

Eksploracja danych Projekt

Marcin Fabrykowski

28 czerwca 2013

Spis treści

1	Opis problemu	3
2	Klasyfikator	3
3	Wykorzystywane technologie	3
4	Działanie programu	3

1 Opis problemu

Celem projektu jest opracowanie klasyfikatora drzewiastego do klasyfikacji wina na podstawie jego parametrów.

Jako zbiór do nauki, został wykorzystany zbiór

<http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

2 Klasyfikator

Klasyfikator jest operatorem potrafiącym, po wcześniejszym 'nauczeniu', przewidzieć klasyfikowaną wartość na podstawie badanych atrybutów. W naszym przypadku wartością klasyfikowaną jest jakość wina.

Podczas tworzenia klasyfikatora, mamy możliwość określenia maksymalnej głębokości drzewa (`max_depth`) oraz ilości iteracji przycinania drzewa (`m_pruning`).

3 Wykorzystywane technologie

Projekt został napisany w języku Python.

Do działania, wymagane są następujące biblioteki:

- SciPy
- Orange
- pygraphviz

Orange jest biblioteką wspomagającą operacje związane z eksploracją danych. Udostępnia ona narzędzie do przeprowadzania klasyfikacji, regresji, asocjacji oraz kategoryzacji. Orange pozwala na wyeksportowanie danych o drzewach w formacie `.dot`, które mogą zostać sparsowane przez program `graphviz`. `Pygraphviz` jest bindingiem do `graphviza` i pozwala na generowanie diagramów z plików `.dot`.

4 Działanie programu

Na początku następuje sprawdzenie czy w systemie znajdują się wymagane moduły oraz załadowanie ich.

Następnie następuje wczytanie danych wejściowych z pliku `csv` do tabeli systemu Orange.

Po wczytaniu danych, następuje wygenerowanie klasyfikatora dla wczytanych

danych oraz ustawionych parametrów tego klasyfikatora.

Mając wygenerowane drzewo klasyfikujące, zapisujemy je do pliku w formacie .dot.

Wygenerowany plik .dot zostaje przetworzony przez moduł pygraphviz w celu wygenerowania pliku graficznego PNG przedstawiającego otrzymany klasyfikator.

W ostatnim kroku obliczamy statystyki dla naszego klasyfikatora