FEATURE SELECTION METHODS BASED ON MINIMIZATION OF CPL CRITERION FUNCTIONS

Tomasz Łukaszuk

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

Abstract: The feature selection is a method of data analysis commonly used as a preliminary step in the techniques of classification and pattern recognition. It is particularly important in situations when data are represented in high-dimensional feature space. Examples of these are collections of bioinformatics data, particularly data obtained from DNA microarrays. The paper presented two methods of feature selection based on minimizing the CPL criterion function: basic SEKWEM/GENET method, in which the selection of features is done in conjunction with the construction of a linear classifier separating objects from different decision classes, and the RLS method extending the primary method by linear separability relaxation stage in order to obtain a subset of features with better generalization ability. The results of the SEKWEM/GENET and RLS methods were confronted with the results obtained from other common feature selection methods in application to the state of the art microarray data sets.

Keywords: feature selection, CPL criterion function, SEKWEM/GENET algorithm, RLS method

METODY SELEKCJI CECH BAZUJĄCE NA MINIMALIZACJI FUNKCJI KRYTERIALNYCH TYPU CPL

Streszczenie Selekcja cech jest metodą analizy danych powszechnie stosowaną jako wstępny krok w technikach klasyfikacji czy rozpoznawania wzorców. Ma ona szczególne znaczenie w sytuacji gdy dane reprezentowane są w wysoko wymiarowej przestrzeni cech. Przykładem takich danych są zbiory bioinformatyczne, a w szczególności dane uzyskane na podstawie mikromacierzy DNA. W pracy przedstawione zostały dwie metody selekcji cech bazujące na minimalizacji funkcji kryterialnych typu CPL: podstawowa metoda SEKWEM/GENET, w której selekcja cech dokonywana jest w połączeniu z budową liniowego klasyfikatora separującego obiekty z różnych klas decyzyjnych, oraz metoda RLS rozszerzająca podstawową metodę o etap relaksacji liniowej separowalności w celu uzyskania

podzbioru cech o lepszych zdolnościach generalizacji. Wyniki metod SEKWEM/GENET i RLS zostały także skonfrontowane z wynikami uzyskanymi z innych popularnych metod selekcji cech w zastosowaniu do „benchmarkowych" zbiorów danych mikromacierzowych.

Słowa kluczowe: selekcja cech, funkcja kryterialna typu CPL, algorytm SEKWEM/GENET, metoda RLS

Artykuł zrealizowano w ramach pracy badawczej S/WI/2/2008.