BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net ***Prediction model***

Volume 6(7)

Robust consensus clustering for identification of expressed genes linked to malignancy of human colorectal carcinoma

Gatot Wahyudi, Ito Wasito*, Tisha Melia, Indra Budi

Faculty of Computer Science, University of Indonesia, Kampus UI Depok 16424, Indonesia; Ito Wasito - Email: ito.wasito@cs.ui.ac.id; Phone: +62 21 786 3419; Fax: +62 21 786 3415; *Corresponding author

Received May 24, 2011; Accepted June 13, 2011; Published June 23, 2011

Abstract:
Previous studies have been conducted in gene expression profiling to identify groups of genes that characterize the colorectal carcinoma disease. Despite the success of previous attempts to identify groups of genes in the progression of the colorectal carcinoma disease, their methods either require subjective interpretation of the number of clusters, or lack stability during different runs of the algorithms. All of which limits the usefulness of these methods. In this study, we propose an enhanced algorithm that provides stability and robustness in identifying differentially expressed genes in an expression profile analysis. Our proposed algorithm uses multiple clustering algorithms under the consensus clustering framework. The results of the experiment show that the robustness of our method provides a consistent structure of clusters, similar to the structure found in the previous study. Furthermore, our algorithm outperforms any single clustering algorithms in terms of the cluster quality score.

Background:
In the past decade, gene expression analysis has been applied on a colorectal carcinoma data in order to identify groups of genes that characterize each stage during the progression of this particular disease or provide clues for the possibility of malignancy. Colorectal carcinoma has three common stages called the TNM system, which stands for Tumor/Node/Metastasis system. Previous study in gene expression analysis for colorectal carcinoma predominantly focuses on identifying these three groups: presence of tumor, lymph node metastasis, and distant metastasis **[1]**. Existing approaches utilize a single clustering algorithm in order to identify the groups of colorectal carcinoma during gene expression profiles analysis. One of the most popular clustering methods for unveiling underlying features of gene expression profiles is hierarchical clustering **[2]**. The shortcoming of using hierarchical clustering analysis is that it needs subjective interpretation to determine the number of clusters produced; a consequence that arises from the fact that hierarchical clustering lacks valid statistical evaluation measures. Gaussian Mixture clustering is another powerful clustering method that offers promising results for identifying differential gene expression linked to malignancy of human colorectal carcinoma **[1]**. Unfortunately, the last method has lack of stability in finding the cluster structures.

Consensus clustering has emerged as a powerful method for improving both robustness as well as stability of unsupervised classification solution **[3]**. In its early development, a consensus clustering algorithm was built by performing consensus among multiple runs of a single clustering algorithm while using a re-sampling technique **[4]**. Another different approach of consensus clustering was to find a consensus among different clustering algorithms, coupled with using simulated annealing to introduce a small change in each run, to find the best consensus clusters solution **[5]**. Other approaches built consensus clustering by combining partitions generated by weak clustering algorithms **[3]**.

We propose a consensus clustering algorithm that employs different parametric clustering algorithms with a similar property: they are all centroid-based clusterings. Our consensus strategy employs a majority vote scheme that uses cluster validation techniques to examine the cluster quality. Centroid-based clusterings require the initialization of initial centroid, which may lead to a different clustering result in each different run. Hence, we introduce an approach to find a consensus initial centroid. The consensus initial centroid is then used as the initial parameter of each centroid-based clustering algorithm that participates in the consensus clustering framework. By applying a systematic selection to find a consensus initial centroid, we are able to produce stable and robust clusters among different runs of the algorithm.

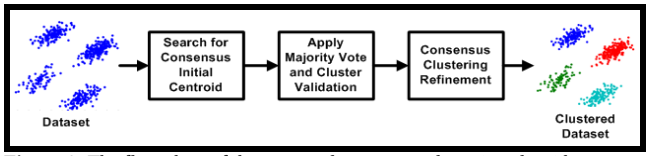

Figure 1: The flow chart of the proposed consensus clustering algorithm

Methodology:
Our proposed method enhances existing consensus clustering schemes as it produces robust and distinctive clusters. The robust property indicates that it produces stable result. That is, clusters structure is kept consistent in each run of the consensus clustering. We implement three centroid-based clustering algorithms to participate in the consensus clustering: K-means, K-medoids, and Gaussian Mixture Model (GMM). The proposed consensus clustering consists of three steps as shown in **Figure 1**. The first step in the proposed consensus clustering algorithm is to search for a consensus initial centroid. This step is prerequisite to go further as we use three centroid-based clustering as the basis

of our algorithm, where each needs an initial centroid configuration as a parameter. We use systematic selection on the furthest objects to generate initial centroid candidates. Finally, K-means clustering is used to evaluate each initial centroid candidate. The initial centroid candidate that is able to produce clusters with minimum sum distance is then selected as the consensus initial centroid. Since the consensus initial centroid is obtained by a systematic selection process rather than a random process, it is able to guarantee the structure of resulting clusters to be similar among clustering algorithm participants, and the result of the clustering is consistent for each run. The second step is to apply majority vote strategy to build consensus among the result of single clustering algorithms. The majority vote strategy determines the cluster of each single element by considering the majority number of clustering algorithms, which agree to put the element in the same cluster. The rest of elements that have insufficient agreement during majority vote are then assigned based on the result of certain single clustering algorithm, which is able to produce better cluster validation score. We use Davies-Bouldin index as the cluster validation technique. The consensus clustering refinement step evaluates the clusters of consensus clustering result. The refinement procedure does cluster reassignment of any element with Silhouette index less than a certain threshold, to another cluster which is able to produce better cluster quality score. The cluster quality score is measured by R2 method as shown in Equation (3) in the **Supplementary material [6]**. The refinement procedure is then repeated until the number of elements under the threshold is unchanged.

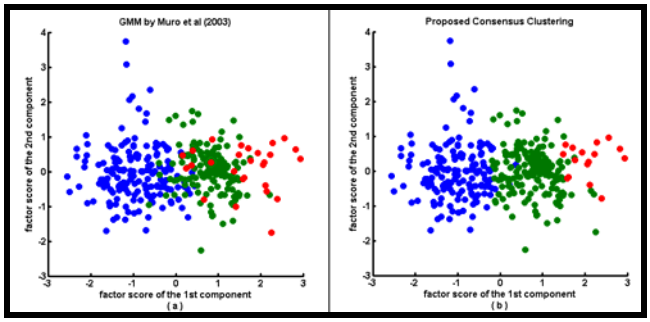

Figure 2: Clustering result comparison on human colorectal carcinoma (a) Gaussian Mixture Model by Muro *et al* **[1]** (b) Proposed consensus clustering

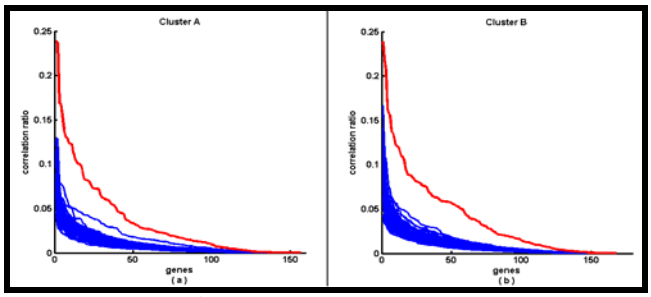

Figure 3: Correlation of gene expressions with cancer phenotype. Vertical axis represents the correlation ratio of the different between tumor and normal tissues. (a) Genes within cluster A (b) Genes within cluster B.

Discussion:

We employ our proposed consensus clustering method on a real gene expression dataset of human colorectal carcinoma provided by Muro *et al*. **[1]**. This dataset consists of 341 selected informative genes out of 1536 genes from 111 samples (100 cancerous and 11 normal). Since the human colorectal carcinoma dataset contains missing values, we perform data imputation to fill the missing value using k-nearest neighbor method with k = 15. Using our consensus clustering method, where parameters are set at: k = 3, r = 6, t = 10, and ts = 0.25, the resulting clusters are similar to the result of Gaussian Mixture Model by Muro *et al*. **[1]** and Wasito *et al*. **[7]**. The outcome consists of two clusters with a large number of genes (i.e. the blue and green clusters) and one cluster with a small number of genes (i.e. the red cluster). For further discussion, we label the blue cluster as cluster A, the green cluster as cluster B, and the red cluster as cluster C. Based on the measurement of the tightness of clusters using R2 score, our consensus clustering method has R2 score of 28.5421 out of 100, which outperforms the result by Muro *et al*. **[1]** that scored of 18.1277 out of 100. A higher value of R2 score implies a higher quality of clusters. Moreover, the visualization of cluster in 2-D graph, **Figure 2**, shows that the result of our consensus clustering is more separable as it contains less overlapping elements between clusters. To assess the usefulness of our algorithm, we perform Correlation Ratio (CR) analysis on cluster A and B with the cancer clinical parameters. Cluster C, which contains small number of genes, is analyzed using a different correlation technique. There are three clinical parameters used in this analysis: the presence of tumor or normal tissues, presence or absence of distant metastasis, and presence or absence of lymph node metastasis. The results of correlation analysis are shown in **Figure 3, 4, 5, 6**.

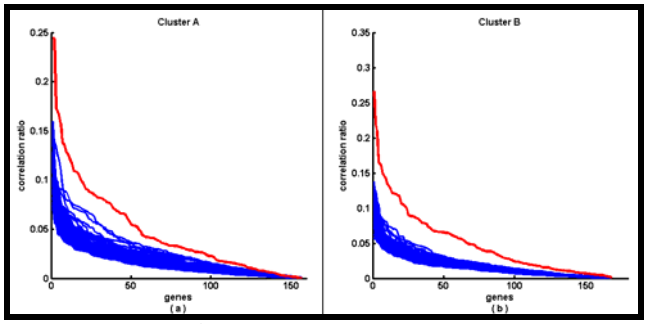

Figure 4: Correlation of gene expressions with cancer phenotype. Vertical axis represents the correlation ratio of the presence or absence of distant metastasis. (a) Genes within cluster A (b) Genes within cluster B.

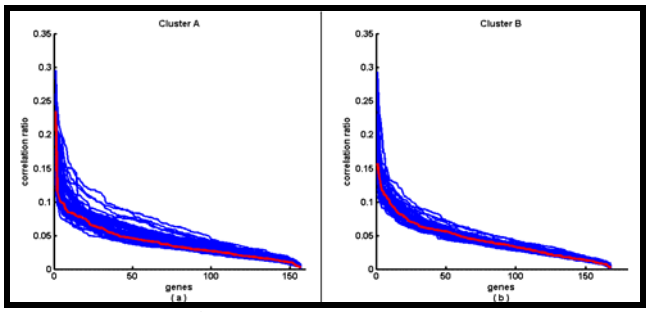

Figure 5: Correlation of gene expressions with cancer phenotype. Vertical axis represents the correlation ratio of the presence or absence of lymph node metastasis. (a) Genes within cluster A (b) Genes within cluster B.

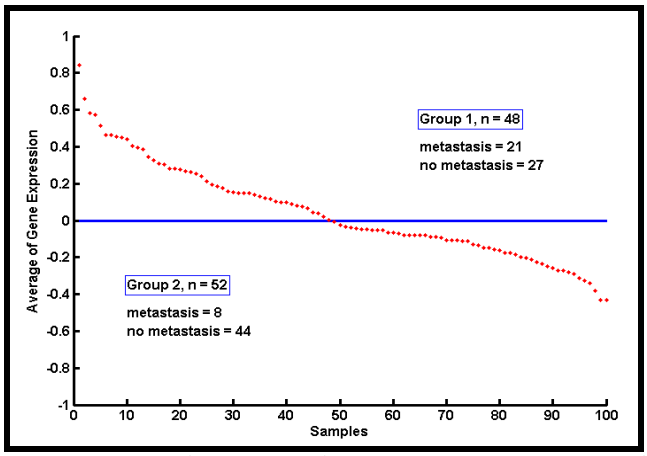

Figure 6: Linkage of the clusters of expressed genes to the existence of distance metastasis in cluster C.

Figure 3 & 4 as the result of correlation ratio analysis on cluster A and B show that both clusters have significant correlation with the first two clinical parameters, which are the presence or absence of tumor and presence or absence of distant metastasis. However, cluster A and B have no correlation to lymph node metastasis as shown in **Figure 5**. On the other hand, cluster C, which consists of 16 genes that are listed in Tumor Classifier (TCL) genes, appears to correlate with the existence of tumor. We calculate the average value of 12 informative genes out of 16 in cluster C, sort them, and split these

genes into two groups. The first group has an average gene expression level of greater than zero while the second group has an average gene expression level of lower than zero. The plot is shown in **Figure 6**, which shows that cluster C correlates to the third colorectal carcinoma clinical parameter (i.e. the distant metastasis).

Conclusion:

In this work, we explore the combination of systematic selection of initial centroid, majority vote scheme, and cluster validation technique to build a stable and robust consensus clustering method. The proposed method successfully combines and improves the performance of centroid-based clustering algorithms used in the consensus. The proposed method has robust property that is able to produce consistent cluster structures and memberships for each run.

Acknowledgement:

This work was supported by a research grant from University of Indonesia, No. 2548/H2.R12/PPM.001.01/2010

References:

[1] Muro S *et al. Genome Biol.* 2003 **4**: R21 [PMID: 12620106]
[2] Eisen MB *et al. Proc Natl Acad Sci U S A.* 1998 **95**: 14863 [PMID: 9843981]
[3] Topchy A *et al. IEEE Trans Pattern Anal Mach Intell.* 2005 **27**: 1866 [PMID: 16355656]
[4] Monti S *et al. Machine Learning* 2003 **52**: 91
[5] Swift *et al. Genome Biol.* 2004 **5**: R94 [PMID: 15535870]
[6] Hastie T *et al. Genome Biol.* 2000 **1**(2): RESEARCH0003 [PMID: 11178228]
[7] Wasito I *et al. Bioinformation* 2007 **2**: 175 [PMID: 18305825]

Edited by P Kangueane

Citation: Wahyudi *et al.* Bioinformation 6(7): 279-282 (2011)

BIOINFORMATION open access

Supplementary Material:

→ (1)

$$V_B = \frac{1}{p}\sum_{j=1}^{p}(\bar{x}_j - \bar{x})^2$$ Between Variance

p is the number of cluster; x_j is mean of gene expression in cluster j; x is mean of overall gene expression in all cluster

→ (2)

$$V_T = \frac{1}{kp}\sum_{i\in S_k}\sum_{j=1}^{p}(x_{ij} - \bar{x})^2$$ Total Variance

k is the number of genes in a cluster; p is the number of cluster; S_k is expression level of cluster k; x_{ij} is the expression level of gene i in sample j; x is mean of overall gene expression in all cluster

→ (3)

$$R^2 = 100\frac{V_B}{V_T}$$ R square

A large number of R^2 implies a tight cluster of coherent genes

→ (4)

$$(CR_i)^2 = \frac{\sum_{c=1}^{C} n_c((\sum_{j\in J_c} x_{ij})/n_c - \bar{x}_i)^2}{\sum_{j=1}^{M}(x_{ij} - \bar{x}_i)^2}$$ Correlation Ratio

n_c is the number of genes; in a particular class J_c; x_{ij} is the expression level of gene i in sample j; and x_i is the average expression level of gene i