Universitas Ahmad Dahlan

Jurnal Ilmiah Matematika
Vol. 8, No. 2, October 2021, pp. 48-55
ISSN 2774-3241
http://journal.uad.ac.id/index.php/Konvergensi

Analysis of diabetes mellitus gene expression data using two-phase biclustering method

Rahmat Al Kafi [a,1,*], **Alhadi Bustamam** [a,2], **Wibowo Mangunwardoyo** [a,3]

[a] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia
[1] rahmat.alkafi@sci.ui.ac.id; [2] alhadi@sci.ui.ac.id*; [3] wibowo.mangun@ui.ac.id
*Correspondent Author

KEYWORDS

Clustering
Singular Value Decomposition
K-Means
Silhouette

ABSTRAK

The purpose of this research is to find bicluster from Type 2 Diabetes Mellitus genes expression data which samples are obese and lean people using two-phase biclustering. The first step is to use Singular Value Decomposition to decompose matrix gene expression data into gene and condition based matrices. The second step is to use K-means to cluster gene and condition based matrices, forming several clusters from each matrix. Furthermore, the silhouette method is applied to determine the number of optimum clusters and measure the accuracy of grouping results. Based on the experimental results, Type 2 Diabetes Mellitus dataset with 668 selected genes produced optimal biclusters, with six biclusters. The obtained biclusters consist of 2 clusters on the gene-based matrix and 3 clusters on the sample-based matrix with silhouette values, respectively, are 0.7361615 and 0.7050163.

Introduction

Data analysis method consists of various approaches. One type of descriptive data analysis is cluster analysis. Cluster analysis can be used to group some objects based on the similarity of their characteristics. Most of the standard clustering literature focuses on one-way clustering. One-way clustering assumes that objects have characteristics across rows or columns, meaning that the objects in rows are grouped by the similarity in columns or variables in columns are grouped by the similarity in rows.

In contrast, two-way clustering analysis or bicluster analysis assumes that certain objects only have characteristics on a particular row or column. Identifying subgroups of rows and subgroups of columns interrelated is an objective in two-way clustering. Thus, the bicluster analysis needs to be considered as the method of data analysis. There are various biclustering methods, one of which is the two-phase biclustering method. Several researchers have developed and modified this method and have implemented their methods to different kinds of data. For example, in Indonesia Cahyaningrum et al. (2017) [1] combined Spectral Clustering and PAM, Frisca et al. (2017) [2] combined Spectral Clustering and KMeans Clustering, and both Cahyaningrum and Frisca implemented their method to Microarray Data of Carcinoma. combined K-Means Clustering and Plaid Model and implemented their combined method to Microarray Data of Carcinoma and Adenoma Tumor [3]. Formalidin et al. (2018) combined SVD and Hybrid Clustering, and Puspa et al. (2018) combined The χ-Sim Co-Similarity Measure and K-means Partition Clustering, and both Formalidin and Puspa implemented their method to

 ISSN 2774-3241

Microarray Data of Lymphoma [4].

This research proposed the two-phase biclustering method that combined SVD and K-Means Clustering for finding the biclusters in Type 2 Diabetes Mellitus gene expression data whose samples are obese and lean people. The first phase is to use the matrix method of decomposition of Singular Value Decomposition (SVD) which transforms the matrix into two matrices based on genes and samples that are suitable for analyzing and modelling gene expression data. The second phase is the two-way clustering process using the K-means that form m clusters from the set of gene and n clusters from the set of a sample. This research is expected to produce submatrices that classify genes and samples simultaneously.

BASIC THEORIES

Log-Mean Centering Normalization

The values of gene expression obtained from microarray observations are very high. These values are due to the difference in the patient's body condition. Therefore, normalization is required to reduce the large values of gene expression data as well as accelerating computational time. Let A denote the $d \times N$ size matrix that represents a gene expression data. Then the Log-Mean Centering Normalization for the matrix A is given by Equation (1)

$$\hat{A} = A - \text{mean}(\log(A)) \tag{1}$$

where $\hat{A}$ is the matrix after normalization and A is the matrix before normalization [5].

Singular Value Decomposition

Let $A \in \Re^{d \times N}$ is a gene expression data matrix with rank r. Singular Value Decomposition (SVD) of A is decomposition of matrix A into a product of matrices

$$A = U\Sigma V^T = \sum_{i=1}^{r} u_i \lambda_i v_i^T \tag{2}$$

where r is the rank of matrix A, U and V respectively are matrixes with size $d \times r$ and $N \times r$ with an orthonormal columns $(U^T U = V^T V = I_r)$, Σ is a diagonal matrix with size $r \times r$ where the main diagonal elements λ_i are the eigen values of matrix $A^T A$ $(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r)$, and the columns of V are the eigen vectors of $A^T A$. The columns of U are obtained from $u_i = \frac{1}{\lambda_i} v_i$, $i = 1, 2, \ldots, r$ with u_i being the i-th column of U, v_i being the i-th column of V, and λ_i being the i-th eigen value of $A^T A$ [6].

From Equation (2), we can construct further equation as,

$$\begin{aligned} A &= \sum_{i=1}^{r} u_i \lambda_i v_i^T \\ &= \sum_{i=1}^{r} (u_i \sqrt{\lambda_i})(\sqrt{\lambda_i} v_i^T) \\ &= \left[u_1\sqrt{\lambda_1}, u_2\sqrt{\lambda_2}, \cdots, u_r\sqrt{\lambda_r}\right]\left[\sqrt{\lambda_1} v_1, \sqrt{\lambda_2}\, v_2, \cdots, \sqrt{\lambda_r}\, v_r\right]^T \\ &= GC^T \end{aligned}$$

Usually not all columns of matrices G and C are needed to reconstruct gene expression pattern with reasonable accuracy. We may use a truncated SVD expression as illustrated in Figure 1, and the following theorem is the property of truncated SVD.

 ISSN 2774-3241

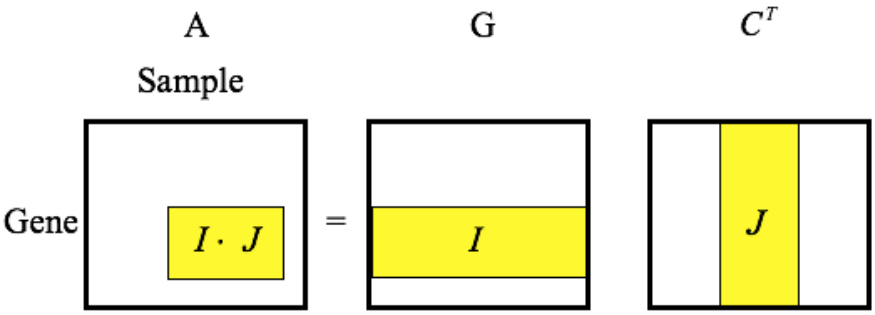

Figure 1. The truncated SVD of A

Given SVD of matrix A with size $d \times N$ as seen in Equation (2). If $l \leq r$ where $r = \text{rank}(A)$, then

$$A^{(l)} = \sum_{i=1}^{l} u_i \lambda_i v_i^T \tag{3}$$

Equation (3) shows how the dimension of matrix A can be reduced and gives us the information that analysis of singular values by reducing the dimension of the characteristic mode matrix C can also describe the original pattern of gene expression data [7].

K-Means Clustering

K-Means clustering is one of the partitioning methods that aims to partition n observations into two or more clusters in which each observation belongs to the cluster with the nearest mean. This following is the steps of the K-means algorithm [8]:

1. Determine the number of cluster k and centroid.
2. Determine the distance of each object to the centroids using the Euclidean distance by using Equation (4)

$$A^{(l)} = \sum_{i=1}^{l} u_i \lambda_i v_i^T \tag{4}$$

where d_{ik} is the distance of object i and centroid k, n is a dimension of feature, x_i is the coordinate of an object i, c_k is the coordinate of k-th centroid.

3. Group the objects based on a minimum distance.
4. Determine new centroid.
5. Repeat procedure in point 1, 2, 3 and 4 until no object moves its assigned group.

Silhouette Method

The silhouette method is used to measure the average of similarity between each cluster and its most similar one. The formula of this method is given by Equation (5)

$$s(i) = \frac{y(i) - x(i)}{\max\{x(i), y(i)\}} \tag{5}$$

where $x(i)$ is the average of the dissimilarity (Euclidean distance) value of the i-th data to all members of the group containing data i and $y(i)$ is the smallest average of the i-th data dissimilarity to all members of the group not including data i. In this study, the mean of the silhouette $\bar{s}(k)$ is used to determine the number of the group. The highest $\bar{s}(k)$ indicates that the obtained clusters are the representative cluster to demonstrate the actual condition [9].

 ISSN 2774-3241

Gene Selection Method

In statistics, an outlier means an observation that is far removed from other observations [10][11]. In other words, it is an extreme value where this value need to be examined, given that there is important information that influences decision making. Gene selection, as well as the case of an outlier, will select genes that have important information by taking extreme genes. The existence of this gene selection is because the dataset of gene expression has thousands or even tens of thousands of genes with relatively few samples, whereas some of the genes have less variable expression values. Genes that do not have this extreme value will be wasted, considering that they do not have important information or in other words reduce noise data [12].

The gene selection method used in this research is the relative deviation and absolute deviation. Here are the explanations of each deviation [13]:

1. Relative Deviation
   This method is defined as the absolute value of the ratio between the maximum and minimum values of gene expression. The selected genes are genes that have relative deviation greater than the threshold δ. Mathematically written as follows:
   $$\left|\frac{\max x(i)}{\min x(i)}\right| \geq \delta,$$
   where $x(i)$ represents gene expression value on row i. Genes that have relative deviation less than the threshold δ are considered genes that are noise and need to be deleted.
2. Absolute Deviation
   This method is defined as the absolute value of range of the maximum and minimum values of gene expression. The selected genes are genes that have absolute deviation greater than the threshold θ. Mathematically written as follows:
   $$|\max x(i) - \min x(i)| \geq \theta,$$
   where $x(i)$ represents gene expression value on row i. Genes that have relative deviation less than the threshold θ are considered genes that are noise and need to be deleted.

Method

The two-phase biclustering method in this research is applied on gene expression data of Diabetes Melitus which has 14,063 genes and 18 samples and is processed in R software. First, the data is selected using relative and absolute deviation as [13] did in Carcinoma data expressions, that is genes with $\left|\frac{\max x(i)}{\min x(i)}\right| \geq \delta$ and $|\max x(i) - \min x(i)| \geq \theta$ are selected, leaving a total of k genes (with $k < 14{,}063$) and 18 samples depending on the chosen threshold. In the preprocessing step, the data is normalized using Log Mean Centering, producing the data between -8,00 and 8,00. In the first phase, the normalized data is decomposed to become two based matrices by using SVD. The two based matrices are gene-based and sample-based matrices. By applying Theorem 1, we get the reduced form of the original SVD in order to get the information about the main genes that correlate to the samples. In the second phase, we applied the K-means - Silhouette to the gene-based and the 4 sample-based matrices obtaining m clusters of genes and n clusters of samples. Then, we combine them to form $m \times n$ biclusters.

In summary of our discussion so far, our new approach algorithm is given in the following flowchart figure 2.

 ISSN 2774-3241

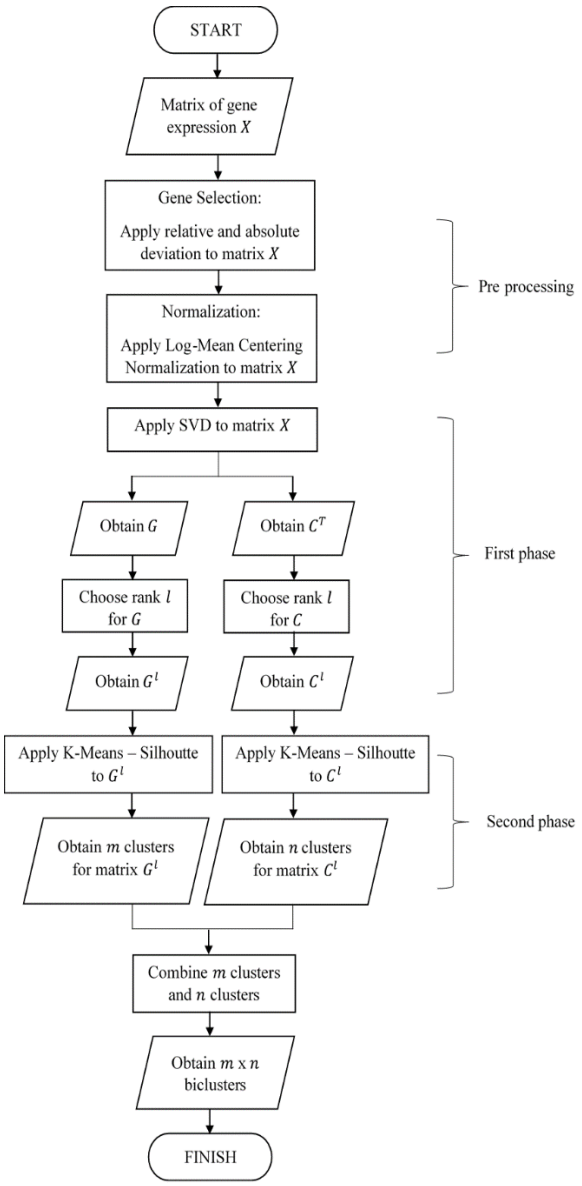

Figure 2. Flowchart of Two-Phase Biclustering Method

Results and Discussion

In our study, Type 2 Diabetes Mellitus gene expression data is considered. The data is obtained from the official website of National Center for Biotechnology Information [14][15]. There are 14,063 kinds of genes and 18 samples of microarray data in the form of a matrix. The 18 samples consisting of 5 lean women not affected by DM disease, 1 obese male not affected by DM disease, 3 obese women not affected by DM disease, 1 obese man controlled by DM disease, 4 obese women were controlled by DM disease, and 2 women and 2 obese men who were under control were affected by the disease.

First, the data is reduced by using relative and absolute deviation with results obtained is presented in Table 1 and Table 2. We can see from the table that the number of genes selected on different thresholds does not affect the number of clusters for both matrix C.

 ISSN 2774-3241

TABLE 1. SELECTED GENES AND SILHOUETTE INDEX FOR THE GENE-BASED MATRIX G.

Threshold $(\frac{\theta}{\delta})$	*Number of Selected Genes*	*Silhouette Index*	*Number of Clusters*
120,000/130	36	0.738518	2
90,000/110	69	0.7523805	2
60,000/70	235	0.7630927	2
20,000/50	668	0.7361615	2
10,000/30	1571	0.6965364	2
5,000/20	2996	0.6554177	2

TABLE 2. SELECTED GENES AND SILHOUETTE INDEX FOR SAMPLE-BASED MATRIX C.

Threshold $(\frac{\theta}{\delta})$	*Number of Selected Genes*	*Silhouette Index*	*Number of Clusters*
120,000/130	36	0.590976	2
90,000/110	69	0.4557817	3
60,000/70	235	0.6802374	2
20,000/50	668	0.7050163	3
10,000/30	1571	0.743409	2
5,000/20	2996	0.7613011	2

The best number of a bicluster is obtained for threshold 20.000/50, with six biclusters. The number of genes and conditions in each bicluster is given in Table 3, heatmap of original biclusters from 668 selected genes figure 3.

TABLE 3. THE NUMBER OF GENES AND CONDITIONS IN EACH BICLUSTER.

Bicluster	*Number of Genes*	*Number of Conditions*
1	351	2
2	351	8
3	351	8
4	317	2
5	317	8
6	317	8

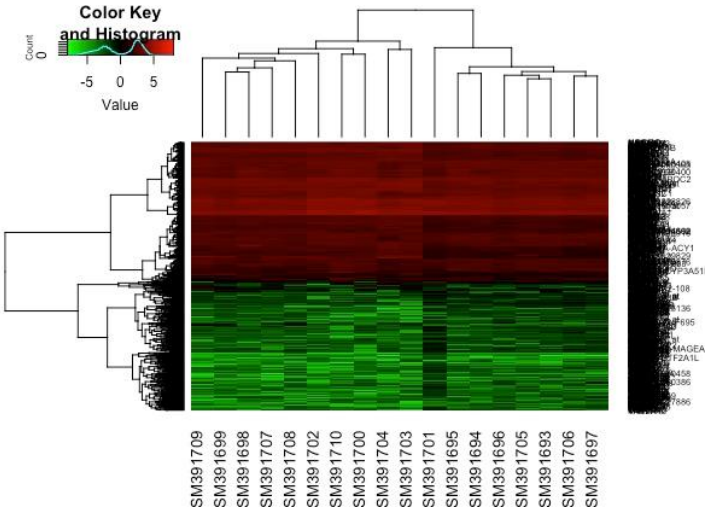

Figure 3. The heatmap of original biclusters from 668 selected genes.

 ISSN 2774-3241

Conclusion

In this paper, gene expression data of Type 2 Diabetes Mellitus could be clustered among genes and conditions through Two-Phase Biclustering method which shows the important result in biclustering gene expression data. In the first step, gene expression data is transformed into two global matrices using SVD. The second step is by using K-Means-Silhouette to get the best m×n original biclusters. Our results show that Type 2 Diabetes Mellitus gene expression data produces six biclusters as the optimal biclusters. The discovered biclusters can be used to analyze the samples that are likely to contract Type 2 Diabetes Mellitus disease. Also, analysis of samples from the six biclusters show that obese people are likely to get Type 2 Diabetes Mellitus disease. This result has a high potential to aid medical practitioners in the follow up of a disease suffered by the patient.

ACKNOWLEDGEMENT

This research is supported by the Indonesia Ministry of Research and Higher Education (Kemenristekdikti), with PDUPT research grant scheme 2019.

References

[1] R. D. Cahyaningrum, A. Bustamam, and T. Siswantining, "Implementation of spectral clustering with partitioning around medoids (PAM) algorithm on microarray data of carcinoma," 2017, p. 020007, doi: 10.1063/1.4978976.

[2] Frisca, A. Bustamam, and T. Siswantining, "Implementation of spectral clustering on microarray data of carcinoma using k-means algorithm," 2017, p. 020008, doi: 10.1063/1.4978977.

[3] G. Ardaneswari, A. Bustamam, and D. Sarwinda, "Implementation of plaid model biclustering method on microarray of carcinoma and adenoma tumor gene expression data," *J. Phys. Conf. Ser.*, vol. 893, p. 012046, Oct. 2017, doi: 10.1088/1742-6596/893/1/012046.

[4] A. Bustamam, S. Formalidin, and T. Siswantining, "Clustering and analyzing microarray data of lymphoma using singular value decomposition (SVD) and hybrid clustering," 2018, p. 020220, doi: 10.1063/1.5064217.

[5] W.-H. Yang, D.-Q. Dai, and H. Yan, "Finding Correlated Biclusters from Gene Expression Data," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 4, pp. 568–584, Apr. 2011, doi: 10.1109/TKDE.2010.150.

[6] I. Jolliffe, "Principal Component Analysis," in *Encyclopedia of Statistics in Behavioral Science*, Chichester, UK: John Wiley & Sons, Ltd, 2005.

[7] O. Alter and G. H. Golub, "Singular value decomposition of genome-scale mRNA lengths distribution reveals asymmetry in RNA gel electrophoresis band broadening," *Proc. Natl. Acad. Sci.*, vol. 103, no. 32, pp. 11828–11833, Aug. 2006, doi: 10.1073/pnas.0604756103.

[8] I. Bin Mohamad and D. Usman, "Standardization and Its Effects on K-Means Clustering Algorithm," *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 17, pp. 3299–3303, Sep. 2013, doi: 10.19026/rjaset.6.3638.

[9] L. Kaufman and P. J. Rousseeuw, *An introduction to cluster analysis*. John Wiley and Sons, Incorporated, 1990.

[10] G. S. Maddala, "Introduction to Econometrics," *Introd. to Econom. (2nd ed.). New York MacMillan*, pp. 88–96, 1992.

[11] A. Bustamam, S. D. Puspa, and T. Siswantining, "Implementation of co-similarity measure on microarray data of lymphoma using K-means partition algorithm," 2018, p. 020222, doi: 10.1063/1.5064219.

[12] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Stat. Assoc.*, vol. 97, no. 457, pp. 77–

87, Mar. 2002, doi: 10.1198/016214502753479248.

[13] T. R. Golub *et al.*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science (80-.).*, vol. 286, no. 5439, pp. 531–537, Oct. 1999, doi: 10.1126/science.286.5439.531.

[14] NCBI, "Diabetes Melitus Data: Obese patients with and without type 2 diabetes: liver," 2009. https://www.ncbi.nlm.nih.gov/sites/GDSbrowser accessed on March 2020.

[15] Y. Shao, H. Shao, M. S. Sawhney, and L. Shi, "Serum uric acid as a risk factor of all-cause mortality and cardiovascular events among type 2 diabetes population: Meta-analysis of correlational evidence," *J. Diabetes Complications*, vol. 33, no. 10, p. 107409, 2019, doi: https://doi.org/10.1016/j.jdiacomp.2019.07.006.