Drzewa decyzyjne są klasą algorytmów należącą do dziedziny uczenia nadzorowanego. Stosowane są zazwyczaj do zagadnień klasyfikacji, jednak umożliwiają również modelowanie ciągłej zmiennej zależnej. Celem działania drzew decyzyjnych jest podział przestrzeni danych na jak najbardziej homogeniczne ze względu na zmienną zależną obszary (nazywane liśćmi drzewa). Podział ten dokonywany jest za pomocą reguł decyzyjnych (węzłów decyzyjnych) opartych na zmiennych niezależnych.

Jeżeli modelowana jest zmienna ilościowa model określa się jako drzewo regresyjne. Jako wynik działania algorytmu zwracana jest średnia wartość zmiennej celu z obszaru, który został wskazany przez ciąg reguł decyzyjnych. W przypadku modelowania zmiennej jakościowej, liście składają się z obserwacji należących do różnych klas zmiennej zależnej. Liczebności te pozwalają obliczyć prawdopodobieństwo klasyfikacji do każdej z kategorii. Inaczej mówiąc, reguły decyzyjne przypisują rekordom wejściowym prawdopodobieństwa przynależności do klas modelowanej zmiennej, zgodnie z wyznaczonym liściem. W tej sytuacji, sterowanie progiem odcięcia umożliwia klasyfikację obserwacji do odpowiedniej kategorii.

Na przestrzeni lat stworzono wiele algorytmów budowy drzew decyzyjnych. Najpopularniejsze z nich przedstawiono w Tabeli 1.

Tabela 1. Algorytmy tworzenia drzew decyzyjnych

NAZWA	ROK	AUTORZY	RODZAJ DRZEWA
CLS	1996	Hunt,Marin, Stone	binarne
ACLS	1982	Paterson, Niblett	binarne
ID3	1983	Quinlan	dowolne
CART	1984	Brieman, Friedman Olshen, Stone	binarne
ASSISTANT	1985	Kononenko	binarne
ID4	1986	Schlimmer, Fisdher	dowolne
PLS	1986	Rendell	dowolne
C4	1987	Quinlan	dowolne
GID 3	1988	Chengf, Fayyad,Irani	dowolne
ID5	1989	Utgoff	dowolne
LMDT	1991	Brodley, Utgoff	binarne, wielowymiarowe
CHAID	1993	SPSSInc.	dowolne
IND	1993	Bruntine, Caruana	dowolne
SADT	1993	Heat,Kasif,Salzberg	binarne, wielowymiarowe
SE-LEARN	1993	Rymonn	dowolne
OC1	1994	Murthy	binarne, wielowymiarowe

Źródło: Ćwik, Mielniczuk, 2009

Wiele praktycznych implementacji stworzono w oparciu o algorytm CART (Classification and Regression Trees). Metoda ta przewiduje wyznaczanie binarnych reguł decyzyjnych z wykorzystaniem 2 kryteriów podziału: Giniego i podziału na dwie części (Breiman, Friedman, Olshen, Stone, 1984). Celem kryteriów podziału jest wyznaczanie kolejnych reguł decyzyjnych w taki sposób, aby otrzymane ostatecznie liście cechowały się jak największą homogenicznością ze względu na zmienną celu. W przypadku reguł binarnych, węzeł nadrzędny (parent node) dzielony jest tylko na dwa mniejsze węzły podrzędne (children nodes).

Kryterium Giniego opiera się na miarze jednorodności określanej jako indeks Giniego, który w m-tym węźle drzewa przy K klasach zmiennej celu definiuje się następująco:

$$G_m = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$$

Oznaczenie p_{mk} oznacza prawdopodobieństwo przyjęcia k-tej klasy w m-tym węźle przez znajdujące się w nim obserwacje. Jeżeli jedna z klas dominuje, tzn. prawdopodobieństwo jej wystąpienia jest zbliżone do 1, indeks Giniego przyjmuje wartość bliską 0. Zakłada się węzeł nadrzędny m_p i węzły podrzędne m_l oraz m_r, do których obserwacje z węzła nadrzędnego trafiają kolejno z prawdopodobieństwami P_l i P_r. Podział węzła nadrzędnego dokonuje się według wartości zmiennej niezależnej maksymalizującej kryterium Giniego o postaci:

$$\Delta G_m = G_{m_p} - P_l G_{m_l} - P_r G_{m_r}$$

Kryterium podziału na dwie części (twoing rule) skupia się na podziale danych na dwie możliwie równe części, jednak uwzględnia również jednorodność klas zmiennej celu w stworzonych węzłach. Maksymalizowana miara ma następującą formułę:

$$TR_{m_p} = \frac{P_l P_r}{4}\left(\sum_{k=1}^{K} |p_{m_l k} - p_{m_r k}|\right)^2$$

Inne algorytmy tworzenia drzew decyzyjnych tj. ID3 i C4.5 stosują również kryteria podziału oparte na entropii:

$$D_m = -\sum_{k=1}^{K} p_{mk} \log_2(p_{mk})$$

Niekontrolowany podział drzew decyzyjnych może w skrajnym przypadku doprowadzić do otrzymania liści z pojedynczymi obserwacjami. Tak zbudowane drzewo idealnie dopasowuje się do danych treningowych, jednak wykazuje słabe zdolności predykcyjne na zbiorach walidacyjnych. Model traci zdolność uogólniania, czyli występuje zjawisko przeuczenia modelu. Podstawową metodą kontroli wzrostu drzewa jest wyznaczenie warunków ograniczających takich jak:

- Minimalna liczba obserwacji w węźle przed podziałem
- Minimalna liczba obserwacji w liściu po podziale
- Maksymalna głębokość (liczba poziomów podziału) drzewa
- Maksymalna liczba liści
- Ograniczenie liczby zmiennych biorących udział w procesie podziału

Nałożenie ograniczeń na proces podziału drzewa decyzyjnego jest łatwe, jednak nie pozwala automatycznie dobrać optymalnych wartości warunków. W tym celu stworzono grupę

metod mających na celu jak najlepszą redukcję drzewa (tree pruning), czyli wybór poddrzewa charakteryzującego się najwyższą wartością miary dopasowania do danych. W praktyce często stosuje się algorytm cost complexity pruning, który wprowadza do modelu parametr złożoności drzewa. Przebieg procesu przycinania drzewa regresyjnego metodą cost complexity wygląda następująco (Gareth i wsp., 2013):

1. Zbudować rozbudowane drzewo T_0 z wykorzystaniem podziału binarnego i warunków ograniczających.
2. Dla kolejnych wartości parametru złożoności α, przypisać poddrzewo $T \subset T_o$ dla którego minimalizowana jest wartość:

$$\sum_{m=1}^{|T|} \sum_{x_i \in T_m} \left(y_i - \hat{y}_{T_m}\right)^2 + \alpha|T|$$

   Gdzie |T| oznacza liczbę liści w poddrzewie T, y_i rzeczywistą wartość zmiennej celu dla obserwacji x_i, a $\hat{y}_{T_m}$ średnią wartość wszystkich obserwacji w liściu T_m.
3. Wykorzystując K-krotną walidację krzyżową, powtórzyć kroki 1-2 K razy i dla każdej wartości parametru α obliczyć średnią wartość błędu predykcji.
4. Wybrać poddrzewo o parametrze złożoności, dla którego otrzymano najniższy średni błąd predykcji.

Proces ten przebiega analogicznie w przypadku drzew klasyfikacyjnych. Parametr złożoności określa koszt wprowadzenia do drzewa dodatkowego liścia, dlatego nazywany jest również parametrem kosztu.

Ze względu na swoją budowę, drzewa decyzyjne charakteryzują się wieloma zaletami w porównaniu do klasycznych modeli:

- Są łatwe w interpretacji i wizualizacji struktury modelu
- Wykorzystanie predyktorów jakościowych nie wymaga tworzenia sztucznych zmiennych
- Dobrze radzą sobie z brakami danych i wartościami odstającymi
- Poprzez zastosowanie pruningu, otrzymuje się elastyczny model.

Z drugiej strony, drzewa decyzyjne wykazują zazwyczaj gorsze cechy predykcyjne niż inne typy modeli. Rozwiązaniem powyższego problemu jest zastosowanie zaawansowanych metod wykorzystujących drzewa decyzyjne, do których należą m.in. bagging, boosting i lasy losowe. Algorytmy te budują wiele drzew decyzyjnych na podzbiorach danych wejściowych lub z wykorzystaniem części zmiennych niezależnych, a następnie po agregacji i przetworzeniu wyników zwracają ostateczne predykcje. W porównaniu do podstawowych drzew decyzyjnych

cechują się większą stabilnością i możliwościami predykcyjnymi, jednak tracą przejrzysty charakter i łatwą możliwość interpretacji modelu.

Bibliografia

Ćwik, J., Mielniczuk, J. (2009). Statystyczne systemy uczące się - ćwiczenia w oparciu o pakiet R. Oficyna Wydawnicza PW, Warszawa.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). Classification and Regression Trees. New York: Chapman & Hall.

Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.