Ćwiczenie 3. Pojedyncze drzewo decyzyjne typu CART

Głównym celem ćwiczenia jest zbadanie zdolności uogólniania drzewa decyzyjnego typu CART w klasyfikacji danych. W tym celu konieczne będzie zastosowanie funkcji `fitctree`, która tworzy i buduje takie drzewo. Listing 1 przedstawia zastosowanie tej funkcji (linia 18). Przy tak dobranych parametrach (linia 5), kryterium podziału węzła to entropia. Predykcja zbudowanego drzewa realizowana jest za pomocą funkcji `predict` (linia 19). Kolejne kroki listingu umożliwiają znalezienie drzewa o najwyższej dokładności obliczonej w procesie walidacji krzyżowej. W linii 27 oraz 28 generowany jest zbiór reguł oraz rysowane jest najlepsze drzewo decyzyjne.

```
1  clear;
2  load('DBC');
3  X = DBC(:,1:end-1); Y = DBC(:,end);

4  % Parametry CART
5  split = 'deviance'; prune = 'off';

6  % Wygenerowanie indeksów do CV
7  indeksy_cv = crossvalind('Kfold', Y, 10);
8  Accuracy_CV = zeros(10,1);

9  for k = 1 : 10
10   % Indeksy do walidacji i uczenia
11   cv_test_ind = (indeksy_cv == k);
12   cv_train_ind = ~cv_test_ind;
13   % Rekordy do walidacji i uczenia
14   X_Test = X(cv_test_ind,:); Y_Test = Y(cv_test_ind);
15   X_Train = X(cv_train_ind,:); Y_Train = Y(cv_train_ind);
16
17   % Uczenie i walidacja CART:
18   C_Tree{k} = fitctree(X_Train,Y_Train,'SplitCriterion',split,'Prune',prune);
19    Label = predict(C_Tree{k}, X_Test);
20    % Dokładnosc (CV) dla CART
21    Accuracy_CV(k) = sum(Label == Y_Test)/length(Y_Test);
22 end

23 Avr_Accuracy = mean(Accuracy_CV);
24 [max,ind_max] = max(Accuracy_CV);
25 ind = ind_max(1);

26 % Podanie reguł i narysowanie struktury
27 view(C_Tree{ind});
28 view(C_Tree{ind}, 'mode', 'graph');
```

Listing 1. Zastosowanie algorytmu drzewa CART w klasyfikacji danych w aplikacji Matlab

Zdolność uogólniania drzewa należy wyznaczyć na podstawie kryterium dokładności. Parametry jakie trzeba zmieniać w ćwiczeniu to kryterium podziału węzła. Oprócz entropii, należy wykorzystać indeks Giniego w następujący sposób: `split = 'gdi';`. Konieczne jest również uwzględnienie innych parametrów drzewa: `MaxNumSplits`, `MinLeafSize` oraz `MinParentSize`. Rezultaty analizy należy przedstawić w formie uśrednionych dokładności walidacji krzyżowej (`Avr_Accuracy`) W każdym przypadku konieczne jest dołączenie:

- zbioru reguł w zrozumiałej składni (instrukcja `view(C_Tree{ind})` prezentuje zapis pośredni);
- model wyrysowanego drzewa.