Jurnal
Teknologi

Full Paper

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

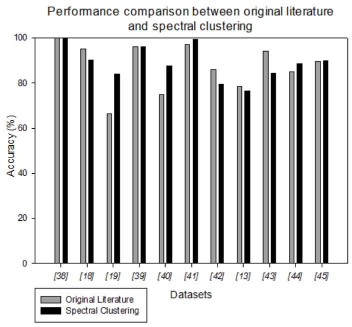
Ang Jun Chin[a], Andri Mirzal[b], Habibollah Haron[a]*

[a]Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia
[b]Computer Science Department, College of Arts and Applied Sciences, Dhofar University, Salalah, Oman

Article history
Received
29 January 2015
Received in revised form
31 March 2015
Accepted
1 August 2015

*Corresponding author
habib@utm.my

Graphical abstract



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 appropriated 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

Profil ungkapan gen adalah terkenal untuk aplikasi yang luas dan pencapaian dalam penemuan dan analisis penyakit, terutama dalam penyelidikan kanser. Kelompok spektrum adalah kukuh terhadap ciri-ciri yang tidak berkaitan dan ia sesuai untuk analisis ungkapan gen. Walau bagaimanapun, penyelidikan sebelum ini menunjukkan bahawa perbandingan prestasi dengan kaedah kelompok lain adalah terhad dan hanya beberapa set data mikrotatasusunan dianalisis dalam setiap kajian. Dalam kajian ini, kami menunjukkan penggunaan kelompok spektrum dalam mengenal pasti jenis-jenis kanser atau sub-jenis daripada profil ungkapan gen mikrotatasusunan. Kelompok spectrum digunakan dalam sebelas set data mikrotatasusunan dan prestasi pengelompokan dibandingkan dengan keputusan di kesusasteraan. Berdasarkan keputusan, secara keseluruhan kelompok spektrum mengatasi keputusan yang sepadan dalam kesusasteraan agak sedikit. Kelompok spektum juga boleh menawarkan prestasi kelompok yang lebih stabil kerana ia menghasil nilai sisihan piawai yang lebih kecil. Selain itu, prestasi kelompok spektrum ini mengatasi enam kaedah yang digunakan berbanding sebelas data set dari kesusasteraan. Oleh itu, boleh dinayatakan bahawa kelompok spektrum adalah satu kaedah yang boleh dipercayai dalam mengenal pasti jenis-jenis kanser atau sub-jenis bagi set data ungkapan gen mikrotatasusunan.

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 using of the spectral clustering in cancer identification are (1) the spectral clustering is naturally a non-linear clustering method, (2) it is robust to irrelevant features which the gene expression data always contains many of these features, and (3) there is still lack of works that explore the possibility of using the spectral clustering in cancer identification.

In this study, we demonstrate the use of the spectral clustering for cancer types or subtypes identification. This method has some benefits compared to the above methods, e.g., (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 [21,22], the 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 methods 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 R^l to 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.

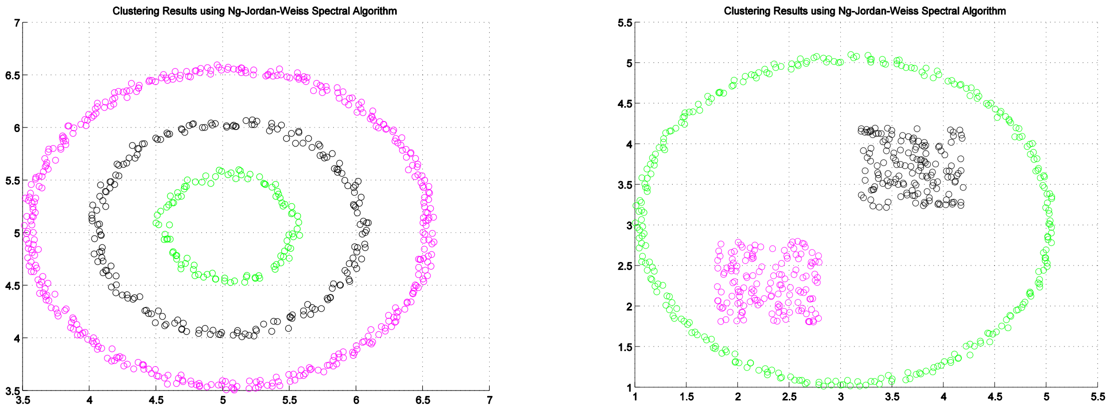


Figure 1 Clustering linearly inseparable data points using the spectral clustering algorithm; points in the same cluster are plotted using the same color

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).

Tritchler et al.(2005) [57] demonstrated the gene clustering based on the spectral bi-partitioning method by using two gene expression data set: leukemia and cutaneous malignant melanoma. The experimental results showed that the spectral clustering outperformed hierarchical clustering and k-means. Higham et al.(2007) [58] compared the performance of normalized and un-normalized spectral clustering by using three microarray data: leukaemia, brain tumours and lymphoma. The authors concluded that the normalized spectral clustering is superior to the un-normalized version in term of sensitivity and feature similarity. Thurlow et al.(2010) [59] combined the spectral clustering with Gene Ontology analysis to reveal the aspects of head and neck squamous cell carcinoma (HNSCC). A recent study [60] developed a new recursive K-means spectral clustering method (ReKS) for disease gene expression data.

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 1 Data set descriptions

Data set	Microarray Type	Tissue	Total samples	No. of classes	Samples per class	No. of gene	Classes
Alizadeh *et al.* (2000) [38]	cDNA	Blood	62	3	42, 9, 11	2093	Diffuse large B-cell lymphoma (DLBCL), Follicular lymphoma (FL), Chronic lymphocytic leukemia (CLL)
Armstrong *et al.* (2002) [18]	Oligonucleotide	Blood	72	3	24, 20, 28	2194	Acute lymphoblastic (ALL), Acute myelogenous leukemia (AML), MLL translocation (MLL)
Bredel *et al.* (2005) [19]	cDNA	Brain	50	3	31, 14, 5	1739	Glioblastomas (GBM), Oligodendroglial morphology(OG), Astrocytomas (A)
Chowdary *et al.* (2006) [39]	Oligonucleotide	Breast, Colon	104	2	62, 42	182	Breast (B), Colon (C)
Dyrskjot *et al.* (2003) [40]	Oligonucleotide	Bladder	40	3	9, 20, 11	1203	Tumor stage TA, T1, T2+
Gordon *et al.* (2002) [41]	Oligonucleotide	Lung	181	2	31, 150	1626	Malignant pleural mesothelioma (MPM), Adenocarcinoma (ADCA)
Nutt *et al.* (2003) [42]	Oligonucleotide	Brain	21	2	14,7	1377	Classic glioblastomas (CG), Classic oligodendrogliomas (CO)
Pomeroy *et al.* (2002) [13]	Oligonucleotide	Brain	34	2	25,9	857	Classic medulloblastomas (CMD), Desmoplastic medulloblastomas (DMD)
Risinger *et al.* (2003) [43]	cDNA	Endometrium	32	2	13, 19	1771	Serous papillary (PS) , Endometrioid (E)
Su *et al.* (2001) [44]	Oligonucleotide	Multi-tissue	174	10	26, 8, 26, 23,12, 11, 7, 27, 6, 28	1571	Prostate (PR), Breast (BR), Lung (LU), Ovary (OV), Colorectum (CO), Kidney (KI), Liver (LI), Pancreas (PA), Bladder/ureter (BL), Gastroesophagus (GA)
West *et al.* (2001) [45]	Oligonucleotide	Breast	49	2	25,24	1198	Estrogen-receptor-positive (ER+) , Estrogen-receptor-negative (ER-)

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 A_{ij} 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:

Logarithmic scale: $x_{ij} \leftarrow \log(x_{ij})$ and

Normalized scale: $x_{ij} \leftarrow \frac{x_{ij}-\min(\mathbf{x}_i)}{\max(\mathbf{x}_i)-\min(\mathbf{x}_i)}$

where x_{ij} is entry (*i,j*) of *X*, log denotes the natural logarithm, and min(x_i) and max(x_i) 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 $x_{ij} \le 0$, 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 multitude orders, then the logarithmic scale will be used. If there are not many 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, $S = \{s_1, \dots, s_n\}$
Initialize:
        Sigma (σ), 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
       $A_{ij} = \exp\left(-\frac{\left\|s_i - s_j\right\|^2}{2\sigma^2}\right)$
   if $i \neq j$, and $A_{ii} = 0$ ($\sigma$ s a variable and used to control distances between the data points.).
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 $x_1, x_2, \dots, x_k$ of $L$ (chosen to be orthogonal to each other in the case of repeated eigenvalues).
5. Form matrix $X = [x_1 x_2 \dots x_k] \in R^{n \times k}$ 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., $Y_{ij} = X_{ij}/\left(\sum_j X_{ij}^2\right)^{1/2}$.
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[46]:

$$Accuracy = \frac{1}{N}\sum_{r=1}^{R} \max_{s} c_{rs}$$

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

Data set	Original Literature		Spectral Clustering		
	Clustering Method	%Accuracy	%Accuracy	Sigma ()	Scaling
Alizadeh *et al*. (2000) [38]	Hierarchical clustering	100	100	1	Normalization
Armstrong *et al*. (2002) [18]	Principal Component Analysis	95	90.28	16001	Non-scaling
Bredel *et al*. (2005) [19]	Principal Component Analysis	66.55	84	1.41	Normalization
Chowdary *et al*. (2006) [39]	Hierarchical Clustering	96	96.15	34	Logarithmic
Dyrskjot *et al*. (2003) [40]	Hierarchical Clustering	75	87.5	6001.5	Non-scaling
Gordon *et al*. (2002) [41]	Bayesian Regression Model	97	99.45	9	Non-scaling
Nutt *et al*. (2003) [42]	K-nearest neighbor model	86	79.31	258	Non-scaling
Pomeroy *et al*. (2002) [13]	Self-Organizing maps	78.3	76.47	525	Non-scaling
Risinger *et al*. (2003) [43]	Hierarchical Clustering	94	84.38	1.31	Logarithmic
Su *et al*. (2001) [44]	Support Vector Machine	85	88.5	10	Logarithmic
West *et al*. (2001) [45]	Bayesian Regression Model	89.47	89.80	16.8	Logarithmic
Average ± standard deviation		87.48 ± 10.53	88.71 ± 7.632		

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

[1] Dudoit, S., Fridlyand, J., and Speed, T. P. 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J Amer. Statist. Assoc.* 97(457): 77-87.

[2] Cozzini, A., Jasra, A., Montana, G. 2013. Model-based Clustering with Gene Ranking Using Penalized Mixtures of Heavy-tailed Distributions. *J Bioinformatics and Computational Biology*. 11.

[3] Qabaja, A., Jarada, T., Elsheikh, A., Alhajj, R. 2014. Prediction of Gene-Based Drug Indications Using Compendia of Public Gene Expression Data Aand Pubmed Abstracts. *J Bioinformatics and Computational Biology*.

[4] Sharan, R., Elkon, R., and Shamir, R. 2002. Cluster Analysis and Its Applications to Gene Expression Data. In Mewes, H.-W., Seidel, H., and Weiss, B., Editors. *Bioinformatics and Genome Analysis*, number 38 in Ernst Schering Research Foundation Workshop. 83-108.

[5] Sneath, P. H. A. and Sokal, R. R. 1973. Numerical Taxonomy: the Principles and Practice of Numerical Classification.

[6] Wei, D., Jiang, Q., Wei, Y., Wang, S., 2012. A novel Hierarchical Clustering Algorithm for Gene Sequences. *BMC Bioinformatics*. 13(1): 174.

[7] Liang, Y., Diehn, M., Watson, N., Bollen, A. W., Aldape, K. D., Nicholas, M. K., Lamborn, K. R., Berger, M. S., Botstein, D., Brown, P. O., and Israel, M. A. 2005. Gene Expression Profiling Reveals Molecularly and Clinically Distinct Subtypes of Glioblastoma Multiforme. *Proc. of the National Academy of Sciences of the United States of America.* 102(16): 5814-5819.

[8] Liu, Q., Zhao, Z., Li, Y.-X., Li, Y. 2012. Feature Selection Based on Sensitivity Analysis of Fuzzy ISODATA. *Neurocomputing.* 85: 29-37.

[9] Xu, R., Damelin, S., Nadler, B., Wunsch II, D.C. 2010. Clustering of High-dimensional Gene Expression Data with Feature Filtering Methods and Diffusion Maps. *Artificial Intelligence in Medicin.* 48: 91-98.

[10] Zhang, S., Wong, H.-S., Shen, Y., Xie, D. 2012. A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity. *IEEE/ACM Trans. Comput. Biol. Bioinformatic.* 9(4): 1257-1263.

[11] Xie, J., Wang, C. 2011. Using Support Vector Machines with a Novel Hybrid Feature Selection Method for Diagnosis of Erythemato-Squamous Diseases. *Expert Systems with Applications.* 38(5): 5809-5815.

[12] George, G. V. S., Raj, V. C. 2011. Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification using Gene Expression Profile. *International Journal of Computer Science & Engineering Survey*. 2(3): 16-27.

[13] Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*. 415(6870): 436–442. [Online]. From: http://www.broadinstitute.org/mpr/CNS/ [Accessed on 13 July 2015].

[14] Chang, R.-I., Chu, C.-C., Wu, Y.-Y., Chen, Y.-L. 2010. Gene Clustering by Using Query-Based Self-organizing Maps. *Expert Systems with Application*. 37(9): 6689-6694.

[15] Wirth, H., Loffler, M., Bergen, M. von, Binder, H. 2011. Expression Cartography of Human Tissues Using Self Organizing Maps. *BMC Bioinformatics*. 12(1): 306.

[16] Takahashi, M., Hayashi, H., Watanabe, Y., Sawamura, K., Fukui, N., Watanabe, J., Kitajima, T., Yamanouchi, Y., Iwata, N., Mizukami, K., Hori, T., Shimoda, K., Ujike, H., Ozaki, N., Iijima, K., Takemura, K., Aoshima, H., Someya, T. 2010. Diagnostic Classification of Schizophrenia by Neural Network Analysis of Blood-based Gene Expression Signatures. *Schizophrenia Research.* 119(1): 210-218.

[17] Zainuddin, Z., Ong, P. 2011. Reliable Multiclass Cancer Classification of Microarray Gene Expression Profiles Using an Improved Wavelet Neural Network. *Expert Systems with Applications.* 38(11): 13711-13722.

[18] Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. 2002. MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. *Nature Genetics*. 30(1): 41–47. [Online]. From: http://www.broadinstitute.org/mpr/publications/projects/Leukemia/expression_data.txt [Accessed on 13 July 2015].

[19] Bredel, M., Bredel, C., Juric, D., Harsh, G. R., Vogel, H., Recht, L. D., and Sikic, B. I. 2005. Functional Network Analysis Reveals Extended Gliomagenesis Pathway Maps and Three Novel MYC-Interacting Genes in Human Gliomas. *Cancer Research*. 65(19): 8679–8689. [Online]. From: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2223 [Accessed on 13 July 2015].

[20] Ma, S., Dai, Y. 2011. Principal Component Analysis Based Methods in Bioinformatics studies. *Brief Bioinform.* 12(6): 714-722.

(21] Lee, H. and Singh, R. 2012. Unsupervised Kernel Parameter Estimation by Constrained Nonlinear Optimization for Clustering Nonlinear Biological Data. *IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM) 2011*. 1-6.

[22] Weston, J., Elisseeff, A., Scholkopf, B., and Tipping, M. 2003. Use of the Zero Norm with Linear Models and Kernel Methods. *The Journal of Machine Learning Research*. 3: 1439-1461.

[23] Shi, J. and Malik, J. 2000. Normalized Cuts and Image Segmentation. *IEEE Trans on Pattern Analysis and Machine Intelligence*. 22(8): 888-905.

[24] Ng, A., Jordan, M., and Weiss, Y. 2002. On Spectral Clustering: Analysis and An Algorithm. *Advances in Neural Information Processing Systems.* 2: 849-856.

[25] Yu, S. and Shi, J. 2003. Multiclass Spectral Clustering. *Proc. 9th IEEE Int. Conf. on Computer Vision*, 2003. 1: 313-319

[26] Luxburg, U. V. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing*. 17(4): 395-416.

[27] Malik, J., Belongie, S., Leung, T., and Shi, J. 2001. Contour and Texture Analysis for Image Segmentation. *Int. J of Computer Vision*. 43(1): 7-27.

[28] Kannan, R., Vempala, S., and Vetta, A. 2004. On clusterings: Good, Bad and Spectral. *J ACM (JACM)*. 51(3): 497-515.

[29] Luxburg, U. V., Belkin, M., and Bousquet, O. 2008. Consistency of Spectral Clustering. *The Annals of Statistics*. 36(2): 555-586.

[30] Luxburg, U. v., Bousquet, O., and Belkin, M. 2004. Limits of Spectral Clustering. *In Neural Information Processing Systems (NIPS)*. 857-864.

[31] Moran, G., Stokes, C., Thewes, S., Hube, B., Coleman, D. C., and Sullivan, D. 2004. Comparative Genomics Using Candida Albicans DNA Microarrays Reveals Absence and Divergence of Virulence-associated Genes in Candida Dubliniensis. *Microbiology*. 150(10): 3363-3382.

[32] Leung, Y. F. and Cavalieri, D. 2003. Fundamentals of cDNA Microarray Data Analysis. *Trends in Genetics*. 19(11): 649-659.

[33] Peterson, L. E. 2013. *Classification Analysis of DNA Microarrays.* 1 edition. John Wiley & Sons.

[34] Shalon, D., Smith, S. J., and Brown, P. O. 1996. A DNA Microarray System for Analyzing Complex DNA Samples Using Two-color Fluorescent Probe Hybridization. *Genome Research*. 6(7): 639-645.

[35] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. 1999. Expression Profiling using cDNA Microarrays. *Nature Genetics.* 21: 10-14.

[36] Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. 1999. High Density Synthetic Oligonucleotide Arrays. *Nature Genetics*. 21: 20-24.

[37] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., and Brown, E. L. 1996. Expression Monitoring By Hybridization to High-density Oligonucleotide Arrays. *Nature Biotechnology*. 14(13): 1675-1680.

[38] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. 2000. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*. 403(6769): 503–511. [Online]. From: http://llmpp.nih.gov/lymphoma/analysis.shtml [Accessed on 13 July 2015].

[39] Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., and Mazumder, A. 2006. Prognostic Gene Expression Signatures Can Be Measured in Tissues Collected in Rnalater Preservative. *The Journal of Molecular Diagnostics*. 8(1): 31-39. [Online]. From: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3726 [Accessed on 13 July 2015].

[40] Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., and Orntoft, T. F. 2003. Identifying Distinct Classes of Bladder Carcinoma Using Microarrays. *Nature Genetics*. 33(1): 90-96. [Online]. From: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE089 [Accessed on 13 July 2015].

[41] Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. 2002. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. *Cancer Research*. 62(17): 4963-4967. [Online]. From: http://www.chestsurg.org/publications/2002-microarray.aspx [Accessed on 13 July 2015].

[42] Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., Deimling, A. v., Pomeroy, S. L., Golub, T. R., and Louis, D. N. 2003. Gene Expression-based Classification of Malignant Gliomas Correlates Better With Survival Than Histological Classification. *Cancer Research*. 63(7): 1602-1607. [Online]. From: http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=82 [Accessed on 13 July 2015].

[43] Risinger, J. I., Maxwell, G. L., Chandramouli, G. V. R., Jazaeri, A., Aprelikova, O., Patterson, T., Berchuck, A., and Barrett, J. C. 2003. Microarray Analysis Reveals Distinct Gene Expression Profiles Among Different Histologic Types of Endometrial Cancer. *Cancer Research*. 63(1): 6-11. [Online]. From: http://home.ccr.cancer.gov/risingerdata1102/ [Accessed on 13 July 2015].

[44] Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., and Hampton, G. M. 2001. Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures. *Cancer Research*. 61(20): 7388-7393. [Online]. From: http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm [Accessed on 13 July 2015].

[45] West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Marks, J. R., and Nevins, J. R. 2001. Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles. *Proc. National Academy of Sciences*. 98(20): 11462-11467. [Online]. From: http://bioinformatics.rutgers.edu/Static/Supplements/CompCancer/datasets.htm [Accessed on 13 July 2015].

[46] Gao, Y. and Church, G. 2005. Improving Molecular Cancer Class Discovery Through Sparse Non-negative Matrix Factorization. *Bioinformatics*. 21(21): 3970-3975.

[47] Kim, H. and Park, H. 2007. Sparse Non-Negative Matrix Factorizations Via Alternating Non-negativity-constrained Least Squares For Microarray Data Analysis. *Bioinformatics*. 23(12): 1495-1502.

[48] Dhillon, I. S., Guan, Y., and Kulis, B. 2004. Kernel k-means: Spectral Clustering and Normalized Cuts. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04. 551-556.

[49] Long, B., Zhang, Z. M., Wu, X., and Yu, P. S. 2006. Spectral Clustering for Multi-type Relational Data. *Proc. 23rd Int. Conf. on Machine Learning, ICML* '0. 585-592.

[50] Alzate, C. and Suykens, J. A. K. 2010. Multiway Spectral Clustering with Out-of-Sample Extensions Through Weighted Kernel PCA. *IEEE Trans on Pattern Analysis and Machine Intelligence*. 32(2): 335-347.

[51] Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. 2003. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Research*. 13(4): 703-716.

[52] Dhillon, I. S. 2001. Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning. Proc. 7th ACM SIGKDD Int. Conf. *Knowledge Discovery and Data Mining*, KDD '01. 269-274.

[53] Speer, N., Frohlich, H., Spieth, C., and Zell, A. 2005a. Functional Grouping of Genes Using Spectral Clustering And Gene Ontology. *Proc. IEEE Int. Joint Conf. on Neural Networks*, 2005. IJCNN '05. 1: 298-303.

[54] Speer, N., Spieth, C., and Zell, A. 2005b. Spectral Clustering Gene Ontology Terms to Group Genes by Function. In Casadio, R. and Myers, G., editors, *Algorithms in Bioinformatics*. 1-12.

[55] Alzate, C. and Suykens, J. A. K. 2006. A Weighted Kernel PCA Formulation with Out-of-Sample Extensions for Spectral Clustering Methods. *Int. Joint Conf. on Neural Networks*, 2006. IJCNN '06. 138-144.

[56] Pelckmans, K., Van Vooren, S., Coessens, B., Suykens, J., and De Moor, B. 2006. Mutual Spectral Clustering: Microarray Experiments Versus Text Corpus. *Proc. workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology*. 55-58.

[57] Tritchler, D., Fallah, S., and Beyene, J. 2005. A Spectral Clustering Method for Microarray Data. *Computational Statistics & Data Analysis*. 49(1): 63-76.

[58] Higham, D. J., Kalna, G., and Kibble, M. 2007. Spectral Clustering and Its Use in Bioinformatics. *J Computational and Applied Mathematics*. 204(1): 25-37.

[59] Thurlow, J. K., Murillo, C. L. P., Hunter, K. D., Buffa, F. M., Patiar, S., Betts, G., West, C. M. L., Harris, A. L., Parkinson, E. K., Harrison, P. R., Ozanne, B. W., Partridge, M., and Kalna, G. 2010. Spectral Clustering of Microarray Data Elucidates the Roles of Microenvironment Remodeling and Immune Responses in Survival of Head and Neck Squamous Cell Carcinoma. *J Clinical Oncology*. 28(17): 2881-2888.

[60] Huang, G. T., Cunningham, K. I., Benos, P. V., CHENNUBHOTLA, C. S. 2013. Spectral Clustering Strategies for Heterogeneous Disease Expression Data. *Pacific Symposium on Biocomputing*. 212-223.