ABSTRACT

RWO-SAMPLING AND RANDOM OVERSAMPLING METHOD IN HANDLING DATASETS IMBALANCE PROBLEM IN CLASSIFICATION TO BINARY LOGISTIC REGRESSION ALGORITHM

by

Windyaning Ustyannie
15/388511/PPA/04950

The imbalance dataset is a condition when one class has a higher percentage than the other so it can affect the accuracy. One method in data mining that can be used to handle the problem of imbalance dataset that is logistic regression classification method. Logistic regression proved to be a powerful and easy to use classification but has under fitted disadvantages if used on data with unbalanced classes. The class imbalance can be overcome by using the data-level approach of resampling (oversampling). RWO-sampling is an oversampling method that has less than optimal performance in synthetic data generation in the discrete attribute.

The method used in this research was the RWO-sampling method using random replicate approach for synthetic data generation on the discrete attribute. The result of the research can handle the problem of class imbalance, RWO-sampling method with random replicate approach showed better accuracy than the RWO-sampling method with roulette and ROS approach. The accuracy value for RWO-Sampling method with roulette and RWO-Sampling approach with random replicate approach has increased to an average of 7.06 ± 4.13 of each dataset. As for the comparison with the ROS method has increased an average of 3.7 ± 3.2 of each dataset. Furthered, to test the problem of underfitting in logistic regression showed that oversampling method was better than non-oversampling with an increase of accuracy value reaching an average of 2.3 ± 4.37 of each dataset.

Keywords : *imbalanced dataset*, RWO-*sampling*, *Random Oversampling, Logistic Regression*

INTISARI

METODE RWO-SAMPLING DAN RANDOM OVERSAMPLING UNTUK MENANGANI MASALAH KETIDAKSEIMBANGAN DATASET PADA KLASIFIKASI ALGORITMA REGRESI LOGISTIK BINER

Oleh

Windyaning Ustyannie
15/388511/PPA/04950

Ketidakseimbangan dataset adalah kondisi ketika salah satu kelas memiliki presentase distribusi data lebih tinggi dibandingkan dengan kelas lainnya sehingga dapat mempengaruhi tingkat akurasi. Salah satu metode dalam data mining yang dapat digunakan untuk klasifikasi dalam masalah ketidakseimbangan dataset yaitu metode klasifikasi regresi logistik. Regresi logistik terbukti menghasilkan klasifikasi yang *powerfull* dan mudah digunakan tetapi memiliki kelemahan *underfitting* jika digunakan pada dataset yang tidak seimbang. Ketidakseimbangan kelas dapat diatasi dengan menggunakan pendekatan level data yaitu metode *resampling* (*oversampling*). RWO-*sampling* merupakan metode *oversampling* yang memiliki kinerja kurang optimal dalam pembangkitan data sintetik pada *attribute descrete*.

Metode yang digunakan dalam penelitian adalah metode RWO-*sampling* menggunakan pendekatan *random replicate* untuk pembangkitan data sintetik pada *attribute descrete*. Hasil dari penelitian dapat menangani masalah ketidakseimbangan kelas, metode RWO-*sampling* dengan pendekatan *random replicate* menunjukkan akurasi yang lebih baik dibandingkan dengan metode RWO-*sampling* dengan pendekatan *roulette* dan ROS. Nilai akurasi untuk metode RWO-*Sampling* dengan pendekatan *roulette* dan RWO-*Sampling* dengan pendekatan *random replicate* mengalami peningkatan mencapai rata-rata 7.06$\pm$4.13 dari setiap *dataset*. Sedangkan untuk perbandingan dengan metode ROS mengalami peningkatan mencapai rata-rata 3.7$\pm$3.2 dari setiap *dataset*. Selanjutnya untuk pengujian masalah *underfitting* dalam regresi logistik menunjukkan metode *oversampling* lebih baik daripada non-*oversampling* dengan kenaikan nilai akurasi mencapai rata-rata 2.3$\pm$4.37 dari setiap *dataset*.

Kata kunci : *imbalanced dataset*, RWO-*sampling*, *Random Oversampling,* Regresi Logistik