Metode Imputasi Missing Values pada Data Ekspresi Gen Pasien Leukemia Limfoblastik Akut Menggunakan Algoritma Biclustering Berbasis Local Pearson Correlation Measure dan Imputasi Least Square (NCBI- LPCM) = Novel Correlation Based Imputing Technique Using Biclustering Based Local Pearson Correlation Measure for Missing Values on Gene Expression Data from Patient with Acute Lymphoblastic Leukemia

Kinanty Tasya Octaviane, author

Abstrak

Teknologi DNA microarray menghasilkan data ekspresi gen yang dapat digunakan untuk membantu berbagai pemecahan masalah dalam dunia kesehatan. Data ekspresi gen merupakan matriks berukuran besar berisi gen dan kondisi eksperimental yang tak jarang mengandung missing values dan outlier. Data yang mengandung missing values dapat mengganggu dan membatasi analisis. Untuk mengatasinya, metode komputasi dinilai layak untuk imputasi missing values pada data ekspresi gen sebelum dilakukan analisis lanjutan, terlebih untuk data yang memiliki outlier. Oleh karena itu, pada penelitian ini digunakan metode imputasi missing values NCBI-LPCM untuk mengatasi permasalahan missing values pada data ekspresi gen yang memiliki outlier. Metode NCBI-LPCM menggunakan ukuran korelasi LPCM yang dapat menangani keberadaan outlier untuk pembentukan bicluster dan imputasi least square yang merupakan metode imputasi dengan pendekatan lokal. LPCM mengidentifikasi gen-gen yang memiliki pola korelasi similar sehingga menjadi informasi lokal untuk dasar imputasi. Metode ini diterapkan pada data ekspresi gen pasien Leukemia Limfoblastik Akut pada missing rate 5%, 10%, 15%, 20%, 25%, 30%, dan 35%. Berdasarkan RMSE dan korelasi Pearson, metode NCBI-LPCM lebih baik jika dibandingkan dengan NCBI-SSSim yang juga dapat menangani keberadaan outlier.

.....DNA microarray technology produces gene expression data that can be used to help solve various problems in healthcare. Gene expression data is a large matrix of genes and experimental conditions that often contains missing values and outliers. Data containing missing values can interfere with and limit analyses. To overcome this, computational methods are considered feasible for imputing missing values in gene expression data before further analysis is carried out, especially for data that has outliers. Therefore, in this study, the NCBI-LPCM missing values imputation method was used to overcome the problem of missing values in gene expression data with outliers. The NCBI-LPCM method uses the LPCM correlation measure which can handle the presence of outliers for bicluster formation and least square imputation which is an imputation method with a local approach. LPCM identifies genes that have similar correlation patterns so that they become local information for the basis of imputation. This method was applied to gene expression data of Acute Lymphoblastic Leukaemia patients at missing rates of 5%, 10%, 15%, 20%, 25%, 30%, and 35%. Based on RMSE and Pearson correlation, the NCBI- LPCM method is better than NCBI-SSSim which can also handle the presence of outliers.