Similarity based entropy on feature selection for high dimensional data classification

Jayanti Yusmah Sari, author

Abstrak

Curse of dimensionality merupakan masalah yang sering dihadapi pada proses klasifikasi. Trans-formasi fitur dan seleksi fitur sebagai metode dalam reduksi fitur bisa diterapkan untuk mengatasi masalah ini. Terlepas dari performanya yang baik, transformasi fitur sulit untuk diinterpretasikan ka-rena ciri fisik dari fitur-fitur yang asli tidak dapat diperoleh kembali. Di sisi lain, seleksi fitur dengan proses komputasinya yang sederhana bisa mereduksi fitur-fitur yang tidak diperlukan dan mampu me-representasikan data untuk memudahkan pemahaman terhadap data. Pada penelitian ini diajukan metode seleksi fitur baru yang berdasarkan pada dua pendekatan filter, yaitu similarity (kemiripan) dan entropi untuk mengatasi masalah data berdimensi tinggi. Tahap awal metode ini adalah meng-hitung nilai similarity antara fitur dengan vektor kelas dari 6 data berdimensi tinggi. Kemudian diperoleh nilai similarity maksimum yang digunakan untuk menghitung nilai entropi untuk setiap fitur. Fitur yang dipilih adalah fitur yang memiliki nilai entropi lebih tinggi daripada entropi rata-rata seluruh fitur. Fuzzy k-NN diterapkan untuk tahap klasifikasi data hasil seleksi fitur. Hasil percobaan menunjukkan bahwa metode yang diajukan mampu mengklasifikasi data berdimensi tinggi dengan rata-rata akurasi 80.5%.

.....Curse of dimensionality is a major problem in most classification tasks. Feature transformation and feature selection as a feature reduction method can be applied to overcome this problem. Despite of its good performance, feature transformation is not easily interpretable because the physical meaning of the original features cannot be retrieved. On the other side, feature selection with its simple com-putational process is able to reduce unwanted features and visualize the data to facilitate data understanding. We propose a new feature selection method using similarity based entropy to over-come the high dimensional data problem. Using 6 datasets with high dimensional feature, we com-puted the similarity between feature vector and class vector. Then we find the maximum similarity that can be used for calculating the entropy values of each feature. The selected features are features that having higher entropy than mean entropy of overall features. The fuzzy k-NN classifier was im-plemented to evaluate the selected features. The experiment result shows that proposed method is able to deal with high dimensional data problem with mean accuracy of 80.5%.