Uczenie się Maszyn
Projekt "Zastosowanie drzew decyzyjnych do klasyfikacji miejsc rozcięcia w sekwencji DNA"

Dokumentacja wstępna

Igor Markiewicz

1 Spis treści

2 Przedstawienie problemu

Zadaniem w miniejszym projekcie jest klasyfikacja przykładów na dwie klasy : pozytywną i negatywną przy użyciu drzew decyzyjnych. Przykładami są fragmenty sekwencji DNA w których dla przykładów pozytywnych istnieje rozcięcie między częścią niekodującą (intronem) a częścią kodującą białko (eksonem). Rozcięcia te są dwóch rodzajów :

1. Donory
2. Akceptory

W przypadku miniejszego projektu zadanie będzie składać się z dwóch rozłącznych części. Osobnego budowania, uczenia i badania dla zbioru przykładów zawierających tylko rozcięcia typu donorowego (oraz przykłady negatywne) i analogicznie dla zbioru przykładów zawierającego rozcięcia typu tylko akceptorowego (wraz z przykładami negatywnymi).

3 Wybór języka programowania

Jako język programowania, który będzie użyty w części implementacyjnej został wybrany C++. Do jego głównych zalet należą :

- Wszechstronność i duże możliwości
- Bardzo dobre dostosowanie do implementacji algorytmów
- Oprogramowanie typu Open Source
- Popularność i dobra dokumentacja

4 Analiza zbioru danych

Zbiór danych składa się z dwóch plików :

- spliceDTrainKIS - zbiór donorów i przykładów negatywnych (5256 przykładów, ciągi znakowe o długości 15)
- spliceATrainKIS - zbiór akceptorów i przykładów negatywnych (5788 przykładów, ciągi znakowe o długości 90)

W obu plikach w pierwszej linii występuje liczba informująca o pozycji rozcięcia dla przykładów pozytywnych, licząc od lewej strony. Następnie w kolejnych liniach, na przemian podawana jest klasyfikacja oraz przykład którego dotyczy. Jednym z zadań będzie zaprojektowanie parsera danych, służącego do ich wczytywania, obrabiania i zapisywania. Każdy znak w ciągu znakowym przykładu jest jedną z liter reprezentujących zasady azotowe : A(adenina), C(cytozyna), G(guanina) lub T(tymina).

5 Metoda uczenia

Jako metodykę budowy drzewa decyzyjnego, został wybrany algorytm typu TDIDT (*Top-Down Induction of Decision Tree*) :

funkcja buduj-drzewo(P, d, S)
argumenty wejściowe:

- P - zbiór przykładów etykietowanych pojęcia c
- d - domyślna etykieta kategorii
- S - zbiór możliwych testów

zwraca: drzewo decyzyjne reprezentujące hipotezę przybliżającą c na zbiorze P;
1: **jeśli** kryterium-stopu(P,S) **to**
2: utwórz liść **l**;
3: $d_{\mathbf{l}}$:= kategoria(P,d);
4: **zwróć l**;
5: **koniec jeśli**
6: utwórz węzeł **n**;
7: $t_{\mathbf{n}}$:= wybierz-test(P,S);
8: d := kategoria(P,d);
9: **dla wszystkich** $r \in R_{t_{\mathbf{n}}}$ **wykonaj**
10: **n**[r] := buduj-drzewo$(P_{t_{\mathbf{n}}r},\ d,\ S - \{t_{\mathbf{n}}\})$;
11: **koniec dla**
12: **zwróć n**

$d_{\mathbf{l}}$ - kategoria d liścia **l**
$t_{\mathbf{n}}$ - związany z węzłem **n** test t
$R_{t_{\mathbf{n}}}$ - zbiór możliwych wyników testu t dla węzła **n**
$P_{t_{\mathbf{n}}r}$ - zbiór przykładów z P dla których dla testu t i węzła **n** są stowarzyszone z wynikiem r
$\mathbf{n}[r]$ - węzeł lub liść potomny do którego prowadzi z węzła **n** gałąź odpowiadająca wynikowi r

5.1 Kryterium stopu

Kryterium stopu daje wynik pozytywny, gdy zachodzi przynajmniej jeden z następujących przypadków :

- Zbiór przykładów P jest pusty
- Zbiór testów S jest pusty
- Zbiór przykładów P zawiera przykłady wyłącznie jednej kategorii

5.2 Kategoria

$$\text{kategoria}(P, d) = \begin{cases} d, & \text{gdy } P = \emptyset \\ \arg\max_{d'} \left|P^{d'}\right|, & \text{w pozostałych przypadkach} \end{cases}$$

5.3 Rodzaj testu

Jako możliwe testy traktujemy pozycje liter w ciągu znakowym przykładu. Dla każdego testu i każdego węzła, zbiór możliwych wyników jest taki sam : $R_{t_\mathbf{n}} = R = \{A, C, G, T\}$. Możemy więc utożsamić testowanie przykładu z testowaniem jego atrybutu i stwierdzamy że mamy doczynienia z nominalnym testem tożsamościowym.

5.4 Wybór testu

Jako kryterium wyboru testu przyjmujemy podejście oparte o teorioinformacyjną entropię. Wybieramy ten test, który dla aktualnego zbioru przykładów spowoduje największy przyrost informacji (czyli ten który ma najmniejszą entropię). Entropię przykładów ze zbioru P ze względu na wynik r testu t określamy jako :

$$E_{tr}(P) = \sum_{d \in C} -\frac{\left|P_{tr}^{d}\right|}{|P_{tr}|} \lg \frac{\left|P_{tr}^{d}\right|}{|P_{tr}|}$$

Natomiast entropię ważoną przykładów ze zbioru P dla testu t określamy jako :

$$E_t(P) = \sum_{r \in R_t} \frac{|P_{tr}|}{|P|} E_{tr}(P)$$

P_{tr}^{d} - zbiór przykładów z P mających dla testu t wartość r i kategorię d
P_{tr} - zbiór przykładów z P mających dla testu t wartość r

W przypadku wyrażenia typu $\lg \frac{0}{a}$ gdzie $a \neq 0$ przyjmujemy że jest ono równe 0, podobnie jak dla $\frac{0}{0} \lg \frac{0}{0}$
Dla równych entropii ważonych, wybieramy pierwszą obliczoną.

6 Przykład

Niech zbiór testowy T będzie określony następująco :

id	x	c
1	TGAT	0
2	ACGC	0
3	ACCT	1
4	ACTT	1
5	ATAG	0
6	ACGT	1

1. Pierwsze wywołanie funkcji konstruującej drzewo ma postać :

$$\text{buduj-drzewo(T, 0, } \{1,2,3,4\}\text{);}$$

gdyż w zbiorze przykładów liczba elementów o kategorii pozytywnej i negatywnej jest taka sama, więc jako kategorię domyślną w takiej sytuacji arbitralnie przyjmujemy 0. Ponad to zbiór możliwych testów tożsamościowych jest równy zbiorowi wszystkich atrybutów.
2. Mamy $P = T = \{1,2,3,4,5,6\}$
3. Kryterium stopu nie jest spełnione
4. Tworzony jest nowy węzeł dla którego nalezy wybrać test
5. Wybór testu:

$$E_{1,A} = -\frac{2}{5}\lg\frac{2}{5} - \frac{3}{5}\lg\frac{3}{5} \approx 0.97$$
$$E_{1,C} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{1,G} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{1,T} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$

$$E_{2,A} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{2,C} = -\frac{1}{4}\lg\frac{1}{4} - \frac{3}{4}\lg\frac{3}{4} \approx 0.81$$
$$E_{2,G} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$
$$E_{2,T} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$

$$E_{3,A} = -\frac{2}{2}\lg\frac{2}{2} - \frac{0}{2}\lg\frac{0}{2} = 0$$

$$E_{3,C} = -\frac{0}{1}\lg\frac{0}{1} - \frac{1}{1}\lg\frac{1}{1} = 0$$
$$E_{3,G} = -\frac{1}{2}\lg\frac{1}{2} - \frac{1}{2}\lg\frac{1}{2} = 1$$
$$E_{3,T} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$

$$E_{4,A} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{4,C} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$
$$E_{4,G} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$
$$E_{4,T} = -\frac{1}{4}\lg\frac{1}{4} - \frac{3}{4}\lg\frac{3}{4} \approx 0.81$$

$$E_1 = 0.97 \cdot \frac{5}{6} \approx 0.81$$
$$E_2 = 0.81 \cdot \frac{4}{6} = 0.54$$
$$E_3 = 1 \cdot \frac{2}{6} \approx 0.33$$
$$E_4 = 0.81 \cdot \frac{4}{6} = 0.54$$

Jako test zostaje wybrany atrybut nr 3.

6. Kategorią domyślną pozostaje 0.
7. Następuje wywołanie rekurencyjne
- **A)** buduj-drzewo($\{1, 5\}$, 0, $\{1,2,4\}$);
  - **1.** Mamy $P = \{1,5\}$
  - **2.** Kryterium stopu jest spełnione (przykłady tej samej kategorii - 0)
  - **3.** Tworzony jest liść **l** z etykietą 0
  - **4.** Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia
- **C)** buduj-drzewo($\{3\}$, 0, $\{1,2,4\}$);
  - **1.** Mamy $P = \{3\}$
  - **2.** Kryterium stopu jest spełnione (przykłady tej samej kategorii - 1)
  - **3.** Tworzony jest liść **l** z etykietą 1
  - **4.** Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia
- **G)** buduj-drzewo($\{2, 6\}$, 0, $\{1,2,4\}$);

2. Mamy $P = \{2,6\}$
3. Kryterium stopu nie jest spełnione
4. Tworzony jest nowy węzeł dla którego należy wybrać test
5. Wybór testu:

$$E_{1,A} = -\frac{1}{2}\lg\frac{1}{2} - \frac{1}{2}\lg\frac{1}{2} = 1$$
$$E_{1,C} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{1,G} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{1,T} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$

$$E_{2,A} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 1$$
$$E_{2,C} = -\frac{1}{2}\lg\frac{1}{2} - \frac{1}{2}\lg\frac{1}{2} = 0$$
$$E_{2,G} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{2,T} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$

$$E_{4,A} = -\frac{0}{0}\lg\frac{0}{0} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{4,C} = -\frac{1}{1}\lg\frac{1}{1} - \frac{0}{1}\lg\frac{0}{1} = 0$$
$$E_{4,G} = -\frac{0}{0}\lg\frac{1}{1} - \frac{0}{0}\lg\frac{0}{0} = 0$$
$$E_{4,T} = -\frac{0}{1}\lg\frac{0}{1} - \frac{1}{1}\lg\frac{1}{1} = 0$$

$$E_1 = 1 \cdot \frac{2}{6} \approx 0.33$$
$$E_2 = 1 \cdot \frac{2}{6} \approx 0.33$$
$$E_4 = 0 \cdot 0 = 0$$

Jako test zostaje wybrany atrybut nr 4.

6. Kategorią domyślną ustalamy jako 0.

7. Następuje wywołanie rekurencyjne

A) buduj-drzewo($\emptyset$, 0, $\{1,2\}$);

1. Mamy $P = \emptyset$

2. Kryterium stopu jest spełnione (pusty zbiór przykładów)

3. Tworzony jest liść **l** z etykietą 0

4. Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia

C) buduj-drzewo($\{2\}$, 0, $\{1,2\}$);

1. Mamy $P = \{2\}$

2. Kryterium stopu jest spełnione (przykłady tej samej kategorii - 0)

3. Tworzony jest liść **l** z etykietą 0

4. Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia

G) buduj-drzewo($\emptyset$, 0, $\{1,2\}$);

1. Mamy $P = \emptyset$

2. Kryterium stopu jest spełnione (pusty zbiór przykładów)

3. Tworzony jest liść **l** z etykietą 0

4. Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia

T) buduj-drzewo($\{6\}$, 0, $\{1,2\}$);

1. Mamy $P = \{6\}$

2. Kryterium stopu jest spełnione (przykłady tej samej kategorii - 1)

3. Tworzony jest liść **l** z etykietą 1

4. Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia

T) buduj-drzewo($\{4\}$, 0, $\{1,2,4\}$);

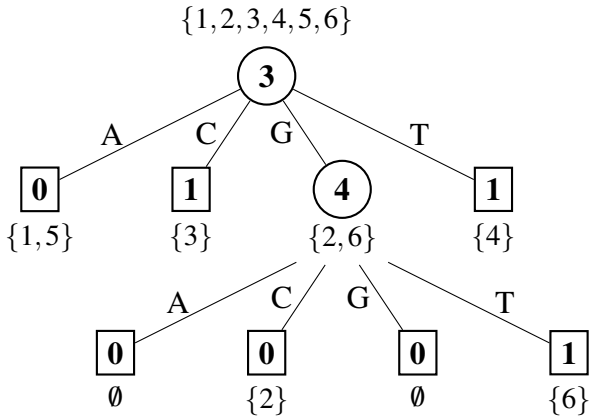
1. Mamy $P = \{4\}$

2. Kryterium stopu jest spełnione (przykłady tej samej kategorii - 1)

3. Tworzony jest liść **l** z etykietą 1

4. Następuje powrót z wywołania rekurencyjnego i zwrot utworzonego liścia

8. Następuje zakończenie procesu budowy drzewa decyzyjnego



7 Podział na zbiór uczący i testujący

Każdy ze zbiorów wejściowych zostanie podzielony w sposób losowy na dwa rozłączne podzbiory, z czego jeden będzie stanowił zbiór uczący a drugi zbiór testowy, służący do określenia błędu rzeczywistego w eksperymentach. Losowanie zapewni wprowadzenie elementów statystyki i najprawdopodobniej poprawi proces uczenia.

8 Plan eksperymentów

- **Badanie błędu rzeczywistego w zależności od ilości przykładów przeznaczonych do zbioru uczącego i testowego.**
  Jako przykładową procedurę można przyjąć np: testowanie w dziewięciu punktach pomiarowych (10 %, 20 %, ... , 90 %) Dodatkowo w każdym z nich możemy w celu polepszenia statystyki wywołać kilka-kilkanaście razy budowanie i testowanie drzewa, a otrzymane wyniki uśrednić.

- **Badanie wpływu przycinania drzewa decyzyjnego metodą przycinania redukującego błąd na błąd rzeczywisty.**
  Jako przykładowy algorytm przycinania drzewa decyzyjnego, wybieramy przycinanie rozrośniętego drzewa (post-pruning) :

  funkcja przytnij-drzewo($\mathbb{T}$, P)
  argumenty wejściowe:

  – $\mathbb{T}$ - drzewo do przycięcia
  – P - zbiór przycinania

  zwraca: drzewo decyzyjne $\mathbb{T}$ po przycięciu;

  1: **dla wszystkich** węzłów **n** drzewa $\mathbb{T}$ **wykonaj**
  2: zastąp **n** liściem **l** z etykietą większościowej kategorii w zbiorze $P_{\mathbb{T},\mathbf{n}}$ jeśli nie powiększy to szacowanego na podstawie zbioru P błędu rzeczywistego drzewa $\mathbb{T}$;
  3: **koniec dla**;
  4: **zwróć** $\mathbb{T}$;

W tym punkcie każdy zbiór wejściowy zostanie podzielony w sposób losowy na trzy, rozłączne podzbiory odpowiadające za uczenie, przycinanie oraz testowanie. Wielkość zbioru przeznaczonego do przycinania będzie obiektem badań.

Jako metodykę badań przyjmujemy podanie błędu rzeczywistego (liczonego jako liczba pomyłek w stosunku do liczby pomyłek i liczby dobrze zaklasyfikowanych przykładów), confusion matrix oraz krzywą ROC.

9 Bibliografia

[1] Wykłady do przedmiotu Uczenie się Maszyn

[2] P.Cichosz *Systemy Uczące się*, Wydanie Drugie, Wydawnictwo Naukowo-Techniczne, Warszawa 2000, 2007