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 4

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 **3**

2 Analiza skupień w programie Weka **3**
2.1 Interfejs klasteryzacji modułu *Explorer* 3
2.2 Ocena jakości grupowania 4

3 Zadania do rozwiązania **5**
3.1 Grupowanie hierarchiczne 5
3.2 Grupowanie kombinatoryczne i probabilistyczne 6

1 Wstęp

Niniejsza instrukcja dotyczy zagadnienia analizy skupień (inaczej grupowania lub klasteryzacji). Uczenie w tym przypadku odbywa się w sposób nienadzorowany czyli bez wykorzystywania informacji a priori o przynależności wektorów danych treningowych do określonych kategorii. Celem takiej analizy jest wykrycie naturalnej struktury zbioru danych złożonego z odrębnych podgrup, wewnątrz których wektory cech wykazują duże podobieństwo względem siebie. Zaletą uczenia nienadzorowanego jest obiektywizm wnioskowania. Algorytm dokonujący podziału zbioru danych wykorzystuje jedynie miarę podobieństwa wektorów (najczęściej odległość euklidesową), abstrahuje on jednocześnie od subiektywnego oznaczania klas.

Istotnym problemem przy dokonywaniu nienadzorowanej klasyfikacji danych jest określenie prawidłowej liczby skupień. Informacja ta, podobnie jak wektor etykiet kategorii, albo nie jest dostępna a priori, albo — dla zachowania obiektywizmu wnioskowania — musi pozostać niejawna w trakcie procesu uczenia. Jednym z możliwych rozwiązań jest wykonanie klasteryzacji w sposób interaktywny, gdzie użytkownik zapoznaje się z częściowym wynikiem analizy (np. drzewem hierarchicznym), a następnie podejmuje decyzję o liczbie klasterów. Alternatywnie, procedura uczenia wykonywana jest kilkakrotnie, za każdym razem z założeniem innej liczby skupień. Poszczególne wyniki grupowania zostają poddawane ocenie przy użyciu wybranej miary jakości (np. kryterium MDL). Ostateczna liczba klasterów wynika z najlepiej ocenionego podziału zbioru danych.

Oczekiwane efekty kształcenia:

→ Znajomość wybranych algorytmów analizy skupień zaimplementowanych w programie Weka.

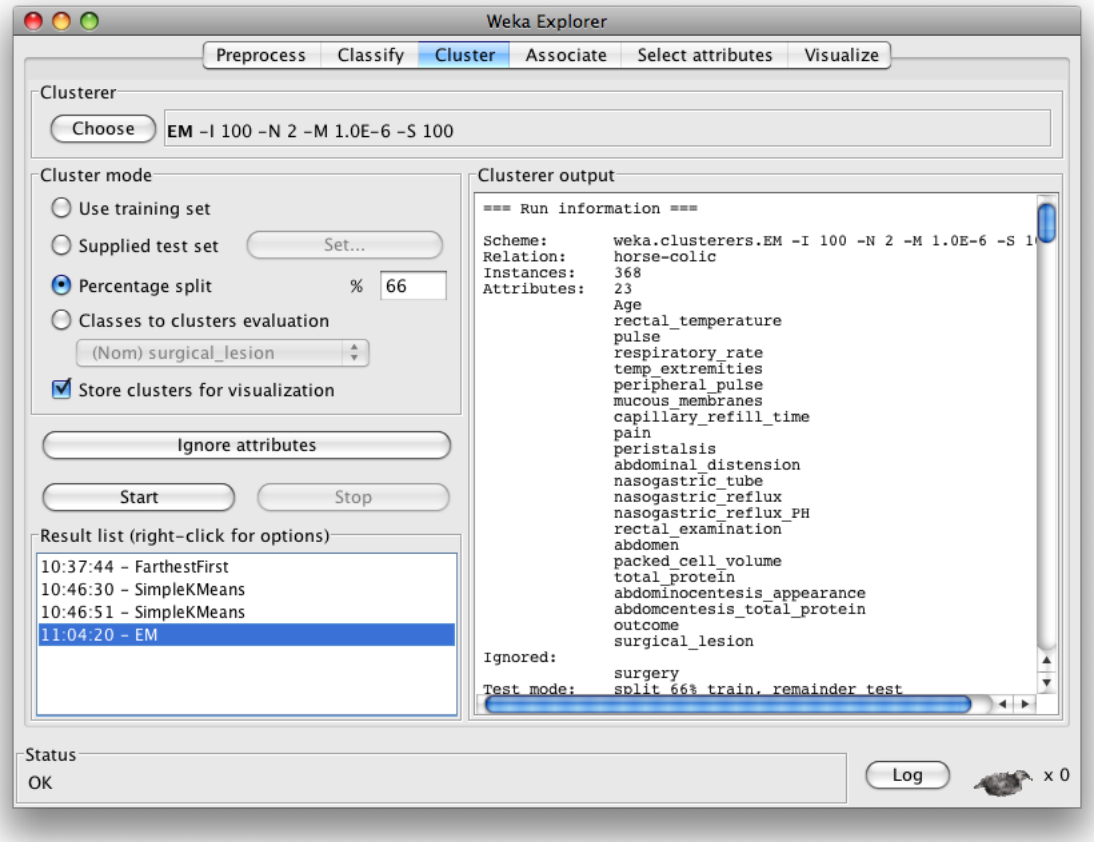
→ Umiejętność automatycznego i półautomatycznego wyznaczania liczby skupień.

2 Analiza skupień w programie Weka

2.1 Interfejs klasteryzacji modułu *Explorer*

Program Weka udostępnia szereg algorytmów grupowania wektorów danych — zarówno podstawowych (np. algorytm k–średnich), jak i bardziej zaawansowanych, zorientowanych na analizę skupień w zasobach dużych baz danych (takich jak np. algorytmy DBScan lub OPTICS). Wykonanie klasteryzacji przeprowadza się za pośrednictwem zakładki *Cluster* (zob. rys. 1) modułu *Explorer*.

Podobnie jak w przypadku klasyfikacji nadzorowanej wybór konkretnej metody grupowania umożliwia przycisk *Choose*. Jego naciśnięcie powoduje wyświetlenie okna pokazanego na rys. 2. Testowanie wyuczonego modelu klasyfikatora (panel *Cluster mode*) można przeprowadzić na zbiorze uczącym (opcja *Use training set*), odrębnym lub wydzielonym ze zbioru treningowego zbiorze testowym (odpowiednio opcje *Supplied test set* lub *Percentage split*), albo poprzez porównanie wyniku klasteryzacji z sugerowanym wektorem etykiet kategorii,



Rysunek 1: Interfejs klasteryzacji w module *Weka Explorer*

o ile taki jest zawarty w zbiorze danych (opcja *Classes to clusters evaluation*). Uruchomienie procedury grupowania odbywa się poprzez naciśnięcie przycisku *Start*.

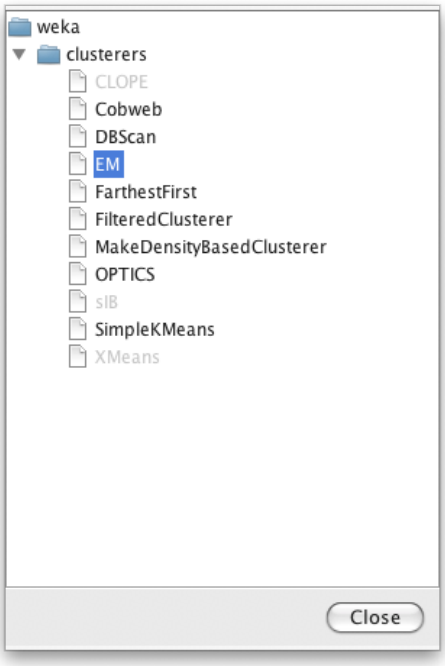
2.2 Ocena jakości grupowania

Brak etykiet kategorii wektorów danych uniemożliwia ocenę klasyfikatora na podstawie błędu klasyfikacji. Konieczne jest zatem zastosowanie innej miary pozwalającej oszacować wiarygodność dokonanego podziału zbioru danych na klastery. Często stosowanym rozwiązaniem jest kryterium MDL, którego postać określa zasada minimalnej długości opisu (ang. *minimum description length*):

$$MDL(\Theta) = -\ell(\mathbf{X}|\Theta) + \frac{n_p}{2},$$

gdzie $\mathbf{X}$ oznacza zbiór wektorów danych, Θ — zbiór parametrów modelu klasyfikatora (takich jak centra skupień oraz odchylenia standardowe wektorów znajdujących się wewnątrz nich), ℓ — logarytmiczne prawdopodobieństwo a posteriori zbioru danych (ang. *log likelihood*), natomiast n_p — liczbę parametrów modelu (zależną m.in od liczby skupień).

Należy zawuważyć, iż tylko niektóre algorytmy grupowania prowadzą do wyznaczenia wszystkich elementów zbioru Θ, niezbędnych do obliczenia kryterium MDL. Dzieje się tak w przypadku algorytmów, w których przyjmuje się założenie obowiązywania probabilistycznego modelu klasterów. Przykładem metody, której wyniki wprost nadają się do oceny za pomocą miary MDL, jest algorytm grupowania *EM*. Inne popularne podejście do klasteryzacji — algorytm k–średnich — pozwala uzyskać jedynie część parameterów zbioru Θ (położenia centrów skupień). Jednakże Weka zawiera mechanizm, który przekształca rezultaty dowolnego algorytmu grupowania do postaci, w której obliczenie kryterium MDL staje się możliwe. W tym celu z listy widocznej na rys. 2 należy wybrać metodę *MakeDensityBasedClusterer*, a następnie w jej opcjach wywołania wskazać właściwy algorytm analizy skupień. Wówczas wynik klasteryzacji będzie obejmował również wartość logarytmicznego prawdopodobieństwa danych ℓ. Wyznaczenie ostatecznej wartości kryterium MDL (uwzględniającej składnik zależny od liczby parametrów modelu) pozostaje już jednak w gestii użytkownika.



Rysunek 2: Algorytmy klasteryzacji dostępne w programie Weka

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 Grupowanie hierarchiczne

Zadanie 1

- Pobierz z serwera zdalnego zbiór danych `zoo.arff` i otwórz go w programie Weka.
- Za pomocą przycisku *Choose* w zakładce *Cluster* wybierz algorytm grupowania hierarchicznego *Cobweb*.
- Za pomocą przycisku *Ignore attributes* wskaż cechy, które powinny być pominięte przy obliczaniu podobieństwa między wektorami danych.
- Wybierz tryb testowania na zbiorze uczącym zaznaczając opcję *Use training set*. Wykonaj procedurę grupowania. Jaki jest rozmiar utworzonego drzewa oraz ile skupień zostało wyróżnionych?

- Otwórz okno dialogowe opcji wywołania algorytmu *Cobweb*. Ustaw zmienną *saveInstanceData* na *True*, a następnie znajdź taką wartość parametru odcięcia *cutoff*, aby liczba wydzielonych klasterów wyniosła 7.
- Kliknij prawym klawiszem myszy na liście wyników klasteryzacji (znajdującej się w lewym dolnym rogu okna *Explorera*) na pozycji skojarzonej z podziałem zbioru danych na 7 skupień. W menu kontekstowym wybierz polecenie *Visualize tree* i przeanalizuj strukturę utworzonego drzewa hierarchicznego.

Zadanie 2

- Wczytaj zbiór danych `iris.arff`. Ile wynosi liczba klasterów wydzielonych przez algorytm *Cobweb* przy aktualnych wartościach parametrów wywołania?
- Zmodyfikuj wartości parametrów odcięcia oraz precyzji tak, aby liczba klasterów wynosiła 3.
- Ustaw tryb testowania klasyfikatora poprzez porównanie wyniku klasteryzacji z wektorem etykiet kategorii zaznaczając opcję *Classes to clusters evaluation*. Ponownie wykonaj grupowanie. Ile wynosi błąd klasyfikacji?
- Wywołaj menu kontekstowe skojarzone z ostatnim wynikiem grupowania i wybierz polecenie *Visualize cluster assignments*. Przeanalizuj uzyskany wykres rozproszenia wektorów danych.

3.2 Grupowanie kombinatoryczne i probabilistyczne

Zadanie 3

- Pobierz plik ze zbiorem danych `breast-w-noclass.arff`, z którego usunięto wektor etykiet kategorii. Za pomocą przycisku *Choose* wybierz algorytm grupowania *EM*. Zakładając, że aktywne są ustawienie domyślne, algorytm ten próbuje automatycznie wyznaczać liczbę klasterów (parametr wywołania *numClusters* jest równy -1). Wykonaj klasteryzację przy zaznaczonej opcji testowania na zbiorze uczącym. Ile wynosi logarytmiczne prawdopodobieństwo utworzonego modelu? Ile skupień zostało wydzielonych w zbiorze danych i jaka jest struktura poszczególnych grup wektorów?
- Przejdź do zakładki *Visualize*. Przesuń suwak *Jitter* maksymalnie w prawo i naciśnij przycisk *Update*. Ile skupień można wyróżnić na podstawie wzrokowej analizy poszczególnych wykresów rozproszenia wektorów danych?
- Wczytaj plik `breast-w.arff` o podobnej zawartości co poprzednio pobrany zbiór danych, tym razem jednak uzupełniony o etykiety kategorii. Zaznacz opcję testowania *Classes to clusters evaluation*. W opcjach wywołania algorytmu *EM* ustaw parametr *numClusters* na spodziewaną wartość liczby klasterów. Wykonaj grupowanie. Jaki jest błąd klasyfikacji i jaka jest jego struktura?

Zadanie 4

- Pobierz plik `glass.arff`. Za pomocą przycisku *Choose* wybierz algorytm *MakeDensityBasedClusterer*, a następnie owtórz okno dialogowe opcji jego wywołania. Wskaż algorytm *SimpleKMeans* jako właściwą metodę grupowania. Wybierz tryb testowania na zbiorze uczącym. Wykonaj klasteryzację dla następujących założeń dotyczących liczby klasterów (parametr *numClusters* w metodzie *SimpleKMeans*): 2, 3, 4, 5.
- Oblicz kryterium MDL dla każdego przypadku. Jaka liczba klasterów jest najbardziej prawdopodobna według przyjętej miary jakości grupowania?

Zadanie 5

- Powtórz ćwiczenia z poprzedniego zadania dla zbioru danych `breast-w.arff`.

Zadanie 6

- Pobierz zbiór danych `Dermatology.arff`. Porównaj efektywność algorytmów *SimpleKMeans* oraz *EM* w zastosowaniu do zbiorów *Dermatology* i *Iris*. Załóż, że prawdziwa liczba skupień jest znana.

Zadanie 7

- Pobierz zbiór danych `Distances.cvs`. Wybierz algorytm grupowania *SimpleKMeans*. W opcjach wywołania algorytmu zmień parametr funkcji odległości *dontNormalize* na wartość *True*. Zaznacz opcję testowania *Classes to clusters evaluation* i wykonaj grupowanie.
- Zmień miarę odległości na funkcję *Manhattan* (suma wartości bezwględnych różnic atrybutów). Podobnie jak w poprzednim przypadku, zablokuj normalizację przestrzeni atrybutów. Wykonaj grupowanie.
- Powtórz procedurę grupowania dla obu algorytmów, tym razem z parametrem *dontNormalize* ustawionym na wartość *False*.
- Porównaj błędy klasyfikacji uzyskane w poszczególnych doświadczeniach.

Łódź 2009

Zredagowane w LaTeX-u