01.06.2020 r.

Metody Odkrywania Wiedzy

Projekt - dokumentacja

Autorzy: Katarzyna Muter, Maciej Rosoł

Temat projektu: Konstruktywna indukcja w zadaniach klasyfikacji sekwencji DNA - należy stworzyć nowe atrybuty poprzez łączenie już istniejących, po czym z powstałego zbioru (pierwotnych i wtórnych) atrybutów należy dokonać selekcji (wyboru) zestawu tych najlepszych przez filtrowanie z różnymi funkcjami oceny przydatności (*ang. attribute selection filters*)

Opis bazy danych

W pliku *spliceDTrainKIS.dat* znajdują się dane do problemu rozpoznawania donorów, a w pliku *spliceATrainKIS.dat* znajdują się dane do problemu rozpoznawania akceptorów. W pierwszej linii każdego z nich jest napisane, na której pozycji (licząc litery od lewej strony) we fragmentach sekwencji jest granica pomiędzy intronem a eksonem. Plik *spliceDTrainKIS.dat* zawiera 5256 sekwencji DNA o długości 15 nukleotydów, z czego 1116 przykładów jest oznaczonych jako pozytywne (1), a 4140 przykładów – jako negatywne (0). Plik *spliceATrainKIS.dat* natomiast zawiera 5788 sekwencji DNA o długości 90 nukleotydów, z czego 1116 przykładów jest oznaczonych jako pozytywne (1), a 4672 przykładów – jako negatywne (0).

Wyznaczenie cech z sekwencji DNA

Cechy zostały utworzone na podstawie sekwencji DNA zawartych w pliku wejściowym. Początkowymi atrybutami są wystąpienia danych liter na danych pozycjach w sekwencji. Na każdej pozycji może wystąpić jedna z czterech liter: A, C, T lub G. W pliku *spliceDTrainKIS.dat* sekwencje DNA mają długość 15 nukleotydów, w związku z tym dla tego pliku wyliczono 60 cech określających wystąpienie danego nukleotydu na danym miejscu w sekwencji (4 nukleotydy razy 15 miejsc). Fragmenty DNA w pliku *spliceATrainKIS.dat* mają długość 90 nukleotydów, co za tym idzie dla tego pliku wyliczonych cech opisujących pozycje było 360. Następne atrybuty określają ile razy dana litera znajduje się w sekwencji. Występują cztery takie cechy – ilość zliczeń dla każdej z liter. Kolejnymi cechami są: procent danej litery po lewej stronie od miejsca rozcięcia (jaki procent liter po lewej stronie od miejsca rozcięcia stanowi dana litera) i procent danej litery po prawej stronie od miejsca rozcięcia. Miejsce rozcięcia (granica pomiędzy intronem a eksonem) w pliku *spliceDTrainKIS.dat* znajduje się na pozycji 7, a w pliku *spliceATrainKIS.dat* miejsce rozcięcia znajduje się na pozycji 68. Wyznaczono cztery cechy określające jaki procent wszystkich liter po lewej stronie od miejsca rozcięcia stanowi dana litera oraz cztery cechy określające jaki procent wszystkich liter po prawej stronie od miejsca rozcięcia stanowi dana litera. Następne atrybuty określają czy dane dwie litery występują na dwóch danych pozycjach. W pliku *spliceDTrainKIS.dat* sekwencje DNA mają długość 15 nukleotydów, więc dla tego pliku wyznaczono 3360 cech. Wynika to z tego, że jedna z czterech liter może znajdować się na jednej z 15 pozycji, a druga z czterech liter może znajdować się na jednej z pozostałych 14 pozycji oraz z faktu, że na pierwszej danej pozycji może znajdować się jedna z czterech liter i na drugiej danej pozycji również może znajdować się jedna z czterech liter. W pliku *spliceATrainKIS.dat* sekwencje DNA mają

długość 90 nukleotydów, więc dla tego pliku wyznaczono 128160 cech. Ostatnie atrybuty określają czy na danej pozycji występuje jedna z czterech liter albo na drugiej danej pozycji występuje jedna z czterech danych liter – np. czy na pozycji 4 znajduje się litera „A" lub na pozycji 5 znajduje się litera „C". W pliku *spliceDTrainKIS.dat* sekwencje DNA mają długość 15 nukleotydów, więc dla tego pliku wyznaczono 3360 cech. Wynika to z tego, że na pierwszej danej pozycji (jednej z 15) może znajdować się jedna z czterech liter i na drugiej danej pozycji (jednej z pozostałych 14) może znajdować się jedna z czterech liter. W pliku *spliceATrainKIS.dat* natomiast sekwencje DNA mają długość 90 nukleotydów, więc dla tego pliku wyznaczono 128160 cech. Wynika to z faktu, iż na pierwszej danej pozycji (jednej z 90) może znajdować się jedna z czterech liter i na drugiej danej pozycji (jednej z pozostałych 89 pozycji) może znajdować się jedna z czterech liter.

Po wyznaczeniu macierzy cech, zastosowano trzy metody selekcji cech: na podstawie współczynnika korelacji Pearsona, metodą F-score (wyznaczenie wartości metryki Fishera) oraz metodą selekcji cech za pomocą t-testu.

Zastosowane metody selekcji cech

1. **Współczynnik korelacji Pearsona**

Dla celów selekcji cech wyznacza się współczynnik korelacji Pearsona wybranej cechy z klasą, który przyjmuje następującą postać:

$$R(i) = \frac{\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{m}(y_k - \bar{y})^2}}$$

Gdzie:

i – numer cechy,

m – liczba próbek,

x_{ki} – k-ta próbka i-tej cechy,

$\bar{x}_i$ – średnia wartość i-tej cechy

y_k – klasa k-tej próbki,

$\bar{y}$ – wartość średnia wektora klas.

Współczynnik korelacji przyjmuje wartości z przedziału [-1, 1], przy czym:

- im wartość bliższa 1 tym zależność jest silniejsza i dodatnia (jeżeli x rośnie to y rośnie),
- im wartość bliższa -1 tym zależność jest silniejsza i ujemna (jeżeli x rośnie to y maleje),
- współczynnik równy 0 oznacza brak związku liniowego pomiędzy zmiennymi.

Korzystając z obliczonych wartości współczynnika korelacji Pearsona dla każdego atrybutu, jako najbardziej istotne wybierane są atrybuty z jego najwyższymi, co do modułu, wartościami.

2. **Metryka Fishera**

Wartość metryki Fishera v_i dla i-tego atrybutu w przypadku klasyfikacji binarnej przedstawiona jest za pomocą równania:

$$v_i = \frac{(\mu_i^+ - \mu_i^-)^2}{(\sigma_i^+)^2 + (\sigma_i^-)^2}$$

Gdzie:

μ_i^+ - średnia wartość cechy x_i dla próbek należących do klasy 1,

$$\mu_i^+ = \frac{1}{N^+}\sum_{j=1}^{N+} x_{ij}$$

N^+ - liczba próbek należących do klasy 1,

$(\sigma_i^+)^2$ - wariancja wartości cechy x_i dla próbek należących do klasy 1:

$$(\sigma_i^+)^2 = \frac{1}{N^+ - 1}\sum_{j=1}^{N+} (x_{ij} - \mu_i^+)^2$$

Wartości μ_i^- i oraz $(\sigma_i^-)^2$ dla próbek należących do klasy 0 wyznaczane są analogicznie.

Korzystając z metryki Fishera wyznaczana jest wartość v_i dla każdego atrybutu, a następnie jako najbardziej istotne wybierane są atrybuty z najwyższymi wartościami.

3. Ocena atrybutów za pomocą t-testu

Jest to test statystyczny, sprawdzający czy wartości i-tego atrybutu różnią się między klasami. W teście tym wyznacza się dwa parametry: wartość t oraz liczbę stopni swobody oznaczoną jako $d.f.$:

$$t = \frac{\mu_i^+ - \mu_i^-}{\sqrt{\frac{(\sigma_i^+)^2}{N^+} + \frac{(\sigma_i^-)^2}{N^-}}}$$

$$d.f. = \frac{(\frac{(\sigma_i^+)^2}{N^+} + \frac{(\sigma_i^-)^2}{N^-})^2}{\frac{(\sigma_i^+)^4}{(N^+)^2(N^- - 1)} + \frac{(\sigma_i^-)^4}{(N^-)^2(N^+ - 1)}}$$

Następnie korzystając z rozkładu t-Studenta wyznaczony jest poziom istotności (*p-value*). Jeżeli wartość *p-value* jest mniejsza niż 0.05, przyjmuje się, że wartości atrybutu różnią się statystycznie znacząco pomiędzy klasami. Przy selekcji cech za pomocą t-testu, dla każdego atrybutu wyznacza się wartość istotności p. Atrybuty charakteryzujące się najniższym poziomem istotności uważane są za najważniejsze w analizowanym zadaniu.

Wyniki klasyfikacji dla poszczególnych algorytmów uczenia maszynowego oraz zestawów cech, ocena jakości selekcji cech (dokładność, czułość, specyficzność)

Ocenę jakości selekcji cech przeprowadzono dokonując klasyfikacji danych przy pomocy trzech klasyfikatorów: drzewa decyzyjnego, lasu losowego oraz metody wektorów nośnych.

Dla metody wektorów nośnych jako jądro zastosowano wielomian 5 rzędu, natomiast metoda lasu losowego i drzewa decyzyjnego została zastosowana z domyślnymi parametrami.

Dla pliku *spliceDTrainKIS.dat* policzono w sumie 6792 cech. Dla każdej z metod selekcji cech wybrano 10 % najlepszych atrybutów ze wszystkich wyznaczonych (680 cech). Następnie na ich podstawie dokonano klasyfikacji. Dane podzielono na dane trenujące – 80 % wszystkich próbek i dane testowe – 20 % wszystkich próbek. Macierze błędów utworzonych modeli przedstawiono w poniższych tabelach.

Wyniki klasyfikacji – las losowy

Pearson		Klasa rzeczywista	
		1	0
Klasa predykowana	1	190	30
	0	22	810

F-score		Klasa rzeczywista	
		1	0
Klasa predykowana	1	194	26
	0	23	809

t-test		Klasa rzeczywista	
		1	0
Klasa predykowana	1	159	61
	0	18	814

Las losowy		Dokładność	Czułość	Specyficzność
Dane testowe	Pearson	0,9506	0,9643	0,8962
	F-score	**0,9534**	**0,9689**	0,8940
	t-test	0,9249	0,9303	**0,8983**

Wyniki klasyfikacji – SVM

Pearson		Klasa rzeczywista	
		1	0
Klasa predykowana	1	191	29
	0	29	803

F-score		Klasa rzeczywista	
		1	0
Klasa predykowana	1	194	26
	0	26	806

t-test		Klasa rzeczywista	
		1	0
Klasa predykowana	1	180	40
	0	24	808

SVM		Dokładność	Czułość	Specyficzność
Dane testowe	Pearson	0,9449	0,9651	0,8682
	F-score	**0,9506**	**0,9688**	0,8818
	t-test	0,9392	0,9528	**0,8824**

Wyniki klasyfikacji – drzewo decyzyjne

Pearson		Klasa rzeczywista	
		1	0
Klasa predykowana	1	185	27
	0	35	805

F-score		Klasa rzeczywista	
		1	0
Klasa predykowana	1	194	26
	0	33	799

t-test		Klasa rzeczywista	
		1	0
Klasa predykowana	1	165	28
	0	55	804

Drzewo decyzyjne		Dokładność	Czułość	Specyficzność
Dane testowe	Pearson	0,9411	0,9583	**0,8726**
	F-score	**0,9439**	**0,9685**	0,8546
	t-test	0,9211	0,9360	0,8549

Ranking najlepszych 10 cech dla różnych metod selekcji cech dla różnych metod selekcji cech.

Pearson:

```
litera 1: G na miejscu: 8 lub litera 2: A na miejscu: 12
litera 1: T na miejscu: 9 i litera 2: G na miejscu: 3
litera 1: G na miejscu: 6 i litera 2: A na miejscu: 14
litera 1: A na miejscu: 9 i litera 2: G na miejscu: 1
litera 1: A na miejscu: 6 lub litera 2: G na miejscu: 14
litera 1: G na miejscu: 6 i litera 2: G na miejscu: 15
litera 1: A na miejscu: 5 i litera 2: G na miejscu: 13
litera 1: C na miejscu: 6 i litera 2: A na miejscu: 12
litera 1: C na miejscu: 9 i litera 2: G na miejscu: 15
litera 1: G na miejscu: 10 i litera 2: A na miejscu: 2
```

F-score:

```
litera 1: A na miejscu: 11 lub litera 2: A na miejscu: 10
litera 1: G na miejscu: 12 i litera 2: A na miejscu: 11
litera 1: G na miejscu: 7 lub litera 2: A na miejscu: 11
litera 1: G na miejscu: 7 lub litera 2: G na miejscu: 12
```

```
litera 1: A na miejscu: 10 lub litera 2: A na miejscu: 11
litera 1: A na miejscu: 11 i litera 2: G na miejscu: 12
litera 1: A na miejscu: 11 lub litera 2: G na miejscu: 7
litera 1: G na miejscu: 12 lub litera 2: G na miejscu: 7
litera 1: A na miejscu: 10 lub litera 2: G na miejscu: 12
litera 1: G na miejscu: 12 lub litera 2: A na miejscu: 10
```

t-test:

```
litera 1: C na miejscu: 5 lub litera 2: G na miejscu: 12
litera 1: A na miejscu: 6 lub litera 2: A na miejscu: 11
litera 1: A na miejscu: 6 lub litera 2: G na miejscu: 12
litera 1: G na miejscu: 7 lub litera 2: A na miejscu: 10
litera 1: G na miejscu: 7 lub litera 2: A na miejscu: 11
litera 1: G na miejscu: 7 lub litera 2: G na miejscu: 12
litera 1: G na miejscu: 7 lub litera 2: T na miejscu: 13
litera 1: A na miejscu: 10 lub litera 2: G na miejscu: 7
litera 1: A na miejscu: 10 lub litera 2: A na miejscu: 11
litera 1: A na miejscu: 10 lub litera 2: G na miejscu: 12
```

Dla pliku *spliceATrainKIS.dat* policzono w sumie 256692 cech. Dla każdej z metod selekcji cech wybrano 0,3% najlepszych atrybutów ze wszystkich wyznaczonych (771 cech). Następnie na ich podstawie dokonano klasyfikacji. Dane podzielono na dane trenujące – 80% wszystkich próbek i dane testowe – 20% wszystkich próbek. Macierze błędów utworzonych modeli przedstawiono w poniższych tabelach.

Wyniki klasyfikacji – las losowy

Pearson		Klasa rzeczywista	
		1	0
Klasa predykowana	1	83	129
	0	7	939

F-score		Klasa rzeczywista	
		1	0
Klasa predykowana	1	164	48
	0	27	919

t-test		Klasa rzeczywista	
		1	0
Klasa predykowana	1	174	38
	0	30	916

Las losowy		Dokładność	Czułość	Specyficzność
Dane testowe	Pearson	0,8826	0,8792	**0,9222**
	F-score	0,9352	0,9504	0,8586
	t-test	**0,9413**	**0,9602**	0,8529

Wyniki klasyfikacji – SVM

Pearson		Klasa rzeczywista	
		1	0
Klasa predykowana	1	43	169
	0	4	942

F-score		Klasa rzeczywista	
		1	0
Klasa predykowana	1	147	65
	0	23	923

t-test		Klasa rzeczywista	
		1	0
Klasa predykowana	1	107	105
	0	15	931

SVM		Dokładność	Czułość	Specyficzność
Dane testowe	Pearson	0,8506	0,8479	**0,9149**
	F-score	**0,9240**	**0,9342**	0,8647
	t-test	0,8964	0,8986	0,8770

Wyniki klasyfikacji – drzewo decyzyjne

Pearson		Klasa rzeczywista	
		1	0
Klasa predykowana	1	88	124
	0	49	897

F-score		Klasa rzeczywista	
		1	0
Klasa predykowana	1	147	65
	0	44	902

t-test		Klasa rzeczywista	
		1	0
Klasa predykowana	1	147	65
	0	44	902

Drzewo decyzyjne		Dokładność	Czułość	Specyficzność
Dane testowe	Pearson	0,8506	0,8786	0,6423
	F-score	**0,9059**	**0,9328**	**0,7696**
	t-test	**0,9059**	**0,9328**	**0,7696**

Ranking najlepszych 10 cech dla różnych metod selekcji cech dla różnych metod selekcji cech.

Pearson:

```
litera 1: T na miejscu: 62 lub litera 2: G na miejscu: 35
litera 1: G na miejscu: 65 lub litera 2: A na miejscu: 32
litera 1: C na miejscu: 62 lub litera 2: G na miejscu: 36
litera 1: G na miejscu: 65 lub litera 2: G na miejscu: 32
litera 1: G na miejscu: 61 lub litera 2: G na miejscu: 36
litera 1: G na miejscu: 65 lub litera 2: G na miejscu: 31
litera 1: G na miejscu: 63 lub litera 2: G na miejscu: 35
litera 1: G na miejscu: 65 lub litera 2: G na miejscu: 33
litera 1: A na miejscu: 63 lub litera 2: G na miejscu: 35
litera 1: G na miejscu: 65 lub litera 2: A na miejscu: 33
```

F-score:

```
litera 1: A na miejscu: 65 lub litera 2: G na miejscu: 68
litera 1: G na miejscu: 68 lub litera 2: A na miejscu: 65
litera 1: G na miejscu: 65 lub litera 2: G na miejscu: 68
litera 1: G na miejscu: 68 lub litera 2: G na miejscu: 65
litera 1: G na miejscu: 66 lub litera 2: G na miejscu: 68
litera 1: G na miejscu: 68 lub litera 2: G na miejscu: 66
litera 1: G na miejscu: 64 lub litera 2: G na miejscu: 68
litera 1: G na miejscu: 68 lub litera 2: G na miejscu: 64
litera 1: A na miejscu: 66 lub litera 2: G na miejscu: 68
litera 1: G na miejscu: 68 lub litera 2: A na miejscu: 66
```

t-test:

```
litera: G na miejscu: 68
litera 1: G na miejscu: 68 i litera 2: A na miejscu: 69
litera 1: G na miejscu: 68 lub litera 2: C na miejscu: 69
litera 1: G na miejscu: 68 lub litera 2: T na miejscu: 69
litera 1: G na miejscu: 68 lub litera 2: G na miejscu: 69
litera 1: G na miejscu: 68 lub litera 2: A na miejscu: 70
litera 1: G na miejscu: 68 lub litera 2: C na miejscu: 70
litera 1: G na miejscu: 68 i litera 2: G na miejscu: 70
litera 1: A na miejscu: 69 i litera 2: G na miejscu: 68
litera 1: G na miejscu: 70 i litera 2: G na miejscu: 68
```

Wnioski

Dla pliku *spliceDTrainKIS.dat* największą dokładność i czułość (równą odpowiednio 95,34 % i 96,89 %), uzyskano przy pomocy modelu lasu losowego dla cech wybranych przy użyciu metody F-score. Największą specyficzność (równą 89,83 %) uzyskano natomiast również przy zastosowaniu modelu lasu losowego, jednakże przy użyciu metody t-test.

W przypadku analizy danych z pliku *spliceATrainKIS.dat* najwyższą dokładnością i czułością (równymi odpowiednio 94,13 % i 96,02 %) charakteryzował się również model lasu losowego utworzony na podstawie cech wyselekcjonowanych metodą t-test. Najwyższa specyficzność została uzyskana również dla modelu lasu losowego, jednakże przy zastosowaniu do selekcji cech wartości współczynnika korelacji Pearsona.

Dla dwóch problemów klasyfikacji i trzech zastosowanych algorytmów uczenia maszynowego, pięć na sześć razy największą dokładność i czułość otrzymano dla modeli utworzonych na cechach wybranych przy użyciu metody F-score. Wyniki klasyfikacji przy zastosowaniu wszystkich trzech metod selekcji cech nie odbiegają od siebie w znacznym stopniu. Wszystkie otrzymane wyniki dokładności są powyżej 85 %.

Na podstawie rankingu dziesięciu najbardziej istotnych cech dla wszystkich trzech metod selekcji cech, można wnioskować, iż istotne dla problemu klasyfikacji fragmentów DNA jest budowa fragmentu w okolicy miejsca rozcięcia.