„Indukcja drzew z danych z wykorzystaniem algorytmu C4.5”

Przewodnik do ćwiczenia nr 2 dla studentów PiESI w ramach przedmiotu „Analiza i Eksploracja Danych” (październik 2003)

Autor: *Jerzy Stefanowski* Instytut Informatyki, Politechnika Poznańska.

Zadanie 1. Generowanie drzewa decyzyjnego.

Zapoznaj się z instrukcją obsługi programu *C*4.5 (menu programu - pozycja *Help*). W programie otwórz plik o nazwie *golf*. Z wykorzystaniem edytora tekstowego obejrzyj reprezentacje plików *golf.nam* oraz *golf.dat*. Wygeneruj drzewo decyzyjne (menu *Tree/Generate*) dla standardowych wartościach parametrów. Dokonaj analizy struktury wygenerowanego drzewa. Odpowiedz na pytania:

1. Jaka jest struktura drzewa? Liczba węzłów?, Liczba liści?, Ile jest możliwych ścieżek „decyzyjnych” wychodzących z korzenia drzewa? Jak wygląda zestaw warunków z najdłuższej ścieżki?
2. Czy mechanizm redukcji/upraszczania drzewa (ang. *pruning*) dokonał jakichkolwiek modyfikacji struktury drzewa?
3. Jakie są wyniki klasyfikowania obiektów za pomocą drzewa? Jak odczytać poziom błędów (ang. *errors*) z macierzy pomyłek (ang. *confusion matrix*)?

Zadanie 2. Klasyfikowanie nowych obiektów.

Dla drzewa wygenerowanego w zadaniu 1 dokonaj klasyfikowania nowych obiektów (opcja „konsultowania” z menu programu *Tree*/*Consult*).

1. Dokonać klasyfikacji wybranych przykładów ze zbioru uczącego *golf.tst*. Czy w rezultacie użycia drzewa popełnia sie jakieś błędy klasyfikacji?
2. Dokonać klasyfikacji przykładów nie pochodzących ze zbioru uczącego (stwórz ich opis samodzielnie lub według uwag prowadzącego). Sprawdź czy występują jakieś błędy klasyfikacji oraz czy przewidywania drzewa zgodne są z intuicją?
3. Dokonać klasyfikacji przykładów z niekompletnym opisem oraz później przykładów, dla których wartości atrybutów są nieprecyzyjne. Mogą to być przykłady charakteryzujące się następującym opisem:

Stan nieba	*Temperatura*	*Wilgotność*	*Wiatr*
słońce	brak danych	brak danych	nie
słońce: 0.9 pochm.: 0.1 deszcz: 0.1	19-20	80-85	tak: 0.9 nie: 0.1
słońce: 0.8 pochm.: 0.1 deszcz: 0.0	20-25	brak danych	tak: 0.7 nie: 0.3

Zadanie 3. Testowanie parametrów procesu generowania drzewa decyzyjnego (w szczególności zbadanie różnicy pomiędzy stosowaniem miar oceny „gain” / „info-gain” oraz tworzeniem drzew binarnych)

Realizację zadania rozpocznij od zapoznania się ze zbiorem przykładów podanym przez prowadzącego. Przeanalizuj strukturę wewnętrzną pliku i sprawdź czy między wartościami atrybutów warunkowych i decyzyjnych występują jakieś zależności. Uruchom generację drzewa przy pomocy programu C4.5 dla dwóch wartości parametru *Criterion*: „Info Gain” oraz „Info Gain Ratio”.

1. Sprawdź czy w postaci wygenerowanych drzew występują jakieś różnice? Opowiedz na pytanie: użycie, której z miar oceny wydaje się być korzystniejsze?
2. Przebadać generacje drzew dla obu poprzednich wartości parametru *Criterion* przy włączonej opcji *Subsetting* (wymuszającej tworzenia drzewa binarnego). Przeanalizuj otrzymane drzewa.

3. Można przy tej okazji obejrzeć i przeanalizować macierz pomyłek dla zbioru uczącego i testującego. Który z zestawów parametrów wydaje się najkorzystniejszy ze względu na wielkość drzewa i ocenę trafności klasyfikowania przykładów testowych?

Zadanie 4. Poszukiwanie właściwego stopnia uproszczenia drzew klasyfikujących.

Celem zadania jest sprawdzenie, w jakim stopniu parametr sterujący uproszczeniem drzewa w systemie C4.5 wpływa na jego zdolności klasyfikacyjne. Ocena skuteczności klasyfikowania powinna być dokonywana za pomocą opcji oceny krzyżowej (10-*fold cross validation*). Zaleca się wykonanie wykresów ilustrujących podstawowe zależności między badanymi parametrami. Do analizy należy wybrać pliki z przykładami wskazane przez prowadzącego.

1. Przeprowadzić serię eksperymentów oceny drzew decyzyjnych wygenerowanych systemem C4.5 zmieniając wartość parametru *Pruning confidence level* od 0.05 do 0.5 z krokiem 0.05 i sporządzić wykres zależności pomiędzy wartością zmienianego parametru a średnią trafnością (lub błędem)) klasyfikowania drzew pełnego i uproszczonego na zbiorze testującym,
   Uwaga! Podczas eksperymentu proszę ustalić wartości wszystkich innych parametrów w systemie C4.5 na standardowe.
2. Wykonaj także wykres ilustrujący zależność średniego błędu klasyfikacji w zależności od średniego rozmiaru drzewa.
3. Przeprowadź dyskusję otrzymanych wyników zwracając uwagę na zjawisko przeuczenia. Sprawdź, jaka jest różnica średniego błędu lub trafności klasyfikowania między drzewami pełnym a uproszczonym. Spróbuj określić, jaka jest najdogodniejsza wartość parametru poziomu ufności procedury upraszczającej. Sprawdź, czy domyślna wartość tego poziomu jest równie korzystna.
4. Przeprowadzić serię eksperymentów oceny skuteczności klasyfikacyjnej drzew decyzyjnych zmieniając w systemie C4.5 wartość parametru *Prepruning* (ograniczającym minimalną liczność przykładów w węźle) od 2 do 10 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, średnią estymatą błędu dla drzewa uproszczonego. Ustalić wartości wszystkich innych parametrów na standardowe. Oceń, jak zmienia się wartość błędu klasyfikacji w zależności od zmiany tego parametru. Czy drzewo uproszczone powyższą technika jest skuteczniejszym klasyfikatorem niż pełne drzewo?

Zadanie 5. Metoda pośrednia generowania zbioru reguł (C4.5rules).

1. Skorzystaj z implementacji systemu C4.5 – opcja generowania reguł. Tzn. przeprowadź najpierw indukcję pełnego drzewa, a później skorzystaj z opcji transformacji do zbioru reguł z wykonaniem operacji upraszczania.
2. Wykorzystaj w pierwszej kolejności zbiór uczący golf. Przeanalizuj otrzymana listę reguł i zinterpretuj jej strukturę. Porównaj ją z pełnym drzewem, z którego powstała – sprawdź czy odzwierciedla wszystkie ścieżki występujące w drzewie.
3. Przeprowadź klasyfikacje wybranych przykładów testowych.
4. Następnie przeanalizuj zbiór uczący zawierający większą liczbę przykładów. Zbadaj wpływ na tworzona listę reguł opcji generacji, tj. tzw. „pruning confidence level" i "Fisher's significance test" (interpretacja podana w dodatku A.1.).

Zadanie 6. Porównanie skuteczności klasyfikacyjnej drzew i reguł decyzyjnych.

Wykorzystując technikę 10-krotnej oceny krzyżowej znajdź najefektywniejsze klasyfikatory regułowe i drzewiaste dla wybranych zbiorów danych. Użyj minimum trzech różnych zbiorów (pamiętaj o zastosowaniu odpowiednich formatów plików w zależności od implementacji). Trafności lub błędy klasyfikowania rejestrów zarówno jako wynik średni z odchyleniem standardowym, jak i indywidualnie w każdej z 10 prac zbiór uczący-testujący. Porównaj otrzymane wyniki dla różnych klasyfikatorów – możesz posłużyć się narzędziami statystycznymi dla porównaniu różnicy między wynikami średnimi.