Metode Bicluster-Based Bayesian Principal Component Analysis dan Robust Least Squares Estimation dengan Principal Components (bi-BPCA-RLSP) untuk Imputasi Missing Value pada Data Ekspresi Gen = Bicluster-Based Bayesian Principal Component Analysis and Robust Least Squares Estimation with Principal Components (bi-BPCA-RLSP) for Missing Values Imputation on Gene Expression Data

Alya Fadhilah Putri Banyu Nur Inayah, author

Abstrak

Data ekspresi gen adalah data yang menyajikan tingkat ekspresi gen pada kondisi yang berbeda-beda. Analisis data ekspresi gen microarray sangat penting karena memiliki aplikasi luas pada bidang biologi, medis, dan lain-lain. Dalam melakukan analisis data ekspresi gen, sebagian besar algoritma analisis membutuhkan data matriks yang lengkap. Sayangnya, beberapa data mungkin hilang karena kerusakan gambar, debu, dan kesalahan eksperimental. Oleh karena itu, metode imputasi missing value diperlukan untuk melakukan pemulihan pada data yang hilang tersebut. Penelitian ini mengembangkan suatu metode imputasi missing value, yaitu bicluster-based Bayesian principal component analysis dan robust least squares estimation dengan principal components (bi-BPCA-RLSP). Metode bi-BPCA-RLSP merupakan metode pengembangan dari bicluster-based robust least squares estimation dengan principal components (bi-RLSP). Pada metode bi-RLSP, tahap praimputasi untuk memperoleh matriks komplit sementara dilakukan dengan menggunakan metode row average. Namun, metode row average dinilai kurang baik dalam menggambarkan struktur keseluruhan data karena hanya menggunakan informasi dari baris yang mengandung missing value. Oleh karena itu, penelitian ini melakukan penggantian metode row average menjadi BPCA. BPCA menggunakan informasi korelasi dari seluruh data sehingga lebih baik dalam menggambarkan struktur keseluruhan data. Metode bi-BPCA-RLSP diterapkan pada data ekspresi gen garis sel kanker serviks dengan missing rate 1%, 5%, 10%, 15%, 20%, 25%, dan 30%. Performa metode bi-BPCA-RLSP diukur dengan menggunakan nilai normalized root mean squared error (NRMSE) dan dibandingkan dengan metode bi-RLSP. Hasil penelitian menunjukkan bahwa kinerja bi-BPCA-RLSP lebih baik daripada bi-RLSP karena nilai NRMSE pada bi-BPCA-RLSP lebih rendah dibandingkan bi-RLSP untuk setiap missing rate.

.....Gene expression data is data that presents the level of gene expression under different conditions. Analysis of microarray gene expression data is very important because it has wide applications in the fields of biology, medicine, and others. In analyzing gene expression data, most of the analytical algorithms require a complete data matrix. Unfortunately, some data may be lost due to image corruption, dust, and experimental errors. Therefore, the missing value imputation method is needed to recover the lost data. This study developed a missing value imputation method, namely bicluster-based Bayesian principal component analysis and robust least squares estimation with principal components (bi-BPCA-RLSP). The bi-BPCA-RLSP method is a development method of bicluster-based robust least squares estimation with principal components (bi-RLSP). In the bi-RLSP method, the pre-imputation stage to obtain a temporary complete matrix is carried out using the row average method. However, the row average method is considered poor in describing the overall structure of the data because it only uses information from rows containing missing

values. Therefore, this study replaced the row average method by BPCA. BPCA uses correlation information of all data so that it describes better the overall structure of the data. The bi-BPCA-RLSP method was applied to gene expression data of cervical cancer cell lines with missing rates of 1%, 5%, 10%, 15%, 20%, 25%, and 30%. The performance of the bi-BPCA-RLSP method was measured using the normalized root mean squared error (NRMSE) and compared with the bi-RLSP method. The results showed that bi-BPCA-RLSP performed better than bi-RLSP because the NRMSE value of bi-BPCA-RLSP was lower than bi-RLSP for each missing rate.