

## Zadanie «Selekcja»

nr IKU \_\_\_\_\_

liczba punktów \_\_\_\_\_ / 10

**Za zadanie można otrzymać 10 punktów.**

Klasyfikacja to rodzaj algorytmu, który przydziela obserwacje do jednej ze zdefiniowanych klas, bazując na cechach (atrybutach) tej obserwacji. Przykładem jednego z najbardziej znanych zbiorów danych dotyczących klasyfikacji jest zbiór Iris. Zbiór ten zawiera cztery cechy i nazwę klasy:

długość działki kwiatu	szerokość działki kwiatu	długość płatka kwiatu	szerokość płatka kwiatu	nazwa gatunku (klasa)
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3,0	1,4	0,2	Iris-setosa
7,0	3,2	4,7	1,4	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,3	3,3	6,0	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica

Do rozwiązywania problemu klasyfikacji używane są tzw. klasyfikatory, które po przeprowadzeniu procesu uczenia na danych treningowych są w stanie lepiej bądź gorzej przyporządkowywać nowe, nieznane wcześniej obserwacje do jednej z klas. Aby ocenić skuteczność klasyfikatora korzysta się z dodatkowego zbioru danych – zbioru testowego, do którego należą obserwacje niebiorące udziału w procesie uczenia, ale dla których przynależność do klas jest znana. Po sklasyfikowaniu takich obserwacji porównuje się klasę, do której dana obserwacja została przyporządkowana przez klasyfikator z klasą rzeczywistą danej obserwacji. Jeśli są one takie same, to mówi się, że obserwacja została dobrze sklasyfikowana. Zawsze dąży się do maksymalizacji dobrze sklasyfikowanych obiektów. Wynikiem tego procesu jest liczba (skuteczność klasyfikacji), która określa stosunek liczby dobrze przyporządkowanych obserwacji do liczby wszystkich obserwacji (dlatego wartość ta jest w zakresie od 0 do 1).

W wielu przypadkach zbiory danych zawierają setki, tysiące, a czasem nawet miliony cech (atrybutów). Wiele z nich nie ma żadnego wpływu na skuteczność klasyfikacji, a czasem wprowadza jedynie szum informacyjny, który nawet obniża skuteczność. Z tego względu stosuje się metody selekcji cech, których celem jest zredukowanie liczby cech bez obniżania skuteczności klasyfikacji.

## Zadanie

Należy zaprojektować algorytm rozwiązujący problem selekcji cech. Rozwiązanie należy podać w pseudokodzie, C/C++, Pascalu lub Javie. Założyć, że do dyspozycji są:

- liczba wszystkich cech (liczbaCech),
- zbiór danych treningowych (zbiórTreningowy),
- zbiór danych testowych (zbiórTestowy),

- na obu zbiorach można wykonać następujące operacje:
  - `dodajCechę(zbiór, X)` – dodaje cechę o numerze  $X$  do wskazanego zbioru,
  - `usuńCechę(zbiór, X)` – usuwa cechę o numerze  $X$  ze wskazanego zbioru,
  - `dodajWszystkieCechy(zbiór)` – dodaje wszystkie cechy do wskazanego zbioru,
  - `usuńWszystkieCechy(zbiór)` – usuwa wszystkie cechy ze wskazanego zbioru,
  - `kopiujZbiór(zbiór)` – kopiuje wskazany zbiór danych wraz z aktualnym zestawem cech,
- algorytm klasyfikujący, który można uczyć zbiorem treningowym i sprawdzać skuteczność klasyfikacji zbiorem testowym:
  - `uczKlasyfikator(zbiór)` – przeprowadza proces uczenia wskazanym zbiorem,
  - `obliczSkutecznośćKlasyfikacji(zbiór)` – zwraca skuteczność klasyfikacji wskazanego zbioru.

W rozwiązaniu oceniana będzie poprawność oraz jego optymalność. Inne operacje niż wcześniej wymienione nie są dopuszczalne.