Merge Method for Shape-based Clustering in Time Series Microarray Analysis

Irene Barbero[1], Camelia Chira[1,2], Javier Sedano[1,3], Carlos Prieto[4], José R. Villar[5], and Emilio Corchado[6]

[1] Instituto Tecnológico de Castilla y León, Burgos, Spain, irene.barbero@itcl.es, camelia.chira@itcl.es, javier.sedano@itcl.es
[2] Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania
[3] Department of Electromechanical Engineering, University of Burgos, Spain
[4] Instituto de Biotecnología de León, Spain, carlos.prieto@unileon.es
[5] University of Oviedo, Gijón, Spain, villarjose@uniovi.es
[6] University of Salamanca, Spain, escorchado@usal.es

Abstract. A challenging task in time-course microarray data analysis is to combine the information provided by multiple time series in order to cluster genes meaningfully. This paper proposes a novel merge method to accomplish this goal obtaining clusters with highly correlated genes. The main idea of the proposed method is to generate a clustering, starting from clusterings created from different time series individually, that takes into account the number of times each clustering assemble two genes into the same group. Computational experiments are performed for real-world time series microarray with the purpose of finding co-expressed genes related to the production and growth of a certain bacteria. The results obtained by the introduced merge method are compared with clusterings generated by time series individually and averaged as well as interpreted biologically.

Keywords: microarray analysis, time series, clustering, merge methods

1 Introduction

Nowadays, microarray technology provides the possibility to measure gene expression levels of thousands of genes. Some difficulties to face with this type of data are the cost and the high dimensionality. Microarray data can be analysed from a static viewpoint [1, 2, 3, 4], ignoring the temporal information of time series, or from a temporal perspective [5, 6, 7, 8]. In this paper, we focus on temporal microarray data analysis.

A premise widely accepted when analysing microarray data is that genes showing similar behaviour in their expression levels, i.e. co-expressed, are possibly functionally related [9]. Hence, the task of clustering genes based on time series is extremely important. Clustering time series microarray data has been intensively investigated and many algorithms have been proposed to carry out this task without any clear overall best performing method. These methods can be classified in point-wise distance based, feature based clustering and model based clustering [6]. Point-wise distance methods compute the clusterings according to an objective function based on a distance measure between gene pairs, feature based clustering methods detect some features among the expression profiles and model based clustering methods create models capable of describing the data. In this paper, we engage two feature based clustering methods with the purpose of merging their results for different time series. These clustering methods are shape index clustering [10] and an extension of it introducing the correlation with an output called Shape Output Clustering.

A merge method is proposed to blend clusterings created independently from time series in order to get a new more meaningful clustering. In the first phase, several clusterings of the same genes are produced using the shape clustering methods based on individual time series data. In the second phase, one single clustering is obtained by merging the groups resulted from applying the same shape-based clustering method to different time series data. The results show how the proposed method is able to generate more restrictive and highly correlated gene clusters. Moreover, this method can help in future work to overcome the problem of detecting significant genes that promote a certain event or variable.

The paper is structured as follows: section two describes in detail the shape-based clustering methods; section three explains thoroughly the merge clustering method proposed; section four discusses the experimental results obtained for the real-world time series microarray data considered, giving also a biological perspective of the results and section five contains the conclusions of paper.

2 Shape-based Clustering Methods

The grouping of genes based on values from individual time series is performed using two shape-based clustering methods: Shape-based Clustering (SC) and Shape Output Clustering (SOC). Both methods rely on modelling the change of expression value between consecutive time points using a shape index. The first method engaged is described in [10] where it is used as the first step of a clustering methodology for time series microarray data based on similar rate of change and modulation patterns in gene expression profiles. SOC extends the functionality of SC by further taking into account the gene correlation with the output. The main idea of SOC is to group two genes together if they follow a similar pattern of changes in gene expression over time and with regard to the output.

Let S be the number of time points considered (samples available). For each sample $i, i = 1 \ldots S$, the value of the current time point is given by t_i and the corresponding gene expression level is denoted by x_i. Let g_step be the rate of change in the gene profile calculated at each time interval as the difference between the two consecutive gene expression levels x_{i+1} and x_i divided to the difference between the time points. Similarly, the rate of change in the output y_step is computed using the value of output instead of the gene expression level.

The values of g_step and y_step ($step$ for short) are used to decide how significant is the change in gene expression level from one time point to the next and how correlated with the change in the output at the same time interval. This is achieved through the use of a threshold ψ which indicates the level of acceptable difference between two consecutive values. Based on the relationship between the step values and the threshold, g_class and y_class ($class$ for short) levels are assigned to each time interval indicating the rate of change in the gene/output pattern. If $step \in (-\psi, \psi)$, the level of change is not significant and the gene/output category $class$ has a 'no change' (*stable*) meaning associated. If $step \geq \psi$ (respectively $step \leq -\psi$) then we have an increase (respectively decrease) of the gene expression level.

Once all g_class and y_class values are assigned, the g_index corresponding to a gene is calculated following Eq. 1 for SC, respectively Eq. 2 for SOC (l is the number of different categories that can be assigned to a gene).

$$g\_index^{SC} = \sum_{i=1}^{S-1} l^i * g\_class(t_i, t_{i+1}) \tag{1}$$

$$g\_index^{SOC} = \sum_{i=1}^{S-1} l^i * g\_class(t_i, t_{i+1}) * y\_class(t_i, t_{i+1}) \tag{2}$$

Two genes are placed in the same group if they have the same g_index value. In SOC, the rate of change in the output level is allowed to directly influence the gene shape index and, consequently, the gene clustering process. The final result is a set groups of genes having the same g_index and therefore similar shape of change and output correlation.

3 Merging Shape-based Clusterings

The proposed merge method combines the information provided by multiple different time series clustered separately by either SC or SOC method. The aim is to create a combination of all of them resulting in a more meaningful clustering. In the context of the considered microarray problem, this merge method has been applied to three time series but it can be generalised to any number of time series.

Let N be the number of genes and S the number of time series in the microarray data. The input of the merge method is represented by the clusterings created by SC/SOC. Let C be a matrix that encompasses this information: c_{ij} is the cluster assigned to a gene g_j when applying SC/SOC to the time series i for $i = 1...S$, $j = 1...N$. The merge method computes a value denoted by v_{ij} for each pair of genes (g_i, g_j) such that $i < j$. These values are saved in a triangular matrix V.

The main steps of the merge shape index clustering algorithm are as follows:

(i) Assign a value to each possible combination of two genes (all v_{ij} of $V, i < j$) depending on the number of time series that group together that pair of genes

according to Eq.(3).

$$\begin{aligned} v_{ij} &= \frac{1}{S}\textstyle\sum_{s=1}^{S} d(c_{si},c_{sj}) \\ d(c_{si},c_{sj}) &= \begin{cases} 1 & \text{if } c_{si}=c_{sj} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

(ii) Obtain the unique values of v_{ij} for all $i < j$ in descending order, which are denoted by $\alpha = \{\alpha_1, \alpha_2, \alpha_3 \ldots, \alpha_k\}$, so that α_1 is the greater value and α_k the smallest.

(iii) Generate the merged clustering M which consists in a list of clusters, initially empty, as outlined in the pseudocode below. The process consists in creating first clusters of genes that were detected as co-expressed by the clustering in all time series. In the second step, clusters are formed with the genes that were detected as co-expressed by the clustering in all time series except one and which are not formed in the previous step. This process continues iteratively until the last clusters formed correspond to genes that are considered to be co-expressed only by the clustering in one time series.

Merge Clustering Procedure:

```
$M = \emptyset$
$genesConsidered = \{g_1, g_2, g_3, ..., g_N\}$
for each $\alpha_l$, $l = 1...k$
    $genesConsidered = genesConsidered - M$
    for each $v_{ij}$ such that $g_i, g_j \in genesConsidered$
        if $v_{ij} >= \alpha_l$
            if $g_i \notin M$ and $g_j \notin M$
                create new group $M_{new} = \{g_i, g_j\}$;
                add $M_{new}$ to $M$
            elseif $g_i \notin M$ but $\exists u \in [1, size(M)] \mid g_j \in M_u$
                $M_u = M_u \cup \{g_i\}$
            elseif $g_j \notin M$ but $\exists u \in [1, size(M)] \mid g_i \in M_u$
                $M_u = M_u \cup \{g_j\}$
            elseif $\exists p \in [1, size(M)] \mid g_j \in M_p$ and
            $\exists q \in [1, size(M)] \mid g_i \in M_q$
                create new group $M_{new} = M_p \cup M_q$
                remove $M_p$ and $M_q$ from $M$
                add $M_{new}$ to $M$
            endif
        endif
    endfor
endfor
$M_0 = \{g_1, g_2, g_3, ..., g_N\} - M$
```

The result is a set of clusters of co-expressed genes denoted by $M = \{M_1, M_2, M_3, \ldots, M_m\}$. Every cluster has an associated α value that created it. This α value will provide information about the number of times that the genes of the group were clustered together by all time series, so that the greater α, the more correlated the genes are. Furthermore, the merge method also generates a cluster M_0 which contains the unclustered genes considered not to be correlated with any other.

4 Computational Experiments and Results

The microarray design was performed with the software eArray 5.0 and has provided the capacity to measure the expression of 8848 genes. Biological samples were extracted as time series of 12 time points that belonged to three different cell lines. The microarray dataset was preprocessed prior to applying a clustering method and combining the clusterings using the proposed merge method.

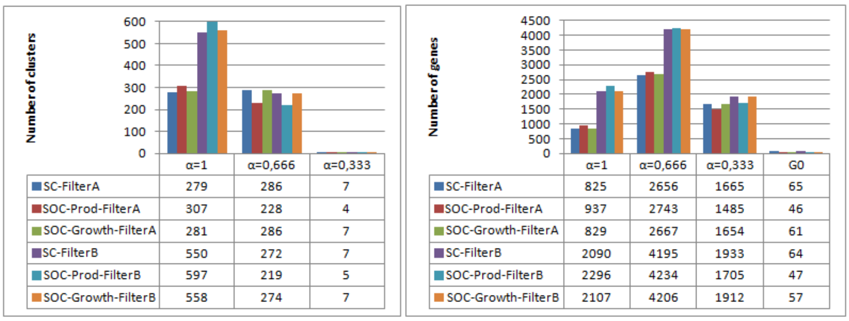
A normalization process was performed with the limma package [11]. Median and none background correction and Cyclic Loess normalization methods were applied in order to determine the best input for the implemented methods. Furthermore, a filtering phase with the aim of eliminating noisy genes was engaged. Two different filters were used as follows: (i) *FilterA* removes those genes for which the difference between the maximum and minimum value taken through the samples is greater than 1.5, and (ii) *FilterB* removes those genes for which difference between the maximum and minimum value taken is less than 0.75 and its standard deviation is less than 0.25. Eventually, FilterA selected 5211 genes and FilterB 8282 genes (out of the total 8848 genes).

The analysis of results compares the proposed method with the outcomes of SC and SOC methods individually applied over the time series as well as applied to the mean of the three time series.

4.1 Merge Clustering Results

To perform the experiments, the three time series (denoted by series1, series2 and series3) have been used independently to generate three different clusterings and averaged (called Mean) to produce one single clustering. Over these data, the SC and SOC methods are applied with threshold $\psi = 0.08$. SOC can use one of the two outputs available in the considered microarray data: production (SOC-Prod) or growth (SOC-Growth). The proposed merge method has been applied over the clusterings created from the time series independently.

Analysing the results of the merge method, the largest number of clusters have an associated α value of 1 and only few clusters emerged from an $\alpha = 0.333$ value (corresponding to genes correlated only in one of the three time series). Nevertheless, if we consider the amount of genes classified within all clusters created according to each α value, by far, the greatest number of genes corresponds to $\alpha = 0.666$ (see Fig. 1 right). Generally, the size of clusters corresponding to $\alpha = 0.333$ and $\alpha = 0.666$ is high while their number is really low. However, the remaining groups contain rather few genes compared to the size of the groups generated by each time series and mean. It should be noted that the merge method considers very few genes as uncorrelated with other genes (genes



	α=1	α=0,666	α=0,333
SC-FilterA	279	286	7
SOC-Prod-FilterA	307	228	4
SOC-Growth-FilterA	281	286	7
SC-FilterB	550	272	7
SOC-Prod-FilterB	597	219	5
SOC-Growth-FilterB	558	274	7

	α=1	α=0,666	α=0,333	G0
SC-FilterA	825	2656	1665	65
SOC-Prod-FilterA	937	2743	1485	46
SOC-Growth-FilterA	829	2667	1654	61
SC-FilterB	2090	4195	1933	64
SOC-Prod-FilterB	2296	4234	1705	47
SOC-Growth-FilterB	2107	4206	1912	57

Fig. 1. Number of clusters (left) and number of genes (right) obtained.

in M_0) compared to the rest of clusterings. The lower the α value, the more probable is to get large groups and the more different the time series, the larger the obtained clusters. Generalizing, we can say that the proposed merge method generates a few more clusters compared to the mean clustering results and less than clusterings obtained by SC/SOC for individual time series.

Regarding the gene profiles, meaningful groups are obtained applying the merge method over the considered clustering methods. A cluster of four genes obtained by the merge method is depicted as an example in Fig. 2. As it can be seen, the merge method groups those genes that have similar genes profiles over time for all time series. The same four genes have been observed clustered in FilterB-SOC-Growth-Mean but included in a larger cluster of 131 genes.

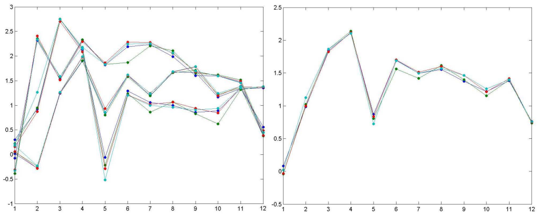


Fig. 2. Genes profiles of a cluster with four genes obtained applying the merge method over the three clusterings created from the time series using either SC or SOC. Profiles for each time series (left) and the mean of the gene values (right) are depicted.

On the other hand, Fig. 3 shows a cluster of 3 genes created by $\alpha = 0.666$ where it can be clearly observed that the correlation between genes is a bit less strict than in previous clusters created by $\alpha = 1$. In this case, series2 considers the three genes as uncorrelated while series1 and series3 classify the three genes in the same group. Unlike the other example, the clustering over the mean of the time series considers that one of the genes is not correlated with any one and the other genes are classified in separate clusters.

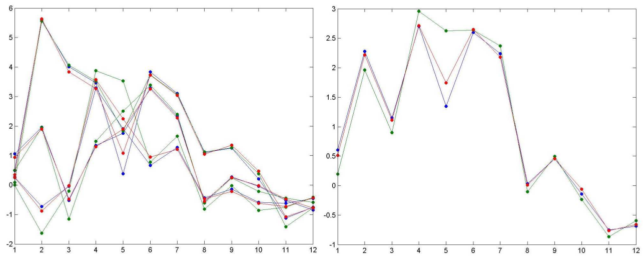


Fig. 3. Genes profiles of a cluster of 3 genes obtained applying the merge method over the three clusterings created from the time series using either SC or SOC. The gene profiles of each time series (left) and the mean of the gene profiles (right) are depicted.

4.2 Biological Perspective

The method proposed has reached the final objective of clustering genes in groups with similar expression profiles. It has been seen that production genes were clustered with others which have similar expression and their expression profiles follow a common pattern which is coherent with the production rate.

Moreover, it has been shown how some genes have different expression values between the three time series. This is not a usual case in these time series because they are well synchronized, but it is interesting to consider that some genes could have high expression variations during time. These genes can be related to tight regulation processes in which the expression changes are quite frequent. In order to deal with these profiles, it has been shown how the series can be analyzed without previous merging. In addition, this approach could be interesting for the analysis of bad synchronized or noisy samples, in order to avoid potential errors in the clustering calculation process.

5 Conclusions and Future Work

A clustering merge method was proposed to combine the information provided by different time series for time-course microarray analysis. The results obtained indicate that the merge method is able to generate valuable clusters containing genes highly correlated with regard to all time series compared to the clustering methods applied over the averaged or individual time series. It has been shown how the mean of the series is unable to find any relationship between genes despite very similar expression profiles with regard to some of the time series.

The proposed method can be useful in further analysis of time series microarrays for detecting important genes that promote a certain event or variable. Because the merge method provides an indicator of the correlation between the genes of the generated clusters, it is possible to analyze groups that have obtained a low indicator of correlation in order to detect those genes with high expression variations over time. This type of genes, characterized by frequent expression changes, are not regular in synchronized time series but they are likely to be fairly common in noisy samples. Therefore, the proposed merge method can be employed for detecting genes related to tight regulation and for analysis of poorly synchronized or noisy samples.

Acknowledgements. This work was supported by a grant of the Agencia de Inversiones y Servicios de Castilla y León (record CCTT/10/BU/0002 and CCTT/10/LE/0001) and the Spanish Ministry of Science and Innovation [PID 560300-2009-11]. Carlos Prieto was supported by a Juan de la Cierva programme (JCI-2009-05444) of the Ministry of Science and Innovation (Spain).

References

[1] Chien-Pang Lee, Wen-Shin Lin, Yuh-Min Chen, and Bo-Jein Kuo. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. *Expert Systems with Applications*, 38(5):4661–4667, May 2011.

[2] Huawen Liu, Lei Liu, and Huijie Zhang. Ensemble gene selection by grouping for microarray data classification. *Journal of Biomedical Informatics*, 43(1):81–87, February 2010. PMID: 19699316.

[3] Yu Wang, Igor V Tetko, Mark A Hall, Eibe Frank, Axel Facius, Klaus F X Mayer, and Hans W Mewes. Gene selection from microarray data for cancer classification–a machine learning approach. *Computational Biology and Chemistry*, 29(1):37–46, February 2005. PMID: 15680584.

[4] Jun S. Wei, Braden T. Greer, Frank Westermann, Seth M. Steinberg, Chang-Gue Son, Qing-Rong Chen, Craig C. Whiteford, Sven Bilke, Alexei L. Krasnoselsky, Nicola Cenacchi, Daniel Catchpoole, Frank Berthold, Manfred Schwab, and Javed Khan. Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer Research*, 64(19):6883 –6891, October 2004.

[5] Norma Coffey and John Hinde. Analyzing time-course microarray data using functional data analysis - a review. *Statistical Applications in Genetics and Molecular Biology*, 10, 2011. peer-reviewed.

[6] Ritesh Krishna, Chang-Tsun Li, and Vicky Buchanan-Wollaston. A temporal precedence based clustering method for gene expression microarray data. *BMC Bioinformatics*, 11(1):68, 2010.

[7] Sung-Gon Yi, Yoon-Jeong Joo, and Taesung Park. Rank-based clustering analysis for the time-course microarray data. *Journal of Bioinformatics and Computational Biology*, 7(1):75–91, February 2009. PMID: 19226661.

[8] John Storey, Wenzhong Xiao, Jeffrey Leek, Ronald Tompkins, and Ron Davis. Significance analysis of time course microarray experiments. *UW Biostatistics Working Paper Series*, August 2004.

[9] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005. PMID: 16162296.

[10] Sieu Phan, Fazel Famili, Zoujian Tang, Youlian Pan, Ziying Liu, Junjun Ouyang, Anne Lenferink, and Maureen Mc-Court O'connor. A novel pattern based clustering methodology for time-series microarray data. *International Journal of Computer Mathematics*, 84(5):585–597, May 2007.

[11] G.K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–73, 2003.