Soft Computing (2020) 24:14265–14274
https://doi.org/10.1007/s00500-020-04793-z

METHODOLOGIES AND APPLICATION

Efficient attribute selection technique for leukaemia prediction using microarray gene data

D. Santhakumar[1] · S. Logeswari[2]

Published online: 24 February 2020

Abstract

The recent advancements in today's medical sciences regarding Data Analytics have made it possible for the use of efficient techniques for analysis. For prognosis, diagnosis and cancer treatment, a microarray-based gene expression profiling is considered. Informative genes causing cancer are determined through the deoxyribonucleic acid microarray technique. Dimensionality is the utmost concern while working with multi-dimensional data analysis which acts as a barrier in extracting information from a dataset which leads to costly computational complexity. Thus, an imperative task in the selection of relevant features in the analysis of cancer microarray datasets is crucial towards effective classification. This work focuses on variable selection techniques by utilizing effective correlation for attribute selection along with ant colony optimization. The criterion of a given classifier is maximized through wrapper-based attribute selection, and so it needs efficient searching techniques in finding optimal feature combinations. A new wrapper-based selection technique which uses ant lion optimization (ALO) in finding optimal feature set is proposed in this work which maximizes classification performance. The natural shooting procedure of ant lions is imitated in the proposed ALO algorithm. Support vector machine technique was utilized for the classification of chosen marker genes.

Keywords Microarray · Gene expression data · Machine learning · Attribute selection · Effective correlation-based attribute selection (ECFS) · Ant colony optimization (ACO) · Ant lion optimization (ALO) · Support vector machine (SVM)

1 Introduction

Scientists use lesser time to measure gene expression in large numbers using the DNA microarray technology; thus, diagnosis of disease on the basis of examination of genes might take some months. In such cases, if the disease had been diagnosed earlier, treatment could have been started at an earlier stage itself, so that the chances of curing it are more. If gene expression microarray data using computational process had been used, the disease could have been diagnosed at an earlier stage itself. Moreover, microarray data can determine the role of certain genes which are answerable for the diseases growth (Babu and Sarkar 2016).

Cancer is the most feared disease of mankind. Conventional techniques of diagnoses of cancer are based on its clinical and morphological structure. Several environmental factors may cause this disease including synthetic chemicals, radiation, smoking and so on which may lead to mutation of genes. Many other unknown causes may also be present which might be detected through the selection of the informative genes that contribute to the phenotype or symptoms, which can increase the chance of disease identification and prediction of the type of cancer. Early diagnosis of cancer is possible through the use of microarray data which helps physicians in suggesting improved healthcare plan for patients thus enhancing their life (Bhola and Tiwari 2015).

The most common malignancy among children is acute lymphoblastic leukaemia (ALL). Recently, in the developed countries, the survival rates have gone up to over 90%

Communicated by V. Loia.

✉ D. Santhakumar
dsanthakumar2@gmail.com

S. Logeswari
slogesh76@gmail.com

[1] Computer Science Engineering, CK College of Engineering and Technology, Cuddalore, India

[2] Computer Science Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India

due to risk-adapted treatments. But then, prognosis is poor in 20% of children which makes ALL the foremost reason of cancer mortality among paediatric disorders. It is quite challenging to classify patients on the basis of appropriate risk groups for proper ALL management. Poor disease courses can be mitigated through stratifying chemotherapeutic treatment over initial detection of related results (Pan et al. 2017).

Many prognostic factors for ALL have been identified through previous group-level investigations such as age at diagnosis, white blood cell (WBC) counts and gene fusions. Also, decline risk for patients at initial treatment can be identified through immune-phenotype, percentage level of immature cells referred to as bone marrow lymphoblast (BML) which was recorded on 15th day and 33rd day, and level of minimal residual disease (MRD). There is quiet dearth of tools for the clinicians to evaluate the risk of ALL relapse in initial cure.

While looking after from data analytics perspective, with the rise in data dimensionality, there is a multi-fold increase in the remarkable size of data needed for consistent and reliable analysis. This was referred to as “curse of dimensionality” by Bellman considering the issues arising through dynamic optimization. In the most unavoidable situations, the issue in multi-dimensional datasets is to project the data into lesser number of attributes/features which can preserve the evidence up to a certain possible extent. The most fitting example for small sample problem is microarray data. Every identical data point has the possibility of having 4.5 lakhs gene probes (variables), and maximum of data points are processed involving increased computational cost (Hira and Gillies 2015).

When there is significant growth in the dimensionality of dataset, it results in increased struggle in significantly evidencing the outcome statistically because of the lack of meaningful data in the dataset. Large datasets “large *p*, small *n*” problem (here *p* and *n* represent the count of variables and samples, respectively) remains prone to overfitting. Small fluctuations can be mistaken by over-fitted model for significant variance in data leading to classification errors. This issue can increase because of noisy attributes. In a dataset, noise can be denoted as “measured variable error from the variance” that leads to measurements error or natural variation (Aziz et al. 2017)-based error.

Noisy data can affect machine learning algorithms. There should be reduction in noise to a certain extent in order to avoid complexities that are not needed in the inferred models and enhance the efficiency of the algorithm. Noise can be classified into two types: when there is error in attribute value (missing or wrongly measured variables), attribute noise is caused by samples that are labelled to belong in one or more classes or misclassifications. As there is an increase in dimensionality, there is also an increase in computational cost, generally in an exponential fashion.

In pattern recognition, machine learning, statistics and data mining, feature selection is a well-known method in which reduction in dimensionalities can take place. Based on certain criteria, this method chooses a subset of pertinent attributes from the features belonging to the original set. Generally, feature selection is considered in the areas where thousands of features make a dataset but the size of the sample is quite small (e.g., data pertaining to gene expression). Gene selection is the feature selection criterion which finds its application in data of gene expression. It is crucial to have gene selection as there are many non-relevant, noisy, unwanted expressions and is quite effective in the detection of premature tumour and cancer detection primes in the direction of dependable cancer prognosis or diagnosis and improved medical action. There can be unlabelled, fully labelled or partially labelled gene expression data. Using the above-mentioned biological designs, prediction of classes can be discovered.

Conventionally, feature selection consists of three stages: filter, embedded, wrapper and hybrid approach. There is evaluation of every individual in filter approach. This can find its application easy to high-dimensional datasets; they are less complex having the classifier independent. Attributes including *t* test, information gain, minimum redundancy maximum relevance (mRMR) and Euclidean distance are very familiar. Here, features that are having best statistical score are chosen. In approaches pertaining to filter feature selection, the classifier’s performance and interdependency do not have any role and so it is a known fact that there will be deficiency in the classifiers presentation (Chandra and Gupta 2011).

The basis for feature evaluation in wrapper techniques is classifier performance. The classification of wrapper approach includes a deterministic and stochastic approach. The techniques which are included under stochastic approach include arbitrary hill climbing (HC), ant colony and genetic algorithm (GA), while deterministic approaches include sequential forward selection (SFS) and sequential backward elimination (SBE). As regards these approaches, the classifier achievement is high; however, the complexity for search space is maximum for issues by thousands of features which will cause higher complexity of time. Some advantages of ant colony optimization (ACO) are: it gives a rapid discovery of good solutions and guarantees the convergence, inherent parallelism and efficiency for travelling salesman problem.

The properties are models are taken into consideration with regard to embedded approaches (Sharbaf et al. 2016) where most salient attributes are chosen. In this group of methods decision tree and neural networks are considered, but these techniques have a high computational complexity. Cancer classification and gene selection are grouped under SVM and recursive feature elimination (SVM-RFE). These issues cannot be overcome by the techniques suggested above. Thus, the literature proposes ensemble approaches. Hybrid prototype is used for feature selection in these approaches and there is integration of results. Two types of feature selections SVM-RFE and mRMR are hybridized here.

With regard to machine learning, classification is an important task. It is difficult to find the useful genes without prior knowledge. Generally, a huge volume of genes gets its entry into the dataset which includes genes that are significant, insignificant and irrelevant, though insignificant and irrelevant genes are not useful in classification. Selection of genes is crucial with regard to cancer. Minimizing the number of relative and informative genes which are more predictive in the process of classification is the aim of gene selection. A subset of genes is chosen through an optimization algorithm with the maximum classification data from the original gene microarray data. The most repeatedly used gene selection technique is classified as filter, wrapper and embedded methods, whereas the optimal gene chosen problem is considered as non-deterministic polynomial (NP)-hard problem. Thus, these issues can be solved through heuristic approaches such as inspired evolutionary algorithms.

Feature selection using ALO technique is being used in this work in the selection of gene subset from cancer microarray data. The rest of the study is organized into the sections given below: Literature review is done in section two, while section three involves the different techniques used in this work. Detailed discussions of outcomes are mentioned in section four and section five of the study.

2 Related works

A two-stage local dimension reduction approach was proposed by Guo et al. (2017). Insignificant redundant data were removed through a novel L1-regularized feature selection technique and for choosing the significant biomarkers in the first step. In the next step, implementation of partial least squares (PLS)-based feature selection was done to extract synthesis features which replicate selective features for sorting. An empirical work was done which reflected the suitability of the proposal with ten microarray datasets that were widely used and outcome showed its effectiveness and competitiveness with ten microarray datasets which resulted in more effective and competitive outcomes than outcomes obtained through four state-of-the-art approaches. Results of the study showed that the St. Jude dataset depicts that this technique can be applied successfully to the evaluation for subtype prediction in microarray data analysis and discovery of gene co-expressions.

In high-dimensional microarray data, gene selection and (Ang et al. 2016; Pyingkodi and Thangarajan 2018) estimation of gene coefficients and sparse logistic regression with the help of L1-norm were done simultaneously, though the performance of L1-norm cannot be effective when the correlation between genes is high. While addressing this issue, Algamal (2017) proposed an efficient sparse logistic regression (ESLR). Gene expression data on extensive application using high dimensions show that this technique can choose genes with high correlation. With regard to classification accuracy and Youdens index, a competitive performance is done by ESLR compared to three other methods.

Analytic hierarchy process (AHP) was utilized to build a prototype with more objective in the study proposed by Lv et al. (2016) which concentrates on classification accuracy than the gene count. Multi-objective estimation of distribution algorithm (MOEDA) which is a multi-objective heuristic algorithm, an enhancement on marginal distribution algorithm (UMDA) was put forth to solve the prototype. ‘Higher and Fewer Rule’ which helps to evaluate and sort entities and ‘Forcibly Decrease Rule’ which is considered to produce possible entities through increased precision in classification with decreased genes. This technique was tested on microarray datasets belonging to both binary and multi-class.

A novel hybrid technique was proposed on the basis of clustering, and particle swarm optimization (PSO) was formulated by Han et al. (2015) for selection of gene and microarray data classification. In this technique, genes are divided first into a specific number of clusters through *K*-means technique. As there is more redundancy in every cluster, mRMR technique is used in reducing the redundancy of the genes that are clustered. Further gene selection was done using PSO from the remainder of clustered genes. PSO was used here because it has better performance generalization, faster convergence compared to other learning algorithms for NNs; ELM is used in the evaluation of candidate gene subsets chosen by PSO, and sample classification was done. Less redundant genes were chosen, and prediction accuracy was increased.

A study was performed by Chaudhari and Agarwal (2018) on quantum PSO with elitist breeding (EBQSO) on gene datasets. For in depth searching and classification applications with genetic datasets elitist is not considered until now. This work uses EBQPSO algorithm on gene datasets to classify cancer. Supervised and unsupervised learning approaches are used in the algorithm, i.e. SVM, J48 and NN. The study’s outcome shows that EBQPSO is better compared to PSO and quantum PSO (QPSO) with regard to recall and accuracy.

Bonilla-Huerta et al. (2016) studied a compound structure with binary phases for selection of gene and gene classification of DNA microarray data. In its initial phase, preliminary gene selection was done using Multiple Fusion Filter where five conventional statistical techniques were combined; in the next step, appropriate gene subsets were chosen by the use of embedded GA, Tabu Search (TS) and SVM with the analysis of frequency of individual genes to different gene subsets. Finally, embedded approach was used to evaluate the genes and to achieve last appropriate small gene subset with the highest presentation. This technique was tested on four DNA microarray datasets.

A hybrid feature selection algorithm was suggested by Lu et al. (2017) which merges mutual information maximization (MIM) and adaptive genetic algorithm (AGA). The obtained outcome shown here is that the MIMAGA-identification technique had reduced gene depression's dimension significantly and also redundancies in classification were removed. The highest classification accuracy is provided by the reduced gene expression dataset in comparison with the traditional feature selection algorithms. Four various classifiers were also applied in demonstrating the strength of the proposed MIMAGA-Selection process.

A new random forests-based feature selection technique was put forth by Yao et al. (2015) which adopts the design of stratifying feature space which puts together generalized sequence backward and forward searching strategies. Features were ranked through a random forest variable importance score, and feature subsets evaluating function were classified using various classifiers. Five microarray expression datasets were examined through the proposed technique which includes nervous, breast, DLBCL, leukaemia, and prostate, and their SVM classifier accuracies were 91.67%, 85%, 100%, 95.24% and 91.67%, respectively. The results of the study showed that the technique not only enhanced classification accuracy but reduced computation time of feature selection process.

A two-stage hybrid mechanism for the classification of cancer was suggested by Jain et al. (2018) combining correlation-based feature selection (CFS) with enhanced binary PSO (iBPSO). Pretty low-dimensional combinations of prognostic genes were considered in the classification of biological examples of binary and multi-class cancers. The early convergence of local optimum of conventional BPSO was also controlled by iBPSO. In the proposed model, eleven microarray datasets that represent distinct types of cancers were evaluated. Study results helped in comparison of seven other popular techniques and the model-conveyed better results with regard to accuracy in classification and total number of genes chosen for the process. Specifically, up to 100% classification precision was got for 7/11 datasets.

A new effective technique for diagnosis prediction was put forth by Gao and Liu (2018) aimed at diagnosis prediction. The dimension complexity of microarray data and the definite classifier was considered; SVM technique on the basis of recursive feature elimination (SVM-RFE) was applied in the optimal gene subset selection. Ultimately, a combined technique of fruit fly optimization and PSO was suggested to optimize the classifier (least square-SVM parameters). These helped in achieving a combined effort of disease diagnosis, and a precision of 100% was gotten with just four features. Various techniques were also used to compare the proposed method. The study indicates that best prediction performance was obtained through the proposed technique.

3 Methodology

In the area of science and technology, the advent of machine learning has had specific impact. This has helped in the invention of innovative tools in the healthcare sector, medical and biomedical industry as a great support to medical practitioners. Disease prediction is the developing field in biotechnology and the field of medicine, and so it is the main focus in processing the given input data in identifying the type of disease. A critical role is played by human genome sequence, and this can balance the way clinical practice is taking place. Genome sequencing provides understanding of disease mechanisms and also is a primary factor in the development of new drugs in the near future. Moreover, genome sequencing will become a part of regular diagnosis which leads to disease prevention rather than cure as identification of diseases will be done at the early stage of illness. Because of the increased size of genome sequence, a crucial role is played by machine learning in analysing the data and ultimately predicting the disease. For precise classification of phenotypes, a small subset of genes were chosen out of thousands of genes in microarray data. Genes are ranked by various techniques based on their differential expression among phenotypes, and top-ranked genes are picked up. However, irrelevant features are present in the data so selected. This section involves attribute selection based on CFS, ACO, ACO and SVM as classifiers.

3.1 Leukaemia dataset

The leukaemia dataset was taken from a collection of leukaemia patient samples reported by Golub et al. (1999). This dataset frequently serves as a benchmark for microarray analysis techniques. It includes a gene expressions corresponding to acute lymphoblast leukaemia (ALL) and acute myeloid leukaemia (AML) samples from bone marrow and peripheral blood. The dataset comprised 72 samples: 49 samples of ALL and 23 samples of AML. Each sample is measured over 7129 genes.

3.2 Effective correlation-based attribute selection (ECAS)

Correlation-based heuristic evaluation function had been carried out for ECFS scores rather than evaluating through scores and ranks to individual features. As the microarray feature space is generally large, a best first-search heuristic is carried upon by ECFS which considers the usefulness of individual features in the prediction of class. Thus, a subset is chosen by ECFS for class with greatest correlation and for features which are least correlation (Alshamlan 2018). A matrix of correlation is computed by ECFS first from the training data samples. Formerly, heuristics allocates a score for the subset of variables which is computed with the help of (1):

$$\mathrm{Merit}_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \tag{1}$$

Here, $Merit_s$ denotes the heuristic merit of a variable subset S that contains k attributes, $\overline{r_{cf}}$ represents the overall average interrelationship between classes and attributes, and $\overline{r_{ff}}$ is the average interrelationship between attributes.

An empty feature set is the starting point ECAS, and the subset that has the top most merit is identified throughout examination. While solving the issue, CAS technique is used to select the genes which can be correlated with specific cancer classes. As there is a huge microarray feature, best first-search heuristic is used by CAS which considers the gain obtained through the individual attribute in the class prediction.

The ultimate idea to apply ECAS gene selection technique is to identify the early microarray dataset through highly correlated subset of genes. ECAS filtering technique is used to process the initial gene expression profile. The individual gene is evaluated and sorted on the basis of a single CAS criterion as detailed in this section. A new subset by name CAS dataset is created to give high classification accuracy with highly correlated genes. In case, if S genes are present in microarray datasets, after the application of ECAS filter technique, there is reduction in the gene count to m genes which has the highest interrelation between genes and class and the lowest interrelation between gene and gene.

3.3 Attribute selection using ant colony optimization

M. Dorigo along with his colleagues introduced ACO early in 1999 (Dorigo and Di Caro 1999) which was a nature-based metaheuristic that is useful for solving hard combinatorial optimization (CO) problem. ACO is a metaheuristics category that achieves best outcomes to typical CO problems along with a prescribed extent in computational time. The natural ants possess a skill to find shortest paths by dropping of pheromone as they move on; every individual ant has a predefined direction based on this pheromone chemical. Evaporation of chemical substance takes place with duration which leads to less pheromone in less travelled paths. Thus, logically the route with the highest rate of travelling will be considered as a shortest path and this path will be strengthened and there will be diminishing of others till every ant tracks the direct footpath (Rasmy et al. 2012). Generally, to any combinatorial problem an ACO algorithm can remain practical to the maximum possible level and can be defined as:

- The problem should be represented in a proper format. The problem description should be denoted as a graph that contains combinations of nodes, and there is edge between nodes.
- Heuristic desirability (η) of edges should be present. In the graph, there should be a “goodness” measure of paths in between the nodes.
- Feasible solutions should be constructed. An appropriate mechanism should be present whereby there is efficient creation of possible solutions.
- Updating criteria for pheromone. While updating edges based on pheromone levels, a corresponding evaporation rule must be followed. Methods typically choose n best ants and then select their footpaths.
- Probabilistic transition rule is one which regulates the possibility to traverse for an ant between the nodes in the graph.

An ACO-suitable problem can be reformulated from a feature selection task. A graph should represent the ACO, and nodes signify features through edges among them. Consideration is given that the ant is at node ‘a’ currently and can choose the variable to include next to its traversal path. Based on the transition rule, Feature b is chosen next and then c and d are chosen consecutively. Once the traversal reaches the dth state, the present subset $\{a; b; c; d\}$ can fulfil the ending condition of the traverse. Its travel is terminated, and the feature subsets are outputted as a candidate for reduction in data.

Any subset evaluation function can be used for a proper heuristic attraction of travelling between features. The probabilistic transition rule can form the edge pheromone levels and heuristic desirability of traversal (Sara et al. 2019) which denote the ant's probability at *I*th attribute which travels to *j*th attribute at time *t* in (2):

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{\left[\tau_{ij}(t)\right]^{\alpha}.\left[\eta_{ij}\right]^{\beta}}{\sum_{l \in J_i^k}\left[\tau_{il}(t)\right]^{\alpha}.\left[\eta_{il}\right]^{\beta}} & \text{if } j \in J_i^k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Here, k represents the total count of ants, η_{ij} denotes the attraction of selecting *j*th attribute, while it is situated at *I*th attribute, η_{ij} denotes effectiveness of algorithm enforcement, J_i^k denotes the combinatory nearby nodes of *i*th node which are yet to be travelled by ant K. $\alpha \succ 0$, $\beta \succ 0$ are important factors used to regulate the combinational value of the heuristic information and pheromone rate (here α, β chosen are found mathematically), and $\tau_{ij}(t)$ denotes the entire pheromone that is considered to be virtual on edge (*i*, *j*).

ACO attribute selection starts with the initialization of the amount of ants, k that are projected arbitrarily on the graph. As an alternative, when looking at the graph, the amount of ants and the number of attributes present in the input data are sometime equal; here, individual ant begins to construct its own path at a distinct feature. They travel to edges probabilistically till they reach a travel stop criteria. The subsets that are resultant are collected and evaluated. If there is an optimal subset or the protocol has been implemented a number of times, then the process is stopped over there which conveys that the best feature subset is met. If both the above-mentioned scenarios do take place, there is updating of the pheromone and a different combination of ants is formed and the procedure moves on again. The pheromone on every edge is modernized based on Eqs. (3, 4):

$$\tau_{ij}(t+1) = (1-\rho).\tau_{ij}(t) + \rho \cdot \Delta\tau_{ij}(t) \quad (3)$$

Here:

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{n}\left(\gamma'(S^k)/|S^k|\right) \quad (4)$$

If the edge (*i*, *j*) has been traversed, $\Delta\tau_{ij}(t)$ is 0 otherwise. The value $0 \le \rho \le 1$ is decay constant helps to act out the evaporation of the pheromone, S^k signifies the feature subset originate by ant k. The pheromone is restructured depending on both the measure of the "goodness" of the ant's feature subset γ' and the size of the subset itself.

3.4 Proposed feature selection using ant lion optimizer (ALO)

ALO algorithm takes its inspiration from natures and imitates the foraging behaviour of the larvae of ant lion. Mirjalili (2015) proposed ALO which is an algorithm for stochastic population-based optimization.

In ant lion, the life is classified into two phases like larva and adulthood. Hunting is an important factor in first phase. A set-up is excavated through the larvae of ant lion in the sand which practises a cone through revolving in a spherical motion and the sand is thrown to separate the trap. Its shape is reliant on its level of hunger and the moon's shape. On burrowing, the larvae move to the bottommost of the cone and delays for its prey to get stuck in the hole. The ant lion catches the prey once it detects its location. The sand is petrified wisely to the trap's edge. If slipping of the prey is presented, it is trapped by the jaw, withdrawn under the soil and expended. Then, food scraps are thrown out of the hole. The ant lion's hunting way is defined as:

- There is a random movement of the ants in state space which damages ant lion traps.
- As a mark of the highest fitness, the ant lion builds a large pit.
- Fitness of the ant lion is directly proportional to its fitness.
- In every iteration, ant lion can catch hold of an ant.
- There is adaptive decrease in the range of random walks depending upon the movement of ants following ant lions.
- In any case, if an ant fits better than ant lion, then it is buried into the soil by the ant lion.
- As soon as the recent prey is obtained, the reposition is done by ant lion itself and constructs a pit to progress its catching of the next prey.

Random walks of ants:

Ants travel by updating the spots around the search space at every repetition (Mafarja and Mirjalili 2019) as in Eq. (5).

$$X(t) = [0,\ \text{cumsum}(2r(t1)-1),\ \text{cumsum}(2r(t2) - 1), \ldots, \text{cumsum}(2r(tn)-1)] \quad (5)$$

here cumsum signifies the aggregate sum, n denotes maximum iteration, t means the actual iteration, and r (t) means a stochastic function if a random number is less than 0.5 and 0 otherwise.

The random walks as in Eq. (6):

$$X_i^t = \frac{(X_i^t - a_i) \times (d_i - c_i^t)}{(d_i^t - a_i)} + c_i \quad (6)$$

here a_i and c_i represent the primary bound and secondary bound of random walk of ith variable and c_i^t and d_i^t are the lower bound and upper bound ith variable in *t*th iteration.

Trapping in ant lion's pits: Traps of ant lion affect the ants' random movement. Equations (7 and 8) model this assumption:

$$c_i^t = \text{Antlion}_j^t + c^t \tag{7}$$

$$d_i^t = \text{Antlion}_j^t + d^t \tag{8}$$

where c^t and d^t are two vectors including the lower bound and upper bound of entire variables in iteration t, c_i^t, d_i^t represent the lower bound and upper bound of the ith ant, and Antlion_j^t means the position of the jth ant lion at the tth iteration.

Building a trap: The prototype for selection criteria is the hunting capacity of ant lions. If the fitness of ant lion is more, then the probability of catching an ant is high. In this study, Roulette wheel selection (RWS) is used for the selection of ant lions based on their fitness value.

Sliding ants towards ant lion:

The ant tries to escape as it slips into the pit. As soon as the ant lion identifies the existence of prey, the sand is shot through the centre of the pit (Mafarja et al. 2017). There is a decrease in the random walk of the ant as in Eq. (9, 10).

$$c^t = \frac{c^t}{I} \tag{9}$$

$$d^t = \frac{d^t}{I} \tag{10}$$

where c^t and d^t are vectors signifying the lowest and highest of entire variables at iteration t, where I is a ration, in Eq. (11).

$$I = 10^w \frac{t}{T} \tag{11}$$

Here, t represents the present repetition, T implies maximum repetition, a constant Z sets the boundaries depending on present repetition ($Z = 2$ while $t > 0.1$ T, $Z = 3$ while $t > 0.5$ T, $Z = 4$ while $t > 0.75$ T, $Z = 5$ while $t > 0.9$ T, and $Z = 6$ while $t > 0.95$ T).

Holding target along with pit rebuilding: The target ranges the end of the pit at the final stages of hunting and gets trapped in the jaw of ant lion. The ant is dragged under soil by the ant lion and consumes it. The general belief is that the target is trapped as the ant develops fitter than the ant lion. Here, its position is updated based on the ant's position. Equation (12) models this procedure:

$$\text{Antlion}_j^t = \begin{cases} \text{Ant}_i^t & \text{if } f(\text{Ant}_i^t) < f(\text{Antlion}_j^t) \\ \text{Antlion}_j^t & \text{otherwise} \end{cases} \tag{12}$$

Here, t signifies the present repetition, Antlion_j^t represents the position of ant lion j, and Ant_i^t indicates ant i at the repetition t.

Elite Rule: For every iteration, an ant lion with the utmost fitness is dignified elite. The ant lion specified as elite and the ant lion selected, traces significant move of an ant as in Eq. (13):

$$\text{Ant}_i^t = \frac{R_A^t + R_E^t}{2} \tag{13}$$

where R_A^t and R_E^t denote the significant move between the selected ant lion and the elite ant lion, respectively.

The feature space is examined using the ALO adaptively to attain the best classification (Zawbaa et al. 2015). The initialization process gets started with n random preys. Through hunting ants, n random ant lion positions are set. By a Roulette wheel method, an ant is selected for hunting. Random walks are executed around the elite/best ant lion. Then the ant fits to its location depending on the past two random walks iteratively if the ant attains a better fitness. This algorithm is applied iteratively to attain a better solution. The exploration rate is lowered, and optimization development is existed for comprehensive identification.

Identical solutions are illustrated as vector with regular flow along with related features. It is in the range [0, 1]. The continuous fitness valued outcome is threshold to binary illustration represented by an Eq. (14).

$$y_{ij} = \begin{cases} 0 & \text{If } (x_{ij} < 0.5) \\ 1 & \text{Otherwise} \end{cases} \tag{14}$$

where x_{ij} remains the continuous value of i at jth dimension and y_{ij} stands the distinct projection of outcome vector x.

3.5 Support vector machine (SVM) classifier

Vapnik developed SVM which became very famous because of its interesting features and useful empirical efficiency. Structural risk minimization is performed by SVM overriding the conventional empirical risk minimization, which is employed commonly in the traditional neural network. An upper bound on generalization error is minimized by structural risk minimization which is contradicted with empirical reduction in risk minimizing error on training data (Begum et al. 2016). Due to this compromise of SVM, it is decorated through an increased skill to generalize. Two class issues are handled by SVM successfully. Let a dataset (x_i, y_i), i = 1, 2,……,M, where M is the total number of samples, $y_i = \{1, -1\}$. $x_i \in R$, x_i is a p-dimensional real vector. The forced optimization is represented in Eqs. (15, 16):

Minimize

$$1/2||w||^2 + C\sum_{i=1}^{m} \xi_i \tag{15}$$

Such that

$$y_i(w^T x_i + b) \geq 1 - \xi_i \tag{16}$$

where $\xi_i \geq 0$, and $i = 1, 2, \ldots, M$.

ξ_i is the attribute that calculates the degree of misclassification of the sample x_i. Error penalty is determined by the parameter C, which penalizes the nonzero ξ_i. Weight vector is w; hyperplane bias is denoted by b. The SVM search separates hyperplane together with maximum margin and $\xi_i = 0$, only when the data can be separated linearly. The above optimization issue can be solved through the transfer of problem into equivalent Lagrangian problem in Eq. (17):

Minimize

$$L(w, b, \alpha) = 1/2||w||^2 - \sum_{i=1}^{M} \alpha_i y_i (w_i x_i + b) + \sum_{i=1}^{M} \alpha_i \tag{17}$$

Equation (17) can be determined by partial derivatives of L in terms of B (Vanitha et al. 2015) and in terms of W as in Eqs. (18, 19):

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{i=M} \alpha_i y_i x_i = 0 \tag{18}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{i=M} \alpha_i y_i = 0 \tag{19}$$

Now substituting Eqs. (18, 19) into Eq. (17), the quadratic optimization problem is altered as Eqs. (20, 21) that is maximized regarding α subject to Eqs. (18, 19):

Maximize

$$L(\alpha) = \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=0}^{M} \alpha_i \alpha_j y_i y_j x_i x_j \tag{20}$$

$$\sum_{i=1}^{i=M} \alpha_i y_i = 0 \tag{21}$$

where $0 \leq \alpha_i \leq C$.

The Karush–Kuhn–Tucker (KTT) “complementarity” states the solution as in Eq. (22).

$$\alpha_i [y_i (w x_i + b) - 1] = 0 \tag{22}$$

For i, there will be either $\alpha_i^* = 0$ or $y_i (w x_i + b) = 1$. The training data vector xi corresponding to $\alpha_i^* \neq 0$ are called the support vectors (SVs). An optimal extrication hyperplane is shown in Eq. (23):

$$f(x, \alpha_i^*, b^*) = \sum_{i \in sv}^{M} \alpha_i^* y_i (x_i . x) + b^* \tag{23}$$

The decision for testing data vector z is shown in Eq. (24):

$$h(z, \alpha_i^*, b^*) = \mathrm{sgn}\left(\sum_{i}^{M} \alpha_i^* y_i (x_i . z) + b^* \right) \tag{24}$$

4 Results and discussion

Leukaemia dataset is used for the experiment. Dataset with 72 leukaemia samples with 22 ALL, 18 MLL and 26 acute myeloid leukaemia (AML). Total number of genes is 12,582. In this section, the without feature selection, CFS variable selection, ACO variable selection and ALO variable selection methods are used. Table 1 depicts the summary of outcomes. The accuracy, sensitivity, specificity and F-measure are shown in Figs. 1, 2, 3 and 4.

Figure 1 shows that the ALO attribute selection has an accuracy of 14.28%, CAS attribute selection 8.7% and ACO attribute selection 6.89% according to without attribute selection.

Figure 2 shows that the ALO attribute selection is obtained with increased average sensitivity of 14.68%, CFS feature selection 8.94% and ACO feature selection 7.02% according to without feature selection.

Figure 3 shows that the ALO attribute selection provides better average specificity of 7.84%, CFS attribute selection 4.91% and ACO attribute selection 3.78%, according to without attribute selection.

Figure 4 shows that the ALO attribute selection has higher average F-measure of 14.32%, CFS attribute selection 8.64% and ACO attribute selection 6.89%, according to without attribute selection.

5 Conclusion

Expression levels are measured through DNA microarray technology with 1000 genes at the same time, which provides great chance for diagnosis of cancer and prognosis. Tens of thousands are exceeded through the gene count, while the collective number of available elements is often no more than 100. Thus, it is crucial and significant in performing gene selection for the purpose of classification. Prediction accuracy of classifiers is enhanced through a good subset with discriminative genes, and computational cost is saved with decreased dimension of data. This work presents ACO, CAS, and ALO-based attribute selection techniques. The main aim is to maximize the precision of classification and mitigate the number of genes that are informative. Heuristic information for ACO is ALO-based attribute selection algorithm, performance of classifier and length of the feature vector selected. So, optimal feature subset can be chosen without prior knowledge of features. So, CAS filter methods are adopted as pre-processing step for ALO algorithm in improving the speed and classification accuracy presentation. Additionally, so as to remove the genes that are irrelevant and filter the genes that are considered to be noisy which decreases the computational

Table 1 Summary of results

Performance measures	Without attribute selection	CAS attribute selection	ACO attribute selection	ALO attribute selection
Accuracy	78.79	83.33	84.85	90.91
Sensitivity—ALL	0.8636	0.9091	0.8636	0.9545
Sensitivity—MLL	0.7222	0.7778	0.8333	0.8889
Sensitivity—AML	0.7692	0.8077	0.8462	0.8846
Specificity—ALL	0.9167	0.8974	0.925	0.9512
Specificity—MLL	0.8864	0.9318	0.9111	0.9362
Specificity—AML	0.8421	0.8947	0.9189	0.9737
F-measure—ALL	0.8636	0.8696	0.8636	0.9333
F-measure—MLL	0.7222	0.8	0.8108	0.8649
F-measure—AML	0.7692	0.8235	0.8628	0.92

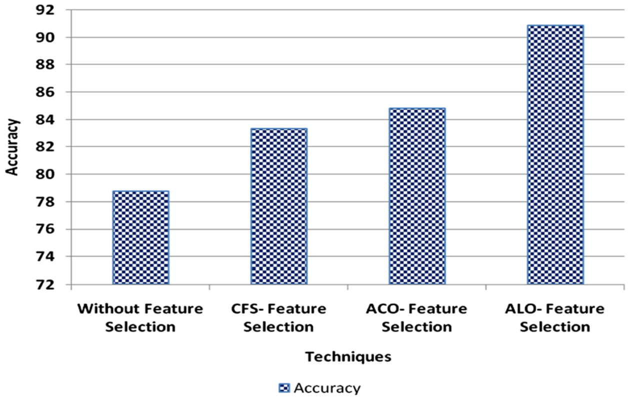

Fig. 1 Accuracy for ALO attribute selection

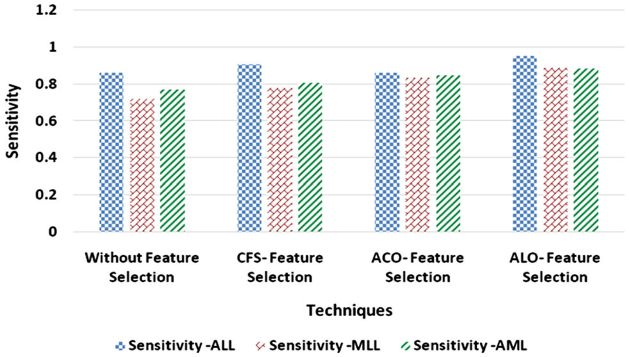

Fig. 2 Sensitivity for ALO attribute selection

complication for ALO algorithm and SVM classifier. It is worthy to mention that SVM-based classifier ALO attribute selection has better precision of 14.28% without feature selection, 8.7% for CAS attribute selection and 6.89% for ACO attribute selection.

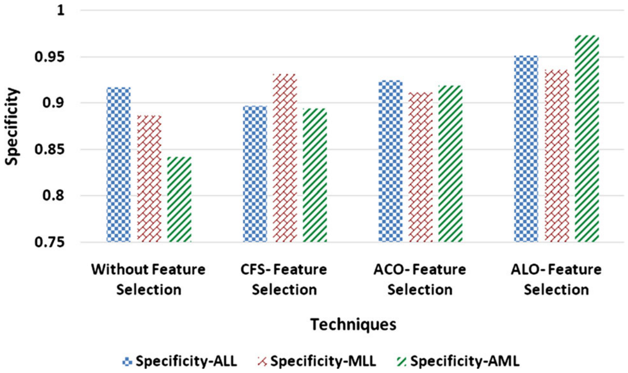

Fig. 3 Specificity for ALO attribute selection

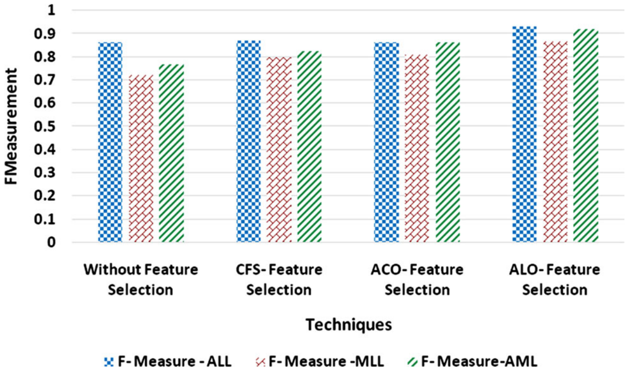

Fig. 4 F measure for ALO attribute selection

Compliance with ethical standards

Conflict of interest All author states that there is no conflict of interest.

Humans and animals rights Humans/animals are not involved in this work.

References

Algamal Z (2017) An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression. Electron J Appl Stat Anal 10(1):242–256

Alshamlan HM (2018) Co-ABC: correlation artificial bee colony algorithm for biomarker gene discovery using gene expression profile. Saudi J Biol Sci 25:895–903

Ang JC, Mirzal A, Haron H, Hamed HNA (2016) Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinform 13(5):971–989

Aziz R, Verma CK, Srivastava N (2017) Dimension reduction methods for microarray data: a review. AIMS Bioeng 4(1):179–197

Babu M, Sarkar K (2016) A comparative study of gene selection methods for cancer classification using microarray data. In: 2016 second international conference on research in computational intelligence and communication networks (ICRCICN). IEEE, pp 204–211

Begum S, Chakraborty D, Sarkar R (2016) Identifying cancer biomarkers from leukemia data using feature selection and supervised learning. In: 2016 IEEE first international conference on control, measurement and instrumentation (CMI). IEEE, pp 249–253

Bhola A, Tiwari AK (2015) Machine learning based approaches for cancer classification using gene expression data. Mach Learn Appl Int J MLAIJ 2(3/4):1–12

Bonilla-Huerta E, Hernández-Montiel A, Morales-Caporal R, Arjona-López M (2016) Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 13(1):12–26

Chandra B, Gupta M (2011) An efficient statistical feature selection approach for classification of gene expression data. J Biomed Inform 44(4):529–535

Chaudhari P, Agarwal H (2018) Improving feature selection using elite breeding QPSO on gene data set for cancer classification. In: Intelligent engineering informatics. Springer, Singapore, pp 209–219

Dorigo M, Di Caro G (1999) Ant colony optimization: a new meta-heuristic. In: Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. no. 99TH8406), vol 2. IEEE, pp 1470–1477

Gao X, Liu X (2018) A novel effective diagnosis model based on optimized least squares support machine for gene microarray. Appl Soft Comput 66:50–59

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene monitoring. Science 286:531–537

Guo S, Guo D, Chen L, Jiang Q (2017) A L1-regularized feature selection method for local dimension reduction on microarray data. Comput Biol Chem 67:92–101

Han F, Yang S, Guan J (2015) An effective hybrid approach of gene selection and classification for microarray data based on clustering and particle swarm optimisation. Int J Data Min Bioinform 13(2):103–121

Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinform. https://doi.org/10.1155/2015/198363

Jain I, Jain VK, Jain R (2018) Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl Soft Comput 62:203–215

Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z (2017) A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 256:56–62

Lv J, Peng Q, Chen X, Sun Z (2016) A multi-objective heuristic algorithm for gene expression microarray data classification. Expert Syst Appl 59:13–19

Mafarja MM, Mirjalili S (2019) Hybrid binary ant lion optimizer with rough set and approximate entropy reducts for feature selection. Soft Comput 23:6249–6265

Mafarja M, Eleyan D, Abdullah S, Mirjalili S (2017) S-shaped vs V-shaped transfer functions for ant lion optimization algorithm in feature selection problem. In: Proceedings of the international conference on future networks and distributed systems. ACM, p 14

Mirjalili S (2015) The ant lion optimizer. Adv Eng Softw 83:80–98

Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, Liang H (2017) Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep 7(1):7402

Pyingkodi M, Thangarajan R (2018) Informative gene selection for cancer classification with microarray data using a metaheuristic framework. Asian Pac J Cancer Prevent APJCP 19(2):561–564

Rasmy MH, El-Beltagy M, Saleh M, Mostafa B (2012) A hybridized approach for feature selection using ant colony optimization and ant-miner for classification. In: 2012 8th international conference on informatics and systems (INFOS). IEEE, pp. BIO-211

Sara VJ, Belina S, Kalaiselvi K (2019) Ant colony optimization (ACO) based feature selection and extreme learning machine (ELM) for chronic kidney disease detection. Int J Adv Stud Sci Res 4(1)

Sharbaf FV, Mosafer S, Moattar MH (2016) A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics 107(6):231–238

Vanitha CDA, Devaraj D, Venkatesulu M (2015) Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Comput Sci 47:13–21

Yao D, Yang J, Zhan X, Zhan X, Xie Z (2015) A novel random forests-based feature selection method for microarray expression data analysis. Int J Data Min Bioinform 13(1):84–101

Zawbaa HM, Emary E, Parv B (2015) Feature selection based on antlion optimization algorithm. In: 2015 third world conference on complex systems (WCCS). IEEE, pp 1–7