Lekcja 9: Indukcja drzew decyzyjnych

S. Hoa Nguyen

1 Materiał

a) Algorytm indukcji drzewa decyzyjnych *Buduj-drzewo (T,D)*

- **Krok 1**: Jeśli *Kryterium-stopu (T,D)* to
  utwórz liść l;
  wyznacz $l.kategoria$; $D := l$;
- **Krok 2**: Wyznacz najlepszy test t;
- **Krok 3**: Niech t dzieli zbiór $T = T_1 \cup T_2 \cup ... \cup T_k$;
- **Krok 4**: Dla $i = 1...k$ {
  Buduj-drzewo (T_i,D_i);
  $D.syn_i := D_i$}

b) Kryterium stopu i ustalenie etykiet

c) Rodzaje testów

- Dla atrybutów symbolicznych: *testy tożsamościowe*, *testy równościowe*
- Dla atrybutów ciągłych i porządkowych: *testy nierównościowe*

d) Kryterium wyboru testu

- *Przyrost informacji (Entropia)*

e) Kryterium przycinania drzewa

- *Przycinanie podczas tworzenia drzewa (Pre-pruning)*
- *Przycinanie po utworzeniu drzewa (Post-pruning)*

f) Znane algorytmy indukcji drzew decyzyjnych *ID3 (dla atrybutów symbolicznych)* i *C45 (dla atrybutów mieszanych)*

2 Zadania podstawowe

Zadanie 1. W tablicy danych *Federer-Nadal-Results.xls* są wyniki pojedynek między dwoma czołowymi tenisistami świata. Zastosować drzewo decyzyjne do przewidywania wyniku meczu z następującymi parametrami [*evening, master, hard*].

a) Proponować formę testu dla atrybutów.

b) Wyznaczyć dla każdego atrybutu najlepszy test, zakładając, że rodzaj testu jest *tożsamościowy* i *Entropia* jest stosowana jako funkcja oceniająca jakości testu. A potem wyznaczyć najlepszy podział (test w korzeniu drzewa decyzyjnego).

c) Przeprowadzić zbiór danych do odpowiedniego formatu systemu Weka, skonstruować drzewo decyzyjne i skorzystać tego drzewa do przewidywania wyniku meczu z następującymi parametrami [*noc, master, hard*].

Zadanie 2. **Generowanie drzewa decyzyjnego**
W systemie Weka otwórz plik o nazwie *weather.arff*. Wygeneruj drzewo decyzyjne dla standardowych wartościach parametrów. Dokonaj analizy struktury wygenerowanego drzewa. Odpowiedz na pytania:

a) Jaka jest struktura drzewa? Liczba węzłów?, Liczba liści?, Ile jest możliwych ścieżek „decyzyjnych" wychodzących z korzenia drzewa? Jak wygląda zestaw warunków z najdłuższej ścieżki?

b) Czy mechanizm przycinania drzewa (ang. *pruning*) dokonał jakichkolwiek modyfikacji struktury drzewa

c) Jakie są wyniki klasyfikowania obiektów za pomocą drzewa? Jak odczytać poziom błędów z macierzy błędów (ang. *confusion matrix*)?

Zadanie 3. Klasyfikowanie nowych obiektów.
Dla drzewa wygenerowanego w zadaniu 2 dokonaj klasyfikowania nowych obiektów.

a) Dokonać klasyfikacji przykładów z niekompletnym opisem oraz później przykładów, dla których wartości atrybutów są nieprecyzyjne. Mogą to być przykłady charakteryzujące się następującym opisem:

x	**Outlook**	**Temprature**	**Humidity**	**Windy**
1	*overcast*	75	85	*yes*
2	*sunny*	–	–	*no*
3	*sunny* : 0.7			*tak* : 0.9
	overcast : 0.2	$75 - 80$	$80 - 85$	*nie* : 0.1
	rainy : 0.1			
4	*sunny* : 0.8			*tak* : 0.7
	overcast : 0.1	$80 - 85$	*brak*	*nie* : 0.3
	rainy : 0.1			

Zadanie 4. Poszukiwanie właściwego stopnia uproszczenia drzew klasyfikujących (2 punkty)
Celem zadania jest sprawdzenie, w jakim stopniu parametr sterujący przycinanie drzewa w algorytmie C4.5 wpływa na jego zdolności klasyfikacyjne. Ocena skuteczności klasyfikowania powinna być dokonywana za pomocą opcji walidacji krzyżowej (*10-fold cross validation*). Zaleca się wykonanie wykresów ilustrujących podstawowe zależności między badanymi parametrami. Do analizy wybierzemy plik *cars.arff*.

a) Przeprowadzić serię eksperymentów oceny drzew decyzyjnych wygenerowanych systemem C4.5 zmieniając wartość parametru *confidence factor* od 0.1 do 0.8 z krokiem co 0.1 i sporządzić wykres zależności pomiędzy wartością zmienianego parametru a średnią trafnością (lub błędem)) klasyfikowania drzew pełnego i uproszczonego na zbiorze testującym

b) Wykonaj także wykres ilustrujący zależność średniego błędu klasyfikacji w zależności od średniego rozmiaru drzewa.

c) Przeprowadzić serię eksperymentów oceny skuteczności klasyfikacyjnej drzew decyzyjnych zmieniając w systemie C4.5 wartość parametru *Prepruning* (ograniczającym minimalną liczność przykładów w węźle) od 1 do 5 z krokiem co 1 i sporządzić wykres zależności pomiędzy wartością zmienianego parametru a średnim rozmiarem drzewa uproszczonego, średnią trafności (błędem) klasyfikowania drzewa uproszczonego na zbiorze testującym. Oceń, jak zmienia się wartość błędu klasyfikacji w zależności od zmiany tego parametru. Czy drzewo uproszczone powyższą techniką jest skuteczniejszym klasyfikatorem niż pełne drzewo?