International Journal of Swarm Intelligence Research
Volume 14 • Issue 1

Breast Cancer Classification With Microarray Gene Expression Data Based on Improved Whale Optimization Algorithm

S. Sathiya Devi, University College of Engineering, Bharathidasan Institute of Technology, Trichy, India*
Prithiviraj K., University College of Engineering, Bharathidasan Institute of Technology, Trichy, India

ABSTRACT

Breast cancer is one of the most common and dangerous cancer types in women worldwide. Since it is generally a genetic disease, microarray technology-based cancer prediction is technically significant among lot of diagnosis methods. The microarray gene expression data contains fewer samples with many redundant and noisy genes. It leads to inaccurate diagnose and low prediction accuracy. To overcome these difficulties, this paper proposes an Improved Whale Optimization Algorithm (IWOA) for wrapper based feature selection in gene expression data. The proposed IWOA incorporates modified cross over and mutation operations to enhance the exploration and exploitation of classical WOA. The proposed IWOA adapts multiobjective fitness function, which simultaneously balance between minimization of error rate and feature selection. The experimental analysis demonstrated that, the proposed IWOA with Gradient Boost Classifier (GBC) achieves high classification accuracy of 97.7% with minimum subset of features and also converges quickly for the breast cancer dataset.

KEYWORDS

Accuracy, Crossover and Mutation, Feature Selection, Gradient Boosting Classifier, Multi Objective Optimization, Support Vector Machine, Whale Optimization Algorithm

1. INTRODUCTION

Cancer is a second dangerous disease that causes 9.6 million deaths worldwide. There are approximately 21. 7 Million people in the world is suffering from cancer by 2030 and predicted 30 million deaths (Aldryan et al., 2018). There are different types of cancer among which breast cancer is the common (prevalent) among females. Nearly one fourth of female population is affected by this cancer irrespective of age factor in India and is common in rural India. The majority of the cancer types can be caused due to either genetic (hereditary) or epigenetic changes and generally 90% of the breast cancer is due to genetic abnormalities. The variations in high penetrance genes such as BRCA1, BRCA2, p53, PTEN, ATM, NBS1, LKB1, etc. can produce genetic abnormalities (Dumitrescu

DOI: 10.4018/IJSIR.317091 *Corresponding Author

& Cotarla, 2005). The common symptoms of breast cancer are (i). a lump in the breast, (ii) blood discharge from the nipple and (iii). shape or texture changes in the nipple or breast. Since the breast cancer cannot be accurately diagnosed by a single clinical test, it requires many tests along with the complete history of the patient and their physical examination. Based on these results, the physician could (i). Identify and confirm the disease, (ii). Consistently monitoring the disease progress and (iii). Schedule for and assess the viability of the treatment. The above mentioned classical diagnosis methods result in uncertain diagnosis and prone to human error. It also requires skilled labors and is time consuming, which causes stress throughout the diagnostic process. Hence the early detection of cancer to reduce the risk of death requires an accurate and reliable diagnosis processes as well as the use of robust tools and techniques.

DNA Microarray technology based cancer prediction is technically significant among lot of diagnosis methods and is used by researchers and clinicians for the past two decades. The recent technologies made the availability of thousands of benchmark gene expression assays through online for microarray data analysis to predict different types of cancerous tumors. The microarray dataset consists of huge number of genes corresponding to small sample size and the genes are highly correlated. The high dimensionality of the genes and small sample size is a challenge for the effective analysis and diagnosis of microarray data resulting in poor diagnosis and prediction accuracy. Hence this paper addresses this issue by proposing an Improved Whale Optimization Algorithm (IWOA) for feature selection and ensemble based classification for breast cancer prediction.

The remaining portion of this paper is organized as follows: section 2 reviews the relevant and significant previous work. The classical Whale Optimization Algorithm (WOA) is described in section 3. The proposed IWOA based feature selection with Support Vector Machine (SVM) and GBC is described in section 4. Section 54 discusses the experimental and result analysis with data set and performance measure. Conclusion is presented in section 6.

2. RELATED WORK

In the microarray dataset, the high ratio between the huge dimension of the genes (features) and the few number of samples resulted in inaccurate and imbalanced cancer prediction. In common, most of the genes in the microarray data are uninformative and redundant. These types of the genes are to be identified with the machine learning technique called as feature subset selection. Though the feature selection techniques not only identify the significant genes, it also improves the classification accuracy. There are three approaches for feature selection: (i) Filter method, (ii) Wrapper method and (iii) Hybrid method. In the filter method, the feature importance is measured with properties of the dataset and order the features based on the relevance score (feature importance score). This method is simple and fast and not considering the correlation among the features. The wrapper method generally incorporates any predefined classification algorithm to search for and select the relevance features. This method considers the feature dependencies and computationally intensive and slower. The hybrid method is the combination of filter and wrapper methods. This method, first apply the filter technique to reduce the feature space then use the wrapper method for feature subset selection. Since the wrapper method is expensive, it is proved to be beneficial in finding feature subsets that suit a predetermined classifier (Alshamlan et al., 2015).

A. K. Shukla et. al. (2019) have introduced a hybrid wrapper approach called TLBOSA, which is the combination of Teaching Learning based Optimization (TLBO) and Simulated Annealing (SA) with SVM for gene expression data. It overcomes the exploitation issue and produces better classification accuracy with small subset of genes. To identify the more discriminative subset of genes and to reduce the dimensionality, the Gravitational Search Algorithm (GSA) is combined with TLBO called TLBOGSA has been described in (Shukla et al., 2020). This method achieves higher classification accuracy with less computational cost when compared with six datasets. P. Gunasekhar et. al. (2020) have used six different filter based approaches for biomarker feature selection. From the