APPLICATION OF THE RECURSIVE FEATURE ELIMINATION AND THE RELAXED LINEAR SEPARABILITY FEATURE SELECTION ALGORITHMS TO GENE EXPRESSION DATA ANALYSIS

Joanna Gościk, Tomasz Łukaszuk

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: Most of the commonly known feature selection methods focus on selecting appropriate predictors for image recognition or generally on data mining issues. In this paper we present a comparison between widely used Recursive Feature Elimination (RFE) with resampling method and the Relaxed Linear Separability (RLS) approach with application to the analysis of the data sets resulting from gene expression experiments. Different types of classification algorithms such as K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Random Forests (RF) are exploited and compared in terms of classification accuracy with optimal set of genes treated as predictors selected by either the RFE or the RLS approaches. Ten-fold cross-validation was used to determine classification accuracy.

Keywords: gene expression analysis, feature selection, classification

REKURENCYJNA ELIMINACJA CECH Z WALIDACJĄ ORAZ RELAKSACJA LINIOWEJ SEPAROWALNOŚCI JAKO METODY SELEKCJI CECH DO ANALIZY ZBIORÓW DANYCH ZAWIERAJĄCYCH WARTOŚCI EKSPRESJI GENÓW

Streszczenie Zdecydowana większość znanych metod selekcji cech skupia się na wyborze odpowiednich predyktorów dla takich zagadnień jak rozpoznawanie obrazów czy też ogólnie eksploracji danych. W publikacji prezentujemy porównanie pomiędzy powszechnie

Advances in Computer Science Research, vol. 10, pp. 39-52, 2013.

stosowaną metodą Rekurencyjnej Eliminacji Cech z walidacją (ang. Recursive Feature Elimination - RFE) a metodą stosującą podejście Relaksacji Liniowej Separowalności (ang. Relaxed Linear Separability - RLS) z zastosowaniem do analizy zbiorów danych zawierających wartości ekspresji genów. W artykule wykorzystano różne algorytmy klasyfikacji, takie jak K-Najbliższych Sąsiadów (ang. K-Nearest Neighbours - KNN), Maszynę Wektorów Wspierających (ang. Support Vector Machines - SVM) oraz Lasy Losowe (ang. Random Forests - RF). Porównana została jakość klasyfikacji uzyskana przy pomocy tych algorytmów z optymalnym zestawem cech wygenerowanym z wykorzystaniem metody selekcji cech RFE bądź RLS. W celu wyznaczenia jakości klasyfikacji wykorzystano 10-krotną walidację krzyżową.

Słowa kluczowe: analiza ekspresji genów, selekcja cech, klasyfikacja

Artykuł zrealizowano w ramach projektu badawczego N N519 657940 finansowanego przez Ministerstwo Nauki i Szkolnictwa Wyższego oraz pracy badawczej S/WI/2/2013 Wydziału Informatyki PB.