SPECTRAL CLUSTERING ON GENE EXPRESSION PROFILE TO IDENTIFY CANCER TYPES OR SUBTYPES

Ang Jun Chin\textsuperscript{a}, Andri Mirzal\textsuperscript{b}, Habibollah Haron\textsuperscript{a*}

\textsuperscript{a}Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia
\textsuperscript{b}Computer Science Department, College of Arts and Applied Sciences, Dhofar University, Salalah, Oman

Abstract

Gene expression profile is eminent for its broad applications and achievements in disease discovery and analysis, especially in cancer research. Spectral clustering is robust to irrelevant features which are appropriate for gene expression analysis. However, previous works show that performance comparison with other clustering methods is limited and only a few microarray data sets were analyzed in each study. In this study, we demonstrate the use of spectral clustering in identifying cancer types or subtypes from microarray gene expression profiling. Spectral clustering was applied to eleven microarray data sets and its clustering performances were compared with the results in the literature. Based on the result, overall the spectral clustering slightly outperformed the corresponding results in the literature. The spectral clustering can also offer more stable clustering performances as it has smaller standard deviation value. Moreover, out of eleven data sets the spectral clustering outperformed the corresponding methods in the literature for six data sets. So, it can be stated that the spectral clustering is a promising method in identifying the cancer types or subtypes for microarray gene expression data sets.

Keywords: Cancer, Gaussian kernel, microarray gene expression, spectral clustering, tumor

Abstrak


Kata kunci: Kanser, inti Gaussian, ungkapan gen mikrotatasusunan, kelompok spectrum, tumor
1.0 INTRODUCTION

According to World Cancer Research Fund, the incidence of cancer is increasing from year to year; there was an estimate of 12.7 million cancer cases around the world in the year 2008 and this number is expected to increase to 21 million by the year 2030. A reliable and precise identification of cancers is crucial for successful diagnosis and treatment [1]. The conventional diagnosis of cancer is based on observation on the morphological appearance of tissue specimens under microscope and chemical analysis. These methods are subjective and highly dependent on the experience of pathologists. Gene expression profiling using microarray offers an objective and unbiased approach to identify cancers independent of previous biological knowledge and morphological appearance of the cancers, and also can accurately identify cancer types or subtypes [2,3].

Clustering methods are widely used for identifying cancer types or subtypes from gene expression profiling. A clustering method groups object patterns into homogeneous groups based on some similarity criteria. It is shown that the clustering methods are important instruments in cancer research with various roles including functional annotation, tissue classification, and motif identification [4].

Hierarchical clustering [5] is the first and the most commonly used method for analyzing patterns of gene expression [6-8]. Some other methods such as k-means [9,10], support vector machine (SVM) [11,12], self-organizing map (SOM) [13-15], artificial neural networks (ANNs) [16,17], principal component analysis (PCA) [18-20], and spectral clustering had also been used. Each of these methods has some benefits and drawbacks. For example, hierarchical clustering can use any valid distance measure as the similarity criterion, but has \( O(n^3) \) or \( O(2^n) \) complexity which makes this technique prohibitive for large datasets. Moreover, it may form no explicit cluster due to a flat partition derived afterward (e.g., via a cut through the dendrogram or termination condition in the construction). K-means has a better computational complexity than hierarchical clustering, but it can only be used to cluster linearly separable data sets, depends on the initialization, and does not have uniqueness property. SVM has good performance for cancer classification using gene expression data sets, but it requires extensive training to choose the optimal parameters and cannot be employed in unsupervised manner. SOM is one of the first methods used in cancer clustering research. It is widely used because of the availability of software and the visibility of the clustering results. However it is not specially designed for clustering purpose, requires intensive computational resources, and cannot be used to cluster linearly inseparable data sets. ANNs are also broadly used for cancer classification. However the performances of ANNs depend on the chosen model and the training process to choose the optimal parameters. And even though it is a more complicated technique than SVM, its performance is comparable to SVM.

The spectral clustering is a multi-way clustering technique that is very simple to implement and can be solved efficiently by standard linear algebra methods. PCA is the closest technique to the spectral clustering. The main difference is PCA uses singular vectors and the spectral clustering uses eigenvectors. However, since PCA is not designed for clustering purpose, one must devise a method for inferring clustering assignments from the computed singular vectors. Our main motivations in promoting the use of spectral clustering are (1) it is a multi-way clustering technique in nature so that it is a suitable method for identifying multiple cancer types that are present in the data sets, (2) it uses eigenvectors that can be computed efficiently since there are many highly efficient algorithms available, (3) it has good convergence property, and (4) it has been successfully used in various domains. In addition, because gene expression profile data sets are often linearly inseparable, spectral clustering is a suitable method since it was originally designed to deal with this kind of data sets.

2.0 SPECTRAL CLUSTERING

The spectral clustering is a multi-way clustering technique that makes use of eigenvectors of an affinity matrix induced from the data to perform clustering. Depending on the affinity matrix, the number of eigenvectors, and the algorithm to infer clustering from the eigenvectors, there are some variants of the spectral clustering algorithms proposed in the literature [23-25]. A detailed discussion on the spectral clustering can be found in Luxburg et al. (2007) [26].

The spectral clustering is a popular clustering technique due to its simplicity, intuitiveness, and capability to cluster linearly inseparable data points. Moreover, it also has competitive computational requirements and can give comparable or better clustering results compared to other popular clustering methods [26]. This technique has been successfully used in various domains including machine learning, computer vision, and data analysis [27,50]. Theoretical results on the characteristics and convergence properties of the spectral method have been shown in the previous literature [28-30]. Here we use the spectral clustering algorithm proposed by Ng et al. (2002) [24]. We choose this
algorithm because of its simplicity, intuitiveness and clustering capability which has been reported to be the best among several spectral clustering algorithms [26].

Figure 1 illustrates clustering linearly inseparable data points using the spectral clustering algorithm. As shown the natural clusters of the original data points are nonlinear so that employing k-means directly will produce incorrect cluster assignments. By transforming the original data space in \( \mathbb{R}^l \) to \( \mathbb{R}^k \) by using the eigenvectors, k-means was successful in finding the correct cluster assignments as indicated by the colors of the data points.

![Figure 1](clustering.png)

**Figure 1** Clustering linearly inseparable data points using the spectral clustering algorithm; points in the same cluster are plotted using the same color

### 3.0 RELATED WORKS

The spectral clustering has been successfully used in several application domains including handwriting recognition [48], word-document clustering [49], image segmentation [50], and bioinformatics. In this section, an overview on the works that reported the using of the spectral clustering in cancer clustering is presented.

One of the earliest work that described the using of the spectral clustering for processing microarray data was a work by Kluger et al.(2003) [51]. The authors modified normalized cuts objective function introduced by Dhillon (2001) [52]. They applied the spectral bi-clustering methods to four groups of cancer microarray data sets: lymphoma, leukaemia, breast cancer, and central nervous system embryonal tumours. This method provides not only a division of clusters, but also ranks the degree of membership of genes to respective cluster according to the actual values in the partitioning-sorted eigenvectors.

Speer et al.(2005) [53,54] presented the feature vector representation with spectral clustering for partitioning gene based on Gene Ontology (GO) annotation. Their experiment revealed that the proposed method was able to detect functional clusters of gene and able to distinguish between clusters of genes. Alzate and Suyken(2006) [55] developed a weighted kernel Principal Component Analysis (PCA) formulation to spectral clustering, and Pelckmans et al.(2006) [56] extended the MIN-CUT problem by using mutual spectral clustering (both models include the out-of-sample extension).


Note that even though there are several works that have been reported the using of the spectral clustering in cancer clustering, usually only a few microarray data sets were analysed in each work. And, performance comparisons with various state-of-the-art clustering methods have not been performed in the previous works. In this study, 11 microarray data sets with 8 types of cancer tissues and 6 state-of-the-art clustering methods are involved to evaluate and verify the performance of the spectral clustering.

### 4.0 EXPERIMENTAL DESIGN

This section discusses the experimental design including microarray data set collection, experimental setup, implementation and evaluation measurement.
4.1 Microarray Data Set Collection

A DNA microarray is a 2D array collection of microscopic DNA spots containing a specific DNA probes attached on a solid substrate. This microarray can be used for many purposes including samples characterizations and cancer gene expressions profiling [31-33]. There are several types of DNA microarrays, e.g., complementary DNA (cDNA), oligonucleotide, bacterial artificial chromosomes (BAC), and single nucleotide polymorphism (SNP) microarrays. There are currently two main techniques in microarray technology, cDNA bi-colour glass slide [34,35] and the high-density oligonucleotide array manufactured by Affymetrix GeneChip [36,37], and it seems that these techniques are the most commonly used techniques for profiling cancer gene expression data sets. In this study, a total of 11 cancer data sets that were profiled using either cDNA or oligonucleotide are used to evaluate the performances of the spectral clustering algorithm. The detail description of the data sets is given in Table 1.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Microarray Type</th>
<th>Tissue</th>
<th>Total samples</th>
<th>No. of classes</th>
<th>Sample per class</th>
<th>No. of gene classes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alizadeh et al. (2000) [38]</td>
<td>cDNA</td>
<td>Blood</td>
<td>62</td>
<td>3</td>
<td>42, 9, 11</td>
<td>2093</td>
<td>Diffuse large B-cell lymphoma (DLBCL), Follicular lymphoma (FL), Chronic lymphocytic leukemia (CLL)</td>
</tr>
<tr>
<td>Armstrong et al. (2002) [18]</td>
<td>Oligonucleotid e</td>
<td>Blood</td>
<td>72</td>
<td>3</td>
<td>24, 20, 28</td>
<td>2194</td>
<td>Acute lymphoblastic (ALL), Acute myelogenous leukemia (AML), MLL translocation (MLL)</td>
</tr>
<tr>
<td>Bredel et al. (2005) [19]</td>
<td>cDNA</td>
<td>Brain</td>
<td>50</td>
<td>3</td>
<td>31, 14, 5</td>
<td>1739</td>
<td>Glioblastomas (GBM), Oligodendroglial morphology(OG), Astrocytomas (A)</td>
</tr>
<tr>
<td>Chowdary et al. (2006) [39]</td>
<td>Oligonucleotid e</td>
<td>Breast, Colon</td>
<td>104</td>
<td>2</td>
<td>62, 42</td>
<td>182</td>
<td>Breast (B), Colon (C)</td>
</tr>
<tr>
<td>Dyrskjot et al. (2003) [40]</td>
<td>Oligonucleotid e</td>
<td>Bladder</td>
<td>40</td>
<td>3</td>
<td>9, 20, 11</td>
<td>1203</td>
<td>Tumor stage TA, T1, T2+</td>
</tr>
<tr>
<td>Gordon et al. (2002) [41]</td>
<td>Oligonucleotid e</td>
<td>Lung</td>
<td>181</td>
<td>2</td>
<td>31, 150</td>
<td>1626</td>
<td>Malignant pleural mesothelioma (MPM), Adenocarcinoma (ADCA)</td>
</tr>
<tr>
<td>Nutt et al. (2003) [42]</td>
<td>Oligonucleotid e</td>
<td>Brain</td>
<td>21</td>
<td>2</td>
<td>14, 7</td>
<td>1377</td>
<td>Classic glioblastomas (CG), Classic oligodendrogliomas (CO)</td>
</tr>
<tr>
<td>Risinger et al. (2003) [43]</td>
<td>cDNA</td>
<td>Endometrium</td>
<td>32</td>
<td>2</td>
<td>13, 19</td>
<td>1771</td>
<td>Serous papillary (PS), Endometrioid (E)</td>
</tr>
<tr>
<td>Su et al. (2001) [44]</td>
<td>Oligonucleotid e</td>
<td>Multi-tissue</td>
<td>174</td>
<td>10</td>
<td>26, 8, 26, 23, 12, 11, 7, 27, 6, 28</td>
<td>1571</td>
<td>Prostate (PR), Breast (BR), Lung (LU), Ovary (OV), Colon (CO), Kidney (KI), Liver (LI), Pancreas (PA), Bladder/ureter (BL), Gastroesophagus (GA)</td>
</tr>
<tr>
<td>West et al. (2001) [45]</td>
<td>Oligonucleotid e</td>
<td>Breast</td>
<td>49</td>
<td>2</td>
<td>25,24</td>
<td>1198</td>
<td>Estrogen-receptor-positive (ER+), Estrogen-receptor-negative (ER-)</td>
</tr>
</tbody>
</table>

4.2 Experimental Setup

There are two parameters need to be chosen for each data set: sigma value (σ) and scaling scheme. The sigma value controls how rapidly the affinity Aij falls off with the correlation between two features. A higher sigma value will make the affinity value lower, hence the cluster might be not tight enough; whereas a lower sigma value will increase the affinity, and it will make the clusters ambiguity. As stated in the original work by Ng et al. (2002) [24], the sigma value can be learned directly from the data set. However in this work, we manually assigned the sigma value by considering the distribution of entries in the affinity
matrix obtained by applying the Gaussian kernel and determine the optimal sigma value based on the highest accuracy achieved.

The original sample-by-gene matrix of the gene expression data set may have entries with vastly different scales. In order to bring the data set into a notionally common scale, a scaling scheme needs to be introduced. This study uses either logarithmic or normalized scale as the scaling scheme. The scaling scheme is a common pre-processing step in clustering and classification as it often improves the accuracy of the results [39,42,43]. Given X to be the sample-by-gene matrix, the logarithmic and normalized scales are defined as:

\[
\begin{align*}
\text{Logarithmic scale: } x'_{ij} & = \log(x_{ij}) \\
\text{Normalized scale: } x'_{ij} & = \frac{x_{ij} - \text{min}(x_{ij})}{\text{max}(x_{ij}) - \text{min}(x_{ij})}
\end{align*}
\]

where \(x_{ij}\) is entry \((i,j)\) of \(X\), \(\log\) denotes the natural logarithm, and \(\text{min}(x_{ij})\) and \(\text{max}(x_{ij})\) respectively denote the minimum and maximum value in \(i\)-th row of \(X\). By inspection it is clear that the normalized scale will bring all entries of the data matrix to the range of \([0,1]\). And as \(\log(x_{ij})\) is not defined for \(x_{ij} \leq 0\), when the data set contains such entries, only the normalized scale will be used.

Scaling scheme was chosen by inspecting the scale differences in the entries of the matrix. If the differences are in multidegree orders, then the logarithmic scale will be used. If there are no differences in the scales, then no scaling will be performed. And the normalization scaling is used when the differences are in the medium scale.

4.3 Implementation

All experiments are implemented in Matlab environment running on a laptop with Intel Core i5 @1.70GHz, and 11.9GB of RAM. The following algorithm outlines the spectral clustering algorithm proposed by Ng, et al. [2002] [24].

```
Algorithm: Spectral Clustering [Ng, et al., 2002] [24]

Input:
Microarray Data set, \(S = (s_1,...,s_n)\)
Initialize:
\(
\sigma\) (\(\sigma\) means), Scaling
Start:
1. Data pre-processed with scaling scheme.
2. Construct the affinity matrix \(A \in R^{n \times n}\) using the Gaussian kernel defined by
\[A_{ij} = \exp \left( -\frac{||s_i - s_j||^2}{2\sigma^2} \right)\]
3. Define \(D\) to be a diagonal matrix whose \((i, i)\) element is the sum of the \(A\)'s \(i\)-th row, and construct the Laplacian matrix \(L = D^{-1/2}AD^{-1/2}\).
4. Compute the \(k\) largest eigenvectors \(x_1, x_2, ..., x_k\) of \(L\) (chosen to be orthogonal to each other in the case of repeated eigenvalues).
5. Form matrix \(X = [x_1 x_2 ... x_k] \in R^{n \times k}\) by stacking the eigenvectors in columns.
6. Form matrix \(Y\) from \(X\) by renormalizing each of \(X\)'s rows to have unit length, i.e., \(y_{ij} = x_{ij}/(\sum_{j}x_{ij}^2)^{1/2}\).
7. Cluster each row of \(Y\) into \(k\) clusters via k-means.
8. Assign the original point \(s_i\) to cluster \(j\) if and only if row \(i\) of the matrix \(Y\) was assigned to cluster \(j\).
9. Evaluate the cluster accuracy.
Output:
Accuracy of clustering.
```

A note on clustering robustness of the algorithm. Since the set of eigenvectors of a matrix is unique (up to scaling factor), the only source of non-uniqueness is the use of k-means to infer cluster assignments from the eigenvectors. Because k-means is applied to the reduced subspace where the data points are more clustered and linearly separable than in the original space (the purpose of transforming \(S\) into \(Y\) is to construct such subspace), clustering results in this subspace will be more stable and decisive.

4.4 Evaluation Measurement

There are a few common evaluation measurement used to evaluate the clustering result, for example Dunn Index, Davies-Boldin (DB) Index, Accuracy, Rand Index, and Jaccard Index. However, the original literatures of microarray dataset involved in this study have used Accuracy as their evaluation measurement. Therefore, this study uses the metric Accuracy to evaluate the clustering performance. Accuracy measures the fraction of the dominant class in a cluster and is defined as:

\[
\text{Accuracy} = \frac{1}{n} \sum_{r=1}^{R} \max_{s} c_{rs}
\]

where \(r\) and \(s\) denote the \(r\)-th cluster and \(s\)-th reference class respectively, \(R\) denotes the number of clusters produced by clustering algorithm, \(N\) denotes the number of samples, and \(c_{rs}\) denotes the number of samples in \(r\)-th cluster that belong to \(s\)-th class. The values of Accuracy are between 0 and 1 with 1 indicates a perfect agreement between the reference classes and the clustering results. In machine learning community, this metric is also known as Purity [47].

5.0 RESULTS AND DISCUSSION

This section presents performance evaluation of the spectral clustering algorithm. To get an objective evaluation, the clustering performances of the algorithm are compared to the results reported in the
literature. The results and some details about experimental setup are outlined in Table 2.

The first three columns of Table 2 show the sources of the data sets, the clustering methods used in the original literature, and the Accuracy values obtained in the corresponding study. The last three columns outline the Accuracy values obtained by the spectral clustering algorithm, the sigma values, and scaling schemes used in the corresponding data sets. In summary, the spectral clustering algorithm outperformed the results of literature in six cases, underperformed in four cases, and produced in par result in one case. In average, the spectral clustering algorithm can slightly outperform the results of literature. The spectral clustering also can offer more stable clustering results as it has smaller standard deviation value. Moreover, the average of clustering accuracy improvements in six cases where it gave better results are larger than the average of clustering accuracy reduction in four cases where it failed to outperform the results in the literature (6.03 and 5.175 respectively).

There are two cases in which the spectral clustering algorithm significantly improved the original results, i.e., Bredel et al. (2005) and Dyrskjot et al. (2003). And only in one case the algorithm produced rather unsatisfactory result compared to the original work, i.e., Risinger et al. (2003). However, in this case, the algorithm actually still performed well as the Accuracy is about 84%. The lowest Accuracy offered by the algorithm is in Pomeroy et al. (2002) which is about 76%. But since the original work also reported a low value of 78%, probably this data set is rather hard to cluster. The best result of the algorithm is in Alizadeh et al. (2000), 100%, and is the same with the result of the literature. By considering the results as a whole, it can be stated that the spectral clustering algorithm is a promising method for identifying tumor types from microarray gene expression data sets as it has stable clustering results over all datasets and also in average performed the best compared to various methods used in the original works.

### Table 1 Performance comparison and experimental setup for the spectral clustering algorithm

<table>
<thead>
<tr>
<th>Data set</th>
<th>Original Literature</th>
<th>Spectral Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clustering Method</td>
<td>%Accuracy</td>
</tr>
<tr>
<td>Alizadeh et al. (2000) [38]</td>
<td>Hierarchical clustering</td>
<td>100</td>
</tr>
<tr>
<td>Armstrong et al. (2002) [18]</td>
<td>Principal Component Analysis</td>
<td>95</td>
</tr>
<tr>
<td>Bredel et al. (2005) [19]</td>
<td>Principal Component Analysis</td>
<td>66.55</td>
</tr>
<tr>
<td>Chowdary et al. (2006) [39]</td>
<td>Hierarchical Clustering</td>
<td>96</td>
</tr>
<tr>
<td>Dyrskjot et al. (2003) [40]</td>
<td>Hierarchical Clustering</td>
<td>75</td>
</tr>
<tr>
<td>Gordon et al. (2002) [41]</td>
<td>Bayesian Regression Model</td>
<td>97</td>
</tr>
<tr>
<td>Pomeroy et al. (2002) [13]</td>
<td>Self-Organizing maps</td>
<td>78.3</td>
</tr>
<tr>
<td>Risinger et al. (2003) [43]</td>
<td>Hierarchical Clustering</td>
<td>94</td>
</tr>
<tr>
<td>Su et al. (2001) [44]</td>
<td>Support Vector Machine</td>
<td>85</td>
</tr>
<tr>
<td>West et al. (2001) [45]</td>
<td>Bayesian Regression Model</td>
<td>89.47</td>
</tr>
<tr>
<td>Average ± standard deviation</td>
<td>87.48 ± 10.53</td>
<td>88.71 ± 7.632</td>
</tr>
</tbody>
</table>

### 6.0 CONCLUSION

The using of computational methods for clustering and classification of tumor types from microarray gene expression data sets has been an active research recently. However, there is still lack of works that explore the possibility of using the spectral clustering for this task. Perhaps this is due to the fact that the spectral clustering is relatively a new approach compared to more established methods like hierarchical clustering, SVM, SOM, ANNs, and k-means clustering. The spectral clustering is in fact a suitable choice for identifying tumor types in unsupervised manner since it is designed for clustering linearly inseparable data points which often the cases in the gene expression data sets. Other unsupervised methods like hierarchical clustering, SOM, and k-means clustering, on the other hand, are originally designed for clustering linearly separable data points. In addition, it uses eigenvectors that can be computed efficiently, has good convergence property, and has been successfully used in various application domains.

In particular, we have shown that the spectral clustering algorithm performed well for identifying tumor types compared to various methods reported in the literature. In summary, the spectral clustering outperformed the results in the literature in six cases, underperformed in four cases, and produced in par result in one case. In average, the spectral clustering
slightly outperformed the results in the literature. The spectral clustering also can offer more stable clustering results as the standard deviation value is smaller compared to the standard deviation of other clustering methods. Moreover, the mean of clustering accuracy improvements in six cases (where it gave better results) is larger than the mean of clustering accuracy reduction in four cases (where it failed to outperform the results in the literature). By considering the results as a whole, it can be stated that the spectral clustering algorithm is a promising method for identifying tumor types from microarray gene expression data sets.

Acknowledgement

This work is financed by Zamalah Scholarship provided by Universiti Teknologi Malaysia (UTM) and the Ministry of Higher Education Malaysia.

References


