Jak działa model klasyfikacji *Random Forest*?

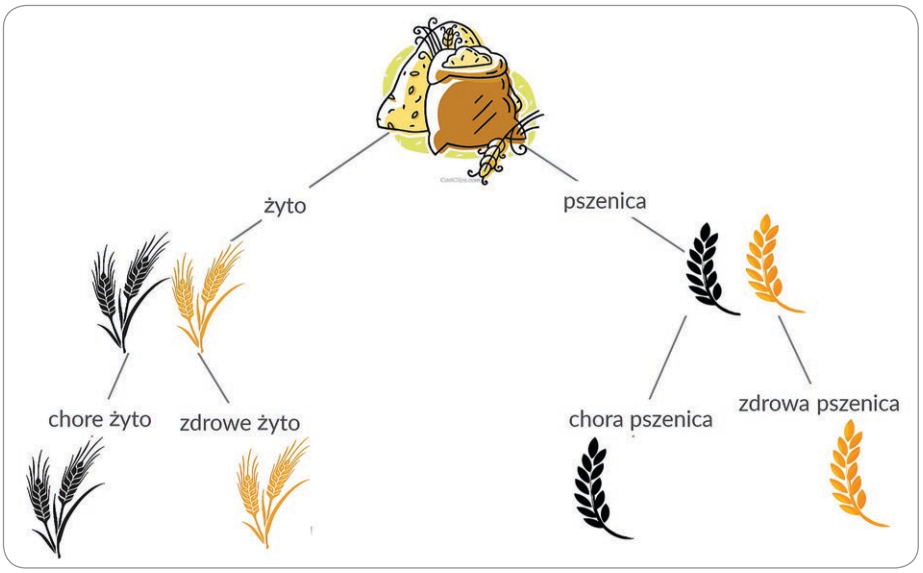
Las losowy (*random forest*) jest szeroko stosowanym modelem uczenia maszynowego dla klasyfikacji i regresji (model regresji nazywa się *Random forest regressor*). Pierwszy algorytm losowych lasów decyzyjnych został opublikowany w 1995 roku przez Tina Kam Ho[1]. Była to implementacja metody klasyfikacji, znanej jako „dyskryminacja stochastyczna", zaproponowanej przez Eugena Kleinberga[2].

Metoda *random forest* jest popularna wśród badaczy uczenia maszynowego, ponieważ jest niezmienna pomimo stosowania skalowania oraz innych przekształceń zbioru zmiennych niezależnych. Metoda ta jest też odporna na nieistotne zmienne niezależne, co pozwala zaoszczędzić czas i ominąć w procedurze badawczej eliminację tych zmiennych. Przy okazji warto wskazać, że metoda *random forest* jest często wykorzystywana właśnie jako metoda tzw. *feature elimination* dla innych modeli *machine learning*. Model *random forest* jest też stosunkowo szybki i oszczędny dla mocy obliczeniowych komputerów, co dodatkowo przysparza mu zwolenników wśród badaczy danych.

Jak działa proces klasyfikacji pojedynczego drzewa decyzyjnego?

Jestem przekonany, że każda osoba, która czyta teraz ten artykuł kilka razy dziennie, nieświadomie przeprowadza proces klasyfikacji oparty na metodzie pojedynczego drzewa decyzyjnego.

Aby zilustrować proste drzewo decyzyjne, posłużymy się przykładem selekcji ziaren zbóż. Zadanie polega na oddzieleniu ziaren zdrowych od tych zainfekowanych pleśnią. Dodatkowo trzeba oddzielić ziarna pszenicy od ziaren żyta.



Typowe drzewo decyzyjne przyjmie postać przedstawioną na poniższym diagramie.

Proces decyzyjny biegnie wzdłuż gałęzi drzewa. Wynikiem procesu klasyfikacji jest jedna określona wartość. Model otrzymuje dane pobrane z worka zawierającego ziarno i na podstawie wielu cech fizycznych takich jak masa, wilgotność, kolor musi zakwalifikować ziarno do odpowiedniej klasy.

Random forest działa na zasadzie mądrości zbiorowej

Losowy las składa się z kilkuset pojedynczych drzew decyzyjnych. Każde ma określoną liczbę gałęzi i liści. Ilości drzew, liści i gałęzi jest określana przez badacza budującego model; ustawienia te nazywane są hiperparametrami.

W przypadku procesu klasyfikacji każde pojedyncze drzewo w lesie generuje pojedynczą prognozę klasy. Klasa, za którą głosowało najwięcej drzew staje się prognozą modelu. Zasada ta przypomina demokratyczne wybory, gdzie wygrywa ten, na kogo oddano najwięcej głosów.

Istnieją jednak różne rodzaje wyborców. Bywają populacje ludzi bardzo podobnych do siebie oraz populacje ludzi różniący się od siebie. Nie od dzisiaj wiadomo, że najtrafniejsze decyzje podejmuje tłum złożony z różnorodnych osób. Powodem tego jest zjawisko wzajemnej eliminacji indywidualnych błędów. W przypadku ludzi bardzo podobnych do siebie istnieje wysokie prawdopodobieństwo, że wszyscy lub większość będzie podążała w jednym kierunku bez możliwości korekty, nawet gdy ten kierunek okaże się błędny.

Podobnie jest w lesie złożonym z drzew decyzyjnych. Podczas, gdy niektóre drzewa mogą się mylić, wiele innych będzie miało rację, więc jako grupa drzewa mogą poruszać się we właściwym kierunku.

Aby model działał dobrze muszą być spełnione dwa warunki:

1 – musi istnieć rzeczywista informacja w zmiennych niezależnych, aby opierający się na nich model klasyfikował lepiej niż przypadkowy rzut monetą,

2 – wyniki klasyfikacji poszczególnych drzew, zarówno błędne jak i trafne, muszą mieć ze sobą niskie korelacje.

Duża liczba nisko skorelowanych ze sobą modeli (pojedynczych drzew) działa podobnie jak społeczność niezależnych od siebie wyborców. Siła tej metody bazuje więc na wzajemnej niezależności poszczególnych drzew decyzyjnych.

Niestety drzewa decyzyjne pracują na tych samych danych, istnieją więc przesłanki, aby nabrać wątpliwości co do ich wzajemnej niezależności. W przypad-

[1] Twin Cam Ho; „Random Decision Forests"; AT&T Bell Laboratories, Murray Hill, New Jersey 07974, USA 1995

[2] Eugene Kleinberg; „Stochastic discrimination", Annals of Mathematics and Artificial Intelligence 2005

ku klasyfikacji, gdzie dane mają postać dyskretną, oczywiście nie możemy mówić o korelacji, lecz o formie zależności, którą opisać można m.in. za pomocą algorytmów chi-kwadrat lub pomiaru pojemności informacji (*mutual information*).

Jak sprawić, aby drzewa decyzyjne stały się różne i niezależne od siebie?

Aby drzewa decyzyjne nie były ze sobą powiązane oraz cechowały się różnorodnością stosuje się jednocześnie dwie proste metody: pakowanie oraz losowość cech.

Pakowanie (*bagging*) polega na tym, że każde drzewo pobiera losowo stałą liczbę k rekordów ze zwracaniem. Załóżmy, że model trenuje na próbce składającej się z k = 12 rekordów. W praktyce stosuje się próbki zawierające setki tysięcy rekordów. Bez użycia metody pakowania każde drzewo pobrałoby do nauki k = 12 takich samych rekordów, co można zapisać jako zbiór: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12].

Przy zastosowaniu metody pakowania drzewa losowo pobierają k=12 rekordów, każdorazowo zwracając je po losowaniu do puli w celu dalszego losowania. Pojedyncze drzewo wykorzystuje więc k=12 rekordów, lecz każde drzewo ma inny zestaw rekordów. W naszym przykładzie zbiór rekordów może wyglądać tak: 1, 2, 3, 4, 4, 4, 4, 5, 7, 10, 10, 12. Jak widać, niektóre rekordy zostały wylosowane nawet cztery razy, inne nie pojawiły się wcale. Oznacza to, że każde drzewo dostaje unikatowy zestaw rekordów do nauki. Sytuację tę można porównać do pojedynczego wyborcy, który ma wiele informacji medialnych, niektóre słyszał wielokrotnie, a o innych nie słyszał nigdy.

Kolejną metodą zwiększającą różnorodność drzew decyzyjnych jest **losowość cech** (*feature randomness*); każde drzewo losowo dostaje określone cechy, które ma wykorzystać przy dokonywaniu klasyfikacji. Tu również możemy posłużyć się przykładem populacji wyborców. Dla niektórych kategorie społeczne takie jak: ekologia czy polityka zagraniczna nie mają znaczenia, a znaczenie mają podatki i praworządność.

Dzięki tym metodom drzewa decyzyjne są wysoce zróżnicowane, co gwarantuje wzmocnienie mądrości zbiorowej. Dzięki różnorodności poszczególnych drzew model *random forest* jest trafniejszy w klasyfikacji od każdego indywidualnego drzewa.

Na koniec należy podkreślić, że w przypadku klasyfikacji niezbędne jest właściwe zbilansowanie zbioru treningowego. Oznacza to, że przy trenowaniu modelu na zbiorze o dychotomicznej charakterystyce wyników model musi mieć podobną liczbę rekordów dla obu klas. W przeciwnym wypadku wszystkie albo większość drzew będzie powtarzało błąd polegający na poparciu klasy większościowej. Dobrym przykładem jest człowiek, którego zadaniem jest selekcjonowanie worków ze zdrowym i spleśniałym ziarnem. Prawdopodobnie zaprzestanie on kontroli i uzna bez sprawdzania, że wszystkie worki zawierają zdrowe ziarno, kiedy dowie się, że worków ze spleśniałym ziarnem jest od 1 do 2% (czyli 1-2 na 100 worków).

Podobnie postąpi każde pojedyncze drzewo z modelu losowego lasu. Dokładność takiego modelu zostanie określona jako 99%, co w oczywisty sposób będzie błędną oceną.

Wojciech Moszczyński