RINGKASAN

OPTIMASI PERFORMA KLASIFIKASI DATA SEL BLAST TAK SEIMBANG MENGGUNAKAN METODE *GRAPH CLUSTERING* BERBASIS DPClusSBO

Muhammad Fikra Adzaky

Dewasa ini sering dijumpai pemanfaatan teknologi *machine learning* dalam bidang medis. Salah satunya adalah penggunaan metode klasifikasi untuk mengenali suatu sampel atau membantu dalam diagnosa penyakit pada sel darah. Metode klasifikasi *machine learning* dapat mengenali kelainan dan perbedaan pada citra objek atau gambar sel darah putih (leukosit). Penyakit leukemia biasa dibedakan menjadi dua jenis yakni *acute myeloid leukemia* (AML) dan *acute lymphoblastic leukemia* (ALL). Sebagai suatu penyakit akut tentunya dalam ketersediaan dataset yang berlabel akan cukup sulit untuk diperoleh. Bila ada, seringkali dataset yang diperoleh adalah data yang tidak seimbang (*imbalance*). Hal ini akan memengaruhi performa klasifikasi dalam beberapa aspek sehingga kurang optimal dalam penggunaannya.

Salah satu upaya menangani kasus ketidakseimbangan data ini adalah dengan melakukan penuruan sampel (*undersampling*) pada salah satu kelas data mayoritas. Teknik ini bertujuan menyeimbangkan distribusi data keseluruhan antara data mayoritas dengan data minoritas. Teknik *undersampling* yang dipilih adalah metode analisis metrik jarak pada *graph clustering* yang dijalankan pada perangkat lunak DPClusSBO.

Penelitian ini menghasilkan metode penanganan untuk meningkatkan performa proses klasifikasi pada data sel blast yang tak seimbang (*imbalance*). Performa yang dinilai bukan hanya dari akurasi namun juga parameter seperti AUC, ROC, F-measure, dan G-means. Berikutnya metode ini dapat diterapkan pada klasifikasi sel blast yang menggunakan metode *support vector machine* (SVM), *k-nearest neighbor* (k-NN), *naïve bayes*, dan *random forest classifier*. Selain itu, harapannya metode pada penelitian ini juga dapat diterapkan pada kasus-kasus data tak seimbang lainnya.

Kata kunci : sel blast, leukosit, klasifikasi, *imbalanced data, graph clustering*, *undersampling*

SUMMARY

PERFORMANCE OPTIMIZATION OF IMBALANCED BLAST CELL DATA CLASSIFICATION USING THE GRAPH CLUSTERING METHODS BASED ON DPClusSBO

Muhammad Fikra Adzaky

It is an everyday thing to see machine learning technology in the medical field. One of them is classification methods to identify a sample or to assist in diagnosing diseases of blood cells. The machine learning classification method can recognize abnormalities and differences in white blood cells (leukocytes) image. There are two types of leukemia, acute lymphoblastic leukemia, and acute myeloid leukemia. Speaking about clinical health labeled datasets, the cost tends to be high and less affordable. On the other hand, the dataset obtained often is imbalanced data. Thus, it will affect classification performance in several aspects.

To address this imbalanced data problem, we will perform undersampling in the majority data class. This technique balances the overall data distribution between majority data and minority data. We use distance metric analysis in graph clustering-based undersampling. This method will be executed in DPClusSBO software.

This research produces a handling method to improve the performance of the classification process on imbalanced blast cell data. Performance is assessed not only from accuracy. There are also parameters such as AUC, ROC, F-measure, and G-means. Furthermore, this method can be applied to blast cell classification using support vector machine (SVM), k-nearest neighbor (k-NN), naïve Bayes, and random forest classifier methods.

Keywords : blast cell, leukocytes, classification, imbalanced data, graph clustering, undersampling