

Abstrak

Sebuah data yang ada saat ini bisa memiliki feature yang banyak. Banyaknya feature yang bisa dimiliki oleh satu objek instance belum tentu merupakan informasi relevan yang dibutuhkan oleh sistem data mining. Feature selection adalah suatu proses memilih subset dari feature/atribut yang relevan dengan menggunakan kriteria tertentu. Dengan melakukan feature selection ini mampu untuk mengurangi jumlah feature yang tidak relevan, menghilangkan redundansi data, dan meningkatkan akurasi learning.

Klasifikasi merupakan salah satu tahapan dalam data mining, yang fungsinya adalah untuk memprediksi keanggotaan atau kelas dari suatu data. Dalam beberapa studi ditunjukkan bahwa sebuah ensemble (himpunan) dari beberapa classifier umumnya lebih akurat dari classifier tunggal. Salah satu cara untuk menghasilkan sebuah ensemble adalah dengan memilih beberapa feature subset yang berbeda dari dataset asli dan untuk setiap feature subset tersebut selanjutnya dilakukan klasifikasi. Pendekatan ini dikenal sebagai ensemble feature selection. Di sini, penulis akan mencoba mengimplementasikan genetic algorithm untuk optimasi feature selection dalam pembentukan ensemble, yaitu Genetic Algorithm-Sequential Ensemble Feature Selection (GA-SEFS). Algoritma feature selection yang konvensional bertujuan untuk menemukan feature subset terbaik, sedangkan ensemble feature selection mempunyai tujuan untuk menemukan himpunan feature subset terbaik yang dapat meningkatkan akurasi dalam klasifikasi.

Dalam GA-SEFS terdapat 6 parameter penting. Parameter ukuran populasi, jumlah generasi, dan offspring tidak berpengaruh secara langsung terhadap akurasi yang dihasilkan dari klasifikasi ensemble. Parameter ukuran ensemble dapat membantu peningkatan akurasi dikarenakan vote feature subset beragam yang mampu membantu meningkatkan akurasi. Parameter alpha dapat membantu memberikan peningkatan akurasi tinggi yang didapat oleh kombinasi 4 parameter diatas (ukuran ensemble, jumlah populasi, jumlah generasi, dan jumlah offspring). Parameter beta dalam percobaan Tugas Akhir ini untuk tiga dataset berbeda ternyata lebih memberikan nilai akurasi yang tinggi pada nilai beta negatif.

Kata Kunci: *feature subset selection, ensemble, genetic search*