Fuzzy Patterns and GCS Networks to Clustering Gene Expression Data

Daniel Glez-Peña¹, Fernando Díaz², Florentino Fdez-Riverola¹, José R. Méndez¹, and Juan M. Corchado³

¹ Department of Computer Science, University of Vigo, Ourense, 32004, Spain
   {dgpena,riverola,moncho.mendez}@uvigo.es
² Department of Computer Science, University of Valladolid, Segovia, 40005, Spain
   fdiaz@infor.uva.es
³ Department of Computer Science, University of Salamanca, Salamanca, 37008, Spain
   corchado@usal.es

Summary. The advent of DNA microarray technology has supplied a large volume of data to many fields like machine learning and data mining. Gene expression profiles are composed of thousands of genes at the same time, representing complex relationships between them. In this context, intelligent support is essential for managing and interpreting this great amount of information. One of the well-known constraints specifically related to microarray data is the large number of genes in comparison with the small number of available experiments. In this situation, the ability of design methods capable of overcoming current limitations of state-of-the-art algorithms is crucial to the development of successful applications. In this chapter we present a flexible framework for the task of feature selection and classification of microarray data. Dimensionality reduction is achieved by the application of a supervised fuzzy pattern algorithm able to reduce and discretize existing gene expression profiles. An informed growing cell structures network is proposed for clustering biological homogeneous experiments starting from the previous simplified microarray data. Experimental results over different data sets containing acute myeloid leukemia profiles show the effectiveness of the proposed method.

6.1 Introduction

The advent of microarray technology has become a fundamental tool in genomic research, making it possible to investigate global gene expression in all aspects of human disease. In particular, cancer genetics based on the analysis of cancer genotypes, provides a valuable alternative to cancer diagnosis in both theory and practice [14]. In recent years, there has been an explosion of methods that analyze gene expression arrays to produce long lists of genes that express differentially in distinct cellular states. Gene expression arrays provide a great amount of valuable biological information, although it represents only a suspicion about the processes taking place within the whole cell.
Recent studies in human cancer have demonstrated that microarrays can be used to develop a new taxonomy of cancer, including major insights into the genesis, progression, prognosis, and response to therapy on the basis of gene expression profiles [26]. The automatic classification of cancer patients has been a promising approach in cancer diagnosis since the early detection and treatment can substantially improve the survival rates. For this task, a number of successful machine learning approaches have been proposed in the literature, including support vector machines (SVM) [4], artificial neural networks (ANN) [34, 21], k-nearest neighbor (k-nn) [25] and hierarchical clustering [1] methods, among others.

Since the number of examined genes in an experiment runs to the thousands, a major problem with the application of existing clustering and classification techniques is the huge number of attributes (genes) in the existing datasets. Gene reduction in microarray data is extremely important because: (i) it generally reduces the computational cost of machine learning techniques, (ii) it usually increases the accuracy of classification algorithms [6] and (iii) it provides clues to researches about genes that are important in a given context (i.e. biomarkers for certain diseases).

In this context, several methods derived from machine learning and multivariate statistical analyses have been applied to gene selection/dimension reduction in the field of microarray data. On the one hand, there are works on applying genetic algorithms [23, 7], wrapper approaches [2], support vector machines [15, 5] or spectral biclustering [24] to achieve significant reduction rates. On the other hand, the utilization of partial least squares (PLS), sliced inverse regression (SIR) or principal component analysis (PCA) have been shown highly useful for classification with gene expression data [6]. Other approaches focus their attention on redundancy reduction and feature extraction [19, 28], as well as the identification of similar gene classes making prototypes-genes [17]. One way or another, the selected method has to pursue two main goals: (i) reduce the cost and complexity of the classifier and (ii) improve the accuracy of the model.

In recent years, the number and variety of applications of fuzzy logic have increased significantly. The applications range from consumer products such as cameras, camcorders, washing machines, and microwave ovens to industrial process control, medical instrumentation, and decision-support systems. But, what is meant by fuzzy logic?. In a narrow sense, fuzzy logic represents a logical system, which is an extension of multi-valued logic. However, from a broader perspective fuzzy logic is almost synonymous with the theory of fuzzy sets, a theory which relates to classes of objects with unsharp boundaries in which membership is a matter of degree [31, 10]. Another basic concept in fuzzy logic, which plays a central role in most of its applications, is that of a fuzzy if-then rule or, simply, fuzzy rule. Although rule-based systems have a long history of use in AI, what is missing in such systems is a mechanism for dealing with fuzzy consequents and fuzzy antecedents. In fuzzy logic, this mechanism is provided by the calculus of fuzzy rules.