Metode iterative bicluster-based bayesian principal component analysis dan least square (bi-BPCA-iLS) untuk imputasi missing values pada data ekspresi gen = Iterative bicluster-based bayesian principal component analysis and least square (bi-BPCA-iLS) for missing values imputation in gene expression data

Yoel Fernando, author

Abstrak

Penelitian biologi dengan menggunakan teknologi microarray menghasilkan data ekspresi gen berbentuk matriks di mana baris adalah gen dan kolom adalah kondisi. Analisis lanjutan dalam data ekspresi gen membutuhkan data yang lengkap. Namun, data ekspresi gen sering kali mengandung nilai hilang atau missing values. Ada berbagai cara untuk mengatasi missing values, antara lain pembuangan gen atau kondisi yang mengandung missing values, pengulangan pengambilan data, dan imputasi missing values pada data ekspresi gen. Pendekatan imputasi missing values awal hanyalah dengan mengisi nilai nol atau rata-rata baris. Namun, pendekatan ini tidak melihat informasi koheren dalam data. Pendekatan imputasi missing values terbagi menjadi empat berdasarkan informasi yang diperlukan pada algoritmanya, yaitu pendekatan lokal, pendekatan global, pendekatan hybrid, dan pendekatan knowledge assisted. Pada penelitian ini peneliti menggunakan algoritma pendekatan lokal dan global. Metode imputasi missing values paling popular untuk pendekatan global adalah Bayesian Principal Component Analysis (BPCA), sedangkan untuk pendekatan lokal adalah Local Least Square (LLS). Pada metode LLS, pemilihan similaritas gen dilakukan dengan teknik clustering dimana seluruh kondisi dalam data digunakan. Kenyataanya, terkadang gen-gen similar hanya dalam beberapa kondisi eksperimental saja. Maka, diperlukan teknik biclustering untuk dapat menemukan subset gen dan subset kondisi yang similar sebagai informasi lokal. Penerapan ide biclustering dalam LLS dinamakan sebagai Iterative Bicluster-Based Least Square (bi-iLS). Salah satu tahapan awal dalam bi-iLS adalah pembentukan matriks komplit sementara yang didapat dengan cara mengisi missing values dengan row average. Namun, row average dinilai kurang bagus karena hanya menggunakan informasi satu baris tersebut. Kekurangan ini diperbaiki dalam penelitian ini. Penggunaan metode BPCA untuk menemukan matriks komplit sementara dinilai lebih baik karena BPCA menggambarkan struktur keseluruhan data. Penggantian row average menjadi BPCA menjadi dasar masalah penelitian ini. Metode iterative Bicluster-based Bayesian Principal Component Analysis dan Least Square (bi-BPCA- iLS) pun diajukan. Penerapan bi-BPCA-iLS terhadap data ekspresi gen yang dihasilkan teknologi microarray terbukti menghasilkan penurunan nilai Normalzied Root Mean Square Error (NRMSE) sebesar 10,6% dan 0,58% secara rata-rata dalam beberapa missing rate (1%, 5%, 10%, 15%, 20%, 25%, dan 30%) jika dibandingkan dengan metode LLS dan bi-iLS.

.....Biological research using microarray technique produce some important gene expression datasets. These data can be expressed as a matrix in which rows are genes and columns are different conditions. Further analysis of these datasets requires a complete dataset or matrix. However, gene expression datasets often contain missing values. There are some ways to handle missing values, such as deletion of genes or conditions that contain missing values, repeat the process of acquiring data, and impute the missing values. Early approaches in missing values imputation are simply to replace missing values with zeros or row

averages, but these methods do not use the coherence inside the data. Later, approaches in missing values imputations are categorized into four groups based on the required information, such as local, global, hybrid, and knowledge assisted approaches. In this paper, local and global approaches are used. Bayesian Principal Component Analysis (BPCA) is a well-known global based method, while the most popular local based method is Local Least Square (LLS). In LLS, selection of similar genes uses clustering technique where all conditions in the data are included. The reality is genes sometimes only correlate under some experimental conditions only. So, a technique that can find subset of genes under subset of experimental conditions for local information is needed. This technique is called biclustering. The usage of biclustering in LLS is called the Iterative Bicluster-based Least Square (bi-iLS). One of the early steps in bi-iLS is to find a temporary complete matrix. Temporary complete matrix is obtained by applying row averages to impute missing values. However, row average cannot reflect the real structure of the dataset because row average only uses the information of an individual row. The missing values in a target gene do not only rely on the known values of its own row. In this research, row average in bi-iLS is replaced with BPCA. The benefit of using BPCA is that it uses global structure of the dataset. This update will be the basic problem of this research. The proposed method is called Iterative Bicluster-based Bayesian Principal Component Analysis and Least Square (bi-BPCA-iLS). This new proposed method is applied to gene expression datasets from microarray technique. It shown a decrease in values of Normalized Root Mean Square Error (NRMSE) about 10.6% from LLS and about 0.58% from bi-iLS based on different missing rates (1%, 5%, 10%, 15%, 20%, 25%, and 30%).