

Feature selection methods based on mutual information for classifying heterogeneous features

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=20447914&lokasi=lokal>

Abstrak

Datasets with heterogeneous features can affect feature selection results that are not appropriate because it is difficult to evaluate heterogeneous features concurrently. Feature transformation (FT) is another way to handle heterogeneous features subset selection. The results of transformation from non-numerical into numerical features may produce redundancy to the original numerical features. In this paper, we propose a method to select feature subset based on mutual information (MI) for classifying heterogeneous features. We use unsupervised feature transformation (UFT) methods and joint mutual information maximization (JMIM) methods. UFT methods is used to transform non-numerical features into numerical features. JMIM methods is used to select feature subset with a consideration of the class label. The transformed and the original features are combined entirely, then determine features subset by using JMIM methods, and classify them using support vector machine (SVM) algorithm. The classification accuracy are measured for any number of selected feature subset and compared between UFT-JMIM methods and Dummy-JMIM methods. The average classification accuracy for all experiments in this study that can be achieved by UFT-JMIM methods is about 84.47% and Dummy-JMIM methods is about 84.24%. This result shows that UFT-JMIM methods can minimize information loss between transformed and original features, and select feature subset to avoid redundant and irrelevant features.

Dataset dengan fitur heterogen dapat mempengaruhi hasil seleksi fitur yang tidak tepat karena sulit untuk mengevaluasi fitur heterogen secara bersamaan. Transformasi fitur adalah cara untuk mengatasi seleksi subset fitur yang heterogen. Hasil transformasi fitur non-numerik menjadi numerik mungkin menghasilkan redundansi terhadap fitur numerik original. Dalam tulisan ini, peneliti mengusulkan sebuah metode untuk seleksi subset fitur berdasarkan mutual information (MI) untuk klasifikasi fitur heterogen. Peneliti menggunakan metode unsupervised feature transformation (UFT) dan metode joint mutual information maximization (JMIM). Metode UFT digunakan untuk transformasi fitur non-numerik menjadi fitur numerik. Metode JMIM digunakan untuk seleksi subset fitur dengan pertimbangan label kelas. Fitur hasil transformasi dan fitur original disatukan seluruhnya, kemudian menentukan subset fitur menggunakan metode JMIM, dan melakukan klasifikasi terhadap subset fitur tersebut menggunakan algoritma support vector machine (SVM). Akurasi klasifikasi diukur untuk sejumlah subset fitur terpilih dan dibandingkan antara metode UFT-JMIM dan Dummy-JMIM. Akurasi klasifikasi rata-rata dari keseluruhan percobaan yang dapat dicapai oleh metode UFT-JMIM sekitar 84.47% dan metode Dummy-JMIM sekitar 84.24%. Hasil ini menunjukkan bahwa metode UFT-JMIM dapat meminimalkan informasi yang hilang diantara fitur hasil transformasi dan fitur original, dan menyeleksi subset fitur untuk menghindari fitur redundansi dan tidak relevan