

Implementasi metode quadratic programming-biclustering (QP_biclustering) untuk imputasi missing values pada data ekspresi gen pasien covid-19 = Implementation of quadratic programming-biclustering (QP_biclustering) method for missing values imputation in gene expression data of covid-19 patients.

Allissa Rahman, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=20527727&lokasi=lokal>

Abstrak

Data ekspresi gen merupakan data yang berisi hasil proses transformasi informasi genetik dari DNA atau RNA menjadi protein (fenotipe) yang disajikan dalam bentuk matriks berisi ekspresi gen (baris) dan kondisi (kolom). Sama halnya seperti jenis data lainnya, data ekspresi gen tidak terlepas dari kemungkinan masalah missing values atau data yang hilang. Untuk menangani masalah keberadaan missing values, diperlukan imputasi, yaitu metode untuk mengisi nilai pada data yang hilang. Metode imputasi yang digunakan pada penelitian ini adalah QP_Biclustering yang merupakan kombinasi antara biclustering dengan permasalahan pemrograman kuadrat yang akan diselesaikan dengan bahasa pemrograman Python. Metode QP_Biclustering diimplementasikan pada data ekspresi gen dari pasien Covid-19 bertipe numerik dengan tipe missing values berjenis Missing Completely at Random (MCAR). Algoritma biclustering yang digunakan sebagai metode dalam penelitian ini adalah hasil modifikasi dari algoritma SwarmBCluster, yaitu constructive heuristic with local search. Metode biclustering ini menggunakan Mean Squared Residue (MSR) dalam proses pembentukan bicluster yang berperan sebagai ukuran kualitas suatu bicluster. Setelah bicluster terbentuk, missing values akan diestimasi menggunakan quadratic programming dengan cara meminimumkan nilai MSR tiap bicluster. Hasil penelitian ini menunjukkan bahwa metode QP_Biclustering memiliki performa cukup baik untuk mengimputasi missing values pada data ekspresi gen pasien Covid-19 yang dapat dilihat dari nilai Normalized Root Mean Square Error (NRMSE) yang lebih rendah dibandingkan dengan metode KNNimpute.

.....Gene expression data contains the results of the transformation process of genetic information from DNA or RNA into protein (phenotype) which is presented in the form of a matrix containing gene expression (rows) and conditions (columns). Just like other types of data, gene expression data can not be separated from the possibility of missing values or missing data problems. To deal with the problem of the presence of missing values, imputation is needed, which is a method to fill in the values in the missing data. The imputation method used in this research is QP_Biclustering which is a combination of biclustering with quadratic programming problems that will be solved using the Python programming language. The QP_Biclustering method is implemented on gene expression data from Covid-19 patients with numeric type and the missing values has Missing Completely at Random (MCAR) type. The biclustering algorithm used as a method in this research is a modified result of the SwarmBCluster algorithm, namely constructive heuristic with local search. This biclustering method uses Mean Squared Residue (MSR) in the bicluster formation process which will be used as a measure of the quality of a bicluster. After the bicluster is formed, the missing values will be estimated using quadratic programming by minimizing the MSR value for each bicluster. The results of this study indicate that the QP_Biclustering method has a fairly good performance to impute the missing values in gene expression data of Covid-19 patients which can be seen from the

NRMSE (Normalized Root Mean Square Error) value which is lower than KNNimpute method.