

ABSTRACT

IMPLEMENTATION SMOTE ALGORITHM AND RANDOMFOREST CLASSIFICATION IN IMBALANCED DATA LYSINE PROTEIN SEQUENCE METHYLATION

By
Annisa Nurwalikadani

One problem that is often encountered when processing data is an imbalance in the number of samples from each class in the data or what is commonly called imbalanced data. In recent years, there have been many studies regarding imbalanced data in the field of bioinformatics, especially Posttranslational Modification (PTM). The case study in this research is the problem of imbalanced data lysine protein sequence methylation. The problem of imbalanced data will certainly affect the classification results. Therefore, a method is needed to deal with this problem, one of which is the SMOTE algorithm that will be used in this research. The purpose of this research is to analyze the performance of the SMOTE algorithm and random forest classification in handling with imbalanced data problems. The lysine protein sequence methylation dataset was obtained from <http://www.uniprot.org/> by searching for the keyword "methylation", has 1000 positive data and 172 negative data and has a sequence length of 15 amino acids. The dataset separation into 2, 80% training data 20% test data and 90% training data 10% test data. The feature extraction used is AA Index, PseACC, Hydrophobicity, and CTD. Then, the extracted data was processed using a random forest classification with ntree parameters 500, 800, and 1000 and mtry 7, 9, and 14. The highest results were obtained in the separation of the dataset 80% training data 20% test data with mtry 14 ntree 500, resulting in 95.65% accuracy, 96.2% sensitivity, 95% specificity, and 91.25% MCC.

Keywords : Post-translational modification, methylation, imbalanced data, feature extraction, random forest, SMOTE

ABSTRAK

IMPLEMENTASI ALGORITME SMOTE DAN KLASIFIKASI RANDOM FOREST PADA IMBALANCED DATA METILASI SEQUENCE PROTEIN LISIN

Oleh
Annisa Nurwalikadani

Salah satu masalah yang sering ditemukan pada saat pengolahan data adalah ketidakseimbangan jumlah sampel dari masing-masing kelas dalam data atau yang biasa disebut dengan *imbalanced data*. Dalam beberapa tahun terakhir, banyak penelitian mengenai ketidakseimbangan data di bidang bioinformatika, khususnya Post translational Modification (PTM). Studi kasus dalam penelitian ini adalah masalah ketidakseimbangan data metilasi *sequence* protein lisin. Masalah ketidakseimbangan data tentunya akan mempengaruhi hasil klasifikasi. Oleh karena itu diperlukan suatu metode untuk mengatasi masalah tersebut, salah satunya adalah algoritme SMOTE yang akan digunakan dalam penelitian ini. Tujuan dari penelitian ini adalah untuk menganalisis kinerja algoritma SMOTE dan klasifikasi *random forest* dalam menangani masalah ketidakseimbangan data. Dataset metilasi urutan protein lisin diperoleh dari <http://www.uniprot.org/> dengan mencari kata kunci "metilasi", memiliki 1000 data positif dan 172 data negatif serta memiliki panjang *sequence* 15 asam amino. Pemisahan *dataset* menjadi 2, yaitu 80% data latih 20% data uji dan 90% data latih 10% data uji. Fitur ekstraksi yang digunakan adalah AA Index, PseACC, Hydrophobicity, dan CTD. Kemudian, data hasil ekstraksi diolah menggunakan klasifikasi *random forest* dengan parameter ntree 500, 800, dan 1000 serta mtry 7, 9, dan 14. Hasil tertinggi diperoleh pada pemisahan dataset 80% data latih 20% data uji dengan mtry 14 ntree 500, menghasilkan akurasi 95,65%, sensitivitas 96,2%, spesifisitas 95%, dan MCC 91,25%.

Kata kunci : *Post-translational modification*, metilasi, data tidak seimbang, fitur ekstraksi, *random forest*, SMOT