Ćwiczenie 3

Selekcja atrybutów systemu decyzyjnego metodą Fishera

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	a16	a17	a18	a19	a20	d
1	2	3	4	5	6	7	8	9	4	2	2	4	4	5	3	2	3	4	3	1
1	2	3	4	5	6	7	8	2	4	2	2	3	4	5	3	2	3	4	3	1
1	2	3	4	5	6	7	8	2	4	2	2	1	4	5	3	2	3	4	3	1
1	2	3	4	5	6	7	8	2	4	2	2	7	4	5	3	2	3	4	3	1
1	2	3	4	5	6	7	8	6	4	4	2	8	4	5	3	2	3	4	3	2
1	2	3	4	5	6	7	8	1	4	4	2	9	4	5	3	2	3	4	3	2
1	2	3	4	5	6	7	8	1	4	4	2	8	4	5	3	2	3	4	3	2
1	2	3	4	5	6	7	8	1	4	4	2	9	4	5	3	2	3	4	3	2

Zadanie do wykonania

1) Tworzymy na pulpicie katalog w formacie Imię_nazwisko, w którym umieszczamy wszystkie pliki związane z ćwiczeniem.

2) Czytamy teorię związaną z selekcją atrybutów metodą Fishera, w razie problemów ze zrozumieniem, analizujemy przykłady na kartce.

3) Generujemy system decyzyjny treningowy za pomocą programu ds_generator.exe.

4) Dla każdego atrybutu warunkowego powstałego systemu, liczymy stopień w jakim ten atrybut separuje poszczególne klasy decyzyjne centralne od pozostałych klas, stosujemy metodę Fishera. Zadanie polega na implementacji algorytmu w wybranym języku programowania.

5) Na koniec wskazujemy cztery atrybuty, które najlepiej separowały klasy centralne od reszty klas i tworzymy z nich nowy system decyzyjny.

5) W przypadku implementacji w języku C++, ułatwieniem może być użycie programu znajdującego się na stronie http://wmii.uwm.edu.pl/∼artem w zakładce Dydaktyka/Sztuczna Inteligencja.

Selekcja atrybutów Metodą Fishera - teoria

Metoda Fishera może być wykorzystana do oszacowania stopnia w jakim dany atrybut separuje pewną wybraną klasę centralną od reszty klas decyzyjnych. Czym stopień separacji jest większy tym klasa centralna jest lepiej odseparowana od pozostałych klas.

Niech będzie dany systemu decyzyjny (U, A, d), gdzie U jest zbiorem obiektów, A

jest zbiorem atrybutów, (dla których wyliczamy stopnie separacji), d jest atrybutem decyzyjnym (diagnozą postawioną przez eksperta),

Dla systemu decyzyjnego (U, A, d), gdzie $U = \{u_1, u_2, ..., u_n\}, A = \{a_1, a_2, ..., a_m\}$, $d \notin A$, wyliczamy stopień separacji atrybutów $a \in A$ dla klas decyzyjnych c_i, $i = 1, 2, ..., k$ w następujący sposób. Przyjmujemy, że,

$$S^{c_i}(a) = \frac{(\overline{C}_i^a - \hat{C}_i^a)^2}{Z_{\overline{C}_i^{a2}} + Z_{\hat{C}_i^{a2}}}, a \in A. \tag{1}$$

gdzie,

$$C_i^a = \{a(u) : u \in U \ and \ d(u) = c_i\}. \tag{2}$$

$$\overline{C}_i^a = \frac{\{\sum a(u) : u \in U \ and \ d(u) = c_i\}}{card\{C_i^a\}}, \hat{C}_i^a = \frac{\{\sum a(v) : v \in U \ and \ d(v) \neq c_i\}}{card\{U\} - card\{C_i^a\}}. \tag{3}$$

$$Z_{\overline{C}_i^{a2}} = \frac{\sum_{a(u) \in C_i^a} (a(u) - \overline{C}_i^a)^2}{card\{C_i^a\}}, Z_{\hat{C}_i^{a2}} = \frac{\sum_{a(v) \in U \setminus C_i^a} (a(v) - \hat{C}_i^a)^2}{card\{U\} - card\{C_i^a\}} \tag{4}$$

Gdy liczenie stopnia separacji $S^{c_i}(a)$ dla wszystkich atrybutów $a \in A$ i klas decyzyjnych c_i dobiegnie końca, atrybuty sortujemy w sposób malejący na podstawie ich stopnia separacji $S^{c_i}(a)$,

$$S_1^{c_1}(a) > S_2^{c_1}(a) > ... > S_{card\{A\}}^{c_1}(a)$$

$$S_1^{c_2}(a) > S_2^{c_2}(a) > ... > S_{card\{A\}}^{c_2}(a)$$

$$\vdots$$

$$S_1^{c_k}(a) > S_2^{c_k}(a) > ... > S_{card\{A\}}^{c_k}(a)$$

Finalnie, wybieramy ustaloną liczbę atrybutów z posortowanej listy stosując następującą procedurę,

Procedura
Dane wejściowe
$A' \leftarrow \emptyset$
$iter \leftarrow 0$
for i=1,2,...,card{A} **do**
  for j=1,2,...,k **do**
    $S^{c_j}(a) = S_i^{c_j}(a)$
    if $a \notin A'$ **then**
      $A' \leftarrow a$
      $iter \leftarrow iter + 1$
      if *iter = ustalona liczba najlepszych atrybutów* **then**
        BREAK

```
        end if
      end if
    end for
    if iter = ustalona liczba najlepszych atrybutów then
      BREAK
    end if
  end for
  return A'
```

Przykład selekcji atrybutów metodą Fishera

Przyjmując, że nasz system decyzyjny (U,A,d) jest postaci,

Tabela 1: System decyzyjny (U, A, d)

	a_1	a_2	a_3	a_4	d
u_1	2	1	2	1	1
u_2	3	2	3	3	1
u_3	1	5	1	2	2
u_4	6	7	3	8	2
u_5	4	5	5	6	3
u_6	5	2	8	3	3

$A = \{a_1, a_2, a_3, a_4\}$, $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$
$d \in C = \{1, 2, 3\}$
$C_1 = \{u_1, u_2\}$, $C_2 = \{u_3, u_4\}$, $C_3 = \{u_5, u_6\}$
$card\{C_1\} = 2$, $card\{C_2\} = 2$, $card\{C_3\} = 2$

Zacznijmy od separacji klasy decyzyjnej 1

$\overline{C}_1^{a_1} = \frac{\{\sum a_1(u):u \in U \ \ and \ \ d(u)=1\}}{card\{C_1\}}$
$\overline{C}_1^{a_1} = \frac{2+3}{2} = 2.5$
$\hat{C}_1^{a_1} = \frac{\{\sum a_1(v):v \in U \ \ and \ \ d(v) \neq 1\}}{card\{U\}-card\{C_1\}}$
$\hat{C}_1^{a_1} = \frac{1+6+4+5}{6-2} = 4$
$Z_{\overline{C}_1^{a}{}^2} = \frac{\sum_{a(u) \in C_1^a} (a(u) - \overline{C}_1^a)^2}{card\{C_1\}}$
$Z_{\overline{C}_1^{a}{}^2} = \frac{(2-2.5)^2+(3-2.5)^2}{2} = \frac{1}{4}$
$Z_{\hat{C}_1^{a}{}^2} = \frac{\sum_{a(v) \in U \setminus C_1^a} (a(v) - \hat{C}_1^a)^2}{card\{U\}-card\{C_1\}}$
$Z_{\hat{C}_1^{a}{}^2} = \frac{(1-4)^2+(6-4)^2+(4-4)^2+(5-4)^2}{6-2} = 3\frac{1}{2}$
$S^{c_1}(a_1) = \frac{(\overline{C}_1^a - \hat{C}_1^a)^2}{Z_{\overline{C}_1^{a}{}^2} + Z_{\hat{C}_1^{a}{}^2}}, a \in A$
$S^{c_1}(a_1) = \frac{(2.5-4)^2}{\frac{1}{4}+3\frac{1}{2}} = \frac{9}{15} = 0.6$

Analogicznie dla kolejnych atrybutów,
$S^{c_1}(a_2) == 3.07273$
$S^{c_1}(a_3) == 0.441441$

$S^{c_1}(a_4) == 1.13084$

Sortujemy atrybuty zależnie od stopnia separacji w sposób malejący,
$S^{c_1}(a_2) == 3.07273$
$S^{c_1}(a_4) == 1.13084$
$S^{c_1}(a_1) = 0.6$
$S^{c_1}(a_3) == 0.441441$

Separujemy klasę 2
$S^{c_2}(a_1) == 0$
$S^{c_2}(a_2) == 3.766923$
$S^{c_2}(a_3) == 1$
$S^{c_2}(a_4) == 0.251282$

Po posortowaniu mamy,
$S^{c_2}(a_2) == 3.766923$
$S^{c_2}(a_3) == 1$
$S^{c_2}(a_4) == 0.251282$
$S^{c_2}(a_1) == 0$

Separujemy klasę 3
$S^{c_3}(a_1) == 0.6$
$S^{c_3}(a_2) == 0.007874402$
$S^{c_3}(a_3) == 6.14894$
$S^{c_3}(a_4) == 0.105263$

Po posortowaniu mamy
$S^{c_3}(a_3) == 6.14894$
$S^{c_3}(a_1) == 0.6$
$S^{c_3}(a_4) == 0.105263$
$S^{c_3}(a_2) == 0.007874402$

Tablica numerów atrybutów najlepiej separujących poszczególne klasy decyzyjne jest postaci
Dla klasy centralnej c_1: a_2 a_4 a_1 a_3
Dla klasy centralnej c_2: a_2 a_3 a_4 a_1
Dla klasy centralnej c_3: a_3 a_1 a_4 a_3

Wybieramy trzy najlepsze atrybuty na podstawie naszej procedury wyboru.

Klasę centralną c_1 najlepiej separuje atrybut a_2 - trafia jako pierwszy atrybut do naszego nowego systemu decyzyjnego,

Klasę centralną c_2 również najlepiej separuje atrybut a_2 - jednak mamy już ten atrybut w nowym systemie.

Klasę centralną c_3 najlepiej separuje a_3 - trafia do nowego systemu i mamy już dwa atrybuty.

Przechodzimy do szukania atrybutu trzeciego, zaczynamy znów od klasy centralnej c_1 i bierzemy kolejny najlepiej ją separujący atrybut czyli a_4, trafia do naszego nowego systemu i kończymy wyszukiwanie, ponieważ mamy ustaloną liczbę atrybutów najlepiej separujących klasy centralne.

Nasz nowy system decyzyjny jest postaci

Tabela 2: System decyzyjny po selekcji cech (U, A, d)

	a_2	a_3	a_4	d
u_1	1	2	1	1
u_2	2	3	3	1
u_3	5	1	2	2
u_4	7	3	8	2
u_5	5	5	6	3
u_6	2	8	3	3