Chapter 7
Wolf–Swarm Colony for Signature Gene Selection Using Weighted Objective Method

Prativa Agarwalla
Heritage Institute of Technology, Kolkata, India

Sumitra Mukhopadhyay
Institute of Radio Physics & Electronics, India

ABSTRACT

Microarray study has a huge impact on the proper detection and classification of cancer, as it analyzes the changes in expression level of genes which are strongly associated with cancer. In this chapter, a new weighted objective wolf-swarm colony optimization (WOWSC) technique is proposed for the selection of significant and informative genes from the cancer dataset. To extract the relevant genes from datasets, WOWSC utilizes four different objective functions in a weighted manner. Experimental analysis shows that the proposed methodology is very efficient in obtaining differential and biologically relevant genes which are effective for the classification of disease. The technique is able to generate a good subset of genes which offers more useful insight to the gene-disease association.

INTRODUCTION

Cancer is a heterogeneous disease which has different stages, classes and subtypes. Early prediction of subtypes and detection of advancement rate of disease can improve the mortality rate and also vital for the course of treatment. In biological terms, cancer can be defined as uncontrolled growth of certain cells due to changes in expression of genes in molecular level. For proper understanding of the disease and categorizing it into different classes, investigation of the changes in genetic expression level is necessary. Selection of relevant genes involved in tumor progression is very essential for the proper medical diagnosis as well as for drug target prediction. Gene expression data (Zhang, Kuljis & Liu, 2008) has a huge impact on the study of cancer classification and identification. It includes the expression levels of thousands of genes, collected from various samples. The expression of a gene in a carcinogenic cell is

DOI: 10.4018/978-1-5225-5852-1.ch007

compared with the expression in normal cell and then through proper analysis microarray gene expression dataset is formed. Proper analysis of the dataset is required as it contains the information regarding the abnormal behavior of a disease gene. But, the high dimensionality of gene microarray datasets makes it challenging to examine and extracting important feature genes from it. Again, the availability of larger number of genes compared to the small number of samples can cause the overfitting issue for classification of samples. Also, the presence of noise and the heterogeneous nature of dataset cause problem in the task of informative feature extraction. It motivates the researchers to apply various statistical and learning based techniques for realizing the useful information content of the dataset. The importance of classifying cancer and appropriate diagnosis of advancement of the disease using those feature genes has led to many research fields, from biomedical to the application of machine learning (ML) methods. The ability of machine learning approaches to detect key features from a huge complex dataset reveals their importance in the field of feature selection from datasets as well as the ability to examine big data framework. So, the modelling of cancer progression and classification of disease by investigating large microarray datasets can be studied by employing learning-based approaches.

Researchers have summarized the microarray data with various statistical approaches (Yang, Parrish & Brock, 2014; Pal, Ray, Cho & Pal, 2016; Arevalillo & Navarro, 2013; Peng, Long & Ding, C., 2005). Those provide fast and scalable output but overlook the feature dependencies of the datasets. Different classifier dependent stochastic bio-inspired algorithms-based learning methodologies are introduced to handle the problem. Recently, numerous hybrid approaches (Hsieh & Lu, 2011; Apolloni, Leguizamón & Alba, 2016) are proving to be very effective where both the statistical filters and the classifiers are implemented along with the optimization algorithms. The use of a particular statistical or classifier dependent objective functions are sometimes not enough for the job of finding out biologically relevant and cancer class identifier gene selection. So, multiple measurement metrics can be utilized for evaluation purpose. Different multi-objective methodologies are proposed in the literature (Mukhopadhyay & Mandal, 2014; Zheng, Yang, Chong & Xia, 2016; Mohamad, Omatu, Deris, Misman & Yoshioka, 2009) where nature inspired evolutionary and swarm algorithms are investigated to obtain a pareto-optimal solution for gene set. Generally, those approaches involve two objectives for the relevant gene selection from microarray dataset and in most of the cases the methodologies are based on a particular type of bio-inspired algorithm. So, it has the limit to produce more promising results compared to any hybrid swarm algorithm approaches. For extracting the most informative genes from the huge gene expression dataset, different measurement indices are to be optimized. Examining two or three objectives may not be sufficient for genuine selection of differentially expressed genes from the cancer data, as the choice of objective function plays an important role in this case. The optimization function should have the property to select the differentially expressed features for different subtype of a disease. Also, it needs to have the ability to detect the proper class of the disease. So, to produce more accurate results for the problem of gene selection from the huge microarray data, authors have involved multiple objectives at a time. As multiple objectives are to be optimized, a weighted fitness function is formed using weighted-objective optimization technique. Weighted factors are involved in mapping those multiple objectives into a one fitness objective to select the required features which are efficient in the classification purpose as well as having differential expressions from disease to disease.

Population based meta-heuristic searching algorithms work efficiently in the course of optimization. One of the popular tools is particle swarm optimization (PSO) (Kennedy, 2011), which is well known for its simplicity. Generally, it is observed that PSO suffers from two issues. It easily loses its diversity. Again, it sticks to local optima having a pre-matured and inaccurate convergence of the employed swarm.