Journal of Fundamental and Applied Sciences

Research Article

ISSN 1112-9867

Special Issue

Available online at http://www.jfas.info

DIFFERENTIAL EVOLUTION ALGORITHM AS FEATURE SELECTION FOR BIOMARKER DISCOVERY

S. A. M. Yusoff[1,*], S. M. Zahari[2], F. H. Mustafa[1], N. A. Rahman[3] and M. H. Abdullah[4]

[1]Faculty of Computer and Mathematical Sciences, UiTM Pulau Pinang, Malaysia

[2]Faculty of Computer and Mathematical Sciences, UiTM Selangor, Malaysia

[3]Faculty of Computer and Mathematical Sciences, UiTM Kelantan, Malaysia

[4]Faculty of Electrical Engineering, UiTM Pulau Pinang, Malaysia

Published online: 30 May 2018

ABSTRACT

The advancement in mass spectrometry technique for proteomic studies has proliferated the discovery of biomarkers from quantitative proteomics pattern. High-throughput data for a given molecule can give rise to a series of inter-related and overlapping peaks in a mass spectrum. The spectrum suffers from high dimensionality data relative to small sample size. Feature selection techniques search parsimonious features through a learning model that exhibits the most accurate results. A computational technique that mimics survival and natural processing known as DE integrated with linear SVM classifier was proposed for feature selection. The comparisons have been made with PSO and ACO algorithms. The proposed feature selection of DE algorithm exhibited accuracy, sensitivity and specificity with 82.2, 80.0 and 84.0% on liver (HCC) datasets respectively and outperformed the PSO and ACO.

Keywords: differential evolution; feature selection; biomarker discovery; classification; bio-inspired.

Author Correspondence, e-mail: syarifah.adilah@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v10i2s.77

1. INTRODUCTION

Recently, protein markers have shown great opportunity in diagnosis and prognosis of diseases. A study done by [1] has shown the rising trend in the proteomics cancer biomarker for the past 10 years. Proteomics biomarkers are based on the idea that the major workhorse of biological system, diseases and other malfunctions may be reflected by the proteomic level. Disturbances in proteome are caused by mutation such as faulty post-translation modification, interference in protein-protein interaction, deleterious effects on pathways and networks and unnatural changes in protein expression.

The tandem mass spectrometry (MS/MS) is mainly used to produce structural information about a compound by fragmenting specific sample ions inside the mass spectrometer and identifying the resulting fragment ions [2]. This information can then be pieced together to generate structural information regarding the intact molecule. Tandem mass spectrometry also enables specific compounds to be detected in complex mixtures on account of their specific and characteristic fragmentation patterns. The advancement in tandem mass spectrometry (MS/MS) which produces high-resolution spectra embarks the in depth study of biomarkers through proteome profiling [3]. Through tandem mass spectrometry, high number of peaks are generated from a single spectrum of sample that represent peptides. Furthermore, they are much better reproducibility between and within machine runs [4], thereby produce predicted model with higher sensitivity and accuracy. Moreover, the spectral resolution from low-level resolution or single fragmentation which produced only parents ions, have no ability to produce specific ions that are close in mass/charge, which can cause multiple specific discreet ions to coalesce into a single peak [5]. High-resolution mass spectrometry analysis remains to be seen as potential method for future clinical diagnostic platform. Anyhow, this high-resolution mass spectrometry generates extremely high-dimensional data of mass spectra [3] consist of tens of thousands points of mass to charge ratio (m/z) of the substance. Each point might depict particular feature of protein or peptide. This huge number of features are relative to small number of samples.

In machine learning research, feature selection analysis and a classifier would accurately distinguish cancer and normal cases from entire thousands of features in spectra, but the classification model does not help in finding specific biomarkers. Therefore, small set of

peaks of mass spectrometry data are used to computationally predict markers with high accuracy [6] and then are considered as panel of biomarkers. On the other hand, the feature selection method for biomarkers discovery are still open for improvement in terms of better accuracy for prediction [7]. Furthermore, the challenge of biomarkers discovery also relies on robustness of the method that should be able to identify markers from different types of dataset. Hence, it is motivating to study on feature selection that is reliable in finding small set of marker over different types of cancer cases.

The Bio-inspired algorithms are any met heuristic algorithms, which are constructed and inspired by nature. The nature phenomenon always finds optimal strategy and addressing interaction among organism ranging from microorganism to fully fledged human being [8]. The algorithms are categorized further into swarm-based, evolution-based and ecology-based [9]. The evolution algorithms are based on genetic adaptation by utilizing iterative process of mutation, reproduce, selection and survival as seen in population. The swarm intelligence is based on collective behaviour of group of simple agents that exhibit decentralized and self-organized mechanism in the foraging process. Meanwhile, the last category of ecology algorithm is based on interaction, either cooperative or competitive of the living organism with the environment such as air, soil and water.

The aim of feature selection purpose in biomarker discovery is to optimize the search for the most parsimonious features that best discriminates disease with normal samples. Due to the fact that mass spectrometry expression data are high dimensional, classification of relevant features will be time consuming. Integrating feature subset selection to the actual classification can effectively reduce calculation time without negatively affecting predictive error rate [10].

This paper is organized as follows: Section 2 elaborates the original proposed algorithm for biomarker discovery, parameters setting and experimental setup; Section 3 discuss the implementation, analysis and result of the proposed feature selection strategies; Section 4 discusses outcome and future recommendation of the study.

2. METHODOLOGY

The major concern of how DE algorithm is constructed and implemented for feature selection

is explained in the next subsections.

2.1. The Differential Evolution in Concept

Differential evolution (DE) is known as one of the simplest and straight forward methods in evolutionary concepts [11]. DE has evolved into a competitive concept of problem solver across many domains since it had been introduced in a technical paper by [12]. The DE retains all evolutionary steps but some major modification focusses to mutation step, which employs the concept of difference on parameter vectors to exploit the search space. The concept of DE is illustrated by Fig.1.

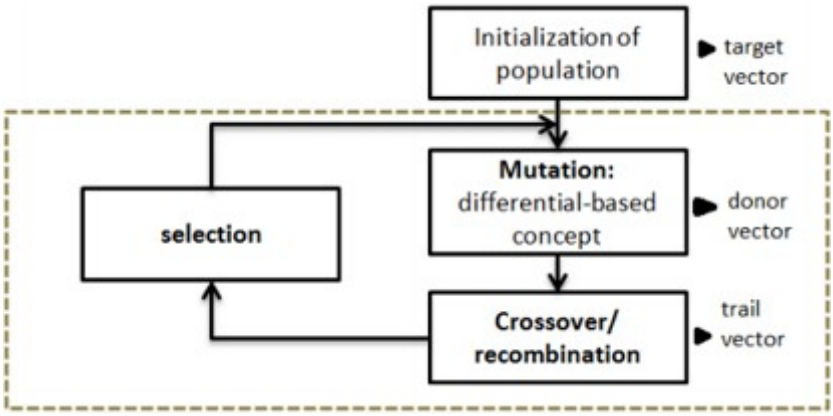

Fig.1. Differential evolution original concept for problem solving

In DE, the series of parameter vector is considered as candidate solutions (population) which are known as genomes or chromosomes that continuously improve their fitness throughout generations. Thus, the current generation, $X_{i,G}$ of the particular population, *i*th vector is best presented as the following Equation (1). Where, i = 1,2,3,4,…,*NP* with *D*-dimension and generation is denoted by *G*.

$$X_{i,G} = \{ x_{1i,G}, x_{2i,G}, \ldots\ldots, x_{Di,G}\} \quad (1)$$

Initially, candidate solutions are chosen randomly across the search space by assuming a uniform random probability distribution applied. Each initial or current generation of genomes is called as target vectors and will generate the new trail vector (offspring), $U_{i,G}$ through mutation process.

2.1.1. Mutation Concept in DE Algorithm

Aforementioned, differential-based mutation is the key differentiating of DE from other types of evolutionary concepts. In order to create a differential mutation vector from a target vector, three district indexes $r1, r2, r3 \in \{1, 2,., NP\}$ are sampled randomly from the current

population and must be different from each other and running index i ($r1$ 6= $r2$ 6= $r3$). Therefore, the *ith* mutant vector is generated based on Equation (2) as follows

$$Vi,G+1 = xr1,G + F(xr2,G - xr3,G) \quad (2)$$

where F is a real and constant factor $\in[0, 2]$ that is responsible in controlling the amplification of the differential variation ($xr2,G - xr3,G$). The mutant vector, $Vi,G+1$ is best known as donor vector. DE algorithm improves their diversification of population and perturbed vectors via crossover operation. The trail vector, $Uji,G+1 = (u1i,G+1,\ u2i,G+1,\ u3i,G+1,..,\ uDi,G+1)$ yields from recombination of components both from donor vectors and target vectors. This process has been illustrated by Fig.2 where $j = 1, 2,., D$.

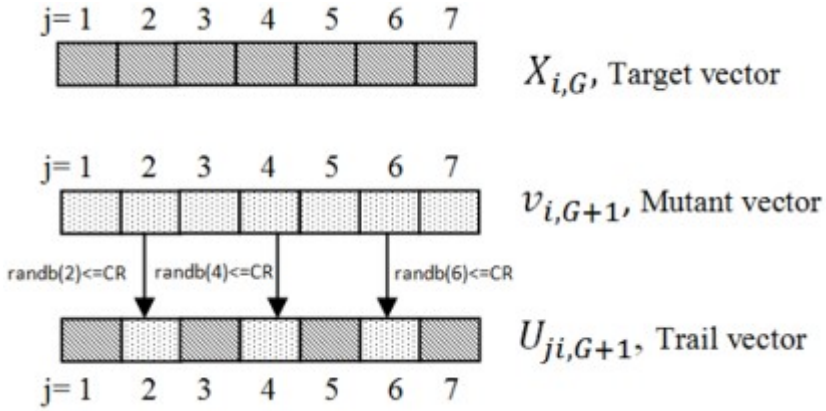

Fig.2. Binomial crossover process

2.1.2. Crossover or Recombination

Generally, the crossover operation relies on *binomial* concept that is performed randomly on each of the D variables by Equation (3) as follows:

$$u_{ji,G+1} = \begin{cases} v_{ji,G+1} & if(randb(j) \le CR) or\ j = rnbr(i) \\ x_{ji,G} & if(randb(j) > CR) or\ j \ne rnbr(i) \end{cases} \quad (3)$$

This binomial crossover performs on any D whenever a randomly generated number [0, 1] is less than or equal to the CR value. Similar to F in mutation process, the CR is a control parameter $\in[0, 1]$ for crossover operation. Meanwhile, *rnbr(i)* is a random chosen index that is used to make sure the trail vectors consist of at least one parameter from mutant vectors.

2.1.3. Selection

Finally, selection process is made to evaluate the quality of trail vector, $Ui\ ,G+1$ compared to current generation. The cost function with smallest value is measured; where if vector $vi\ ,G+1$

yields smaller value than *xi ,G*, then *vi ,G*+1 is set as *Ui ,G*+1 or vice versa. The process will continue until the population ends.

2.2. The Parameters Setting

In order to optimize the searching process, proper parameters setting and tuning has to be considered. The pre-testing was evaluated for 50 runs in order to get the best score for each parameter and was also common in several literatures[13]. Table 1 depicted the parameters setting to generate crossover and scaling factor of DE algorithm as proposed in section 2.1.

Table 1. Parameter setting of crossover and scaling factor

Crossover		**Scaling factor**	
Max crossover rate	1	Max scaling factor	0.5
Median crossover rate	0.5	Median scaling factor	0.4
Min crossover rate	0.2	Min scaling factor	0.3

The comparison of the proposed DE algorithm was made with Particle Swarm Optimisation (PSO) and Ant Colony Optimisation (ACO). Generally, Particle Swarm Optimisation (PSO) is one of the most popular bio-inspired algorithms that has been proposed as feature selection for various problems domain [14]. Along with PSO, Ant Colony Optimisation (ACO) develops as an efficient paradigm of feature selection in the generous domain [15]. Thus, both of algorithms have been accepted and continuously being adapted in many areas of active ongoing research. ACO and PSO have been categorised in the same group of bio-inspired algorithm, emphasizing on population-based search techniques. These ACO and PSO were taken from [16-17]that have described properly the steps taken, thus easy for re-construction of the information for comparison. Prior to the analysis, all algorithms used features from the proposed feature extraction in Fig.3. Therefore, the results depict the performance of the algorithm itself. Table 2 depicted all the parameters setting for both ACO and PSO algorithm.

Table 2. Parameter setting for ACO and PCO algorithms

ACO		PSO	
	Estimation Value		**Estimation Value**
Population size/Number of Ant	50	Number of particles	50
Relative inference of pheromone trial	1	Weighting factor1	0.5
Prior information	1	Weighting factor2	0.2
Evaporation constant	0.1	Constriction factor (inertia weight)	0.5

2.3. Experimental Setup

Hepatocellular carcinoma (HCC) is a type of liver cancer that most commonly occurs among African and Asian people. This kind of cancer frequently happens among men than women and usually caused by scarring of the liver called as cirrhosis. Thisreal world dataset was obtained from the original study [17] that was collected among patients from Cairo, Egypt. The serum samples were collected from three main categories of HCC cases, cirrhosis cases and normal cases. The serum samples were tested clinically using MALDI-TOF techniques and produced spectrums with 23 846 m/z bins after binning process. This study, the goal is to distinguish between HCC and cirrhosis cases.

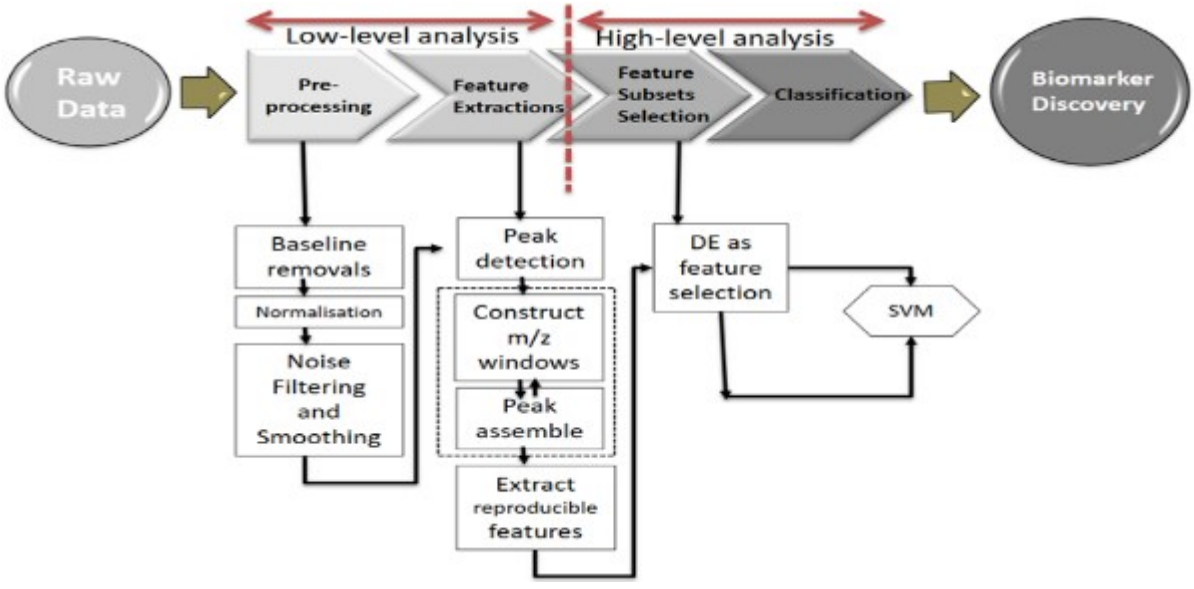

Fig.3. The experimental pipeline for biomarker discovery

The dataset is the real world big data problems. Therefore, careful consideration for the whole setting is crucial. The out samples produced from mass spectrometry analysis using the MALDI-TOF techniques were considered as raw high-dimensionality data. Whereas, the total sample is taken from 129 patients (78 HCC patients, 51 Cirsshosis patients), each sample was binning to 23, 846 with the 919.6650-9980.518 Mass-to-charge ration (M/Z range). Fig.3 depicted the whole process of the experimental setting. The previous mentioned raw samples underwent two levels of analysis that are low-level analysis focusing to cleaning and extracting the potential features and high-level analysis focusing on finding parsimonious features throughout feature selection and classification. This study is focusing to high-level analysis using feature selection that was constructed using Differential Evolution (DE) algorithm and further incorporate with SVM classifier for training and testing the data. The samples were randomly split into 70:30 of training and testing data respectively.

SVM classifier has been used extensively in many application domains because of its flexibility in choosing similarity function, has proficient accuracy and ability of handling larger feature spaces. SVM learn to discriminate between two classes via maximization of margin anlpid work very well in image processing, text and data mining. Therefore, because of stability and simplicity of SVM in most cases lead this study to incorporate SVM with DE algorithm for feature selection. Besides SVM, 4-fold cross validation was incorporated for generalization performance.

3. RESULTS AND DISCUSSION

Implementation of DE features selection for HCC dataset were applied for identification of biomarkers on HCC and cirrhosis cases and followed parameters setting in in subsection 2.2.The most occurrence features selected by DE, ACO and PSO respectively and classified by SVM on training samples for 500 iterations. Comparing these three figures, the method of DE steadily showed many features index was selected to build classification model compared to ACO and PSO. Hence, indicating that DE is suitable to be applied as feature selection. Compared to those figures, number of features selected showed that less features were selected to build classification model in PSO and ACO. Hence, indicating that the exploration and exploitation mechanism in both algorithms do not fit with the data. The searching

mechanism for the both feature selections imposed difficulties in extracting knowledge. The selected features were most equivalently good features for classification. The points that were marked by '*' in Fig.4, 5 and 6 indicated features for building the learning model.

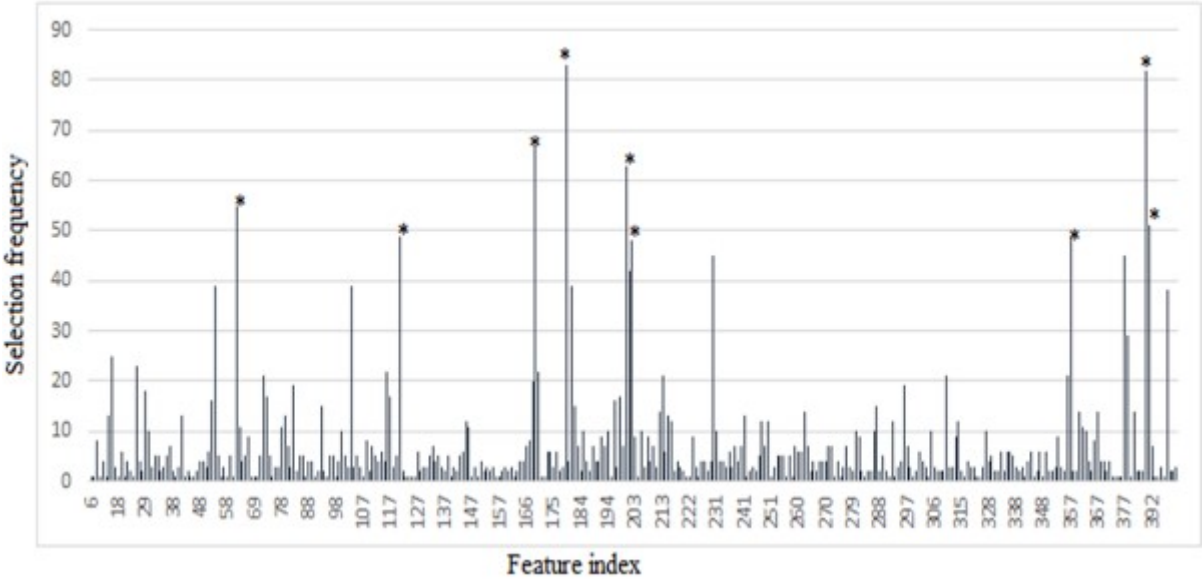

Fig.4. Most occurrence DE features on HCC samples

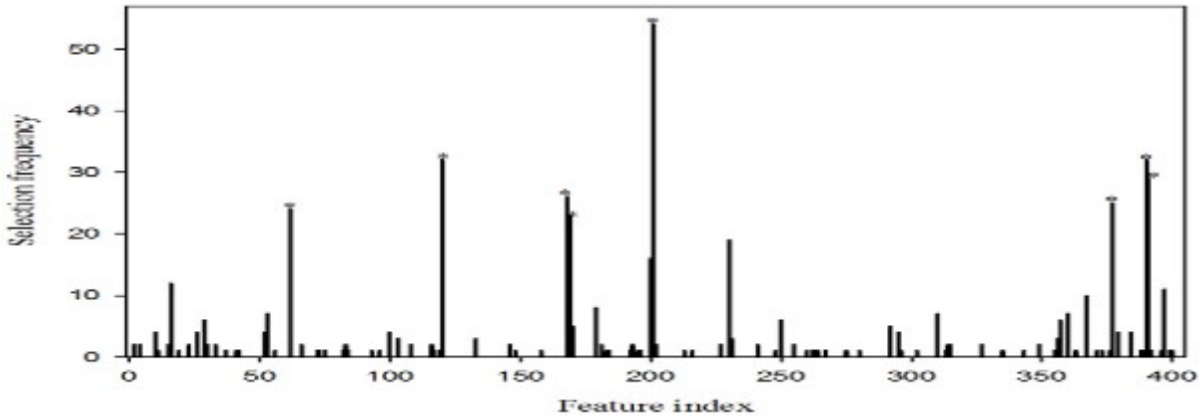

Fig.5. Most occurrence ACO features on HCC samples

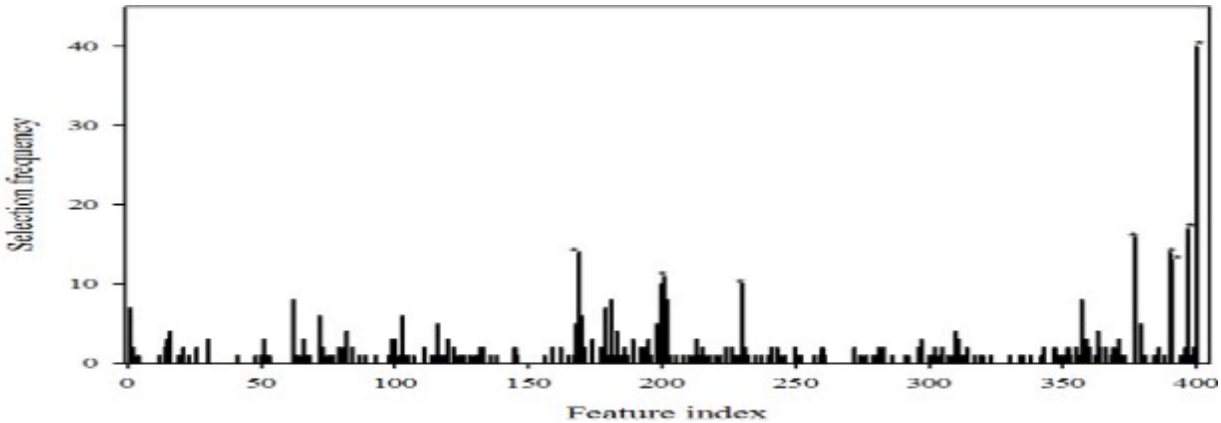

Fig.6. Most occurrence PSO features in HCC samples

The learned model was running independently on testing data. Table 3 showed three different

results generated by three different algorithms. The accuracy, sensitivity and specificity on testing data from DE algorithm again yields the most outstanding result and on the same time proved the stability of the algorithm implemented across different datasets without any parameters changing (re-tuning).

Table 3. Performance comparison between DE, ACO and PSO

Evaluation	DE		ACO		PSO	
	Training	Testing	Training	Testing	Training	Testing
Accuracy	99.05	82.22	99.05	75.56	97.14	80.0
Sensitivity	100	80	100	71.43	96.23	78.95
Specificity	98.15	84.0	98.15	79.17	98.08	80.77

4. CONCLUSION

The research starts with an interest to discover proteomics biomarkers from mass spectrometry data. The feature selection methods select subset of features and evaluate the classification performance. The predictions of biomarkers are based on the most outstanding results produced by the classification model. The outstanding results evaluated from accuracy, sensitivity and specificity of the classification model from parsimonious subset of features. Instead of that, the subset of parsimonious features must show the best discriminative characteristic between healthy and disease cases. The proposed DE algorithm as feature selection was chosen based on the efficiency and simplicity of the searching mechanism in exploring and exploiting new solution the searching area thoroughly. From above comparison in section 3, the proposed DE algorithm potentially obtained excellent classification performance compared to well-known ACO and PSO algorithms due to consistency of producing comparative results. Hence, the implementation showed relevancy of the proposed method and robust for biomarker identification. The proposed algorithm as feature selection mechanism will be enhancedin future into other real world of high dimensionality data to test the robustness of the algorithm. In addition, the proposed methods will also be tested to solve other domain of knowledge for classification purposes.

5. ACKNOWLEDGEMENTS

We wish to express our heartfelt gratitude to the Universiti Teknologi MARA (UiTM) for funded grant under iRAGS (600-RMI/IRAGS5/3(52/2015).

6. REFERENCES

[1] Farina A. Proximal fluid proteomics for the discovery of digestive cancer biomarkers. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2014, 1844(5):988-1002

[2] Parker C E, Borchers C H. Mass spectrometry based biomarker discovery, verification, and validation–Quality assurance and control of protein biomarker assays. Molecular Oncology, 2014, 8(4):840–858

[3] Sajic T, Liu Y, Aebersold R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: Perspectives and clinical applications. PROTEOMICS-Clinical Applications, 2015, 9(3-4):307–321

[4] Conrads T P, Zhou M, III E F P, Liotta L, Veenstra T D. Cancer diagnosis using proteomic patterns. Expert Review of Molecular Diagnostics, 2003, 3(4):411–420

[5] Petricoin E F, Liotta L A. Seldi-tof-based serum proteomic pattern diagnostics for early detection of cancer. Current Opinion in Biotechnology, 2004, 15(1):24–30

[6] Ressom H W, Varghese R S, Drake S K, Hortin G L, Abdel-Hamid M, Loffredo C A, Goldman, R. Peak selection from maldi-tof mass spectra using ant colony optimization. Bioinformatics, 2007, 23(5):619–626

[7] Swan A L, Stekel D J, Hodgman C, Allaway D, Alqahtani M H, Mobasheri A, Bacardit J. A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. BMC Genomics, 2015, 16(1):1-12

[8] Neumann F, Witt C. Bioinspired computation in combinatorial optimization: algorithms and their computational complexity. In ACM15th Annual Conference Companion on Genetic and Evolutionary Computation, 2013, pp. 567–590

[9] Binitha S, Sathya S S. A survey of bio inspired optimization algorithms. International Journal of Soft Computing and Engineering, 2012, 2(2):137–151

[10] Chuang L Y, Chang H W, Tu C J, Yang C H. Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry, 2008, 32(1):29–38

[11] Das S, Suganthan P N. Differential evolution: A survey of the state-of-the-art.IEEE Transactions on Evolutionary Computation, 2011, 15(1):4–31

[12] Storn R, Price K. Differential evolution-A simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization, 1997, 11(4):341–359

[13] Zhao W X, Davis C E. Swarm intelligence based wavelet coefficient feature selection for mass spectral classification: An application to proteomics data. Analytica Chimica Acta, 2009, 651(1):15–23

[14] Tu C J, Chuang L Y, Chang J Y, Yang C H. Feature selection using PSO-SVM. IAENG International Journal of Computer Science, 2007, 33(1):111–116

[15] Ahmed A A. Feature subset selection using ant colony optimization. International Journal of Computational Intelligence, 2005, 2(1) 53–58

[16] Ressom H W, Varghese R S, Drake S K, Hortin G L, Abdel-Hamid M, Loffredo C A Goldman R. Peak selection from maldi-tof mass spectra using ant colony optimization. Bioinformatics, 2007, 23(5):619–626

[17] Ressom H W, Varghese R S, Abdel-Hamid M, Eissa S A L, Saha D, Goldman L, Petricoin E F, Conrads T P, Veenstra T D, Loffredo C A and Goldman R. Analysis of mass spectral serum profiles for biomarker selection. Bioinformatics, 2005, 21(21):4039–4045

How to cite this article:
Yusoff S A M, Zahari S M, Mustafa F H, Rahman N A, Abdullah M H. Differential evolution algorithm as feature selection for biomarker discovery. J. Fundam. Appl. Sci., 2018, *10(2S), 1043-1054.*