Clustering and classifying gene expressions data through Temporal Abstractions

L. Sacchi[1], R. Bellazzi[1], C. Larizza[1], P. Magni[1], T. Curk[2], U. Petrovic[3], B. Zupan[2,3,4]

[1] *Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy*
[2] *Department of Computer Science, University of Ljubljana, Slovenia*
[3] *J. Stefan Institute, Ljubljana, Slovenia*
[4] *Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, USA*

Abstract

This paper describes a new technique for clustering short time series coming from gene expression data.The technique is based on the labeling of the time series through temporal trend abstractions and a consequent aggregation of the series on the basis of their labels.Results on simulated and on yeast data are shown. The technique appears robust and efficient and their results easy to be interpreted.

1. Introduction

The rationale and motivation for applying clustering techniques in bioinformatics research has been recently studied [1]. Within this area, an issue of raising interest is related to the classification and clustering of time series of gene expression data. The methods which have been proposed in the literature can be classified in two broad categories: discriminative or similarity-based approaches [1] and generative or model-based approaches [2]. Rather interestingly, in both cases the a posteriori analysis of the clustering results are often based on both a qualitative assessment of the similarity of the clustered time series, together with speculations on the functional relationships between the clustered genes. In the case of short time series, an alternative choice could be to resort to template matching classification techniques, such that the gene expression profiles may be associated (classified) to the closest temporal profile [3]. Template matching, however, requires that templates are hypothesized or exhaustively generated on the basis of the available data set. For this reason, we resorted to a new technique which dynamically generates temporal templates corresponding to gene expression clusters. Such technique is based on temporal abstractions [4].

2. Method

The method we propose is based on the description of the time course of a variable through a set of consecutive trend temporal abstractions. In this way a numerical variable (i.e. gene expression) is represented through a set of qualitative labels like Increasing, Steady, and Decreasing.

The mechanism for Temporal Abstraction (TA) detection is based on a modification of an algorithm for piecewise linear curve approximation applied in image filtering [5]. The algorithm works as follows:

The first step of the algorithm finds a piecewise linear approximation for each initial time series, in order to consider only significant slope changes in the gene expression. This is performed through two sub-steps: first, within the initial set of points a subset of change points, called dominant points, is found and, second, a least square fitting is performed between dominant points to find a final approximating curve. In order to choose the set of dominant points, we start at the first point of the curve and consider each successive point, and then compute chord length C and arc length S. For instance, for the example in Figure 1, we can compute the chord length C as the distance between the points collected at t1 and t3, while S is the sum of the distances between points collected at t1, t2 and t3. Once S and C are computed, we evaluate $Th = \frac{\sqrt{S^2 - C^2}}{2}$ (1). When *Th* it is greater than a certain threshold, we declare the previous point as dominant; otherwise the algorithm goes on by considering the next point. The same method may be applied a second time on the set of dominant points found, in order to further eliminate some of the points retained in the first step. To find the final approximating curve, we consider couples of neighbouring dominant points and we compute a least squares first order approximation to the points on the original curve between the dominant points. In this way we obtain a piecewise linear curve as an approximation of the original one. In Figure 1a it is possible to see the

dominant points, denoted by a circle, detected by the algorithm.

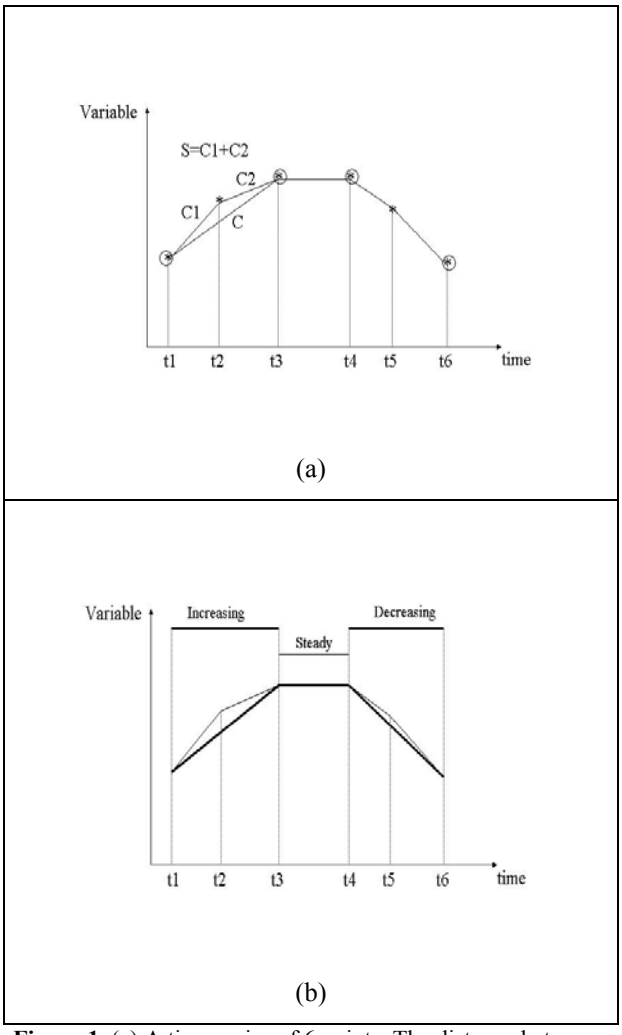

Figure 1. (a) A time series of 6 points. The distance between chord (C) and the arc (S) is used to detect change in slopes. The points for which a change is found are denoted as dominant points (circles). (b) Once the piecewise linear approximation is found the trend abstraction is easily derived through the step detection and interval aggregation.

As a second step, we consider the slope of each piece of the resulting curve and we test its statistical significance. We then associate to each piece a *Steady* TA if its slope is zero or non significant, to an *Increasing* TA if its slope is >0 and with a *Decreasing* Temporal Abstraction if the slope is <0. In this way we are able to label each curve with the sequence of labels of its pieces. We then check the labels of each time interval in order to aggregate consecutive equal elements: if two or more of adjacent elements present the same value, we collapse them in a single element. In the example of Figure 1 b, the set [Increasing Increasing Steady Decreasing Decreasing] will become [Increasing Steady Decreasing]. We can denote this set as the *abstract pattern*.

The third step is to put together temporal series into clusters or classes. A new class is created every time the abstract pattern of the gene differs from that of the ones that have already been classified; in this way, genes with the same abstract pattern are put together in one class.

Let us note that the method proposed can be considered as a template-based matching strategy, in which the templates are not pre-defined, but they are dynamically built on the basis of the abstract features of the gene expression profiles. Rather interestingly, the above described method is very efficient from a computational viewpoint, since, once the abstract pattern is extracted, the classification is readily performed. Moreover, the clusters are self-explanatory, since they are always denoted by their set of temporal abstractions.

The algorithm described above has been validated through a simulated study in which temporal profiles randomly extracted from a set of five templates (each one consisting of 5 points) and corrupted by noise have been blindly re-classified using our approach. The misclassification rate for the simulated data is shown in Table 1, in presence of a fixed threshold for the value *Th* of equation (1). The data are normalized between 0 and 1. Notice that the algorithm works very well when the level of noise is low, while presents problems in discriminating complex patterns when the level of noise increases. Rather interestingly the algorithm is still able to discriminate between increasing and decreasing trends even if the values may span for the 40% of the original signal. If the noise level is known in advance, the overall performance may be improved by tuning the threshold parameter.

Noise (std. dev)	**Misclassification Rate** (%)				
	Increasing	Decreasing	Decreasing Increasing	Increasing Decreasing	Steady Increasing
0	0	0	0	0	0
±0.01	0	0	0	0	16
±0.05	2	0	36	22	46
±0.1	32	20	68	66	54
±0.25	84	86	94	100	78

Table 1. Results on simulated data

The algorithm has been used to classify the data coming from several different gene expression studies on yeast which are publicly available at http://www.transcriptome.ens.fr/ymgv/. Our final goal is to derive a predictive model for the function of unknown genes on the basis of the information coming from multiple experiments. Figure 2 shows some results obtained on the yeast expression data from a study on stress response [6]. Notice that the TA-based clustering finds meaningful

aggregations, providing the user a clear explanation of the clustering results.

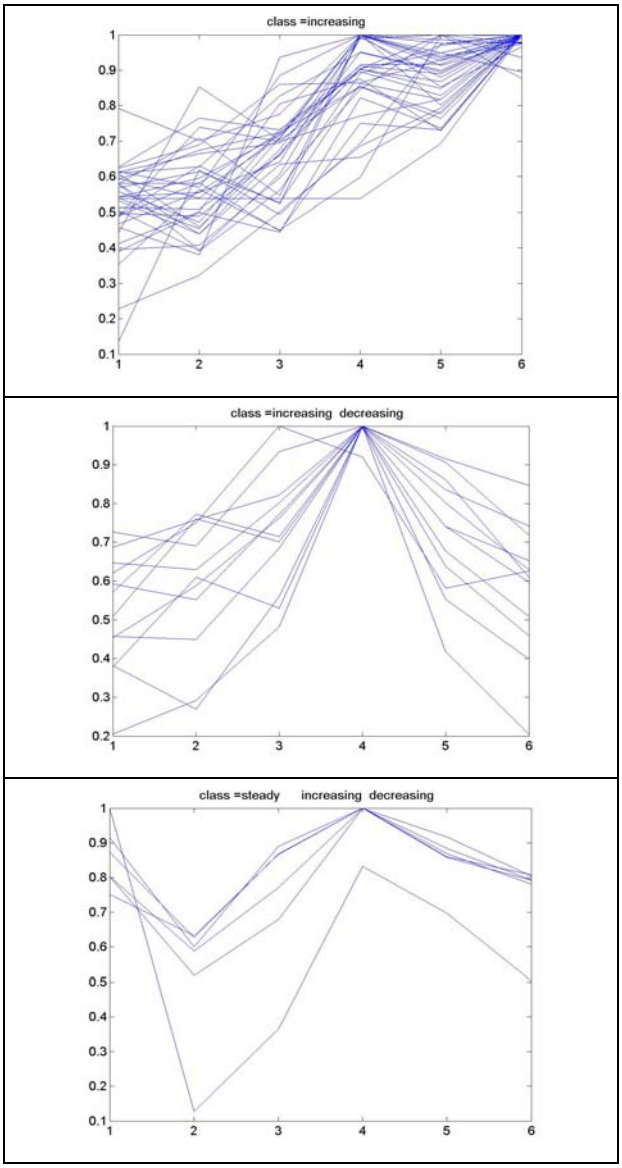

Figure 2. Three Temporal patterns extracted from the yeast data reported in [6]. The results are meaningful and easy to be interpreted.

3. Discussion and future work

The method presented in this paper allows to perform clustering of short time series. It works well in presence of few points, by extracting the abstract behaviour and classifying the series on the basis of the behaviour itself. In order to properly apply the algorithm and to further improve its performance, it is necessary to analyse its limitations. The abstract patterns may aggregate time series with strong dislocations; for example a series of the kind [Increasing Decreasing Decreasing Decreasing] has the same abstract pattern of the series [Increasing Increasing Increasing Decreasing]. This may become unacceptable as the series becomes longer. To this end, we are currently working on an algorithm for the extraction of a taxonomy of abstract patterns, in which the classes are further subdivided into subclasses; the subclasses are made taking into account the information on the synchronous occurrence of Temporal Abstractions. The Temporal Abstraction mechanism is strongly dependent on the choice of the threshold parameter *Th*. In order to derive a general strategy to define such parameter, we have extracted an experimental rule for defining *Th* as a function of the noise on the data after normalization.

A final comment must be made on the potential exploitation of the algorithm within a supervised learning strategy for assessing gene functions on the basis of their temporal information. Given the clustering strategy described above, the idea is to understand if one or more of the abstract behaviours are able to discriminate between the functional roles of the genes. To this end we are evaluating the results obtained on functionally annotated genes, and are considering to combine classifications coming from the multiple experiments.

References

[1] Eisen M, Spellman PT, Botstein D, and Brown PO, Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998 Dec 8;95(25):14863-8.

[2] Ramoni M, Sebastiani P, Cohen P, Bayesian clustering by dynamics, Machine Learning, 2002 47: 91-121.

[3] Komorowski J, Hvidsten TR, Jenssen TK,Tjeldvoll D, Hovig E, Sandvik AK, Laegreid A. Towards Knowledge Discovery from cDNA Microarray Gene Expression Data. PKDD 2000, LNAI 1910, 2000, 470-475.

[4] Shahar, Y.: A Framework for Knowledge-Based Temporal Abstraction, Art. Int. 90 (1997) 79-133.

[5] Horst JA, Beichl I. A Simple Algorithm for Efficient Piecewise Linear Approximation of Space Curves, International Conference on Image Processing (ICIP '97) 3-Vol. 2 October 26 - 29, 1997 WA, DC, 744-747.

[6] Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA. Remodeling of Yeast Genome Expression in Response to Environmental Changes. Mol Biol Cell 2001 Feb;12(2):323-337.