Analisis perbandingan kinerja antara metode imputasi biclustering berbasis Shifting and Scaling Similarity (SSSim) dan euclidean score pada data ekspresi gen kanker usus besar = Performance comparison analysis between of biclustering based Shifting and Scaling Similarity (SSSim) and euclidean score for missing values imputation on colon cancer gene expression data.

Panjaitan, Andreas Pangihutan, author

Abstrak

Kebutuhan data di zaman sekarang semakin meningkat seiring dengan perkembangan teknologi. Penggunaan dataset dengan ukuran besar sudah menjadi keperluan dalam berbagai bidang, termasuk kebutuhan data di bidang bioinformatika, yang dihasilkan melalui teknologi microarray berbentuk matriks berisi gen dan dan kondisi observasi. Sulit untuk menghasilkan data ekspresi gen yang sempurna dan tidak ada kekurangan karena berbagai keterbatasan dalam proses pengumpulan data. Kehadiran nilai hilang atau missing values pada data ekspresi gen adalah hal yang tidak dapat dihindarkan, sehingga dapat mengganggu jalannya proses analisis data lanjutan. Pada penelitian ini, keberadaan missing values pada data diatasi dengan metode imputasi biclustering berbasis Shifting and Scaling Similarity (SSSim) dan imputasi biclustering berbasis euclidean score. Metode imputasi biclustering berbasis SSSim dan imputasi biclustering berbasis euclidean score adalah 2 metode imputasi berbeda yang dikombinasikan dengan konsep biclustering yang berbeda. Kedua metode imputasi biclustering ini menggunakan konsep least square dan pembobotan gen dalam proses imputasinya, serta menggunakan konsep korelasi SSSim dan korelasi euclidean score dalam proses biclustering-nya. Kedua konsep korelasi tersebut memiliki perbedaan prinsip yang saling berkebalikan, di mana korelasi SSSim dapat mendeteksi pola shifting and scaling dalam data ekspresi gen sedangkan korelasi euclidean score tidak dapat mendeteksi pola shifting and scaling. Metode imputasi biclustering berbasis SSSim dan imputasi biclustering berbasis euclidean score diaplikasikan pada data eskpresi gen kanker usus besar dan diukur tingkat performanya bersama dua metode pembanding lain yaitu K-Nearest Neighbor Imputation (KNNimpute) dan column mean impute menggunakan nilai Root Mean Squared Error (RMSE). Berdasarkan penelitian ini, metode imputasi biclustering berbasis SSSim dan imputasi biclustering berbasis euclidean score memiliki tingkat akurasi yang hampir sama, tetapi secara konsisten lebih baik dari metode KNNimpute dan column mean impute pada data dengan missing rate (5%,10%,15%,20% dan 25%).

.....The need for data today is increasing along the technological advances. The use of large data sets has become a necessity in various fields, including the need for data in bioinformatics, which is generated through microarray technology and produce data’s form of a matrix containing genes type and genes observation. It is difficult to produce perfect gene expression data, due to various limitations in the data collection process. The presence of missing values in gene expression data is unavoidable, so it can interfere further analysis. In this research, the presence of missing values was handled by the biclustering based on Shifting and Scaling Similarity (SSSim) and biclustering based on euclidean score for missing values imputation. Biclustering based on Shifting and Scaling Similarity (SSSim) and biclustering based on euclidean score for missing values imputation are 2 different imputation methods combined with

biclustering concepts. This two methods use the least square concept and gene weighting in the imputation process, and use the SSSim and the Euclidean score correlation in the biclustering process. This two correlation concepts have contradictory basic principles, where SSSim correlation can detect shifting and scaling patterns in gene expression data while Euclidean score correlation cannot detect. Biclustering based on Shifting and Scaling Similarity (SSSim) and biclustering based on euclidean score for missing values imputation were applied to colon cancer gene expression data and their performance level was measured by Root Mean Squared Error (RMSE) with two other comparison methods, namely K-Nearest Neighbor Imputation (KNNimpute) and column mean impute. Based on this study, biclustering based on Shifting and Scaling Similarity (SSSim) and biclustering based on euclidean score for missing values imputation has almost the same accuracy level, but consistently better than the KNNimpute method and column mean impute on data with missing rate (5%, 10%, 15%, 20% and 25%).