Krzysztof Sękowski, Jan Szataniak

Klasyfikacja pochodzenia białek drożdży

Celem projektu jest stworzenie systemu zdolnego do odgadnięcia, z którego organellum drożdża pochodzi dane białko.

Zbiór danych treningowych został stworzony przez Kentę Nakai z Uniwersytetu Osakijskiego. Dane zostały w przeszłości użyte do stworzenia systemu eksperckiego rozpoznającego pochodzenie białek drożdży.

Dane są zapisane w pliku tekstowym w kolumnach rozdzielonych zmienną liczbą spacji. Wartości są zeskalowane do przedziału <0-1>, więc możemy pominąć ten etap. Spis kolumn:

- nazwa białka w bazie danych SWISS-PROT
- mcg: wynik testu metodą McGeoch'a
- gvh: wynik testu metodą van Heijnego
- alm: wynik programu predykcyjnego ALOM dla białek transbłonowych
- mit: wynik analizy dyskryminacyjnej zawartości aminokwasów końca aminowego białek mitochondrialnych i niemitochondrialnych
- erl: obecność podłańcucha "HDEL" (wartość binarna)
- pox: PTS końca karboksylowego
- vac: wynik analizy dyskryminacyjnej zawartości aminokwasów końca aminowego białek wakuoli i pozakomórkowych
- nuc: wynik analizy dyskryminacyjnej NLS białek jądrowych i niejądrowych

Pierwszym krokiem było wczytanie wartości. Ze względu na rozdział danych zmienną liczbą spacji, konieczna była obróbka danych do formatu, który jest w stanie odczytać pyspark. Spacje zostały zamienione na pojedyncze tabulacje. Dopiero po zapisaniu zmodyfikowanego pliku, dane zostały wczytane.

Podczas wstępnej analizy poczyniliśmy następujące obserwacje:

- Klasy mają różną liczebność, co może spowodować problemy z klasyfikacją mniej liczebnych przypadków
- Zbiór został ustandaryzowany i nie występują w nim luki

Zbiór został podzielony na dwie grupy - testową i treningową - w stosunku 3:7.

Na zbiorze treningowym wyuczono trzy różne modele klasyfikacji: lasu losowego, Bayesa i regresji logistycznej.

Metoda lasu losowego polega na użyciu wielu drzew losowych. Algorytm polega na utworzeniu drzew o różnych konfiguracjach parametrów i trenowaniu ich tak samo, jak pojedynczego drzewa w procesie nadzorowanym. Wagi parametrów są poprawiane oddzielnie w każdym drzewie. Zakłada się, że statystycznie większość po wytrenowaniu będzie dawała poprawny wynik.

Wielomianowa regresja logistyczna to rodzaj klasyfikacji metodą regresji dla wielu zmiennych. Regresja polega na odnalezieniu wielomianu, który jest bliski punktom należącym do danej klasy. W tym przypadku dopasowywana jest funkcja logistyczna, tj. 1/(1+e^(-(x-m)/s), gdzie m i s to parametry a x to wektor danych.

Naiwny klasyfikator Bayesa to algorytm oparty na dystrybucji Gaussa. Każda cecha jest traktowana osobno i ma taką samą wagę. Model ten używa prawdopodobieństwa przynależności danej wartości do klasy.

Efektywność została sprawdzona poprzez porównanie poprawnego przyporządkowania w zbiorze początkowym z wynikiem uzyskanym przez model dla każdej ktorki. Dokładność lasu losowego jest największa, więc zostanie on użyty w systemie. Można zauważyć, że wszystkie systemy miały problem z dopasowaniem podgrup białek błonowych. Żadna z metod nie rozpoznawała białek wakuoli — drugiej najmniejszej grupy w zbiorze —, lecz dokładność przyporządkowania dla najmniej popularnej grupy retikulum endoplazmatycznego wynosiła ponad 50%. Metoda Bayesa sprawdziła się tylko do wykrywania białek cytoplazmy.

Dwie pozostałe metody miały bardzo zbliżone wyniki a o tym, która z nich sprawdziła się lepiej, decydował zbiór treningowy. Podejrzewamy, że w danych nie występuje rozrzut standardowy, co znacznie zmniejszyłoby dokładność tej metody. Hipoteza została potwierdzona na dwóch klasach. Okazało się, że część danych jest nierównomiernie rozłożona, co prowadzi do błędnego przyporządkowania przez metodę Bayesa.

Model został wykorzystany, by klasyfikować losowo wygenerowane dane drożdży w formacie JSON. Są one wysyłane przez skrypt na temat kafki. Z tematu sczytuje je drugi program, który wczytuje zapisany pipeline z metodą lasu losowego i zwraca wynik klasyfikacji dla wysłanych wartości.