Selekcja cech

Należy otworzyć skonstruowany w ramach poprzedniej listy zbiór danych *XXXXXXL3 1.arff* poprzez udostępnione przez *Weke* GUI. Należy zapoznać się z działaniem modułu selekcji cech (Zakładka *Select Attributes*).

Metody selekcji cech wykorzystujące entropie
Typowymi metodami stosowanymi do selekcji cech są algorytmy wykorzystujące pojecie entropii:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

oraz entropii warunkowej:

$$H(X|Y) = \sum_{y \in Y} p(y) H(X|y)$$

W *Wece* wyróżnić można dwie metody które wykorzystują entropie do oceny istotności atrybutów: **GainRatioAttributeEval**, oraz **InfoGainAttributeEval**. Pierwszy z nich bada istotność atrybutów ze względu współczynnik *GainRatio* definiowany w następujący sposób:

$$GainRatio(Class, Attribute) = \frac{H(Class) - H(Class|Attribute)}{H(Attribute)}$$

natomiast drugi z nich wykorzystuje tzn. *InfoGain*:

$$GainRatio(Class, Attribute) = H(Class) - H(Class|Attribute)$$

Opisane metody wykonują ocenę każdego atrybutu ze względu na przyjęte kryterium niezależnie, dla każdego z atrybutów osobno.

Oba zadania zostaną wykonane na pliku *XXXXXXL3 1.arff*.

Ćwiczenie 1 Należy dokonać dyskretyzacji zmiennych numerycznych z wykorzystaniem filtra pracującego w trybie nadzorowanym. W dalszej kolejności należy zapoznać sie z działaniem filtrów do selekcji cech **GainRatioAttributeEval**, oraz **InfoGainAttributeEval**. Należy wybrać cechy dla których zarówno *GainRatio*, jak i *InfoGain* przyjmują wartości wyższe niż 0.001. Należy uszeregować atrybuty rosnąco względem *GainRatio* i zbiór po procesie selekcji i uszeregowaniu zapisać jako *XXXXXXL4 1.arff*

Ćwiczenie 2. Należy własnoręcznie (bez wykorzystywania klas **GainRatioAttributeEval**, **InfoGainAttributeEval**) zaimplementować metodę **GainRatioAttributeEval** i zweryfikować jej działanie na zbiorze *XXXXXXL3 1.arff*.

Należy zidentyfikować podstawę logarytmu, jaką wykorzystuje implementacja **GainRatioAttributeEval** w *Wece* zadając jej wartość jako parametr programu