

DATA PREPROCESSING IN THE CLASSIFICATION OF THE IMBALANCED DATA

Katarzyna Borowska¹, Magdalena Topczewska²

¹ Student of Faculty of Computer Science, Białystok University of Technology, Białystok, Poland

² Faculty of Computer Science, Białystok University of Technology, Białystok, Poland

Abstract: The article concerns the problem of imbalanced data classification. Two algorithms improving the standard SMOTE method have been created and tested. To measure the distance between objects the Euclidean or the HVDM metric was applied, depending on the number of nominal attributes in a dataset.

Keywords: class imbalance, oversampling, classification

PRZETWARZANIE WSTĘPNE W PROBLEMIE KLASYFIKACJI DANYCH NIEZRÓWNOWAŻONYCH

Streszczenie: Artykuł dotyczy problemu klasyfikacji w przypadku, gdy mamy do czynienia z klasami nie zrównoważonymi. W tym celu stworzone zostały dwa algorytmy poprawiające wyniki uzyskiwane za pomocą standardowego algorytmu SMOTE. Do pomiaru odległości między obiektami zastosowano metrykę euklidesową lub metrykę HVDM, w zależności od liczby cech nominalnych w zbiorze.

Słowa kluczowe: klasy nie zrównoważone, tworzenie nowych obiektów, klasyfikacja

Artykuł w części zrealizowano w ramach pracy badawczej S/WI/2/2013.