

ABSTRACT

Gene expression is a process by which mRNA and eventually protein are synthesised from the DNA template of each gene. Recent advances in microarray technology allow scientists to measure the expression levels of thousands of genes simultaneously and determine whether those genes are active, hyperactive, or silent in normal or cancerous tissues. This technology finally produces gene expression data. Current studies on the molecular level classification of tissue have produced remarkable results and indicated that gene expression data could significantly aid in the development of an efficient cancer classification (Mohamad et al., 2005). However, classification based on the data confronts with more challenges. One of the major challenges is the overwhelming number of genes relative to the number of samples in a data set. Many of the genes are also not relevant to the classification process. Hence, the selection of genes is the key of molecular classification, and should be taken with more attention. The task of cancer classification using gene expression data is to classify tissue samples into related classes of phenotypes, e.g., cancer versus normal (Mohamad et al., 2007). A gene selection process is used to reduce the number of genes used in classification while maintaining an acceptable classification accuracy. Gene selection methods can be classified into two categories. If gene selection is carried out independently from the classification procedure, the methods belong to the filter approach. Otherwise, it is said to follow a wrapper (hybrid) approach. Most previous works have used the filter approach to select genes since it is computationally more efficient than the hybrid approach. However, the hybrid approach usually provides greater accuracy than the filter approach (Mohamad et al., 2005). The application of hybrid approaches using genetic algorithm (GA) with a classifier has grown in recent years. From the previous works, the GA performed well but only on data that have a number of features that is less than 1,000. Multi-objective optimisation (MOO) is an optimisation problem that involves multiple objectives or goals. Generally, the objectives may estimate very different aspects of solutions. Being aware that gene selection is a MOO problem in the sense of classification accuracy maximisation, and gene subset size minimisation. Therefore, this research proposes a multi-objective approach in a hybrid of GA and support vector machine classifier (GASVM) for genes selection and classification of gene expression data. It is known as MOGASVM.