Imputasi Missing Values pada Data Ekspresi Gen Pasien Penderita Leukemia Menggunakan Metode Cosine Similarity dan Normalized Mean Residue Similarity (NMRS) Based Biclustering = Imputation of Missing Values on Leukemia Patient Gene Expression Data using Cosine Similarity and Normalized Mean Residue Similarity (NMRS) Based Biclustering Method

Mush`ab Muzzammil, author

Abstrak

Ekspresi gen adalah proses pembentukan molekul protein dengan cara menguraikan informasi yang terkandung dalam gen. Ekspresi gen dapat diubah menjadi data numerik dengan bantuan teknologi microarray. Penyakit chronic lymphocytic leukemia (CLL) merupakan salah satu penyakit kanker yang terjadi karena pembentukan lymphocytes yang tidak normal pada sumsum tulang. Data ekspresi gen dari pasien CLL dapat diperoleh dengan menggunakan teknologi microarray. Namun, penggunaan teknologi microarray dapat menghasilkan missing values pada data ekspresi gen CLL akibat dari adanya goresan atau debu pada microarray slides. Keberadaan missing values dapat mengakibatkan hasil analisis menjadi bias dan tidak merepresentasikan sifat aslinya. Untuk mengatasi hal tersebut, salah satu pendekatan yang dapat dilakukan adalah dengan melakukan imputasi missing values. Imputasi adalah proses mengisi missing values berdasarkan informasi yang terdapat dalam data. Nilai pada data hasil imputasi diharapkan mendekati nilai dari elemen yang hilang. Proses imputasi menghasilkan data yang lengkap sehingga analisis selanjutnya dapat berjalan dengan baik dan diperoleh hasil yang lebih akurat. Pada penelitian ini dilakukan proses imputasi missing values dengan metode imputasi Cosine Similarity Based Biclustering dan Normalized Mean Residue Similarity (NMRS) Based Biclustering. Metode Cosine Similarity Based Biclustering dan NMRS Based Biclustering melakukan imputasi dengan memanfaatkan analisis biclustering berbasis korelasi cosine similarity dan NMRS. Data yang digunakan untuk melakukan penelitian ini adalah data numerik berupa ekspresi gen pada pasien chronic lymphocytic leukemia (CLL). Kinerja dari metode imputasi pada penelitian ini dievaluasi dengan menghitung korelasi Pearson dari nilai asli pada data awal dengan nilai pada data yang sudah dilakukan imputasi. Hasil evaluasi dari kinerja metode imputasi menggunakan Cosine Similarity Based Biclustering dan NMRS Based Biclustering dibandingkan dengan kinerja metode imputasi K-Means. Berdasarkan hasil penelitian, didapatkan nilai koefisien korelasi Pearson dari metode imputasi menggunakan Cosine Similarity Based Biclustering dan NMRS Based Biclustering untuk missing rate 5%, 15%, 25%, 35% dan 45% memiliki rentang yang lebih tinggi dibandingkan metode imputasi K-Means, dengan sebagian besar nilai korelasi Pearson di atas 0,96. Selain itu metode NMRS Based Biclustering memiliki rentang korelasi Pearson paling tinggi, sehingga dapat dikatakan metode NMRS Based Biclustering menghasilkan nilai imputasi terbaik di antara metode yang digunakan untuk mengisi missing values pada data CLL.

.....Gene expression is the process of forming protein molecules by deciphering the information contained in genes. Gene expression can be converted into numerical data using microarray technology. Chronic lymphocytic leukemia (CLL) is cancer that occurs due to the formation of abnormal lymphocytes in the bone marrow. Gene expression data from CLL patients can be obtained using microarray technology.

However, the use of microarray technology can produce missing values in the CLL gene expression data due to scratches or dust on the microarray slides. The existence of missing values can lead to analysis results being biased and not representing their true nature. To overcome this, one approach that can be taken is to impute missing values. Imputation is the process of filling in the missing values based on the information contained in the data. The value of the imputed data is expected to be close to the value of the missing element. The imputation process produces complete data so that further analysis can run well and obtained more accurate results. In this study, the imputation process for missing values was carried out using the Cosine Similarity Based Biclustering and Normalized Mean Residue Similarity (NMRS) Based Biclustering imputation methods. Cosine Similarity Based Biclustering and NMRS Based Biclustering methods perform imputation by utilizing biclustering analysis based on cosine similarity correlation and NMRS. The data used to conduct this research is numerical data in the form of gene expression in chronic lymphocytic leukemia (CLL) patients. The performance of the imputation method in this study was evaluated by calculating the Pearson correlation of the original value in the initial data with the value in the imputed data. The results of the evaluation of the performance of the imputation method using Cosine Similarity Based Biclustering and NMRS Based Biclustering were compared with the performance of the K-Means imputation method. Based on the results of the study, the Pearson correlation coefficient values obtained from the imputation method using Cosine Similarity Based Biclustering and NMRS Based Biclustering for missing rates of 5%, 15%, 25%, 35% and 45% have a higher range than the K-Means imputation method, with most Pearson correlation values above 0.96. In addition, the NMRS Based Biclustering method has the highest Pearson correlation range, so it can be said that the NMRS Based Biclustering method produces the best imputation value among the methods used to fill in the missing values in CLL data.