ABSTRAK

Penelitian ini berfokus pada penanganan data yang memiliki sebaran data yang tidak seimbang sebagai input proses klasifikasi. Kondisi data seperti ini banyak ditemukan dalam dunia nyata, salah satu contohnya yaitu masalah *churn customer* (konsumen yang tidak loyal) pada industri telekomunikasi atau perbankan dimana sampel data kelas *churner* jumlahnya jauh lebih sedikit (kelas minor) dibandingkan kelas *non-churner* (kelas mayor). Jika tidak ditangani dengan baik, kondisi kelas minor tersebut dapat menghilangkan potensi pendapatan perusahaan dalam jumlah yang besar. Saat ini telah banyak penelitian yang mengembangkan berbagai metode pada tingkat data atau *sampling* untuk menangani masalah tersebut. Namun, permasalahan kualitas data pada data tidak seimbang seperti keberadaan *noise* juga mengganggu proses *sampling* dan dapat berdampak negatif terhadap performa model klasifikasi. Masalah *noise* juga menjadi permasalahan utama pada metode *over-sampling* yang populer seperti *Synthetic Minority Over-sampling Technique* (SMOTE). Oleh karena itu, penelitian ini bertujuan mengembangkan metode *evolutionary hybrid sampling* untuk meningkatkan performa model klasifikasi pada data tidak seimbang dengan melakukan perbaikan kualitas data sebelum dan sesudah proses *balancing* SMOTE. Metode yang diusulkan yaitu Tomek-SMOTE-GA (TSGA) dan fokus pada penyelesaian masalah data *binary class*. Metode TSGA melakukan proses *sample denoising* menggunakan metode *Tomek links* sebelum menerapkan SMOTE, yang kemudian dilanjutkan dengan proses *sample optimization* menggunakan *evolutionary algorithm* setelah SMOTE. *Genetic algorithm* (GA) sebagai salah satu *evolutionary algorithm* digunakan untuk proses optimasi sampel. Selanjutnya, *train set* yang telah dipilih digunakan untuk mengembangkan model klasifikasi dengan lima *classifier*, yaitu *decision tree*, *logistic regression*, *support vector machine*, *k-nearest neighbors*, dan *naive bayes*. Hasil eksperimen dan uji statistik pada 24 dataset tidak seimbang menunjukkan bahwa metode TSGA yang diusulkan, secara signifikan lebih unggul dibandingkan dengan metode sampling *baseline* maupun *state-of-the-art* dalam hal *geometric-mean*, terutama saat menggunakan *decision tree* sebagai *classifier*. Hasil tersebut telah menjawab dua *research question* pada disertasi ini. Penelitian kedepannya dapat dilakukan dengan mengembangkan metode TSGA untuk menangani masalah data tidak seimbang pada dataset *multi-class*, data *time-series*, dan aplikasi *real-time*.

Kata kunci: data tidak seimbang, *noise*, SMOTE, *evolutionary hybrid sampling*, Tomek-SMOTE-GA (TSGA)