Hybrid of Hierarchical and Partitional Clustering Algorithm for Gene Expression Data

Shamini Raja Kumaran[1], Mohd Shahizan Othman[2], Lizawati Mi Yusuf[3]

[1,2,3]Faculty Engineering, School of Computing, Universiti Teknologi Malaysia

E-mail: shamini.rajakumaran@hotmail.com

Abstract. Microarray analysis able to monitor thousands of gene expression data, however, to elucidate the hidden patterns in the data is a complex process. These gene expression data show its imprecision, noise and vagueness due to its high dimensional properties. There are a handful of clustering algorithms have been proposed to extract the important information from the gene expression data. However, identifying the underlying biological knowledge of the data is still hard. To acknowledge these issues, clustering algorithms are used to reduce the data complexity. In this article, hybrid of agglomerative hierarchical clustering and modified k-medoids (partitional clustering) are proposed. Application of the proposed of clustering algorithms to group the genes that have similar functionality which might assist pre-processing procedures. In order to emphasize the quality of the clustering results, cluster quality index (CQI) is determined. Lung and ovary data sets used and the method retrieved a fair clustering with CQI, 0.37 and 0.48 respectively. This research contributes by avoiding biasness toward genes and provide true sense of clustering output using the advantage of hierarchical and partitional clustering methods.

1. Introduction

Clustering of gene expression data able to assist in understanding the gene functions and diagnostics of disease conditions and medical treatment [1]. In fact, clustering algorithms have the strength to extract useful insights from large amount of data. Gene expression data are usually vague and noisy, thus, this scenario allowed the experiments to create insignificant hypotheses. Even though, there are proposed clustering algorithms in market to overcome the mentioned drawbacks in gene expression data, [1] stated that there were no clustering algorithms developed with good performance for all clustering hindrances. Based on gene expression data, clustering steps can be accomplished based on samples and genes. The significance of clustering on samples and genes is important because the cluster of samples able to displays the expression of genes, while, cluster of genes shows the conditions across the gene expressions [2]. In this article, gene-based clustering are used and it is determined through gene expression matrix. Generally, gene expression matrix extracted from microarray analysis contain missing values and noise which require data pre-processing before clustering procedures. In this article, assume that the input gene expression data sets has already been pre-processed. Such gene-based clustering identifies the samples as the objects and the genes as the attributes [3]. The attributes will be discriminated through the strongly correlated informative genes within the genes' distinction and provide weak informative genes in a data. Thus, the main objective will be to group the genes based on similar expressions with similar conditions in a same group and discover the structure of attributes. The process of grouping the genes is named as clustering. There are two main categories in clustering which are hierarchical and partitional method [4]. In order to succeed in identifying the grouping of gene

expression data, this article would like to hybrid hierarchical and partitional clustering methods, focusing mainly on agglomerative and modified k-medoids. The rest of the article is organized as follows: Section 2, the related works. The algorithm behind agglomerative hierarchical clustering and modified k-medoids discussed in Section 3 and Section 4, the experimental results from the gene expression data sets. Finally, Section 5 is the conclusion.

2. Related Works

There has been a few encouraging work of clustering methods in this field for unsupervised learning as depicted in [5,6,8] using microarray data. However, considering the characters of gene expression data, an optimal algorithm required these properties: efficiency, robustness, order insensitivity, dimensionality and interpretability of clustering results. Somehow, the related works had drawbacks and believed to emerge ideas for this research with proposed algorithm. Beginning with [5] used mutual information for feature clustering by reducing the data ambiguities. [5] has also stated that the task of clustering within the data allows the data to exhibit significant in the variability of class. Indeed, every proposed algorithm using hierarchical clustering has its own advantages such as ease in handling the similarity or distance measures and application to any variety of attributes. However, the drawback of [5] is the sensitivity toward the variation between the data points in terms of distance using Euclidean concepts. This limits the motive of mutual clusters as the sample size data increases. Therefore, the idea behind [5] using hybrid of top-down and bottom up can break the mutual clusters with very bad partitions of data. Thus, a better choice of hybrid methods can be used to overcome the disadvantage present in [5] with fair number of clusters and along with minimized joint of data points using Euclidean distances. While [6] proposed partition clustering, k-means for rheumatoid arthritis gene expression data in recent work of 2017. [6] used k-means directly to recognize the relevant group of similar genes and samples. Indeed, the advantage of k-means is computationally efficient compared to hierarchical clustering, if only if, k values were small. However, using only k-means as a clustering method is not advised as k-means can't handle data with different sizes and densities, k always dependent on initial values and distance measures might converge to a constant values [7]. Hybrid clustering methods improvised tremendously from [5,6] into one of recent research by [8] whereby proposed hierarchical agglomerative clustering algorithm (*SilHAC*) with spectral clustering and k-means. Even in this recent research implemented Euclidean distance which shows the importance of this distance measure for clustering methods. However, the major discussion for the proposed algorithm in [8] is the use of k-means in determining the number of clusters. k-means are sensitive to scales which will impact on the final results.

3. Hybrid of Hierarchical and Partitional Algorithm

Initially in agglomerative hierarchical clustering, the objects as individual data points used the correlation of similarity function. Furthermore, at each step, the point merges to the closest pair of clusters till the only one k-medoid left. First, the algorithm carried out agglomerative hierarchical clustering and from the generated clusters through hierarchical process, the algorithm computes the mean value of each clusters as the initial point for the k-medoid to identify the centroid. After these process, the procedure of modified k-medoid will by minimizing the total distance to other data points. Illustration of proposed algorithm as shown in Figure 1 based on [9] and [10].

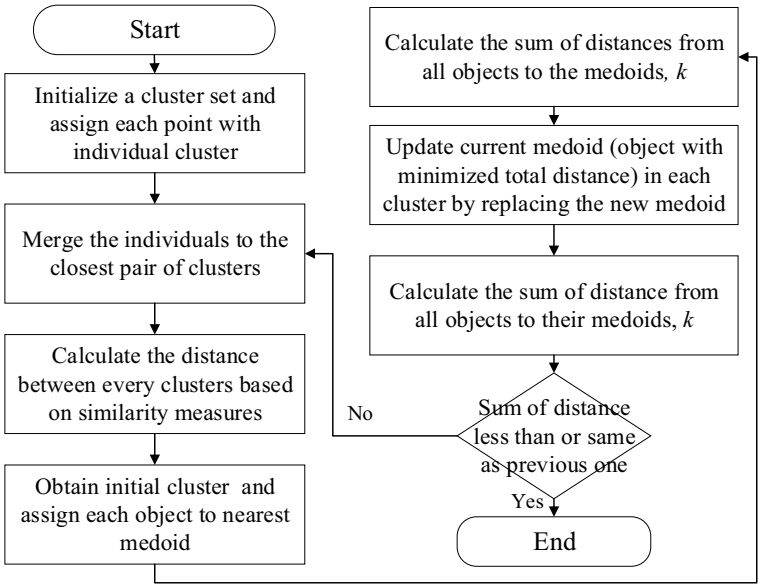


Figure 1. Hybrid of agglomerative hierarchical and modified k-medoid.

As shown in Figure 1, the ideal reason to hybrid both agglomerative and modified k-medoid is due to the disadvantage of basic hierarchical clustering at the stage of termination. Thus, modified k-medoid able to assist agglomerative method and update the medoids in each cluster and minimize the distance before termination compared to non-modified k-medoid that selects the initial medoids randomly and influences the number of iterations. During the clustering of data, grouping the similar genes is the main motive. The measurement used in this research is Euclidean distance. Deriving the Euclidean distance between two data points involves computing the sum of square with square root of the difference through corresponding values. The main advantage of Euclidean distance that the distance between two objects is not affected by the addition of data points which might be outliers [4]. Euclidian distance measured summarized as follows [4]:

$$d = \sqrt{\sum_{i=1}^{v} p_{1i} - p_{2i}} \quad (1)$$

where, the difference between two scores, squared, and summed for v genes. Such distances would be calculated, p_{1i}, p_{2i} where i th expression are feature values of genes. This distance matrix is basically, distance between pair of gene objects.

4. Description of Data sets

Assume that the input gene expression data already been pre-processed and normalized. Table 1 shows the description of lung and Table 2 is the ovary data sets, respectively [11].

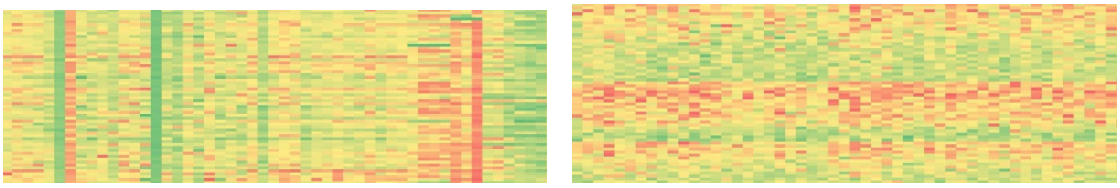
Table 1. Description of lung data set.

Classes	Instances	Genes
Normal	17	12600
Lung Adenocarcinomas	139	
Pulmonary Carcinoids	20	
Squamous Cell Lung Carcinoids	21	
Small Cell Lung Carcinoids	6	
Total Instances	203	

Table 2. Description of ovary data set.

Classes	Instances	Genes
Normal	91	15154
Cancerous	162	
Total Instances	253	

Nevertheless, in order to visualize the data sets, Figure 2 showed the heat maps that present the scattered of pre-processed and normalized lung and ovary data. The heat maps indicate a normalized scale whereby toward green is toward value 1 while toward red is toward value 0 as normalized gene measurements were scaled from 0 to 1.



(a) (b)

Figure 2. Heat maps of (a) Lung (b) Ovary data sets.

5. Experimental Results

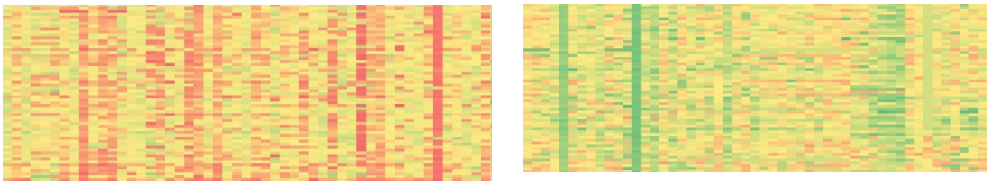
In this section, the article showcased the obtained experimental results through cluster distribution and validated by heatmap visualization and cluster quality index (CQI). Heat maps were used instead of dendograms because this approach ease to visualize and analyze large data sets [12]. Cluster quality index (CQI) considers the result quality of a clustering algorithm in an attempt to identify the grouping of samples or data that best suits the nature of data [13] as shown in previous research such as in [8, 13].Table 3 shows the cluster distribution of lung data set. Using agglomerative and modified k-medoid algorithm, the genes were clustered into two major clusters whereby the first cluster consists of 37.33% of genes and the second cluster consists of 62.26% of genes.

Table 3. Lung data's cluster distribution

Clusters	Clustered genes
1	4755
2	7845
Total genes	12600

For the grouping of genes, the heat maps visualized both clusters of lung data. Figure 3 presents the heat maps of Cluster 1 and Cluster 2 of lung data set. As it can be seen in Figure 3, Cluster 1 consists

more range toward 0 due to presence of more orange and red. While Cluster 2 consists of more yellowish green and toward green, which is ranged toward 1. In addition, to prove the cluster quality, Figure 4 shows the silhouette measure of the cluster quality with CQI = 0.37, a fair clustering.



(a) (b)

Figure 3. Heat maps of lung data set: (a) Cluster 1 and (b) Cluster 2

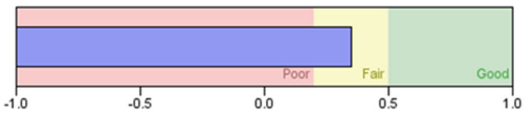


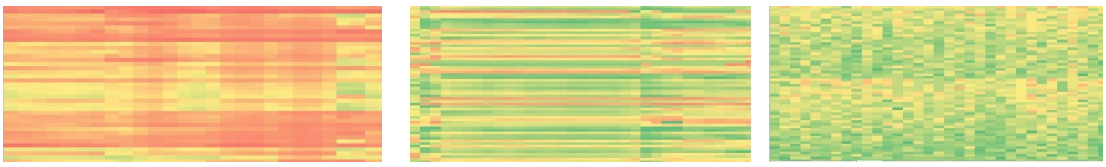
Figure 4. Cluster quality of lung data set.

For ovary data, the clustering of genes is presented in Table 4. There are 3 major clusters for ovary data with Cluster 1 has 30.92% genes, while Cluster 2 covers 9.63% of genes and Cluster 3 consists of 59.45% of genes.

Table 4. Ovary data's cluster distribution.

Clusters	**Clustered genes**
1	4685
2	1460
3	9009
Total genes	15154

Figure 5 visualizes the three grouping of genes from ovary data set. Cluster 1 indicates a range toward 0 as it has more orange and red gene measurements, Cluster 2 has more range toward 1and Cluster 3 has strong green spread on data which indicates the gene measurements clustered more range toward 1.



(a) (b) (c)

Figure 5. Heat maps of lung data set: (a) Cluster 1, (b) Cluster 2, and (c) Cluster 3.

Figure 6 is the cluster quality of ovary data set with CQI value, 0.48, a fair quality result of genes grouping.

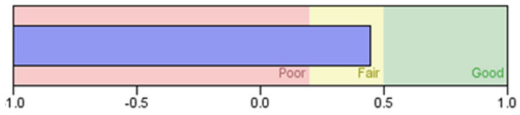


Figure 6. Cluster quality of ovary data set.

6. Conclusion

Gene expression data hide important information due to its imprecision, noisiness and vagueness. As a result, this research proposed hybrid of hierarchical and partitional clustering which can identify a high-quality cluster of genes that are related to the diseases. Based on the experimental works, hierarchical agglomerative and modified k-medoid algorithm able to yield more accurate clustering. This research also validated the clustering genes using heat maps and cluster quality index (CQI). Comparing the result from [8], the number of clusters predicted is also two clusters for lung data, however, the index value obtained is 0.35 proven lower compared to the research output. Thus, this research showed that the advantages of both methods can discover the informative genes.

Acknowledgment

Authors wishing to acknowledge financial assistance or encouragement from School of Graduate Studies, Universiti Teknologi Malaysia (Zamalah Scholarship) and Fundamental Research Grant under the Vote No. 5F207.

References

[1] Geetha T and Arock M 2010 *Int. J. Comp. App.* vol 1(22) pp 92-98
[2] Pirim., B. Ekşioğlu, A.D. Perkins & C. Yüceer 2012 *Comp. & Op. Res*. vol. 39(12) pp 3046-3061
[3] Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F and Adebiyi E 2016 *Bioinformatics and Biology Insights* pp 237-253
[4] Bora D J and Gupta A K 2014 *Int. J. Comp. Sci. and Info. Tech*. vol. 5(2) pp 2501-2506
[5] H.Chipman, R 2006 *Biostatistics* vol. 7(2) pp 286-301
[6] Srivastava S and Joshi N 2014 *Int. J. Comp. Sci. and Mobile Comp.* vol. 3(5) pp 359-364
[7] Sonagar D and Badheka S 2014 *Int. J. Comp. Sci. and Mobile Comp.* vol. 3(10) pp 58-61
[8] Nidheesh N, Abdul Nazeer K A and Ameer P M 2018 A hierarchical clustering algorithm based on silhouette index for cancer subtype discovery from omics data *bioRxiv* pp 1-10
[9] Gandhi G, Srivastava R 2014 *Int. J. Res. Eng. Tech.* pp 150-153
[10] Ahn H and Chang T W 2019 *Sustainability* pp 1-18
[11] Zhu Z, Ong Y S and Dash M 2007 *Patt. Recog.* vol. 40(11) pp 3236-3248
[12] Kuhfled W F 2017 Heat Maps: Graphically Displaying Big Data and Small Tables *SAS Institute Inc., Cary, North Carolina, USA*
[13] Hämäläinen J, Jauhiainen S and Kärkkäinen T 2017 *Algo*. pp 1-14.