INTISARI

IMPLEMENTASI KOMBINASI K-MEANS DAN SMOTE (K-MEANS SMOTE) UNTUK KLASIFIKASI DATA TIDAK SEIMBANG

Oleh

Zulfah Zuhrotunnisa

17/409531/PA/17838

Data merupakan aspek penting dalam proses pengambilan keputusan dalam berbagai bidang. Dalam bidang medis, analisis klasifikasi merupakan salah satu analisis yang sangat umum digunakan dalam pengolahan data. Namun salah satu permasalahan yang sering ditemui pada proses pengolahan data khususnya data untuk analisis klasifikasi yaitu adanya persebaran jumlah observasi di setiap kelompok yang tidak merata (*imbalanced data). Imbalanced data* ini dapat menyebabkan analisis klasifikasi bersifat bias karena adanya kelompok mayoritas dan kelompok minoritas. Untuk mengatasi hal tersebut dapat dilakukan metode *oversampling* data menggunakan SMOTE yang akan membentuk *instance* sintesis pada kelas minoritas. Akan tetapi SMOTE memiliki kekurangan dimana SMOTE dapat mengakibatkan *overfitting* pada antar kelasnya karena *instance* kelas minoritas yang ditambahkan akan berada di bagian kelas mayoritas ataupun sebaliknya. Oleh karena itu, dikombinasikan SMOTE dengan analisis *clustering* KMeans untuk mengatasi kekurangan dari SMOTE sehingga hasil yang diperoleh akan lebih representatif dan menghasilkan performa yang lebih baik. Pada skripsi ini dilakukan analisis klasifikasi Naïve Bayes pada data *haberman* dan *breast cancer* yang dikombinasikan dengan SMOTE dan KMeans SMOTE. Untuk menentukan jumlah *cluster* terbaik digunakan Indeks Davies Bouldin.

Kata kunci : klasifikasi, *imbalanced data*, SMOTE, KMeans SMOTE, Naïve Bayes, Indeks Davies Bouldin.

ABSTRACT

IMPLEMENTATION OF COMBINATION OF K-MEANS AND SMOTE (K-MEANS SMOTE) FOR CLASSIFICATION OF IMBALANCED DATA

by

Zulfah Zuhrotunnisa

17/409531/PA/17838

Data is an important aspect in the decision-making process in various fields. In the medical field, classification analysis is one of the analyzes that is very commonly used in data processing. However, one of the problems that is often encountered in the data processing process, especially data for classification analysis, is the uneven distribution of the number of observations in each group (imbalanced data). This imbalanced data can cause biased classification analysis due to the existence of majority groups and minority groups. To overcome this, a data oversampling method can be used using SMOTE which will form synthetic instances in the minority class. However, SMOTE has a drawback in that it does not look at the data structure and can result in overfitting between classes. Therefore, SMOTE is combined with KMeans clustering analysis to overcome the shortcomings of SMOTE so that the results obtained will be more representative and produce better performance. In this thesis, a Naïve Bayes classification analysis was carried out on Haberman and breast cancer data combined with SMOTE and KMeans SMOTE. To determine the best number of clusters, the Davies Bouldin Index is used.

Keywords: classification, imbalanced data, SMOTE, KMeans SMOTE, Naïve Bayes, Davies Bouldin Index.