INTISARI

METODE SELEKSI FITUR TEKS MENGGUNAKAN CHI SQUARE, ANT COLONY OPTIMIZATION DAN GREY WOLF OPTIMIZER PADA KLASIFIKASI DOKUMEN TEKS

Joan Angelina Widians
NIM 20/471774/SPA/00764

Kategorisasi memberikan kemudahan pencarian informasi yang sesuai dan tepat bagi pengguna. Namun dokumen teks yang berjumlah besar dengan ribuan fitur atau high-dimensional menjadi suatu tantangan dalam proses kategorisasi. Dengan jumlah fitur yang besar membuat model susah untuk memiliki kinerja yang optimal karena semakin banyak jumlah fitur akan membuat ruang pencarian konfigurasi fitur menjadi sangat luas. Penelitian ini mengembangkan metode seleksi fitur pendekatan hybrid filter-wrapper dengan menggabungkan metode Chi-square (CS), Ant Colony Optimization (ACO) dan Grey Wolf Optimizer (GWO). Dataset berupa abstrak artikel ilmiah berbahasa Inggris tentang tanaman obat berlabel antibakteri, antikanker, antifungi dan antiviral yang dikumpulkan melalui Google Scholar dengan waktu publikasi di tahun 2014 hingga 2024.

Pengujian model menggunakan metrik micro F1 score yang dikomparasi dengan ACO, Artificial Bee Colony (ABC), GWO, Random Walk-GWO (RW-GWO), GWO-Whale Optimization Algorithm (GWO-WOA). Pada model ekstraksi fitur TF-IDF dan seleksi fitur CS-ACOGWO diperoleh 85.38% sedangkan CS-ACO 83.02%, CS-ABC 83.06%, CS-GWO 82.84%, CS-RW-GWO 83.53%, CS-GWOWOA 83.21%. Sedangkan seleksi fitur pada ACO 79.76%, GWO 79.53%, ACOGWO 80%, dan tanpa seleksi fitur hanya 76.54%. Dengan demikian, metode usulan CS-ACOGWO dapat mengidentifikasi subset fitur yang terbaik dan relevan dalam *imbalance dataset*, dan model berhasil melakukan klasifikasi multi-label sehingga meningkatkan kinerja yang ditunjukkan dengan micro F1 score sebesar 8.84 dibandingkan tanpa seleksi fitur.

Kata kunci: Ant Colony Optimization, Chi Square, Grey Wolf Optimizer, Klasifikasi teks, Seleksi fitur.

ABSTRACT

TEXT FEATURE SELECTION USING CHI SQUARE, ANT COLONY OPTIMIZATION AND GREY WOLF OPTIMIZER FOR TEXT DOCUMENT CLASSIFICATION

Joan Angelina Widians
NIM 20/471774/SPA/00764

Categorization provides users with the appropriate information. However, large text documents with thousands of features or high-dimensional features become a challenge in the classification process. Many features make it challenging to achieve optimal performance, as the increased number of features significantly widens the feature configuration search space. This study develops a filter-wrapper feature selection method by combining the Chi-square (χ^2), Ant Colony Optimization (ACO), and Grey Wolf Optimizer (GWO) methods. The dataset is in the form of abstracts of English-language scientific articles about medicinal plants labelled antibacterial, Anti-cancer, antifungal and antiviral published in open access collected through Google Scholar with a publication time of 2014 to 2024.

The evaluation model utilizes a Bag-of-Words (BOW) and TF-IDF feature extraction approach. Assessment of the TF-IDF model and CS-ACOGWO yielded an accuracy of 85.38%. In comparison, the BOW and CS-ACOGWO models achieved 83.10% and 76.54%, respectively, without feature selection. The model is compared with five benchmarks: ACO, Artificial Bee Colony (ABC), GWO, Random Walk-GWO (RW-GWO), and GWO-Whale Optimization Algorithm (GWO-WOA). The results of the TF-IDF and CS-ACOGWO models are 85.38%, while CS-ACO achieves 83.02%, CS-ABC 83.06%, CS-GWO 82.84%, CS-RW-GWO 83.53%, and CS-GWOWOA 83.21%. Meanwhile, feature selection achieved 79.76% accuracy on ACO, 79.53% on GWO, 80% on ACOGWO, and 76.54% without feature selection. Furthermore, the CS-ACOGWO feature selection method effectively identifies the most relevant feature subsets in high-dimensional, imbalanced datasets. The model successfully categorizes a text document into multiple labels, thereby increasing performance as indicated by a micro F1 score of 8.84 compared to the score without feature selection.

Keywords: Ant Colony Optimization, Chi Square, Feature selection, Grey Wolf Optimizer, Text classification.