 2: Parametry filtru *AttributeSelection*



Rysunek 3: Widok zakładki *Select attributes*

W przypadku strategii typu filtr jedyną możliwą metodą przeszukiwania (parametr *search*) jest algorytm *Ranker*, który szereguje cechy zgodnie z miarą wskazaną w parametrze *evaluator*. Metoda ta przy domyślnych ustawieniach nie usuwa żadnych elementów z oryginalnego zbioru cech. Aby wymusić redukcję wymiaru przestrzeni atrybutów należy odpowiednio zmodyfikować jeden z parametrów algorytmu *Ranker* — *numToSelect* (wymagana liczba cech) lub *threshold* (wartość graniczna funkcji oceny). Z kolei parametr *evaluator* powinien wskazywać algorytm przeznaczony do oceny pojedynczych atrybutów. Algorytmy te wyróżnia ostatni człon ich nazwy, tzn. *AttributeEval*. Przykładowymi miaramia mogą tu być *ChiSquaredAttributeEval*, *GainRatioAttributeEval* lub *OneRAttributeEval*.

Dla metod typu powłoka zdefiniowano szereg metod przeszukiwania, m.in. *BestFirst* — przeszukiwanie sekwencyjne w przód, *LinearForwardSelection* — przeszukiwanie sekwencyjne z korektą, *RankSearch* — przeszukiwanie sekwencyjne w przód z wstępnym uszeregowaniem cech w rankingu lub *GeneticSearch* — przeszukiwanie za pomocą algorytmu genetycznego. Natomiast funkcje oceny posiadają nazwy, których ostatni człon nazwy zawiera wyrażenie *SubsetEval*, np. *CfsSubsetEval*, *ConsistencySubsetEval* lub *WrapperSubsetEval*.

2.2 Interfejs selekcji cech modułu *Explorer*

Interfejs selekcji cech modułu *Explorer* (rys. 3) posiada interfejs zbliżony wyglądem do poznanych wcześniej interfejsów klasyfikacji oraz klasteryzacji. Przy czym, tak jak w przypadku filtru omówionego wyżej, w górnej części zakładki *Select attrubutes* znajdują się dwie sekcje. W nich należy wskazać (przyciski *Choose*) metodę przeszukiwania przestrzeni atrybutów oraz miarę oceny istotności cech. Zasady wybóru określonych metod w zależności od strategii typu filtr albo powłoka pozostają takie same, jak w przypadku filtru *Attribute selection*.

Ponadto, możliwe jest wykonanie selekcji jednokrornie dla całego zbioru danych treningowych lub z zastosowaniem walidacji krzyżowej (sekcja *Attribute Selection Mode*). Uruchomienie procedury selekcji następuje po naciśnięciu przycisku *start*. Kolejne wyniki obliczeń zapamiętywane są w oddzielnych raportach, do których odwołania znajdują się w panelu po lewej stronie u dołu okna.

3 Zadania do rozwiązania

Uwaga! Zbiory danych, których dotyczą kolejne ćwiczenia umieszczone są na serwerze pod adresem `http://eletel.p.lodz.pl/aklepaczko/open/`.

3.1 Selekcja cech za pomocą filtru *AttributeSelection*

Zadanie 1

- Uruchom program Weka oraz moduł *Explorera*. Pobierz z serwera plik `wine.arff` i wczytaj go do programu.
- Przejdź do zakładki *Classify* i przeprowadź klasyfikację wczytanego zbioru danych za pomocą algorytmu *Logistic* znajdujący się w grupie klasyfikatorów *functions*. Ile wynosi oszacowanie błędu rzeczywistego klasyfikacji przy użyciu metody 10-krotnej walidacji krzyżowej?
- Powróć do zakładki Preprocess. Za pomocą przycisku *Choose* w sekcji *Filter* wybierz filtr *AttributeSelection*. Skonfiguruj filtr tak, aby algorytmem przeszukującym przestrzeń cech (parametr *search* filtru) była metoda *Ranker*, natomiat jakość cech była mierzona za pomocą współczynnika zysku informacyjnego *InfoGainAttributeEval* (parametr *evaluator*). Ustaw wymiar docelowej przestrzeni cech na wartość 5 (parametr *numToSelect* algrytmu *Ranker*. Następnie naciśnij przycisk *Apply*, aby przefiltrować dane. Ponownie wykonaj klasyfikację.
- Powtórz doświadczenia dla dwóch innych metod oceny cech — *OneRAttributeEval* oraz *ReliefFAttributeEval*. Porównaj błędy klasyfikacji uzyskane w poszczególnych przypadkach.

Zadanie 2

- Pobierz i wczytaj do modułu *Explorora* zbiór danych `sonar.arff`. Przed wykonaniem selekcji cech wykonaj klasyfikację za pomocą algorytmów *Logistic* oraz *IB1* (z grupy klasyfikatorów *lazy*). Zanotuj uzyskane błędy klasyfikacji.
- Przejdź do zakładki *Preprocess*. Ponownie wybierz filtr *AttributeSelection*, tym razem jednak jako metodę przeszukiwania przestrzeni cech wskaż algorytm *RankSearch*. w pierwszym etapie selekcji wykonywany jest tu ranking cech za pomocą wybranej miary oceny pojedynczych atrybutów (parametr wywołania *attributeEvaluator* algorytmu *RankSearch*). Następnie poprzez dobór kolejnych cech z rankingu tworzony jest szereg podzbiorów o wzrastającej liczności, poddawanych ocenie za pomocą miary wskazanej w parametrze *evaluator* filtru.
- Spośród wymienionych niżej miar oceny pojedynczych atrybutów oraz podzbiorów cech znajdź kombinację dającą najmniejszy błąd klasyfikatorów *Logistic* oraz *IB1*.

  Miary oceny pojedynczych cech:

  – *ChiSquaredAttributeEval*,
  – *GainRatio*,
  – *OneRAttributeEval*,
  – *ReliefFAttributeEval*.

  Miary oceny podzbiorów cech:

  – *CfsSubsetEval*,
  – *WrapperSubsetEval* z klasyfikatorem *IB1*,
  – *WrapperSubsetEval* z klasyfikatorem *Logistic*.

3.2 Seleckcja cech za pomocą interfejsu *Select attributes*

Zadanie 3

- Pobierz i wczytaj zbiór danych `GR_HW_RA_16_all.csv`.
- Wykonaj klasyfikację wczytanego zbioru dla wszystkich cech przy użyciu algorytmu *IB1*.
- Przejdź do zakładki *Select attributes*. w sekcji *Attribute Evaluator* wybierz algorytm oceny podzbiorów cech *WrapperSubsetEval* z klasyfikatorem IB1. Upewnij się, że w sekcji *Search Method* jest wybrana metoda BestFirst. Uruchom procedurę selekcji cech za pomocą przycisku *Start*.

- Przejdź do zakładki *Preprocess*. Wybierz filtr *Remove* znajdujący się w grupie *unsupervised/attribute*. w opcjach wywołania filtru zmień wartość parametru *invertSelection* na *True*, natomiast w polu *attributeIndices* wpisz indeksy cech wyselekcjonowanych w poprzednim kroku dodając na końcu argument *last* oznaczający w tym przypadku atrybut klasy. Naciśnij przycisk *Apply*. Przejdź do zakładki *Classify* i przeprowadź klsyfikację za pomocą algorytmu *IB1*.
- Powróć do zakładki *Preprocess* i anuluj wykonaną selekcję cech naciskając przycisk *Undo*.
- Wykonaj standaryzację danych za pomocą filtru *Standardize*, który także znajduje się w grupie *unsupervised/attribute*. Powtórz selekcję cech oraz klasyfikację za pomocą tych samych algorytmów, co poprzednio.
- Zmień metodą przeszukiwania przestrzeni cech na algorytm *LinearForwardSelection*. w jego opcjach wywołania ustaw parametr *forwardSelectionMethod* na wartość *Floating forward selection*, parametr *numUsedAttributes* na wartość 250, zaś parametr *performRankig* na wartość *False*. Wykonaj selekcję cech oraz przeprowadź klasyfikację (skorzystaj ponownie z filtru *Remove*).
- Anuluj wykonaną selekcję cech naciskając przycisk *Undo* w zakładce *Preprocess*.
- Powtórz selekcję tym razem wykonując najpierw ranking cech (parametr *performRanking* ustawiony na *True*) i ograniczając przeszukiwanie do przestrzeni 100 najlepszych atrybutów (parametr *numUsedAttributes*).
- Który schemat selekcji okazał się najlepszy biorąc pod uwagę dokładność klasyfikacji? Czy selekcja cech pozwoliła zmniejszyć błąd klasyfikacji w stosunku do wyniku uzyskanego dla oryginalnego zbioru wektorów danych?

Zadanie 4

- Pobierz i wczytaj zbiór danych `GR_HW_RA_16_rand.csv`.
- Wykonaj klasyfikację wczytanego zbioru dla wszystkich cech przy użyciu algorytmu *IB1*.
- Wykonaj selekcję cech dwukrotnie, raz z użyciem strategii przeszukiwania *GreedyStepwise*, a następnie za pomocą metody *LinearForwardSelection*. w obu przypadkach zastosuj algorytm *WrapperSubsetEval* z klasyfikatorem *IB1* do oceny istotności cech.
- Przeprowadź standaryzację cech i powtórz ich selekcję.
- Dla każdego uzyskanego podzbioru cech wykonaj klasyfikację, zarówno dla danych w oryginalnej skali, jak i w przestrzeni ustandaryzowanej. Czy standaryzacja danych pozwoliła w tym przypadku uzyskać lepsze przestrzenie atrybutów?

Zadanie 5

- Pobierz i wczytaj zbiór danych `BA_HW_PS_RA_20_all.csv`.
- Wykonaj klasyfikację wczytanego zbioru dla wszystkich cech przy użyciu algorytmu *IB1*.
- Przejdź do zakładki *Preprocess*. Wybierz filtr *Resample* znajdujący się w grupie *supervised/instance*. w opcjach wywołania filtru zmień wartość parametrów *invertSelection* oraz *noReplacement* na *True*, natomiast parametru *sampleSizePercent* na wartość 20. Naciśnij przycisk *Apply*. Zapisz uzyskany zbiór danych na dysku lokalnym naciskając przycisk *Save...*. Anuluj dokoną filtrację (przycisk *Undo*). Zmień parametr *invertSelection* filtru na wartość *False* i ponownie naciśnij przycisk *Apply*.
- Przejdź do zakładki *Classify*. Zaznacz opcję testowania klasyfikatora za pomocą niezależnego zbioru danych testowych i za pomocą przycisku *Set...* wskaż zapisany poprzednio zbiór danych. Wykonaj klasyfikację z użyciem algorytmu *IB1*.
- Przeprowadź selekcję cech dwukrotnie, w obu przypadkach z użyciem strategii przeszukiwania *LinearForwardSelection* oraz algorytmem oceny istotności cech typu *WrapperSubsetEval*. w pierwszym przypadku jednak zastosuj klasyfikator *IB1*, w drugim zaś *Logistic*.
- Przeprowadź klasyfikację zbioru danych za pomocą algorytmów *IB1*, *Logistic*, a także *SMO* dla każdego uzyskanego podzbioru cech. Porównaj błędy klasyfikacji uzyskane w każdym przypadku.

Zadanie 6

- Powróć do zakładki *Preprocess* i anuluj ostatnią wykonaną selekcję cech naciskając przycisk *Undo*.
- Wybierz filtr *RandomSubset* znajdujący się w grupie *unsupervised/attribute*. w opcjach wywołania filtru zmień wartość parametru *numAttributes* na 0.2. Przeprowadź filtrację zbioru danych.
- Wykonaj klasyfikację dla wszystkich cech przy użyciu algorytmu *IB1*.
- Wykonaj selekcję cech dwukrotnie, raz z użyciem strategii przeszukiwania *GeneticAlgorithm*, a następnie za pomocą metody *LinearForwardSelection*. w obu przypadkach zastosuj algorytm *WrapperSubsetEval* z klasyfikatorem *IB1* do oceny istotności cech. Przeprowadź klasyfikację dla uzyskanych podzbiorów atrybutów.
- Jak oceniasz zdolność prostego algorytmu genetycznego do szybkiego wyznaczania podzbiorów cech znaczących?

Łódź 2009

Zredagowane w LaTeX-u