Modifikasi Metode Synthetic Minority Oversampling Technique (SMOTE) Menggunakan Algoritma Genetika
Untuk Menangani Masalah Imbalanced Dataset
TINCE ETLIN TALLO, Aina Musdholifah, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2018 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

MODIFIKASI METODE *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE* (SMOTE) MENGGUNAKAN ALGORITMA GENETIKA UNTUK MENANGANI MASALAH *IMBALANCED DATASET*

Tince Etlin Tallo
15/388504/PPA/04943

Imbalanced dataset adalah suatu kondisi *dataset* yang memiliki kelas minoritas atau kelas yang mempunyai distribusi *instance* yang jauh lebih sedikit dibanding kelas lainnya. Kondisi *imbalanced* dapat mempengaruhi kinerja algoritma *classifier* standar sehingga hasil klasifikasi lebih cenderung ke kelas yang memiliki jumlah *instance* yang lebih banyak. Terdapat beberapa pendekatan untuk menangani masalah *imbalanced dataset* salah satunya adalah menggunakan metode SMOTE. Metode SMOTE bekerja dengan cara membuat sejumlah *instance* buatan pada daerah antara sebuah *instance* dalam kelas minoritas dengan *instance* minoritas tetangga terdekatnya.

Dalam penerapannya SMOTE dapat mengakibatkan *blind oversampling* sehingga membuat batas antar kelas menjadi tidak jelas. Pada penelitian ini, diterapkan algoritma genetika pada SMOTE untuk mencari dan memilih *instance-instance* yang harus dibangkitkan pada proses *oversampling* sehingga dapat mengatasi masalah *blind oversampling* pada SMOTE yang dapat mengatasi masalah *imbalanced dataset.*

Pengujian dilakukan dengan 2 skenario, yang pertama perbandingan performansi metode SMOTE dan SMOTE-SGA dan yang kedua perbandingan metode yang memakai teknik *oversampling* dan yang tidak memakai teknik *oversampling*. Dengan menggunakan 3 buah *imbalanced dataset* hasil klasifikasi yang diukur menggunakan *G-means* dan *F-Measure* untuk skenario pertama didapatkan *G-means* dan *F-measure* yang lebih baik 18% pada *dataset* Balance Scale, 4.35% pada *dataset* Diabetes dan 7.1% pada *dataset* Transfusion dengan *imbalance ratio* antara 0.99-1. Sedangkan untuk skenario kedua didapatkan nilai *G-means* dan *F-measure* yang lebih baik 65.5% pada *dataset* Balance Scale, 9.8% pada *dataset* Diabetes dan 4.3% pada *dataset* Transfusion.

Kata kunci: *imbalanced dataset*, klasifikasi, SMOTE, algoritma genetika

Modifikasi Metode Synthetic Minority Oversampling Technique (SMOTE) Menggunakan Algoritma Genetika
Untuk Menangani Masalah Imbalanced Dataset
TINCE ETLIN TALLO, Aina Musdholifah, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2018 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

MODIFICATION SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) USING GENETIC ALGORITHM FOR HANDLING IMBALANCED DATASET PROBLEM

Tince Etlin Tallo
15/388504/PPA/04943

An imbalanced dataset is a condition that has a minority class which is a class has far fewer instance distributions than other classes. The imbalanced condition can affect the performance of standard classifier algorithms and then classification results are more likely to majority class. There are several approaches to dealing with the imbalanced dataset, one of which is using the SMOTE method. The SMOTE method works by creating a number of artificial instances on the area between an instance in a minority class with a nearby minority instance of its neighbor.

In its implementation SMOTE can results blind oversampling because the generated instances have the same amount regardless of the distribution of instances and making the boundary between the classes unclear. In this study, a genetic algorithm is applied to SMOTE to search and select instances that must be generated in the oversampling process and to overcome the blind oversampling problem in SMOTE and solved the imbalanced dataset problem.

The test was performed with 2 scenarios, the first comparison of the SMOTE and SMOTE-SGA method of performance and the second comparison of the method using the oversampling technique and not using the oversampling technique. Using the 3 imbalanced dataset classified as measured using G-means and F-Measure for the first scenario, G-means and F-measure were found to be 18% better on the Balance Scale dataset, 4.35% in the Diabetes dataset and 7.1% in the Transfusion dataset with an imbalance ratio between 0.99-1. While for the second scenario we get better G-means and F-measure value of 65.5% on Balance Scale dataset, 9.8% in Diabetes dataset and 4.3% on Transfusion dataset.

Keywords: imbalanced dataset, classification, SMOTE, genetic algorithm