**Algorytmy rozpoznawania obrazów**

**Drzewa decyzyjne w klasyfikacji**

**Piotr Kuboń 252871**

Drzewa decyzyjne to algorytm uczenia nadzorowanego wykorzystywany zarówno do problemów klasyfikacji jak i regresji. Posiada on drzewiastą strukturę, w której można wyróżnić takie elementy jak węzeł główny (zwany korzeniem), węzły wewnętrzne, gałęzie oraz liście.

Wyróżniamy drzewa binarne oraz takie które wychodzących gałęzi mają więcej.

Korzeniem nazywamy początkowy węzeł drzewa, z którego poprzez podziały powstają kolejne węzły potomne. Końcowe węzły, które nie podlegają podziałom nazywamy liśćmi, a linie łączące węzły nazywamy gałęziami

Proces decyzyjny zaczyna się od korzenia. Na podstawie dostarczonych danych uczących, wybierana jest cecha, dzięki której jesteśmy w stanie podzielić zbiór na jednorodne podzbiory.

Najczęściej warunek podziału jest generowany na podstawie jednej cechy, ale istnieją także podziały wielowymiarowe, jednak ze względu na złożoność obliczeniową są rzadziej stosowane.

Proces uczenia drzewa wykorzystuje strategię dziel i zwyciężaj przeprowadzając zachłanne wyszukiwanie w celu znalezienia jak najlepszych punktów podziału w drzewie.

Proces ten jest powtarzany rekurencyjnie, aż do napotkania na kryterium końcowe.

Wyróżnia się następujące kryteria końcowe:

* uzyskania jednorodnego węzła
* dotarcie do zadanej przez parametr głębokości
* uzyskanie pustego węzła
* wyczerpanie możliwych reguł podziałów

Drzewa służące do zadań klasyfikacyjnych, zawierają w liściach informację o tym, która klasa jest najbardziej prawdopodobna. Dąży się przy tym do takiego podziału by kolejne węzły, a co za tym idzie również liście, były ja najbardziej jednorodne.

Uczenie drzewa decyzyjnego wiąże się z ryzykiem przeuczenia modelu. W tym celu wprowadza się reguły zatrzymania oraz przycinanie drzewa.

Przycinanie jest to usuwanie pewnych gałęzi, służące lepszemu generalizowaniu problemu dzięki czemu zapobiegamy nadmiernemu dopasowaniu do danych uczących przez drzewo.

Jednym z algorytmów wykorzystujących drzewa decyzyjne jest algorytm lasu losowego. Polega on na utworzenie zespołu drzew oraz na podejmowaniu decyzji między innymi na podstawie głosowania większościowego.

Klasyfikator ten przewiduje dokładniejsze wyniki niż pojedyncze drzewo decyzyjne, zwłaszcza gdy poszczególne drzewa nie są ze sobą skorelowane.

Wyróżnia się następujące algorytmy uczenia drzew decyzyjnych:

* C4.5
* CART

Algorytm CART jest wykorzystywany do uczenia drzew decyzyjnych w module Scikit-Leadn.

Algorytm ten zazwyczaj wykorzystuje nieczystość Giniego jako funkcję koszu do identyfikacji idealnego atrybutu do podziału. Nieczystość Giniego mierzy, jak często losowo wybrany atrybut jest błędnie klasyfikowany. Podczas oceny za pomocą zanieczyszczenia Giniego, niższa wartość jest bardziej idealna.

Istotną zaletą algorytmu jest jego łatwość interpretacji. Dzięki możliwości wizualizacji struktury drzewa jesteśmy w stanie określić warunki decyzyjne a dzięki temu interpretować wyniki. Algorytm ten działa również na nieprzetworzonych danych, dzięki czemu nie jesteśmy zmuszeni do znacznego obrabiania i przygotowywania danych. Algorytm ten również rzadko popełnia grube błędy.

Natomiast do wad tego algorytmu można zaliczyć między innymi możliwość niestabilnego działania (niewielkie różnice w danych, mogą spowodować wygenerowanie znacznie innych reguł decyzyjnych, a co za tym idzie, znacznie innych drzew). Wyuczenie drzewa wymaga również dużej ilości danych uczących.