INTISARI

MODIFIKASI METODE *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE – EDITED NEAREST NEIGHBORS* (SMOTE-ENN) UNTUK MENINGKATKAN KINERJA ALGORITMA KLASIFIKASI PADA DATASET TIDAK SEIMBANG DAN BERDIMENSI TINGGI

Adi Zaenul Mustaqim
22/498910/PPA/06327

Dataset tidak seimbang ialah kondisi di mana jumlah data suatu kelas terlalu sedikit daripada jumlah data pada kelas lain. SMOTE-ENN, salah satu algoritma *hybrid sampling* untuk mengatasi hal ini. SMOTE membuat data sintetis dari data minoritas hingga seimbang dan ENN menghapus *noise* agar batas kelas jelas.

Salah satu tahap SMOTE-ENN yaitu memilih data minoritas dan mencari tetangga terdekatnya menggunakan *Euclidean distance* untuk membuat data sintetis dan penghapusan *noise*. Sifat *blind oversampling* SMOTE-ENN pada data *noise* akan berpotensi menciptakan *noise* baru. Pada dataset dimensi tinggi, *Euclidean distance* di SMOTE-ENN menyebabkan hilangnya konsep kedekatan yang menyebabkan jarak semua titik hampir sama satu sama lain dan menganggap semua fitur sama penting. Penelitian ini mengusulkan WKSMOTE-RENN (*Weighted k-Nearest Neighbors Synthetic Minority Oversampling Technique – Reverse Edited Nearest Neighbors*) untuk memperbaiki SMOTE-ENN dalam memilih data minoritas untuk di *oversampling* dan penggunaan *Manhattan* yang dibobotkan daripada *Euclidean* untuk mengatasi masalah dataset berdimensi tinggi. Pengujian dilakukan dengan menggunakan algoritma klasifikasi Naive Bayes (NB) dan *Artificial Neural Network* (ANN).

WKSMOTE-RENN diuji menggunakan 7 dataset dengan ukuran evaluasi *accuracy, recall, specificity,* dan *g-mean*. Hasil menunjukan bahwa SMOTE-ENN menghasilkan rata-rata *accuracy* sebesar 85,92%, *recall* 81,49%, *specificity* 89,38%, dan *gmean* 82,24% (NB) dan rata-rata *accuracy* sebesar 88,24%, *recall* 75,40%, *specificity* 94,31%, dan *gmean* 83,61% (ANN). Dibandingkan hasil yang diperoleh SMOTE-ENN, metode WKSMOTE-RENN nyatanya berhasil meningkatkan rata-rata *accuracy* sebesar 1,33%, *recall* 6,97%, dan *gmean* 4,04% (NB) dan rata-rata *accuracy* sebesar 0,55%, *recall* 6,28%, dan *gmean* 1,46% (ANN). Tetapi WKSMOTE-RENN mengalami penurunan *specificity* sebesar 2,51% pada NB dan 1,02% pada ANN.

Kata kunci: *Imbalanced dataset, hybrid sampling,* WKSMOTE-RENN, pembobotan fitur, *Manhattan distance*, dimensi tinggi

ABSTRACT

MODIFIED THE SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE – EDITED NEAREST NEIGHBORS (SMOTE-ENN) METHOD TO IMPROVE THE PERFORMANCE OF THE CLASSIFICATION ALGORITHM ON IMBALANCED AND HIGH DIMENSIONAL DATASETS

Adi Zaenul Mustaqim
22/498910/PPA/06327

An imbalanced dataset is a condition where the amount of data in one class is too little than in another class. SMOTE-ENN, one of the *hybrid* sampling algorithms to overcome this. SMOTE creates synthetic data from minority data until it is balanced and ENN removes *noise* to make class boundaries clear.

One of the stages of SMOTE-ENN is to select minority data and find its nearest neighbors using Euclidean distance to create synthetic data and *noise* removal. The blind oversampling nature of SMOTE-ENN on *noise* data will potentially create new *noise*. On high-dimensional datasets, the Euclidean distance in SMOTE-ENN leads to a loss of the concept of proximity which causes all points to be almost equidistant from each other and considers all features equally important. This research proposes WKSMOTE-RENN (Weighted k-Nearest Neighbors Synthetic Minority Oversampling Technique - Reverse Edited Nearest Neighbors) to improve SMOTE-ENN in the selection of minority data to be oversampled and the use of weighted Manhattan instead of Euclidean distance to overcome the problem of high dimensional datasets. Tests were conducted using Naive Bayes (NB) and Artificial Neural Network (ANN) classification algorithms.

WKSMOTE-RENN was tested using 7 datasets with accuracy, recall, specificity, and g-mean evaluation measures. The results show that SMOTE-ENN produces an average accuracy of 85.92%, recall 81.49%, specificity 89.38%, and gmean 82.24% (NB) and an average accuracy of 88.24%, recall 75.40%, specificity 94.31%, and gmean 83.61% (ANN). Compared to the results obtained by SMOTE-ENN, the WKSMOTE-RENN method was able to increase the average accuracy by 1.33%, recall 6.97%, and gmean 4.04% (NB) and the average accuracy by 0.55%, recall 6.28%, and gmean 1.46% (ANN). However, WKSMOTE-RENN decreased the specificity value by 2.51% in NB and 1.02% in ANN.

Keywords: Imbalanced dataset, *hybrid* sampling, WKSMOTE-RENN, feature weighting, Manhattan distance, high dimensional dataset