Kurs K1 (Wprowadzenie do eksploracji danych) – Laboratorium 6

Selekcja cech – Metody rankingowe

Przykład: Utworzenie rankingu cech dla zbioru „iris" różnymi metodami

Instalujemy potrzebne pakiety i importujemy biblioteki

```
> install.packages("corrplot")
> install.packages("caret")
> install.packages("ggplot2")
> install.packages("mlbench")
> install.packages("e1071")
> install.packages("FSelector")## dla metod opartych na entropii – wymaga
                                                          zainstalowanej javy!
> library(corrplot)
> library(caret)
> library(mlbench)
> library(caret)
> library(FSelector)
```

Wczytujemy dane ze zbioru "iris":

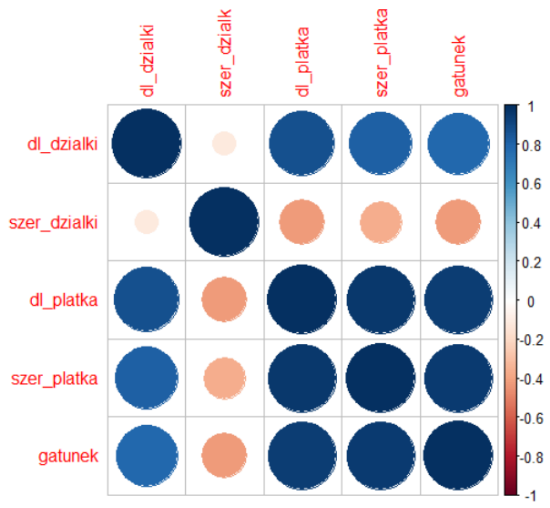
```
> library(datasets)
> data(iris)                              ## wczytanie danych
> colnames(iris)<-c("dl_dzialki","szer_dzialki","dl_platka","szer_platka","gatunek")
                                  ## polskie nazwy cech
> View(iris)
```

Przygotowanie danych (zmienną celu musimy w tym zadaniu przekształcić do typu liczbowego):

```
> iris$gatunek<-as.character(iris$gatunek)
> iris$gatunek[iris$gatunek=="setosa"]<-1
> iris$gatunek[iris$gatunek=="versicolor"]<-2
> iris$gatunek[iris$gatunek=="virginica"]<-3
> iris$gatunek<-as.numeric(as.character(iris$gatunek))
> View(iris)
```

Wyznaczamy macierz korelacji

```
> res.cor <- cor(iris)
> View(res.cor)
> corrplot(res.cor                        ## wykres macierzy korelacji
```



Funkcja drukująca ranking cech na podstawie listy indeksów cech

```
> CreateRanking<-function(w)
{
   weightframe <- data.frame(features = row.names(w), w, row.names=NULL)
   weightframewithranks <- transform(weightframe, importance = rank(-
attr_importance, ties.method = "first"))
   ranks <- weightframewithranks[,-2]
   ranks
}
```

1) Pearson correlation filter

```
> x <- subset(iris, select = -gatunek)
> weights1 <- cor(x, iris$gatunek, method = "pearson") ## orelacja ze znakiem +/-
> print(weights1)
> weights1 <- abs(weights1)                              siła korelacji (bez znaku)
> print(weights1)
> colnames(weights1)<-c("attr_importance") ##dodajemy nazwę kolumny, żeby można
                                          ## było wykorzystać funkcję CreateRanking
> wynik1 <- CreateRanking(weights1)
> wynik1
```

2) FILTER METHOD: Chi-squared filter

```
> weights2 <- chi.squared(gatunek~., iris)
> print(weights2)
> CreateRanking(weights2)
```

3) Information gain

```
> weights3 <- information.gain(gatunek~., iris)
> print(weights3)
> CreateRanking(weights3)
```

4) Gain ratio

```
> weights4 <- gain.ratio(gatunek~., iris)
> print(weights4)
> CreateRanking(weights4)
```

5) Symmetrical uncertainty

```
> weights5 <- symmetrical.uncertainty(gatunek~., iris)
> print(weights5)
> CreateRanking(weights5)
```

6) OneR

```
> weights6 <- oneR(gatunek~., iris)
> print(weights6)
> CreateRanking(weights6)
```

7) Relief

```
> set.seed(7)
> iris$gatunek=as.factor(iris$gatunek)    ##regresja (zamiast klasyfikacji)
> weights7 <- relief(gatunek~., iris, neighbours.count = 10, sample.size = 10)
      #docelowo większy sample.size (ale długie obliczenia)
> print(weights7)
> CreateRanking(weights7)
```

Zadanie do wykonania

1. Sprawdź na zbiorze „iris” działanie kodu R przedstawionego w niniejszym opracowaniu.
2. Następnie wykonaj podobne instrukcje dla zbioru danych *wine* (dla zbioru bez obserwacji brakujących i bez punktów oddalonych). Odpowiedz na poniższe pytania/polecenia. W pliku tekstowym przygotuj odpowiednie informacje (wklej polecenie R oraz wyniki):
   a) Utwórz i wyświetl macierz korelacji.
   b) Oblicz i wydrukuj indeksy oraz rangi cech stosując po kolei siedem metod filtracyjnych.
   c) Utwórz tablicę 2D, gdzie liczba kolumn = 7, a liczba wierszy = liczba cech w zbiorze danych. Zapisz do tablicy rangi cech dla poszczególnych metod „kolumnami” (można wykorzystać funkcję *cbind*), gdzie:
      - nazwy kolumn to nazwy metod rankingowych,
      - nazwy wierszy to nazwy cech.
   d) Do tablicy 2D dodaj nową kolumnę ze średnią wartością rangi dla danej cechy.
   e) Posortuj cechy według wartości średniej rangi.
   f) Przedyskutuj wyniki.