Journal of Technology Management and Business (ISSN: 2289-7224)
Vol 5, No 1, Special Edition, 2018
Managing Epidemic and Natural Disaster using Mathematical Modelling

HYBRID FLOWER POLLINATION ALGORITHM AND SUPPORT VECTOR MACHINE FOR BREAST CANCER CLASSIFICATION

Muhammad Nasiru Dankolo[1], Nor Haizan Mohamed Radzi[2] Roselina Salehuddin[3] Noorfa

Haszlinna Mustaffa[4]

[1,] Department of Computer Science, Universiti Teknology Malaysiya, Johor Bahru, Malaysia

*Corresponding E-mail: nasirdankolo@gmail.com[1]

Abstract

Microarray technology is a system that enable experts to examine gene profile at molecular level for early disease detection. Machine learning algorithms such as classification are used in detection of dieses from data generated by microarray. It increases the potentials of classification and diagnosis of many diseases such as cancer at gene expression level. Though, numerous difficulties may affect the performance of machine learning algorithms which includes vast number of genes features comprised in the original data. Many of these features may be unrelated to the intended analysis. Therefore, feature selection is necessary to be performed in the data preprocessing. Many feature selection algorithms are developed and applied on microarray which including the metaheuristic optimization algorithms. This paper proposed a new technique for feature selection and classification of breast cancer based on Flower Pollination algorithm (FPA) and Support Vector machine (SVM) using microarray data. The result for this research reveals that FPA-SVM is promising by outperforming the state of the earth Particle Swam Optimization algorithm with 80.11% accuracy.

Keywords: *Microarray, Feature Selection, Classification, High Dimensionality*
DOI: https://doi.org/10.30880/jtmb.2018.05.01.006

Received: January 16, 2018 Accepted: January 25, 2018 Published: February 15, 2018

1.0 Introduction

Microarray technology is an influential innovation which can be used for disease detection in bioinformatics particularly in cancer detection and diagnosis (Canul-Reich et al. 2008). The categorisation of the gene expression data is now becomes a central focus to many of researchers in machine learning for bioinformatics data (Tabakhi et al. 2015). Using different gene expression forms with normal expression profile, irregularity could be recognized and treated before it develops abnormalities in the patient (Yang et al. 2008). The major problem in managing microarray data is the size of its dimension and small sample size (Hira et al. 2015). The feature size of microarray data is very vast, which mostly due to the incidence of noisy or unsuitable features that are recorded during the observation, therefore, learning algorithm's performance will significantly be affected if they are to learn on the whole datasets. To address the effect of irrelevant features (genes), feature selection methods are employed. Feature selection works by finding optimal subset of features that can best represent the original features without degradation in performance (Hira et al. 2015). Different kind of feature selection algorithms were proposed to scale down the dimension of the features generated by the microarray including the metaheuristic based search algorithms (Diao and Shen 2015). In this paper, we are going to study different types of metaheuristics algorithms for feature selection applied on microarray data.

2.0 Literature Review

Feature selection has been proven to be effective and efficient in preparing high-dimensional data for data mining and machine learning problem (Y. Wang and Chaib-draa 2016). The objective of this process is to identify and remove irrelevant features from the training dataset. Thus, it increases the performance of prediction algorithms by providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. It attempt to obtain an optimal feature sets from a problem domain while keeping an appropriately high accuracy in representing the original features (Yang et al. 2008). Feature selection process consist of two main components: 1) searching procedure that searches the solution vector, and 2) the evaluation of the searched features. Search methods approach that are widely used in the algorithm include complete, heuristic features (Yang et al. 2008), and random (Gütlein et al. 2009) search. Whilst, the techniques used for feature evaluation are categorized into two either classifiers specific or classifier independent (Gütlein et al. 2009). The classifiers specific requires a learning technique that will be employed to evaluate the quality of feature selection based on the classifier accuracy (Shi et al. 2016). However, the classifier independent theories a classifier independent procedures for evaluating the features importance, this kind of measures include mutual information gain, dependence measure and consistency measure (Wang et al. 2016).

Regardless the approaches used in the algorithm, they are all aimed for searching for the optimal sub-feature (Abdel-Fattah et al. 2016). To this level, exhaustive approaches are applied in different areas, however in bioinformatics, where mostly machine learning algorithms are applied to discover patterns from a very high dimensional data especially the microarray data, the application of exhaustive search strategy in feature selection is often impractical. Another alternative considered by many researchers is applying hill-climbing based approaches. Hill climbing considers adding or removing features one at a time as long as it will lead to improvement of the current solution vector. However, despite Hill-climbing approaches converge at a good solution faster, the challenge to these approaches is the possibilities of convergence at a non-optimal subset and time consumption.

To overcome sub-optimal convergence, the concept of meta-heuristics search algorithms are applied to search for optimal features (genes) that can best represent the entire original features from microarray datasets (Shrivastava et al. 2017). Most of these search systems are nature inspired such as genetic algorithms, genetic programming, simulated annealing, and particle swarm optimisation are exploited which presents different level of fitness. Majority of the metaheuristic search strategies are normally anticipated to handle functional optimisation tasks. To apply these technique in feature selection task, a mapping concepts into the field of feature selection is required.

Generally, metaheuristic algorithm is a higher-level procedure that seeks and generates a sufficiently good solution to optimisation problems. Metaheuristics algorithms generate solution for optimisation problem based on stochastic approach. Thus, metaheuristic systems often are used with less computational cost (Dhaenens 2010). These properties make them potential to solve feature selection problems.

Many meta-heuristics algorithms from the evolutionary and swarm intelligent category were applied in the literature to solve feature selection, however, there is still a room of improvement that need to be made. This is due to the complexity and the nature of the sample collected from gene analysis using microarray and the requirement for finding the optimal solution with minimum computational cost (Dhaenens 2010).

A lot of work on feature selection using PSO and its variants such as Discrete PSO, Binary PSO and Competitive PSO have been done and applied on microarray dataset. (Rathasamuth et al, 2016) proposed a new discrete particle swam optimization for feature selection in binary classification problem. They modified discrete particle swarm optimization (PSO) algorithm for the feature selection task. Their method expresses an adaptive feature selection procedure that can dynamically accounts for the importance and dependence of the features included in the selected feature set.

In 2015, Sina et al propose a new unsupervised gene selection algorithm based on the Ant colony optimization algorithm and the filter approach of feature selection (Tabakhi et al. 2015). The algorithm was developed to minimize the redundancy between genes and maximizing the relevance of genes. In the algorithm, they incorporate a new fitness function that will evaluate the quality of the selected genes (features) which does not depend on any learning algorithm, this makes their algorithm to be part of filter approach. They evaluate the performance of their algorithm by extensively testing it with five publicly available microarray datasets.

Artificial Bee colony (ABC) has been applied for feature selection methods for different problems and data sets including microarray data. In 2016, Hala et al, proposed a new hybrid ABC-SVM feature selection algorithm for multiclass classification task using microarray dataset (Alshamlan et al.2016). To develop the new hybrid algorithm, they made some modifications to the ABC algorithm representation so that it can be used to solve the microarray gene selection problem to measure the classification accuracy for selected genes. They evaluate the performance.

In order to evaluate the quality of feature selection techniques, classification algorithms are applied to learned over the features selected and assess the performance of the classifiers. Improvement in performance of the classification algorithm will indicate if the features selected are truly optimal. Data mining techniques and methods have been used by researchers to classify diseases from the data generated my microarray. K-Nearest Neighbor and Support Vector Machine are the most widely used in classification of high dimensional data.

Flower pollination algorithm (FPA) is inspired from the rules of reproduction process of flowering plants by Yang in 2013 as shown in Fig. 1 below. The flower pollination algorithm (FPA) is mainly for optimization, it is applied in feature selection and other optimization tasks to reduce the dimensionality of the search space. In feature selection task, we need a fitness function that will test the fitness of the return features by the FPA. This fitness function may be classifier dependent like K-Nearest Neighbor (KNN), Support Vector Machine (SVM) that use the classifying accuracy of a feature to decides it relevance or classifier independent like entropy that decides the feature important by calculating the amount of information gained before and after splitting the classes using information theory.

Algorithm (or simply Flower Algorithm)

Objective min or max f(x), x = (x1,x2, ...,xd)
Initialize a population of n flowers/pollen gametes with random solutions
Find the best solution g_* in the initial population
Define a switch probability p ∈ [0, 1]
while (t <Max_Itration)
for i = 1 : n (all n flowers in the population)
if rand < p,
Draw a (d-dimensional) step vector L which obeys a Lévy distribution

Global pollination via $x_i^{t+1} = x_i^t + L\left(x_i^t - g\right)$

else
Draw ϱ from a uniform distribution in [0,1]
Randomly choose j and k among all the solutions

Do local pollination via $L \sim \frac{\lambda\Gamma(\lambda)sin(\pi\lambda/2)}{\pi}\frac{1}{s^{1+\lambda}}, \quad (s >> s_0 > 0)$

end if
Evaluate new solutions If new solutions are better, update them in the population
end for
Find the current best solution g_*
end while

Figure 1: Flower Pollination Algorithm (FPA)

Where $\mathbf{x}_i^{t+1}$ is the i[th] gamete during the t[th] iteration, *e* is the probability function for switching from global to local pollination drawn from normal probability distribution p[0,1] and x_j and x_k are any random flowers

In this research, we propose a new method of feature selection and classification of breast cancer using microarray dataset based on Hybrid flower pollination algorithm (FPA) and support vector machine (SVM).

3.0 Methodology

In this research we proposed a new high-level hybrid method of feature selection and classification of breast cancer using microarray data. The dataset was collected from the National Center for Biotechnology Information (NCBI) which is a well-known public data repository for high dimensional genome dataset. We used breast cancer stroma dataset from the NCBI website that consist of 22,283 features and 47 instances out of which 13 instances are cancerous while 34 instances are non-cancerous.

We apply flower pollination algorithm (FPA) to perform feature optimization on this high dimensional dataset. The result from the FPA algorithm (number of selected features) is then pass to the Support Vector Machine (SVM) for classification of cancer. The performance of the entire algorithm is validated using k-fold cross validation with k=10. The Fig. 2 below show the graphical representation of this model.

On the other hand, we implement PSO-SVM to compare the performance of our model with. The PSO-SVM is implemented using the same specification and environment with FPA-SVM, this will enable us to compare the performance of the two algorithms.

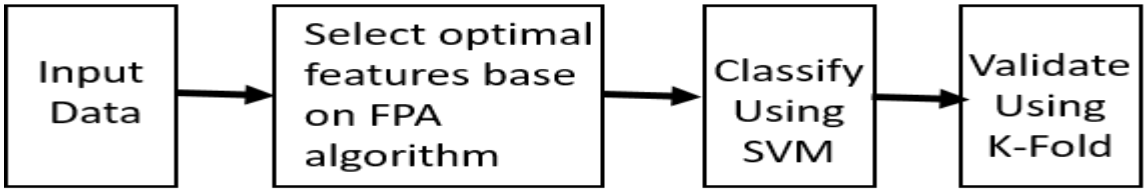

Figure 2: FPA-SVM Model.

4.0 Results and Discussions

The following figures illustrates the results we obtained after running our experiment.

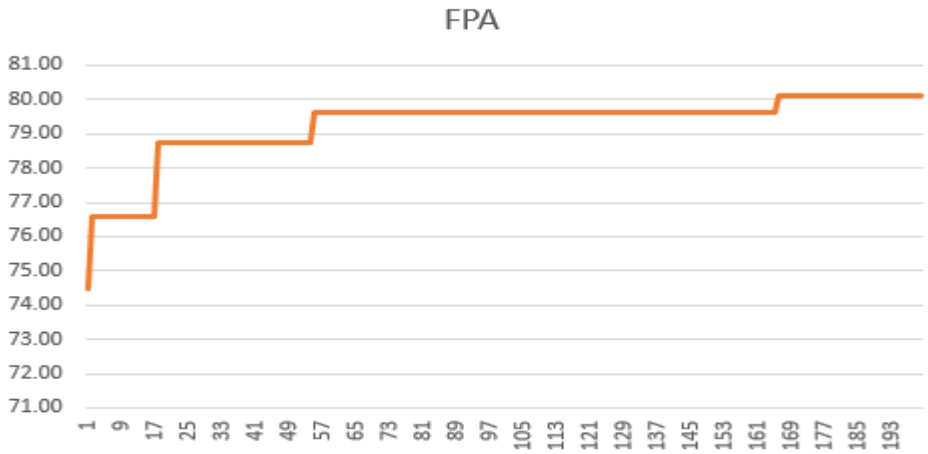

Figure 3: FPA-SVM Result.

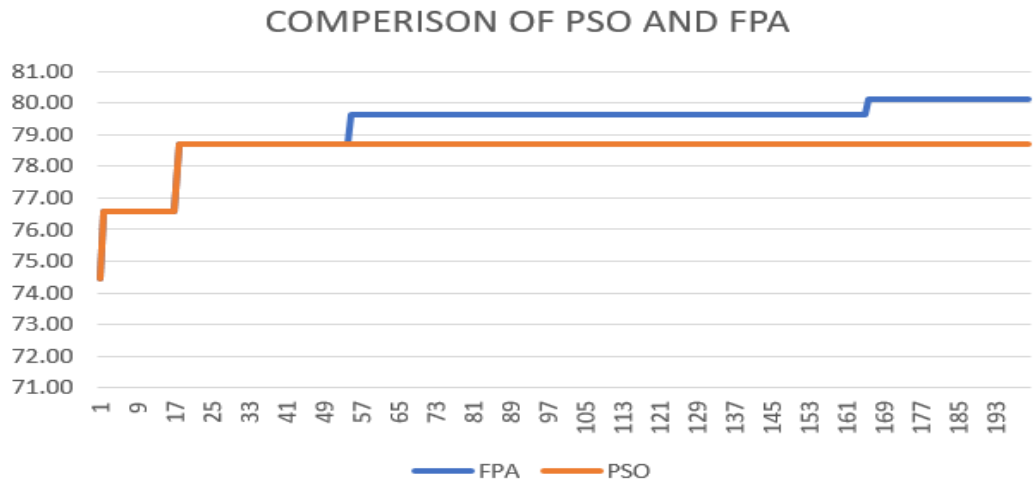

Figure 4: FPA-SVM and PSO-SVM comparison.

The Fig. 4 show the accuracy obtained using FPA-SVM algorithm in 200 iterations. The algorithm obtained an accuracy of 80.11% using the breast stroma dataset.

In Fig. 4 however, it is a comparison between the FPA-SVM and PSO-SVM. The result clearly shows that PSO is trapped at local minima which is the reason for early convergence at suboptimal solution. The result shows that PSO-SVM can attain the maximum accuracy of 78.4% using the breast stroma dataset while FPA-SVM attained an accuracy of 80.11%

5.0 Conclusions

This research has reveals the potentials of flower pollination algorithm for high dimensional feature selection, the performance of FPA-SVM for feature selection and classification of breast cancer using high dimensional microarray dataset is promising by outperforming PSO-SVM with significance improvement in accuracy.

References

Abdel-Fattah Sayed, Safinaz, Emad Nabil, and Amr Badr. 2016. "A Binary Clonal Flower Pollination Algorithm for Feature Selection." *Pattern Recognition Letters* 77: 21–27. http://dx.doi.org/10.1016/j.patrec.2016.03.014.

Alshamlan, Hala M, Ghada H Badr, and Yousef A Alohali. 2016. "ABC-SVM : Artificial Bee Colony and SVM Method for Microarray Gene Selection and Multi Class Cancer Classification." 6(3): 184–90.

Canul-Reich, Juana, Lawrence O. Hall, Dmitry Goldgof, and Steven A. Eschrich. 2008. "Feature Selection for Microarray Data by AUC Analysis." *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*: 768–73.

Dhaenens, Clarisse. 2010. "Metaheuristics for Bioinformatics." : 1–90.

Diao, Ren, and Qiang Shen. 2015. "Nature Inspired Feature Selection Meta-Heuristics." *Artificial Intelligence Review* 44(3): 311–40. http://dx.doi.org/10.1007/s10462-015-9428-8.

Gütlein, Martin, Eibe Frank, Mark Hall, and Andreas Karwath. 2009. "Large-Scale Attribute Selection Using Wrappers." *2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009 - Proceedings*: 332–39.

Hira, Zena M., Duncan F. Gillies, Zena M. Hira, and Duncan F. Gillies. 2015. "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data." *Advances in Bioinformatics* 2015(1): 1–13. http://www.hindawi.com/journals/abi/2015/198363/.

Rathasamuth, Wanthanee, and Supakit Nootyaskool. 2016. "Comparison Solving Discrete Space on Flower Pollination Algorithm, PSO and GA." *2016 8th International Conference on Knowledge and Smart Technology, KST 2016*: 18–21.

Shi, Bingbing et al. 2016. "Recent Advances on the Encoding and Selection Methods of DNA-Encoded Chemical Library." *Bioorganic & Medicinal Chemistry Letters* 27(3): 361–69. http://www.sciencedirect.com/science/article/pii/S0960894X16312926?dgcid=raven_sd_aip_email.

Shrivastava, Prashant et al. 2017. "A Survey of Nature-Inspired Algorithms for Feature Selection to Identify Parkinson's Disease." *Computer Methods and Programs in Biomedicine* 139: 171–79. http://dx.doi.org/10.1016/j.cmpb.2016.07.029.

Tabakhi, Sina, Ali Najafi, Reza Ranjbar, and Parham Moradi. 2015. "Gene Selection for Microarray Data Classification Using a Novel Ant Colony Optimization." *Neurocomputing* 168: 1024–36.

Wang, Lipo, Yaoli Wang, and Qing Chang. 2016. "Feature Selection Methods for Big Data Bioinformatics: A Survey from the Search Perspective." *Methods* 111: 21–31. http://dx.doi.org/10.1016/j.ymeth.2016.08.014.

Wang, Yali, and Brahim Chaib-draa. 2016. "KNN-Based Kalman Filter: An Efficient and Non-Stationary Method for Gaussian Process Regression." *Knowledge-Based Systems* 114: 148–55. http://dx.doi.org/10.1016/j.knosys.2016.10.002.

Yang, Cheng-San, Li-Yeh Chuang, Chang-Hsuan Ho, and Cheng-Hong Yang. 2008. "Microarray Data Feature Selection Using Hybrid GA-IBPSO." *Trends in Intelligent Systems and Computer Engineering* 6: 243–53. http://dx.doi.org/10.1007/978-0-387-74935-8_18.