Clustering Microarray Data with Space Filling Curves

Dimitrios Vogiatzis[1] and Nicolas Tsapatsoulis[2]

[1] Department of Computer Science, University of Cyprus, CY 1678, Cyprus (phone: +357-2289-2749; fax: +357-2289-2701; email: dimitrv@cs.ucy.ac.cy).
[2] Department of Telecommunications Science and Technology University of Peloponnese Greece (email: ntsap@uop.gr).

Abstract. We introduce a new clustering method for DNA microarray data that is based on space filling curves and wavelet denoising. The proposed method is much faster than the established fuzzy c-means clustering because clustering occurs in one dimension and it clusters cells that contain data, instead of data themselves. Moreover, preliminary evaluation results on data sets from Small Round Blue-Cell tumors, Leukemia and Lung cancer microarray experiments show that it can be equally or more accurate than fuzzy c-means clustering or a gaussian mixture model.

keywords: clustering, space filling curve, wavelets, microarray DNA

1 Introduction

Microarray experiments allow the simultaneous study of expression patterns of thousands of genes. Usually, microarray datasets are characterized by a large number of genes across a relatively small number of different experimental conditions [1]. The genes form a data set of a few thousands of vectors, while the experimental conditions (a few tens) constitute the dimensions of each vector. One of the reasons behind microarray experiments is to figure out the genes that have similar biological function, by comparing their expression patterns. An unaided researcher trying to make sense of these data will have a hard time. Clustering is a widely used method to group those genes that have similar expression levels into the same clusters. *Hierarchical clustering* has been widely used in microarray experiments, where smaller clusters are merged to form a hierarchical tree called the dendrogramme [2]. However, the visualisation that is offered by such a method is problematic as thousands of tiny line segments representing the genes can clutter the screen. In *Partition based clustering* the data are split into a fixed number of clusters (either crisp of fuzzy) by optimising an objective function through a series of steps. A representative is the fuzzy c-means clustering [3]. Moreover, in *Grid based clustering* the input space is first quantised into a fixed number of cells and then the clusters are formed out of cells [4]. Finally, in *Density based clustering* the aim is to find high density regions of the data space that are separated from low density regions. High density regions stand for clusters. A widely used density estimation method is through a mixture of gaussian models. Mixture models can be learnt with the expectation maximisation Algorithm (EM). A fast method for dynamically computing a mixture model appeared in [5]. We refer the interested reader to a recent survey of clustering and cluster analysis for gene expression data [6].

The gene clustering method for DNA microarray we propose is based on a four step process. First, we partition (quantise) the input space. Second, we map the multidimensional gene expression vectors onto one dimension, the end result of which is a spatial signal. Then we use one dimensional discrete wavelet transform on the spatial signal to denoise the signal. Finally, we cluster the one dimensional data based on the assumption that cells that are not close belong to different clusters. Also, low density cells represent the boundaries of clusters.

To place our work into context, we would say that it has some elements of partition based clustering and it is also related to a *WaveCluster*, where wavelets are used to cluster data of very large databases. In this method low pass filter are used to remove outliers [7]. It has been shown to be very efficient and to detect arbitrary shaped clusters on benchmark datasets. However, *WaveCluster* has been applied to two dimensional data, whereas our proposal can deal with multidimensional data.

The rest of the paper is organised as follows: In Sect. 2 we introduce the concept of space filling curves, and we also present wavelet denoising. Then in Sect. 3 we present the space filling based clustering method we developed. Experiments and evaluation are presented in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 Space Filling Curves and Wavelets

A space filling curve is a one dimensional curve that can fill an entire plane [8]. There are many space filling curves, in particular we are interested in the *Z-space filling curve*. This curve is a mapping $S : \mathbb{R}^n \rightarrow \mathbb{R}$, which is constructed by interleaving bits from a point's M dimensions into a single dimension. For example, given vector $\mathbf{e} = (v_1, v_2, \ldots v_n)$, with k bits $b_1 \ldots b_k$ used to represent each dimension, with b_1 and b_k being the most and least significant bits respectively. The one dimensional projection of $\mathbf{e}$ is $e' = (b_1^{v_1} b_1^{v_2} \ldots b_1^{v_n} b_2^{v_1} b_2^{v_2} \ldots b_2^{v_n} \ldots b_k^{v_1} b_k^{v_2} \ldots b_k^{v_n})$, where b_i^j denotes the i-th ordered bit from dimension j. In Fig. 1 is depicted a two dimensional version of a Z-curve ordering of the cells in an area. In particular, we can observe a first, second and third order curve. Higher order curves represent a "denser" covering of the input space. The limit of Z-curve is the area that contains the curve.

The Z- curve has the interesting property (easier to visualise in two dimensions, but also holds for more dimensions), that it tends to preserve the locality of the data. That is data that are close together in $\mathbb{R}^n$ tend also to be close in $\mathbb{R}$, which does not hold for row major ordering. The Z-curve can be considered as a spatial signal, which can analysed with signal processing techniques, and in particular wavelets.

From the point of view of mathematics, a function can be represented as an infinite series expansion in terms of a dilated and translated version of a basis function called the *mother wavelet* denoted as $\psi(x)$ and weighted by some coefficient $b_{j,k}$: $f(t) = \sum_{j,k} b_{j,k} \psi_{j,k}(t)$ Normally, a wavelet starts at time $t = 0$ and ends at time N. Instead of time one can consider space (as it is often the case in image analysis). A shifted wavelet, denoted as ψ_{jo}, starts at time $t = k$ and ends at time $t = k + N$. A dilated wavelet w_{j0} starts at time $t = 0$ and ends at time $t = N/2^j$. A wavelet w_{jk} that is dilated j times and shifted k times is denoted as: $\psi_{j,k}(t) = \psi(2^j t - k)$. For practical purposes, we can use the discrete wavelet transform, which removes some of the redun-

dancy found in the continuous transform. In this study we rely on *wavelet shrinkage* for denoising. The shrinkage is based on discarding some of the detail coefficients and then by reconstructing the signal based on the reduced set of coefficients. Moreover, in [9] it has been shown that the wavelet shrinkage method outperforms other methods for denoising signals.

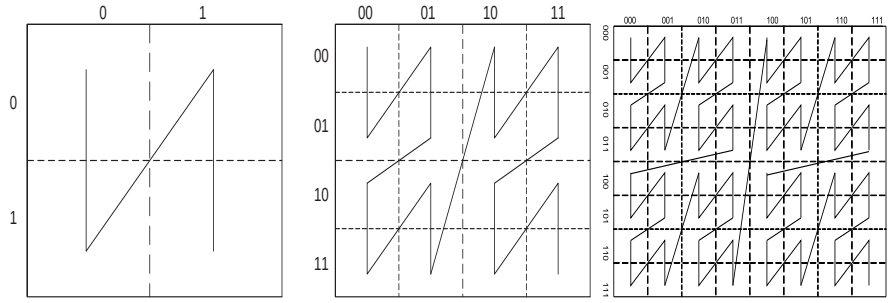

Fig. 1. The Z-space filling curve. The curve in each figure denotes the order of visitation of the cells.

3 Clustering with the Z-curve

The proposed algorithm with the Z-space filling curve, henceforth called Space Filling Curve Clustering (SFCC), accepts as input a matrix of microarray data, and it assigns genes into clusters. The number of clusters are discovered by the algorithm. The steps are summarised as follows: Quantise Input space, Construct space filling curve, Smooth the curve, discover clusters. The algorithm is exposed in Table 1. The rationale of step 3 (i.e. smoothing by denoising) is based on the assumption that the limits of a cluster is marked by a high frequency component. Thus by zeroing the high frequency components we make cluster detection clearer. Step 4, says that points that belong to the same cluster must be close (determined by threshold t_1) and the cell must have a minimum amount of data density (determined by threshold t_2).

For the proposed clustering method the time complexity is the steps it takes to create the curve, to apply wavelet denoising and to cluster the data. Let N_c be the number of cells that contain data points, and N the number of data, consequently $N_c < N$. The time to create the Z-curve is $M \times nb \times N$, where nb is the number of bits used to encode each dimension. Thus $O(M \times nb \times N + N_c log N_c + N)$.

The computational complexity of the FCM algorithm is $O(Nk^2M)$, where N is the number of data, k the number of clusters and M the number of dimensions (experiments) of the data. The important thing to notice is that the complexity of algorithm is quadratic with regard to the number of clusters.

Table 1. Clustering with the Z-curve

1. Quantise each dimension of the input space into equally spaced intervals.
2. Record the number of data in each resulting hyper rectangle
3. Construct a space filling curve $S(i)$ that passes through the created hyper rectangles (cells). Because the cell space is sparsely populated, $S(i)$ is created only for cells that contain data. $i \in \mathbb{N}$ represents the index of the cells, and $S(x)$ represents the number of data points per cell.
4. Smooth the curve by applying wavelet denoising.
5. Cluster the cells (i.e. their indices) that contain data (one dimensional clustering) as follows for $i = 2 \ldots i = length(S)$.
   (a) If $||S(i) - S(i-1)||_2 \leq t_1$ and $S(i) \geq t_2$ then put current cell in the existing cluster.
   (b) else if $||S(i) - S(i-1)||_2 > t_1$ and $S(i) \geq t_2$ then create a new cluster, which becomes the current cluster and put current cell into new clusters
   (c) else this cell is an outlier and ignore it.

Finally, learning gaussian mixtures with the greedyEM algorithm takes $O(N \times k^2)$ steps, under certain conditions the complexity can be reduced to $O(N \times k)$ according to [7].

4 Experiments and Evaluation

We compared the proposed method (i.e. SFCC) with FCM and greedyEM in terms of two validation indices: *figure of merit* [10] and *silhouette* [11]. The figure of merit (FOM) is defined as: $FOM(e) = \sqrt{\frac{1}{N}\|R_c(x,e) - \mu_c(e)\|^2}, \forall c$ where $R_c(x,e)$ represents the e dimension of datum x that belongs in cluster c, μ_c represents the average value of $R_c(x,e)$, N is the number of data (genes), M the number of experiments (dimensions) and e is index in the experiments. The FOM index of the whole clustering is defined as:

$$FOM = \sum_{e=1}^{M} FOM(e) \tag{1}$$

Smaller values of FOM denote better clustering, for the same number of clusters by different algorithms.

The silhouette index for datum x of cluster c is defined as:

$$s_c(x) = \frac{min[b_{\forall c}(x)] - a_c(x)}{max\{a_c(x), min[b_{\forall c}(x)]\}} \tag{2}$$

where $a(x)$ is the average dissimilarity of datum x to the data of the same cluster, and $b(x)$ is the average dissimilarity of a datum x from all the data of another cluster. Dissimilarity can be defined as the eucledian distance. The silhouette index of cluster c is: $S_c = \frac{1}{|c|}\sum_{i=1}^{|c|} s_c(i)$. Finally, the silhouette index of the whole clustering is:

$S = \frac{1}{k}\sum_{j=1}^{k} S_j$, where k is the number of clusters. From the definition if follows that: $s_c(x) \in [-1, 1]$. An $s(x)$ value for datum x close to 1 denotes good clustering, a value close to 0 denotes that the datum belongs to more than one clusters, and a value close to -1 denotes that x belongs to another cluster.

We have used datasets from 3 microarray DNA experiments. The first data set was obtained from "The Microarray Project cDNA Library" `http://research.nhgri.nih.gov/microarray/Supplement/`. The second and third data sets were obtained from the Gene Expression Datasets collection `http://sdmc.lit.org.sg/GEDatasets`. The first data set is about Small Round Blue-Cell tumours (SRBCT), investigated with cDNA microarrays containing 2308 genes, over a series of 83 experiments. The 83 samples included tumour biopsy material and cell lines from 4 different types: Ewing's sarcoma (EWS), rhabdomyo sarcoma (RMS), neuroblastoma (NB) and Burkitt's lymphoma (BL) [12]. The provenance of the second data set stems also is from oligonucleotide microarrays, with a view of distinguishing between acute lymphoblastic leukemia (ALL) and acute meyeloid leukemia (AML). The data set consisted of 72 bone morrow samples from 7130 human genes [13]. The third data set also stems from a microarray experiment and consists of lung malignant pleural mesothylioma (MPM) and adenocarcinoma (ADCA) samples [14]. The data set consists of 181 samples from 12534 human genes. All data sets have been normalised in the $[0, 1]$ region.

Experiments have been performed at Matlab 6.1, with the implementation of FCM from fuzzy toolbox 2.1.1, and wavelet denoising from the wavelet toolbox 2.1. The code for greedyEM was obtained from the author's site `http://www.science.uva.nl/~vlassis/publications`. The code for space filling curves and evaluation was developed by the authors in matlab. In the wavelet based smoothing we employed daubechy of order 2. The wavelet smoothing is achieved by applying 4 levels of decomposition for SRBCT and Leukemia data sets and 8 levels for the Lung set. Then we set the detail coefficients to zero and we reconstructed the signal. Also, the thresholds t_1 and t_2 influence the performance of the algorithm since they define when clusters occur, thus we varied the values of t_1 and t_2 from 0.05 to 0.7 with a step of 0.01. Finally, the quantisation step for each dimension for all data sets has been set to 10. The levels of wavelet decomposition that are used to smooth the signal (i.e. the space filling curve) play a crucial role in the performance of the algorithm. Currently, the number of levels of decomposition are experimentally determined.

In Fig. 2 and we depict the results of evaluating SFCC and comparing it with FCM and greedyEM under the FOM criterion (recall that smaller values indicate better results). The FOM in the case of the space filling curve has been applied to the multidimensional data according to the cluster they belong to. SFCC is depicted with diamonds, FCM with rectangles and greedyEM with small dots. At the first diagramme, corresponding to the SRBC experiments, the SFCC is overall winner for a small number of clusters (2-4). At the middle diagramme (Leukemia) greedyEM is the best method. At the right most diagramme, which corresponds to Lung Cancer, it is shown that the SFCC outperforms FCM or greedy in for most cases (from 5 till 20 clusters).

Finally, in Fig. 3 we present the evaluation of SFCC, FCM and greedyEM with respect to the silhouette validation index (recall that bigger values are better and non

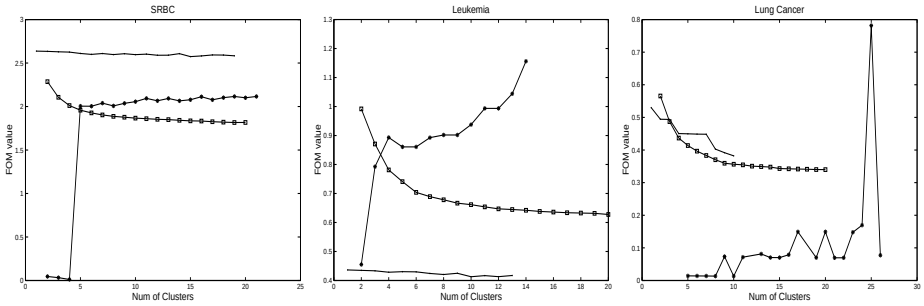

Fig. 2. Evaluation results based on comparing FOM values for the proposed method (star curve), FCM (square curve), greedyEM (small dots curves)

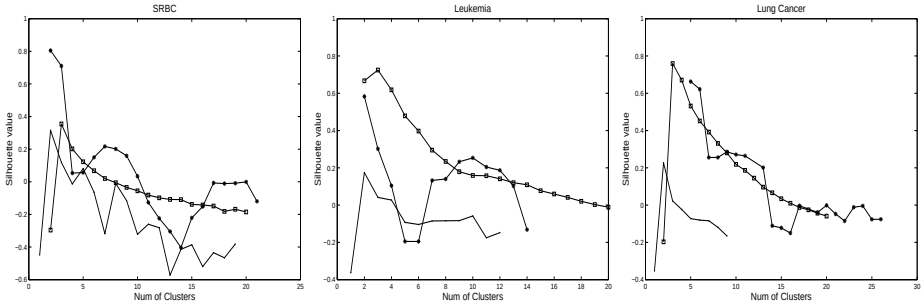

Fig. 3. Evaluation results based on comparing silhouette values for the proposed method (star curve), FCM (square curve), greedyEM (small dots curves)

positive values denote bad clustering). At the leftmost diagramme (SRBC data), the winner is SFCC in most cases. Considering the Leukemia data (middle diagramme) the winner in most cases is the FCM. At the rightmost diagramme (Lung Cancer), SFCC and FCM have a comparable behaviour beyond three clusters, whereas greedyEM is generally worse than all the other methods.

Considering both the FOM and the silhouette validation indices, the SFCC is better or at least equally good as FCM. For the leukemia data, the two indices do not concur about the overall clustering quality of each of the clustering algorithms. In any case, we must recall that even if SFCC is equally or slightly worse than FCM, it is much faster to compute.

5 Conclusions and Future Directions

We have developed an efficient method to cluster genes from DNA microarray experiments. Our method is based on Z-space filling curve which maps multidimensional genes into one dimension and it performs clustering into one dimension which is very efficient in terms of computational complexity. It is important to emphasize that the proposed method actually does not cluster data directly, but it clusters cells into which data belong (after some partitioning). Thus it is independent of the number of data but dependent on the quantisation step. The outcome of the Z-space filling curve is a one dimensional spatial signal which can be processed as described to detect clusters. The algorithm is dependent on two thresholds, the maximum distance between two cells so that they belong to the same cluster and also on the minimum data density of a cell so as not to be considered as outlier. Of course, by clustering cells, we also cluster the data that belong to each cell. Wavelets play an important role, because they constitute a preprocessing step to the actual clustering. With wavelet shrinkage, we can denoising the spatial signal and achieve better results. Thus this paper also contributes in introducing signal processing techniques into multidimensional data. As evaluation, we have employed the FOM and silhouette criteria to compare SPCC with FCM and greedyEM, where we obtained promising results. In any case there can be no clustering method that is panacea. The clustering results will always depend on the data distribution of the samples, on the amount of noise they contain, and on the model the user tries to apply to these data.

For the future, it is important to enhance our evaluation with other measures such as the Partition Coefficient, Dunn's index and the Geometric index in order to check the validity of the derived clusters; the aforementioned indexes have been used in a work related to evaluating clusters in cDNA experiments [15]. Furthermore, all aforementioned evaluation measures are based on statistics, and we need to investigate the biological significance of the discovered clusters. For example, in [16] a clustering experiment is described, where the genes of each cluster are mapped into the functional categories of the Martinsried Institute of Protein Sciences. Then for each cluster $P-$values were calculated to measure the statistical significance of clusters.

Moreover, the space filling curve is of crucial importance in the algorithm and a basic property it must have is to preserve the locality of the data. There is enough

research on such curves and there is evidence that the hilbert curve can achieve better clustering. We need to investigate that on more microarray experiments.

References

1. Macgregor, P., Squire, J.: Application of microarrays to the analysis of gene expression in cancer. Clinical Chemistry **48** (2002) 1170–1177
2. Eisen, M., Spellman, P., Brown, P., Botsetein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Scienes **95** (1998)
3. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
4. Hinneburg, A., Keim, D.: Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. In: Proceedings of the 25th VLDB Conference, Edinburgh, Scotland. (1999)
5. Verbeek, J., Vlassis, N., Kröse, B.: Efficient greedy learning of gaussian mixture models. Neural Computation **15** (2002) 469–485
6. Jiang, D., Tang, C., Zhang, A.: Cluster Analysis for Gene Expression Data: A Survey. IEEE transactions on knowledge and data engineering **16** (2004) 1370–1386
7. Sheikholeslami, G., Chatterjee, S., Zhang, A.: WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal **8** (2000) 289–304
8. Faloutsos, C., Roseman, S.: Fractals for secondary key retrieval. In: 8th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems PODS. (1989) 247–252
9. Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Wavelet shrinkage: Asymptopia? J. R. Statist. Soc. B. **57** (1995) 301–337
10. Yeung, K., Haynor, D., Ruzzo, W.: Validating clustering for gene expression data. Bioinformatics **17** (2001) 309–318
11. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comp. App. Math **20** (1987) 53–65
12. Khan, J., Wei, J., Ringer, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., Meltzer, P.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network. Nature Medicine **7** (2001) 673–679
13. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (1999)
14. Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richard, W., Sugarbaker, D., Bueno, R.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research (2002) 4963–4967
15. Lam, B., Yan, H.: Cluster Validity for DNA Microarray Data using a Geometrical Index. In: Proceedings of the 4th International Conference on Machine Learning and Cybernetics. (2005)
16. Tavazoie, S., Hughes, D., Campbell, M., Cho, R., Church, G.: Systematic determination of genetic network architecture. Nature Genetics **22** (1999) 281–285