Universitas Indonesia Library >> UI - Skripsi Membership

Metode Biclustering Terurut Berbasis k-Nearest Neighbour, Mean Square Residual, dan Jarak Euclidean dalam Imputasi Missing Values pada Data Ekspresi Gen = Sequential Biclustering Method Based on k-Nearest Neighbor, Mean Squared Residual, and Euclidean Distance in Missing Values Imputation on Gene Expression Data

Fenni Amalia, author

Deskripsi Lengkap: https://lib.ui.ac.id/detail?id=20528396&lokasi=lokal

---

Abstrak

Bioinformatika merupakan ilmu yang ditujukan untuk menganalisis informasi biologis. Dalam perkembangan penelitian bioinformatika, data diperoleh salah satunya dengan menggunakan teknologi microarray. Teknologi microarray digunakan oleh lingkup biologi molekuler dalam melihat perbedaan tingkat ekspresi gen dengan cara mengonversi gambar monokromik yang berisi ratusan bahkan ribuan gen dari sampel sel dan menghasilkan data ekspresi gen. Teknologi microarray sering kali menghasilkan data ekspresi gen yang hilang atau tidak terdeteksi akibat adanya kesalahan teknis. Oleh karena itu, diperlukannya suatu metode imputasi pada data untuk mengatasi missing values. Pada penelitian ini, akan dikembangkan suatu metode imputasi yang disebut Biclustering Terurut berbasis k-Nearest Neighbor, Mean Squared Residual, dan Jarak Euclidean. Metode ini merupakan metode imputasi berbasis biclustering dimana bicluster dibentuk berdasarkan suatu kriteria yang melibatkan skor Mean Squared Residue dan Jarak Euclidean. Penggunakan k-Nearest Neighbor sebagai metode pra-imputasi didasarkan pada data ekspresi gen yang sering kali memiliki pola kompleks dan sulit terdeteksi, sehingga perlu pendekatan yang dapat memetakan struktur korelasi pada data. k-Nearest Neighbor mempertimbangkan korelasi pada data microarray dengan menyeleksi kumpulan gen yang memiliki profil ekspresi mirip dengan gen yang ingin diimputasi (gen target). Pada penelitian ini, metode SBi-kNN-MSREimpute diterapkan pada data ekspresi gen pasien penderita COVID-19 yang dilakukan tes rapid harian. Evaluasi kinerja metode SBi-kNN-MSREimpute dilakukan dengan menggunakan NRMSE, dimana hasilnya dibandingkan dengan metode SBi-MSREimpute. Berdasarkan penelitian yang dilakukan, metode SBi-kNN-MSREimpute dinilai lebih baik dibandingkan dengan SBi-MSREimpute untuk setiap missing rate pada tingkatan c berbeda. Nilai c optimal untuk imputasi missing values pada data COVID-19 adalah c = 10% untuk missing rate 25%, 30%, 40% dan c = 15% untuk missing rate 5%, 10%, 15%, 20%, dan 50%. Hasil akhir juga menunjukkan bahwa nilai NRMSE untuk SBi-kNN-MSREimpute relatif stabil bahkan untuk data dengan missing rate tinggi hingga 50%.

.....Bioinformatics is a study designed to analyze biological information. In the development of bioinformatics research, data was obtained using microarray technology. Microarray technology is used by the scope of molecular biology in transposing hundreds and even thousands of genes from cellular samples simultaneously and producing a gene expression data. Microarray technology often produces data that is lost or undetected as a result of technical error. Therefore, an imputation method is needed to address the missing values. In this study, a new imputation method called Sequential Biclustering based k-Nearest Neighbor, Mean Squared Residual, and Euclidean Distance (SBi-kNN-MSRE) will be developed. This method is a biclustering-based imputation method where the bicluster is formed based on a criterion involving Mean Squared Residue and Euclidean Distance. The use of k-Nearest Neighbor as a pre-

imputation method is based on data on gene expression that often has a complex and difficult pattern of detection, so it requires an approach that can map correlation structures on data. K-nearest neighbor considers a correlation on a microarray data by selecting groups of genes that have an expression profile similar to a gene that wants to be imputed (the target gene). In this study, the SBi-kNN-MSRE method was applied to the data on the genes of patients with covid-19 that daily rapid tests were performed. The performance evaluation of the SBi-kNN-MSRE method is done using NRMSE, where the results are compared to the SBi-MSRE method. According to the result, the SBi-kNN-MSRE method performed better than SBi-kNN-MSRE for each missing rate on different c levels. The optimal c value on the covid-19 data is c = 10% for missing rate 25%, 30%, 40% and c = 15% for missing rate 5%, 10%, 15%, 20% and 50%. The results also showed that NRMSE scores