

Metode Bicluster Berbasis k-Nearest Neighbors dan Robust Least Squares Estimation menggunakan Principal Components (bi-KNNRLSP) untuk imputasi Missing values pada Data Ekspresi Gen = Missing values Imputation for Microarray Data Using Bicluster-Based k-Nearest Neighbors and Robust Least Squares Estimation with Principal Components (bi-KNN-RLSP)

Nurul Shabrina, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920521892&lokasi=lokal>

Abstrak

Microarray merupakan salah satu teknologi pada bidang biologi yang memberikan informasi tentang ekspresi gen. Data microarray mentah berupa gambar, yang harus diubah menjadi matriks ekspresi gen dimana baris menunjukkan gen, kolom menunjukkan kondisi eksperimental. Namun, pada praktiknya data microarray banyak ditemukan missing values yang tentunya akan menghambat proses dari analisis datanya. Imputasi merupakan salah satu solusi yang dapat mengatasi adanya missing values pada data microarray. Dengan menggunakan imputasi, nilai missing values yang terdapat pada matriks data diprediksi atau diestimasi sehingga diperoleh matriks data yang lengkap. Metode imputasi yang digunakan pada penelitian ini bernama bi-KNN-RLSP, yang menggunakan konsep biclustering, principal component analysis, dan regresi kuantil. Dalam proses pembentukan biclustering, dibutuhkan matriks lengkap sementara yang diperoleh melalui proses praimputasi dengan KNNimpute. Percobaan bi-KNN-RLSP dilakukan pada data ekspresi gen garis sel kanker serviks dengan menerapkan missing rate yang berbeda, yaitu 1%, 5%, 10%, 15%, 20%, 25%, dan 30% dengan menggunakan parameter $k=10$ pada proses praimputasi KNNimpute. Hasil percobaan tersebut dievaluasi performanya menggunakan normalized root mean squared error. Nilai rata-rata NRMSE pada percobaan yang dilakukan sebanyak lima kali memiliki nilai yang lebih rendah dibandingkan dengan metode bi-RLSP dan row average. Waktu komputasi untuk metode bi-KNN-RLSP dan bi-RLSP tidak jauh berbeda, sehingga dengan waktu yang tidak signifikan berbeda, metode bi-KNN-RLSP dapat menghasilkan nilai NRMSE yang lebih kecil dibandingkan dengan bi-RLSP. Oleh karena itu, dapat dikatakan bahwa modifikasi praimputasi row average pada metode bi-RLSP menjadi KNNimpute dapat menghasilkan performa imputasi yang lebih bagus. Selain itu, diperoleh hasil bahwa nilai NRMSE untuk metode bi-KNN-RLSP meningkat seiring dengan meningkatnya missing rate.

.....Microarray is a technology in biology that provides information about gene expression. The raw microarray data is in the form of images, which must be converted into a gene expression matrix where rows indicate genes, columns indicate experimental conditions. However, in practice, many missing values are found in microarray data, which of course will hinder the process of data analysis. Imputation is one solution that can overcome the missing values in microarray data. By using imputation, the missing values contained in the data matrix are predicted or estimated so that a complete data matrix is obtained. The imputation method used in this study is called bi-

KNN-RLSP, which uses the concept of biclustering, principal component analysis, and quantile regression. In the process of forming biclustering, a temporary complete matrix is needed which is obtained through the pre-imputation process with KNNimpute. The bi-KNN-RLSP experiment was carried out on cervical cancer cell line gene expression data by applying different missing rates, namely 1%, 5%, 10%, 15%, 20%, 25%, and 30% using the parameter $k=10$ in the KNNimpute pre-imputation process. The results of these experiments were evaluated for their performance using the normalized root mean squared error. The average value of NRMSE in the five times experiment has a lower value than the bi-RLSP and row average methods. The computation time for the bi-KNN-RLSP and bi-RLSP methods is not much different, so with the time that is not significantly different, the bi-KNN-RLSP method can produce a smaller NRMSE value compared to bi-RLSP. Therefore, it can be said that the modification of the row average preimputation in the bi-RLSP method to KNNimpute can produce better imputation performance. In addition, it was found that the NMRSE value for the bi-KNN-RLSP method increased along with the increase in the missing rate.