Metode imputasi missing values chronological biclustering dengan basis korelasi pearson, skor mean squared residue, dan jarak euclidean (PCor-MSRE) pada data ekspresi gen = Pearson correlation, mean squared residue score, and euclidean distance (PCor-MSRE) based chronological biclustering missing values imputation method on gene expression data

Silvia, author

Abstrak

Teknologi microarray merupakan analisis terhadap tingkat ekspresi puluhan ribu gen secara paralel untuk melihat perbedaan ekspresi gen. Penelitian microarray menghasilkan suatu nilai yang dirangkum dalam sebuah data yang disebut sebagai data ekspresi gen. Data ekspresi gen umumnya memiliki ukuran yang besar dan penggunaannya luas. Akan tetapi, data ekspresi gen sering mengalami masalah missing values. Data ekspresi gen umumnya mengandung persentase missing values sebesar 10% atau bahkan hingga 90% gen memiliki satu hingga lebih missing values. Salah satu solusi untuk mengatasi adanya missing values adalah dengan menggunakan teknik imputasi. Pada penelitian ini, diajukan metode imputasi missing values Chronological Biclustering dengan basis PCor-MSRE yang berdasarkan pada konsep biclustering. Penentuan anggota bicluster dengan kesamaan sifat co-expressed dan ukuran magnitude dilakukan berdasarkan pada skor Mean Squared Residue (MSR), jarak Euclidean, dan ukuran jarak korelasi Pearson antara masing-masing gen dengan gen yang mengandung missing values. Dilakukan perhitungan skor MSR, jarak Euclidean, dan ukuran jarak korelasi Pearson pada setiap gen, kemudian dipilih k gen yang memberikan skor terkecil untuk masing-masing kriteria. Selanjutnya, dibentuk bicluster yang digunakan untuk mengimputasi nilai observasi yang missing. Metode ini merupakan pengembangan dari metode SBi-MSREimpute yang cocok digunakan pada data ekspresi gen non-time series atau time series. Metode diimplementasikan pada data ekspresi gen lengkapnon-time series GSE142693 mengenai sel tumor 12 pasien Glioblastoma. Pada data GSE142693, dilakukan konstruksi missing values MCAR dengan missing rate sebesar 5%, 10%, 20%, 30%, 40%, 50%, dan 60%. Performa metode diukur dengan skor NRMSE dan korelasi Pearson, kemudian dibandingkan dengan metode SBi-MSREimpute. Berdasarkan pada skor korelasi Pearson, metode Chronological Biclustering dengan basis PCor-MSRE merupakan metode yang cukup baik dibanding SBi-MSREimpute dalam mengimputasi missing values pada data GSE142693 jika missing rate-nya cukup besar (40%, 50% dan 60%) dengan penggunaan nilai yaitu dan. Untuk nilai k yang lebih kecil dari 25, metode Chronological Biclustering dengan basis PCor-MSRE cukup baik digunakan (dibanding SBi-MSREimpute) jika jumlah observasi yang missing sebanyak 50% dan 60%. Performa metode Chronological Biclustering dengan basis PCor-MSRE semakin baik seiring dengan membesarnya nilai k yang digunakan. Artinya, performa metode Chronological Biclustering dengan basis PCor-MSRE dapat dipengaruhi oleh penentuan nilai k di awal.

.....Microarray technology is an analysis of the expression levels of tens of thousands of genes in parallel to see differences in gene expression. Microarray research produces a value that is summarized in a data called gene expression data. Gene expression data are generally large in size and widely used. However, gene expression data often suffer from missing values problems. Gene expression data generally contain a

percentage of missing values of 10% or even up to 90% of genes having one or more missing values. One solution to overcome the missing values is to use the imputation technique. In this research, the method of imputing missing values Chronological Biclustering is proposed on the PCor - MSRE basis which is based on the biclustering concept. Determination of bicluster members with similar co-expressed traits and magnitude measures was carried out based on the Mean Squared Residue (MSR) score, the Euclidean distance, and the measure of the Pearson correlation distance between each gene and the gene containing missing values. The MSR score, Euclidean distance, and Pearson correlation distance measures were calculated for each gene, then k genes were selected that gave the smallest score for each criterion. Next, a bicluster is formed which is used to impute the missing observation values. This method is a development of the SBi-MSRE impute method which is suitable for use in non-time series or time series gene expression data. The method was implemented on the complete non-time series gene expression data GSE142693 regarding tumor cells of 12 Glioblastoma patients. In the GSE142693 data, MCAR missing values were constructed with a missing rate of 5%, 10%, 20%, 30%, 40%, 50%, and 60%. The performance of the method was measured by the NRMSE score and Pearson correlation, then compared with the SBi-MSREimpute method. Based on the Pearson correlation score, the Chronological Biclustering method with PCor - MSRE basis is a method that is quite good compared to SBi-MSRE impute in imputing missing values in GSE142693 data if the missing rate is large enough (40%, 50% and 60%) with the use of namely k=25,k=45,k=65,k=105,k=335, and k=375. For k values less than 25, the Chronological Biclustering method on the basis of PCor - MSRE is quite good to use (compared to SBi-MSRE impute) if the number of missing observations are 50% and 60%. The performance of the Chronological Biclustering method on the PCor - MSRE basis is getting better as the value of k used increases. This means that the performance of the Chronological Biclustering method on the PCor-MSRE basis can be affected by determining the initial k value.