Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 59 (2015) 171 – 177

Procedia Computer Science

International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)

Intelligent Kernel K-Means for Clustering Gene Expression

Teny Handhayani[a], Lely Hiryanto[b]

[a]Computer Science Department, Tarumanagara University, Jl. S. Parman No 1 Gedung R Lantai XI, Jakarta 11440, Indonesia
[b]Computer Science Department, Tarumanagara University, Jl. S. Parman No 1 Gedung R Lantai XI, Jakarta 11440, Indonesia

Abstract

Intelligent Kernel K-Means is a fully unsupervised clustering algorithm based on kernel. It is able to cluster kernel matrix without any information regarding to the number of required clusters. Our experiment using gene expression of human colorectal carcinoma had shown that the genes were grouped into three clusters. Global silhouette value and davies-bouldin index of the resulted clusters indicated that they are trustworthy and compact. To analyze the relationship between the clustered genes and phenotypes of clinical data, we performed correlation (CR) between each of three phenotypes (distant metastasis, cancer and normal tissues, and lymph node) with genes in each cluster of original dataset and permuted dataset. The result of the correlation had shown that Cluster 1 and Cluster 2 of original dataset had significantly higher CR than that of the permuted dataset. Among the three clusters, Cluster 3 contained smallest number of genes, but 16 out of 21 genes in that cluster were genes listed in Tumor Classifier List (TCL).

Peer-review under responsibility of organizing committee of the International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)

*Keywords:*Kernel K-Means; Human Colorectal Carcinomal; Unsupervised Clustering Algorithm; Tumor Classifier List;

1. Introduction

Biologists usually spend almost one year to analyze huge amount of newfound human genes that may contribute to cancer disease. In the field of Bioinformatics, there exist methods that could make the analysis time far more efficient. One of the methods is clustering, which can group genes based on their closest similarity. Hence, it can help find new genes having closest similarity with known highly suspected cancer genes. The Biologists can further

Corresponding author:
Email: tenyh@fti.untar.ac.id, lelyh@fti.untar.ac.id

Peer-review under responsibility of organizing committee of the International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)
doi:10.1016/j.procs.2015.07.544

analyze those new genes to confirm whether they contribute to cancer disease, but in far less amount of genes.

Conventional clustering requires the number of clusters in advance, e.g. K-Means. In certain cases especially in clustering human colorectal carcinoma (cancer genes)[1-6,] it can be difficult to define the number of clusters in advance, and fully unsupervised clustering is the solution[7,8]. Handhayani et. al[9] proposed a fully unsupervised clustering algorithm that combines Intelligent K-Means[7,8] and Kernel K-Means[10]. Intelligent K-Means can be implemented successfully for analyzing genes in colorectal carcinoma disease[11], but it only clusters data on input domain. Kernel K-Means[10], however, can be used for clustering kernel matrix but it still needs to know the number of clusters in advance. The aim of clustering kernel matrix is to analyze non-linearly separable data which most of the human gene expression are in that type of data. More significant information can be gained from Kernel Matrix. Current progress fulfils the first objective of our research, implementing the Intelligent Kernel K-Means proposed in [9] to cluster gene expression of colorectal carcinoma and calculate correlation ratio of the clustered genes with clinical data to further analyze the performance of the algorithm.

2. Intelligent Kernel K-Means

Intelligent Kernel K-Means (IKKM) is a new clustering technique that can be used to cluster the data in the feature space. It is able to cluster kernel-based integration data of gene expression and DNA copy number[9]. Figure 1 depicts steps in fully unsupervised clustering gene expression of human colorectal carcinoma using IKKM into *n*-clusters.

In the preprocessing step, we only used complete gene expression values of 111 tissuesfrom 341 genes, as conducted by Muro et.al[12] and Ma'sum et.al[11]. Using those values, the kernel matrix is generated using Linear Kernel function of Equation (1). The generated kernel matrix is then fed into the clustering step using IKKM.

In the third step, firstly anomalous pattern algorithm is used to find centroid the first two centroids, as follows:

1. Compute Center of Mass (*CoM*) using Equation (2) from Taylor and Cristianini[10],
2. Find object C_1 having the farthest distance from *CoM* using Equation (3) from Taylor and Cristianini[10],
3. Find another object C_2 having the farthest distance from C_1 using Equation (4) from Taylor and Cristianini[10],
4. Calculate distance of other objects against centroid C_1 and C_2.
5. Group the objects; Objects with distance closest to C_1 are labelled as cluster S_1, while the others with distance closes to C_2 are labelled as cluster S_2.

Figure 2 shows illustration of those five steps for the first iteration of the algorithm. Next, Iterative Kernel Anomalous Pattern (see Figure 3 for illustration) is used to find other centroids, as follows:

1. For each cluster S_i, find its new centroid candidate, C_i', by finding the farthest object from its group centroid C_i using Equation (4), where $i = 1,2, \ldots,n\text{-}1$.
2. Optimize distance among the centroids by finding the mean of distance value of each candidate, C_{n_i}, with all group centroids, C_{n_i}= **mean**($d(C_1,C_i')$, $d(C_2,C_i')$, ..., $d(C_{n\text{-}1},C_i')$)
3. The new centroid, C_n = **max**$(C_{n_1}, C_{n_2}, \ldots, C_{n_{n-i}})$.
4. Re-group all the objects according to their nearest distance with one of centroids $\{C_1,C_2, ..,C_n\}$; Hence, there may be cluster of $\{S_1, S_2, .., S_n\}$.
5. Repeat step 1 until there is no object change its cluster.

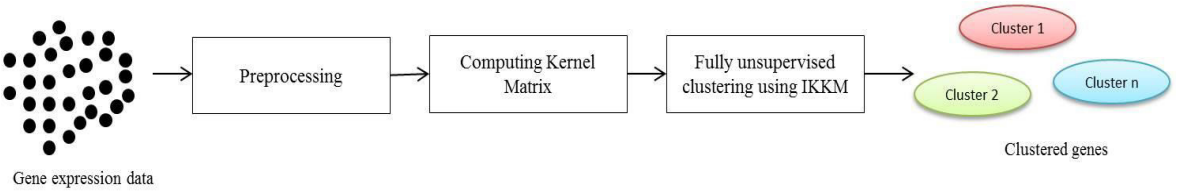

Fig. 1Flowchart for Fully Unsupervised Clustering using Intelligent Kernel K-Means

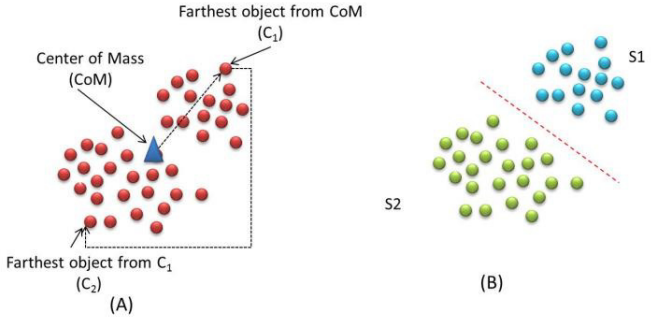

Fig. 2 First Iteration of Anomalous Pattern Algorithm

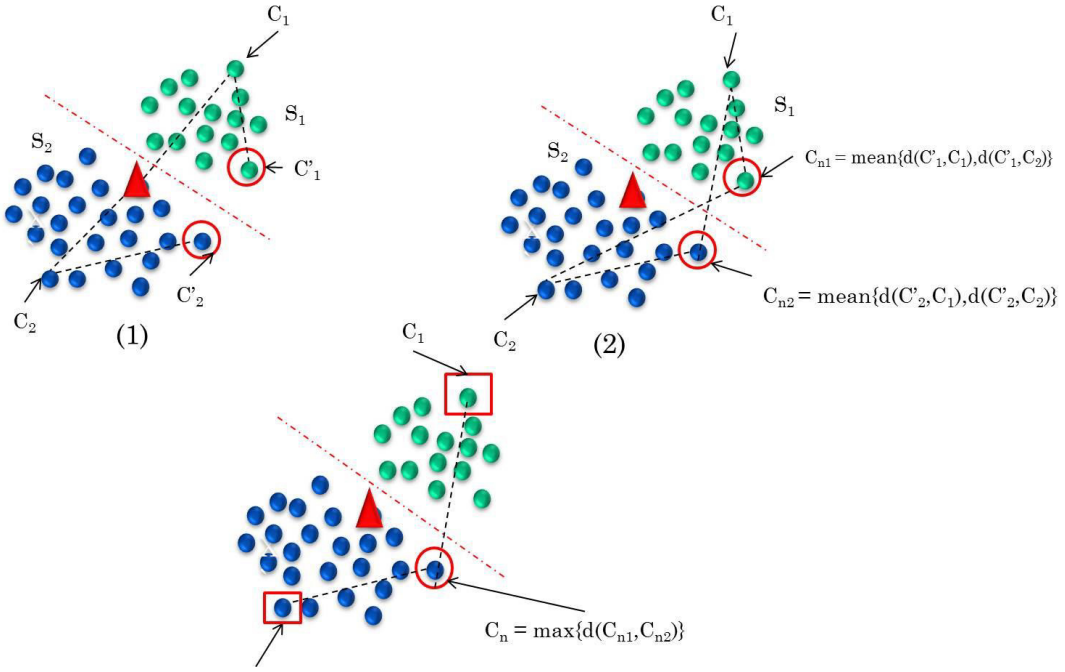

Fig. 3Next Iteration of Anomalous Pattern Algorithm

$$K = X.X^T \tag{1}$$

$$\emptyset(S) = \frac{1}{l}\sum_{i=1}^{l} \emptyset(x_i) \tag{2}$$

$$\|\emptyset(x) - \emptyset_S\|^2 = K(x,x) + \frac{1}{l^2}\sum_{i,j=1}^{l} K(x_i,x_j) - \frac{2}{l}\sum_{i=1}^{l} K(x,x_i) \tag{3}$$

$$\|\emptyset(x) - \emptyset(z)\|^2 = K(x,x) - 2K(x,z) + K(z,z) \tag{4}$$

Notes:

- K is kernel matrix
- X is gene expression data
- $\emptyset(S)$ is center of mass
- l is size of kernel matrix
- x and z are vectors in kernel matrix
- i, j are index vector in kernel matrix.

3. Experimental Design

We used dataset consisting of 1536 genes of human colorectal carcinoma, each gen consists of111 tissues (100 cancer tissues and 11 normal tissues). After preprocessing, we only got 341 representative genes. Not only using the original dataset, we also generated permutated data from the dataset for analysis purpose.

For evaluating the results of IKKM, we used the following techniques:

1. Cluster Visualization using scatter plot
2. Calculating global-silhouette [6,13], and davies-bouldin index[14].
3. Calculating correlation ratio using Equation (5)[12]

$$(CR_i)^2 \equiv \frac{\sum_{c=1}^{C} n_c\left(\left(\sum_{j\in J_c} x_{i,j}\right)/n_c - \bar{x}_i\right)^2}{\sum_{j=1}^{M}\left(x_{i,j}-\bar{x}_i\right)^2} \quad (5)$$

Hence, n_c is the number of genes in class J_c,x_{ij} is the expression level of gene i in sample j,$\bar{x}_i$is the average expression level of gene i.

For correlation evaluation, we focused more on Cluster 1 and Cluster 2. Cluster 3 was evaluated separatelydue to its very small size (21 genes).Correlation was performed between each cluster and each of the three phenotypes, which are distant metastasis, tumor and normal tissue, lymph node metastasis.

4. Result and Discussion

Using 341 genes as input, our algorithm produced 3 clusters. These clusters were then evaluated using the three evaluation techniques as shown in Figure 4, Table 1, and Figure 5 until Figure 7 respectively. We compared the results of IKKM with Intelligent K-Means (IKM) [11] using the first two evaluation techniques

Figure 4 shows the scatter plots of clusters that produced by IKKM (Figure 4a) and IKM (Figure 4b).From the scatter plots, gene objects in the same color are in the same cluster. In Figure 4a, there are three colors, blue is for Cluster 1, green is Cluster 2, and red is Cluster 3. In Figure 4b, on the other hand, there are five colors, hence, five clusters. Comparing the two plots, IKKM can give better clustering. In each cluster of IKKM, we observed that almost all objects that are close to each other are in the same cluster.

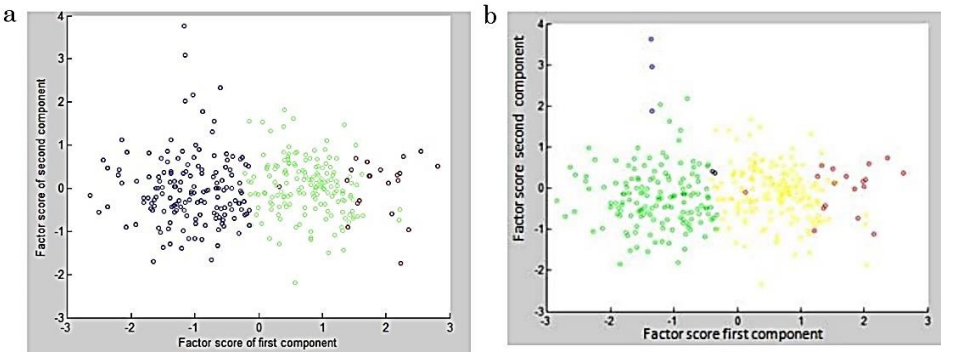

Fig. 4 (a) Scatter plot of clusters IKKM (b) Scatter plot of clusters IKM

In terms of performance, IKKM gave better performance than IKM. Average silhouette of all IKKM clusters gave values > 0, meaning that the clusters were trustworthy, and it was also higher than that of IKM. IKKM clusters were also more compact, since it had half Davies - Bouldin index than that of IKM.

Table 1 Global-Silhouette value, and Davies - Bouldin Index

	Intelligent K-Means [11]	Intelligent Kernel K-Means
The number of clusters	5	3
Global Silhouette	0.072468	0.1453
Davies - Bouldin index	2.2117	1.0261

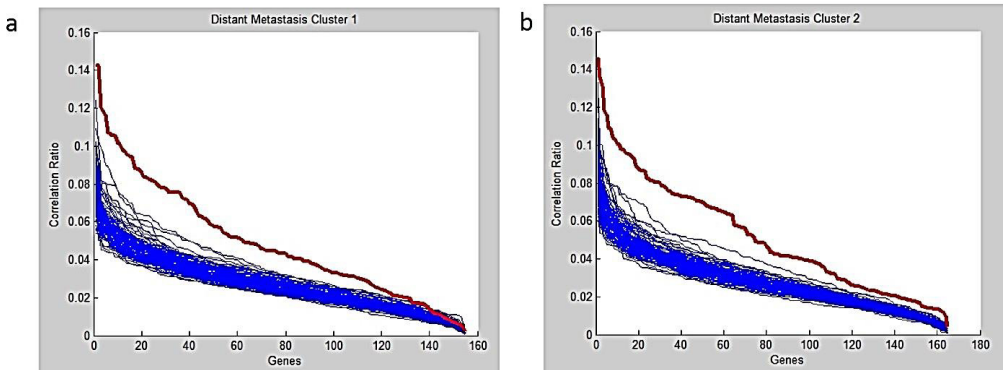

Fig. 5 Correlation of gene expression with distant metastasis phenotype. Vertical axis represents correlation ratio. (a) Genes cluster 1 (b) Genes cluster 2.

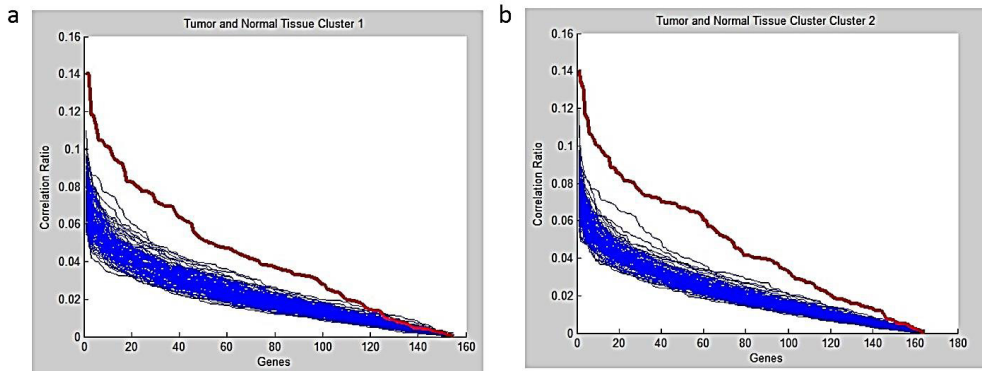

Fig. 6 Correlation of gene expression with tumor and normal tissue phenotype. Vertical axis represents correlation ratio. (a) Genes cluster 1 (b) Genes cluster 2.

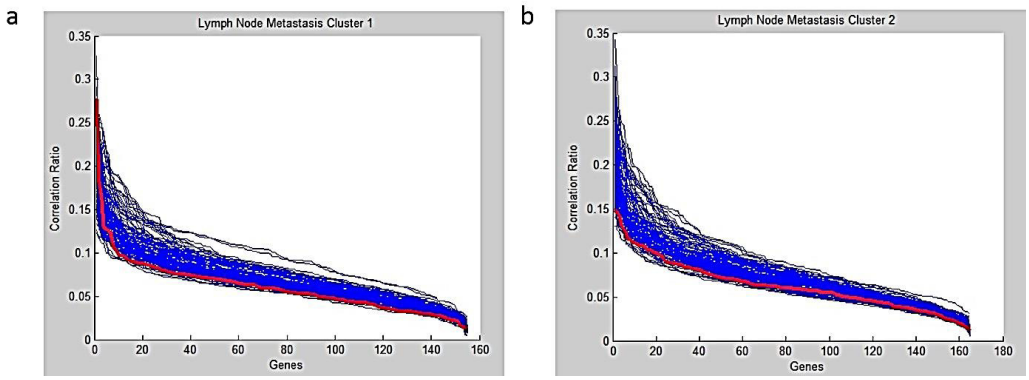

Fig. 7 Correlation of gene expression with lymph node metastasis phenotype. Vertical axis represents correlation ratio. a) Genes cluster 1 (b) Genes cluster 2

We calculated a correlation ratio (CR) to serve as an indicator for correlation between each IKKM cluster with clinical data (distant metastasis, normal and cancer tissues, and lymph node metastasis phenotypes). In each cluster, genes were first sorted by CR value order, and then the CRs of the original dataset were compared with those of permuted dataset. The results of correlation analysis are shown in Figure 5, 6, and 7. The red line is original dataset and the blue lines are permuted dataset. For distant metastasis and normal and cancer tissues phenotypes, the CR values of Cluster 1 and Cluster 2 of original dataset were consistently higher throughout the full range of CRs,

suggesting that the correlation was global for each cluster. We could not identify significant correlation for the lymph node metastasis shown in Figure 7.

Cluster 3 contains only a small number of genes, which were 21 genes. Thus, we analyzed it differently. In this cluster, 16 genes were in the Tumor Classifier List (TCL). Table 2 shows the 24 genes in TCL. Genes marked with * is the 16 genes which were also the member of Cluster 3. Table 3 shows the other 5 genes. We used the same analyzing method with the previous research[6] to show the correlation of genes in Cluster 3 with clinical data. We calculated the average gene expression of 16 representative genes, sorted them and split them into two groups. The first group was genes having positive average gene expression value, and the second group was the genes with the negative average of it. Figure 8 shows correlation of cluster 3 to the distant metastasis.

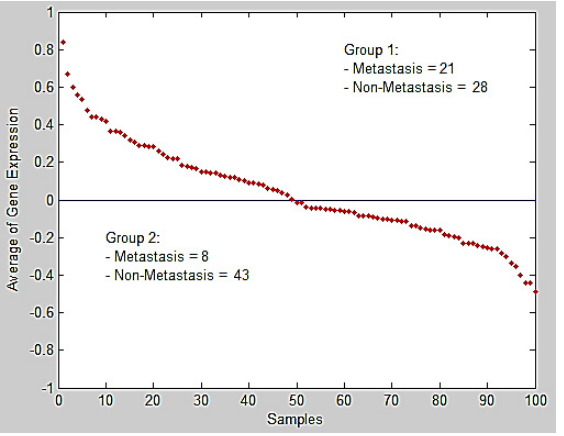

Fig. 8 Linkage of the clusters of expressed genes to the existence of distance metastasis in Cluster 3

Table 2List of TCL (tumor-classifier)

No	GS Number	Accession Number	Symbol	Annotation
1	GS3170	L35240	(*)	Human enigma gene, complete coding sequence
2	GS2892	NM_004368	CNN2(*)	Homo sapiens calponin 2 (CNN2), mRNA
3	GS4015	NM_005433	YES1(*)	Homo sapiensv-yes-1 Yamaguchi sarcoma viral oncogene homolog 1 (YES1), mRNA
4	GS4780	AD001530	(*)	Homo sapiens XAP-5 mRNA, complete coding sequence
5	GS4941	NM_016380	(*)	Homo sapiens differentiation-related protein dif13 (LOC51212), mRNA
6	GS4945	NM_016343	CENPF(*)	Homo sapiens centromere protein F (350/400 kD, mitosin) (CENPF), mRNA
7	GS3387	NM_013317	hT1a-1(*)	Homo sapienshT1a-1 (hT1a-1), mRNA
8	GS3386	NM_003337	UBE2B(*)	Homo sapiens ubiquitin-conjugating enzyme E2B (RAD6 homolog) (UBE2B), mRNA
9	GS4946	NM_002439	MSH3(*)	Homo sapiens mutS (E. coli) homolog 3 (MSH3), mRNA
10	GS3019	NM_003348	UBE2N(*)	Homo sapiens ubiquitin-conjugating enzyme E2N (homologous to yeast UBC13) (UBE2N)
11	GS4022	NM_002433	MOG(*)	Homo sapiens myelin oligodendrocyte glycoprotein (MOG), mRNA
12	GS715	AL096800	(*)	Human DNA sequence from clone RP1-303A1 on chromosome 6
13	GS1102	Y18000	NF2(*)	Homo sapiensNF2 gene
14	GS3002	AL023806	STM2(*)	Human DNA sequence from clone 466P17 on chromosome 6q24
15	GS5239	AL139229	(*)	Human DNA sequence from clone RP4-540A13 on chromosomeXq22.1-22.3
16	GS4163	AC007565	(*)	Homo sapiens chromosome 19, cosmid R27656, complete sequence
17	GS3588	AF131848		Homo sapiens clone 24922 mRNA sequence, complete coding sequence
18	GS4947	NM_018520		Homo sapiens hypothetical protein PRO2268 (PRO2268), mRNA
19	GS1341	AC006165		Homo sapiens clone UWGC:y54c125 from 6p21, complete sequence
20	GS4512	NM_005768	C3F	Homo sapiens putative protein similar to nessy(Drosophila) (C3F), mRNA
21	GS4501	AF261689		Homo sapiens DNA polymerase epsilon p17 subunit gene, complete coding sequence
22	GS6969	AL022316		Human DNA sequence from clone CTA-126B4 on chromosome 22q13.2-13.31
23	GS6493	AF113695		Homo sapiens clone FLB5224 PRO1365 mRNA, complete coding sequence
24	M15990	M15990	yes	Yes

(*) Genes that found in Cluster 3 by Intelligent Kernel K-Means

Table 3Addition genes in Cluster 3

No	GS Number	Accession Number	Symbol	Annotation
1	AA857163	AA857163	Amphiregulin	Amphiregulin (schwannoma-derived growth factor)
2	GS3339	U83246		Homo sapiens copine I mRNA, complete cds.
3	GS3754	AF004429		Homo sapiens D54 isoform (hD54) mRNA, partial cds.
4	GS3683	AK026017		Homo sapiens cDNA: FLJ22364 fis, clone HRC06575.
5	GS3309	AB002135	GPAA1	Homo sapiens mRNA for glycosylphosphatidylinositol anchor attachment 1 (GPAA1), complete cds.

5. Conclusion

Intelligent Kernel K-Means (IKKM) is able to cluster the gene expression of human colorectal carcinoma into three trustworthy and compact clusters. Using the evaluation performance of calculating global-silhouette value and davies-bouldin index, the algorithm outperforms the Intelligent K-Means. Using clinical data of distant metastasis and human tissues, the Cluster 1 and Cluster 2 CR of original data indicates stronger relationship than that of permuted data. Although the number of genes in Cluster 3 is very small, 16 genes out of 21 genes are listed in TCL genes. In our next experiment, we will evaluate further the algorithm by using gene expression of breast cancer.

Acknowledgements

The first author says thank you to Muhammad Anwar Ma'sum for the meaningful and insightful discussion.

References

1. J.A. Berger, S. Hautaniemi, S.K. Mitra, J. Astola. Jointly Analyzing Gene Expression and Copy Number Data in Breast Cancer Using Data Reduction Models. *IEEE/ACM Trans. Computational Biology and Bioinformatics*. 2006; **3**(1):2- 16.
2. R.M. Neve et al. A Collection of breast cancer cell lines for study of functionally distict cancer subtypes. *Journal of Cancer Cell*. 2006; **10**(6): 515-527.
3. R.X. Manezez, M. Boetzer, M. Sieswerda, G-J. B. Ommen, J. M. Boer. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*. 2009; **10**(203): 1-15.
4. C. Xu et al. Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Molecular Cancer*. **9**(143): 1-12.
5. J. Sheng, H.W. Deng, V.D. Calhoun,Y.P. Wang. Integrated Analysis of Gene Expression and Copy Number Data on Gene Shaving Using Independent Component Analysis. *IEEE/ACM Trans. On Computational Biology and Bioinformatics*. 2011; **8**(6): 1568 – 1579.
6. G. Wahyudi, I.Wasito, T.Melia, I.Budi. Robust consensus clustering for identification of expressed genes linked to malignancy of human colorectal carcinoma. 2011; **6**(7): 279-282.
7. B. Mirkin. *Clustering for Data Mining: A Data Recovery Approach*. Boca Raton: ChapmanandHall/CRC; 2005.
8. M.M-T. Chiang, B.Mirkin. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spread. *Journal of Classification Springer*. 2010; **27**: 3-4.
9. T.Handhayani, I.Wasito, M.Sadikin, Ranny. Kernel Based Integration of Gene Expression and DNA Copy Number. In: Proc. of IEEE Advanced Computer Science and Information Systems (ICACSIS). 2013; pp. 303 – 308.
10. J.S-Taylor, N.Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge,UK: Cambridge University Press; 2004.
11. M. A. Ma'sum, I. Wasito, A. Nurhadiyatna. Intelligent K-Means Clustering For Expressed Genes Identification Linked to Malignancy of Human Colorectal Carcinoma. In: Proc. International Conference on Advanced Computer Science And Information System(ICACSIS). 2013; pp. 437-443.
12. S. Muro et al. Identification of Expressed Genes Linked to Malignancy of Human Colorectal Carcinoma by Parametric Clustering of Quantitative Expression Data. *Genome Biology*. 2003.
13. P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational And Applied Mathematics*. 1987; **20**: 53-65.
14. D. Davies, D. Bouldin. A Cluster Separation Measure. " *IEEE Transactions on Pattern Analysis and Machine Intelligence.* PAMI. 1979;**1**(2).