Chapter 27
Cancer Gene Expression Data Analysis Using Rough Based Symmetrical Clustering

Anasua Sarkar
Government College of Engineering and Leather Technology, India

Ujjwal Maulik
Jadavpur University, India

ABSTRACT

Identification of cancer subtypes is the central goal in the cancer gene expression data analysis. Modified symmetry-based clustering is an unsupervised learning technique for detecting symmetrical convex or non-convex shaped clusters. To enable fast automatic clustering of cancer tissues (samples), in this chapter, the authors propose a rough set based hybrid approach for modified symmetry-based clustering algorithm. A natural basis for analyzing gene expression data using the symmetry-based algorithm is to group together genes with similar symmetrical patterns of microarray expressions. Rough-set theory helps in faster convergence and initial automatic optimal classification, thereby solving the problem of unknown knowledge of number of clusters in gene expression measurement data. For rough-set-theoretic decision rule generation, each cluster is classified using heuristically searched optimal reducts to overcome overlapping cluster problem. The rough modified symmetry-based clustering algorithm is compared with another newly implemented rough-improved symmetry-based clustering algorithm and existing K-Means algorithm over five benchmark cancer gene expression data sets, to demonstrate its superiority in terms of validity. The statistical analyses are also performed to establish the significance of this rough modified symmetry-based clustering approach.

DOI: 10.4018/978-1-4666-2518-1.ch027

INTRODUCTION

The progress of microarray technology in the field of cancer research has enabled scientists to measure the molecular signatures of cancer cells. The scientists today monitor the expression levels for differentially expressed cancer genes simultaneously over different time points under different drug treatments (Tusher, 1940). The efficient machine learning classifiers helps in the diagnosis of cancer sub types for patients (Spang, 2003).

Clustering is one unsupervised classification method based on maximum intra-class similarity and minimum inter-class similarity. Historically Eisen et al. (Eisen, 1998) first classified groups of co-expressed genes using hierarchical clustering. Other already proposed clustering, which can be applied for cancer subtype detection are - self-organizing map (SOM) (Spang, 2003), K-Means clustering (Tavazoie, 2001), (Hoon, 2004), simulated annealing (Lukashin, 1999), graph theoretic approach (Xu, 1999), fuzzy c-means clustering (Dembele, 2003) and scattered object clustering (de Souto, 2008). Several other methods like (Maulik, 2009), (Bandyopadhyay, 2010) are also which may be applicable efficiently for cancer subtype detection problem.

While the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition, the membership function of rough sets also enables efficient handling of overlapping partitions. Therefore, recently rough set theory is being used for clustering [4,7,8,10,(Dembele, 2003)(Qin, 2003)]. Hirano and Tsumoto [7,8] proposed an indiscernibility based clustering method that can handle relative proximity. Lingras (Xu, 1999),(Dembele, 2003)(Qin, 2003)] used rough set theory to develop interval representation of clusters. This model is useful when the clusters do not necessarily have crisp boundaries.

The present study focuses on the integration of rough-set theoretic automatic optimal classification with knowledge extraction and the modified symmetry-based clustering method for analyzing cancer gene expression data sets. Clusters are associated with indiscernibility classes containing sample cancer genes that occur in precisely defined intervals or conditions. The most widely used clustering algorithms for microarray gene expression analysis are Hierarchical clustering (Eisen, 1998), K-Means clustering (Tavazoie, 2001), (Hoon, 2004) and SOM (Spang, 2003).

Among these conventional clustering methods, KMeans is an effective partitional clustering algorithm which utilizes heuristic global optimization criteria. Given a set of n points in d-dimensional space, R^d, and an integer k, the problem is to determine a set of k points $(z_1,z_2,\cdots,z_K)$ in R^d, called centroids, for k disjoint clusters $C_1,C_2,\cdots,C_K$, so as to minimize the mean squared distance norm of each data point to its nearest centroid (Jain,1988). The objective is to minimize the following crisp partitioning metric (Bandyopadhyay, 2007):

$$\sum_{i=1}^{K} \sum_{x_j \in C_i} D^2\left(x_j, z_i\right) \tag{1}$$

where $D\left(x_j, z_i\right)$ denotes Euclidean distance of pattern x_j from centroid z_i (Duda,1981). Among different versions of KMeans algorithms, Euclidean norm-based methods find only spherical shape clusters (Hoon, 2004) and the Mahalonbis norm based method finds only ellipsoidal ones. Hence we study other algorithms with different distance norms to detect clusters which are line, ring, polygonal-shaped or hyper-spherical without overlapping (Gath, 1989), (Dave, 1989), (Man, 1994).

Symmetry is considered as an inherent feature for recognition and reconstruction of shapes hidden in any clusters. In (Su, 2001), Su and Chou have proposed a new variation of KMeans algorithm that uses a new symmetry-based distance measure (SBCL). However it has been shown that it fails when inherent symmetry with respect to some intermediate point lies within any symmetrical intra-cluster. To overcome this problem,