International Journal of Swarm Intelligence Research
Volume 11 • Issue 1 • January-March 2020

Gene Expression Dataset Classification Using Artificial Neural Network and Clustering-Based Feature Selection

Audu Musa Mabu, SHUATS, Uttar Pradesh, India
Rajesh Prasad, African University of Science and Technology, Abuja, Nigeria
Raghav Yadav, SHUATS, Uttar Pradesh, India

ABSTRACT

With the progression of bioinformatics, applications of GE profiles on cancer diagnosis along with classification have become an intriguing subject in the bioinformatics field. It holds numerous genes with few samples that make it arduous to examine and process. A novel strategy aimed at the classification of GE dataset as well as clustering-centered feature selection is proposed in the paper. The proposed technique first preprocesses the dataset using normalization, and later, feature selection was accomplished with the assistance of feature clustering support vector machine (FCSVM). It has two phases, gene clustering and gene representation. To make the chose top-positioned features worthy for classification, feature reduction is performed by utilizing SVM-recursive feature elimination (SVM-RFE) algorithm. Finally, the feature-reduced data set was classified using artificial neural network (ANN) classifier. When compared with some recent swarm intelligence feature reduction approach, FCSVM-ANN showed an elegant performance.

KEYWORDS

Artificial Bee Colony, Artificial Neural Network, Clustering, Gene Expression Profile, Particle Swarm Optimization, Support Vector Machine, SVM-Recursive Feature Elimination, Swarm Intelligence

1. INTRODUCTION

A huge quantity of data generation driven the progression of numerous complex strategies and tools aimed at visualization and scrutiny of information. These tremendous measures of data, especially aimed at the biological examination along with explanations, are made accessible by microarray technology Kohbalan, *et al.,* (2013). The microarray technology advent has profited research workers in directing extensive experiments on chiliads of genes via scrutinizing the difference of communications amongst genes Muhammad, (2017). Actually, just few genes are exceptionally connected to a similar example classes. Those genes are alluded to as the information gene. These enclose the samples' classification information Jiang, Xie, *et al.,* (2013). Numerous cases have been established that extensive observing of GE through microarrays is the utmost propitious strategies to

DOI: 10.4018/IJSIR.2020010104

enhance medicinal diagnostics in addition to functional genomics studies Muhammad, (2017). In the uprightness of gene microarray examination, precise categorization of tumor subtypes might progress toward becoming reality, taking into consideration particular treatment that amplifies efficacy, further, limits toxicity Liu, *et al.,* (2007).

Microarray technologies as of late have initiated numerous chances to explore cancer utilizing gene expressions. The essential onus of a microarray data analysis stands to decide a computational model as of specified microarray data which foresee the type of the specified unidentified examples. The accuracy, value, and also strength are imperative components of microarray analysis Hala, *et al.,* (2014). The tumor diagnosis along with classification of GE data stands as a two interesting topics recently. As it may be, GE data contains a chiliads of genes with few samples that makes it tough to examine and process. In addition, it is linearly indivisible, noisy besides being imbalanced Huijuan, *et al.,* (2017). In the preceding decade, a few endeavors are dutiful to the improvement of classification techniques for higher-dimensional GE data started by means of microarray experiments Carlotta and Carlo, (2013). It is obvious that K-means is the most popular clustering algorithm, but can only generate local optimal solution. Swarm optimization clustering algorithms are more advantageous as they perform a globalized search over entire search space. A PSO+K-means algorithm has the ability to search globally, thereby enhancing fast convergence than using conventional K-means algorithm alone. It is promising to generate multi-objective PSO based K-means clustering algorithm that has the ability to cluster both genes and samples simultaneously for GE data Cui and Potok, (2005). The categorization of diverse tumor sorts in GE data is of extraordinary significance in cancer analysis besides drug discovery. Nevertheless, it is intricate attributable to its enormous size. There are many of techniques attainable to assess gene expression profiles. A general trait for these means is picking a subset of genes which is extremely instructive aimed at classification process furthermore to decrease the dimensionality issue of profiles Udhaya, *et al.,* (2014). Dimensionality reduction is especially applicable in bio-informatics research, especially with regards to microarray data, described by moderately little samples in a high-dimensional gene (feature) spaces. Unrelated genes (features) prompt deficient classification accuracy and furthermore include additional troubles in discovering possibly valuable information Amit, *et al.,* (2014).

Microarray technology, designed to screen a chiliads of GE patterns concurrently intended for recognition of ailment development. Managing highly-dimensional data in feature spaces which debases classification proficiency is the troublesome task in examining microarray data Jacophine, *et al.,* (2016). The primary cause is that DNA microarray dataset encloses 10,000 genes together with few experimental samples or remarks concerning cancer. This implies chiliads of genes are insignificant or noisy or else redundant aimed at specific gene assortment along with classification algorithms Hernández-Montiel, *et al.,* (2014). This bane of dimensionality, a noteworthy impediment in machine learning in addition to data mining henceforth feature selection in addition to dimensionality reduction dependably is a dynamic research subject in bioinformatics Jun, *et al.,* (2015). As the commencement of higher throughput systems, as microarrays, the heap of genomic data expanded geometrically, alongside that the requirement aimed at computational techniques that will comprehend those data George, *et al.,* (2016). Discoveries in gene analyses accounts for various symptoms thereby results in discoveries of molecular mechanisms with which proper understanding of genes and gene expressions Sreeja and Vinayan, (2017).

DNA microarray technology that gauges the expression stages in numerous genes concurrently of biological tissues area then creates cancer databases in light of GE data has astonishing potentiality on the research of cancer. Since the customary diagnosis technique aimed at malignancy is imprecise, GE data is generally utilized to recognize cancer biomarkers intently connected with cancer Shuaiqun, *et al.,* (2016). Likewise, an ever rising number of statistical strategies was produced and implement to the malady classification utilizing microarray GE data Pingzhao, *et al.,* (2011). The microarray data technology utilization permits observing simultaneously chiliads of GE levels over numerous cases. GE data assumes a crucial part as a biomarker aiding in an assortment of work, for example, cancer