Formulating new enhanced pattern classification algorithms based on ACO-SVM

Hiba Basim Alwan and Ku Ruhana Ku-Mahamud

Abstract— This paper presents two algorithms that integrate new Ant Colony Optimization (ACO) variants which are Incremental Continuous Ant Colony Optimization ($IACO_R$) and Incremental Mixed Variable Ant Colony Optimization ($IACO_{MV}$) with Support Vector Machine (SVM) to enhance the performance of SVM. The first algorithm aims to solve SVM model selection problem. ACO originally deals with discrete optimization problem. In applying ACO for solving SVM model selection problem which are continuous variables, there is a need to discretize the continuously value into discrete values. This discretization process would result in loss of some information and hence affects the classification accuracy and seeking time. In this algorithm we propose to solve SVM model selection problem using $IACO_R$ without the need to discretize continuous value for SVM. The second algorithm aims to simultaneously solve SVM model selection problem and selects a small number of features. SVM model selection and selection of suitable and small number of feature subsets must occur simultaneously because error produced from the feature subset selection phase will affect the values of SVM model selection and result in low classification accuracy. In this second algorithm we propose the use of $IACO_{MV}$ to simultaneously solve SVM model selection problem and features subset selection. Ten benchmark datasets were used to evaluate the proposed algorithms. Results showed that the proposed algorithms can enhance the classification accuracy with small size of features subset.

Keywords— Support Vector Machine, Ant Colony Optimization, Incremental Continuous Ant Colony Optimization, Incremental Mixed Variable Ant Colony Optimization, Model Selection, Feature Subset Selection, and Pattern Classification.

I. INTRODUCTION

PATTERN classification attaches the input samples into one of the present number of groups through an classifier approach. The approach is found through learning the training data group [1].

Support Vector Machine (SVM) represents supervised machine learning approaches [33]. SVM is an excellent pattern classification approach built on statistical learning approach [2]. The main concept of SVM is to obtain the Optimal Separating Hyperplane (OSH) between positive and negative samples. This can be done through maximizing the margin between two parallel hyperplanes. Once the plane is found, SVM can forecast the classification of unlabeled

Hiba Basim Alwan is with the School of Computing, College of Art and Science, University Utara Malaysia, 06010 Sintok, Kedah, Malaysia, (e-mail: hiba81basim@yahoo.com).
Ku Ruhana Ku-Mahamud is a Professor with the School of Computing, College of Art and Science, University Utara Malaysia, 06010 Sintok, Kedah, Malaysia, (e-mail: ruhana@uum.edu.my).

sample through identifying the side where the sample lies [3].

Solving SVM model selection problem and selecting an optimal feature subset for SVM classifier are two problems that influence the classification accuracy. These problems affect each other because error produced from the feature subset selection phase will affect the values of the SVM parameters [4].

There is no regular methodology that accepts advance approximation of optimal values for SVM parameter. Currently, almost all SVM researches choose these variables experimentally via searching a bounded number of values and preserving those that supplies the lowest amount of errors. This approach requires a grid search through the area of variable values and requires identifying the range of executable solution and best sampling step. This is a tricky task because best sampling step changes from kernel to kernel and grid range may not be simple to identify without advance knowledge of the problem. Furthermore, when hyper parameter exceeds two of the manual prototype chosen, it may become intractable [5]. Approaches such as trial and error, grid search, cross validation, generalization error estimation and gradient descent can be used to find optimal parameter values for SVM. Evolutionary approaches such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) may also be utilized.

Feature Selection (FS) is a process of determining a subset of fields in database and it minimizes the number of fields that appears during data classification [4], [6], and [7]. The main idea behind FS is to select a subset of input variables by deleting features that contain less or no information [8]. FS aims to decrease the dimension of the initial features group by determining the unauthentic features which would eventually supply the best performance under certain classification dataset [9] and [10], and to delete unrelated, unneeded, or noisy features while preserving the richness of the instructive ones [11] and [12]. FS may be considered as an optimization problem which looks out for potential feature subsets which ultimately determines the optimal one [11].

ACO algorithms have been applied to solve SVM model selection problem. These algorithms work through repetitive creation procedures where each procedure directs a dependent heuristic by intelligently mixing various ideas for exploring and exploiting the seek space. The learning fashions are utilized to construct information to efficiently obtain near optimal solutions. ACO algorithms deal with discrete and continuous variables. ACO that deals with continuous variables is considered as a recent research field [13].

Continuous Ant Colony Optimization (ACO_R) uses probability density function instead of discrete probability distribution to determine the direction that an ant should follow. The main disadvantages of ACO_R are the high running time and the application gap with the-state-of-art continuous solvers. To overcome these disadvantages, the $IACO_R$ Algorithm is introduced [14].

$IACO_{MV}$ algorithm that is adopted in this research is based on [14] and it has been used to simultaneously optimize SVM parameters and feature subset selection. Incremental mixed variable Ant Colony Optimization ($IACO_{MV}$) contains two main parts. The first part is related with optimizing SVM continuous parameter as explained in $IACO_R$ and the second part deals with optimizing the discrete variable and related with selecting suitable feature subset.

The rest of the paper is organized as follows. Section II presents a brief introduction to SVM while Section III and Section IV present $IACO_R$ and $IACO_{MV}$ respectively. Section V reviews some previous studies on solving SVM model selection problem as well as to simultaneously solve SVM model selection problem and features subset. And Section VI describes the proposed algorithms. Experimental results are discussed in Section VII while concluding remarks and future work is presented in Section VIII.

II. SUPPORT VECTOR MACHINE

For binary class classification problem, given M training examples where each example is represented through a tuple (x_i, y_i) where $i = 1, ..., M$, $x_i \in R^M$ corresponds to the feature group for the ith example , and $y_i \in \{+1, -1\}$ denoted the class label, SVM need the solution of the following problem [7] and [15]:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{M} \xi_i \qquad (1)$$

Subject to $\quad y_i(w^T \phi(x_i) + b) + \xi_i - 1, \xi_i \geq 0, i = 1, ..., M$

where C represent the penalty of misclassifying the training instances, ξ_i is the positive slack parameter, and ϕ is the mapping function that maps the training instances from input space to higher dimensional feature space. Often Eq. (1) can be addressed through solving the following dual problem [23]:

$$\min_\alpha F(\alpha) = \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \qquad (2)$$

Subject to $\quad 0 \leq \alpha_i \leq C, i = 1, ..., M$
$$y^T \alpha = 0$$

where e is the vector of all ones and Q is an M by M positive semi-definite matrix. The $(i, j)^{th}$ element of Q is given through:

$$Q_{i,j} = y_i y_j K(x_i, x_j) \qquad (3)$$

$$K(x_i, x_j) \equiv \phi^T(x_i)\phi(x_j) \qquad (4)$$

where $K(x_i, x_j)$ is the kernel function, Radial Basis Function (RBF) is a common one and as follow [23]:

$$K(x_i, x) = exp^{(-\gamma\|x_i - x\|^2)} \qquad (5)$$

$\{\alpha_i\}_{i=1}^{i}$ is Lagrange multipliers, and

$$w = \sum_{i=1}^{M} \alpha_i y_i \phi(x_i) \qquad (6)$$

is the weight vector. The classification decision function is [23]:

$$sgn(w^T \phi(x) + b) = sgn(\sum_{i=1}^{M} \alpha_i y_i K(x_i, x) + b) \qquad (7)$$

For multi-class classification problem, there are two major techniques, which are One-Against-One (OAO) and One-Against-All (OAA). In the OAO technique, numbers of binary SVM classifier are built using the following equation [7]:

$$C_2^v = \frac{v(v-1)}{2} \qquad (8)$$

SVM binary classifiers are built and each classifier is trained utilizing two classes. A new x is classified into the majority class that is voted on through all of the decision functions. While in OAA technique, v binary SVM decision functions are built for an v-class problem. The jth ($j = 1, 2,..., v$) decision function is trained through labelling all of the examples in the jth class with positive labels, and all of the examples that are not in the jth class with negative labels. A new x is classified into the class that has the largest decision function [4].

III. INCREMENTAL CONTINUOUS ANT COLONY OPTIMIZATION

Incremental continuous Ant Colony Optimization ($IACO_R$) proposed by [14] and it is built on enhanced ACO_R [13]. It starts with a small size for solution archive defined by a parameter *InitArhiveSize*. This solution archive will be filled with initial solutions which are randomly generated. $IACO_R$ also characterizes a strategy to alternate from the one utilized in ACO_R for choosing the solution that directs the creation of new solutions. The new procedure built on probability parameter $p \in [0, 1]$, which monitors the probability of utilizing just the best solution in the archive as a directing solution. With a probability $1 - p$, all the solutions in the archive are utilized to create new solutions. Once a directing solution is chosen, and a new one is created exactly the same way as in ACO_R, they are compared according to their objective function. The newly created solution will replace the directing solution in the archive if it is better. This replacement mechanism is differing from the one utilized in ACO_R in which all solutions in the archive and all the newly created ones compete. A new solution is appended to them in every growth iterations until a maximum archives sizes, defined by *MaxArchiveSize*, is reached. A parameter *Growth* monitors the percentage at which the archives grow. Fast growth percentage supports seek diversification while slow growth supports intensification. Each time a new solution is appended, it is initialized by using the information from the best solution in the archives. First, a new solution S_{new} is created fully in an arbitrary way and then it is moved in the direction of the best solution in the archives S_{best} utilizing the following formula:

$$\bar{S}_{new} = S_{new} + rand(0,1)(S_{best} - S_{new}) \qquad (9)$$

where *rand(0, 1)* is an arbitrary umber in range [0, 1].

$IACO_R$ involves an algorithm-level diversification strategy to avoid stagnation. The strategy includes restarting the algorithm and initializing the new initial archive with the best-so-far solution. The restart condition is the number of successive iterations, *MaxStagIter*, with a relative solution improvement lower than a certain threshold. An outline of $IACO_R$ is given in Figure 1 [14].

Input: *p*, *InitArhiveSize*, *Growth*, *MaxArchiveSize*, *MaxStagIter*, no. of ants, and Termination criterion
Output: Optimal Value for *C* and *γ*
k = InitArhiveSize
initialize *k* solutions and evaluate it
while Termination criterion not satisfied do
   // Generate new solutions
   if *rand (0,1) < p* then
      for *i = 1* to no. of ants do
         Select best solution
         Sample best selected solution
         if Newly generated solution is better than S_{best} then
            Substitute newly generated solution for S_{best}
         end
      end
   else
      for *j = 1* to *k* do
         Select *S* according to its weight
         Sample selected *S*
         Store and evaluate newly generate solutions
         if Newly generated solution is better than S_j then
            Substitute newly generated solution for S_j
         end
      end
   end
   // Archive Growth
   if current iterations are multiple of *Growth* & *k < MaxArchiveSize* then
      Initialize new solution
      Add new solution to the archive
      k + +
   end
   // Restart Mechanism
   if # (number) of iterations without improving S_{best} = *MaxStagIter* then
      Re-initialize *T* (solution archive) but keeping S_{best}
   end
end

Figure 1: IACO$_R$ Algorithm

IV. INCREMENTAL MIXED-VARIABLE ANT COLONY OPTIMIZATION

Incremental mixed variable ant colony optimization (IACO$_{MV}$) is based on [14] aims to enhance ACO$_{MV}$ introduced by [16]. This variant is similar to ACO$_{MV}$ except for the part related with optimizing continuous variable. In incremental mixed variable ant colony optimization, IACO$_R$ is used to optimize continuous variable instead of using ACO$_R$ as in ACO$_{MV}$.

While for optimizing discrete variable the same procedure used in ACO$_{MV}$ will be used in IACO$_{MV}$ which will be as follows. In standard ACO, solutions are constructed from solution components using a probabilistic rule based on the pheromone values. However, there is no static pheromone value, but only a solution archive in ACO$_{MV}$. As in standard ACO, the construction of solutions for discrete variables is done by choosing the components, that is, the values for each of the discrete decision variables. However, since the static

pheromone values of standard ACO are replaced by the solution archive, the actual probabilistic rule used has to be modified. Similarly to the case of continuous variables, each ant constructs the discrete part of the solution incrementally. For each discrete variable, each ant will choose one of the solution components C^i based on a probability. The probability of choosing the lth value is given by:

$$o_l^i = \frac{w_l}{\sum_{r=1}^{c} w_r} \tag{10}$$

where w_l is the weight associated with the lth available value. It is calculated based on the weights w and some additional parameters:

$$w_l = \frac{w_{jl}}{u_l^i} + \frac{q}{\eta} \tag{11}$$

$$w_{jl} = \frac{1}{qk\sqrt{2\pi}} e^{-\frac{(l-1)^2}{2q^2k^2}} \tag{12}$$

where k is the size of solution archive, and q is the algorithm's parameter to control diversification of search process.

The final weight, w_l, is hence a sum of two components. The weight, w_{jl}, is calculated according to Eq. 12, where the j_l is the index of the highest quality solution that uses value v_l^i for the ith variable. In turn, u_l^i is the number of solutions using the value v_l^i for the ith variable in the archive. Therefore, the more popular the value v_l^i is, the lower is its final weight. The second component is a fixed value (i.e., it does not depend on the value v_l^i had chosen): η is the number of values v_l^i from the c_i available that are unused by the solutions in the archive, and q is the same parameter of the algorithm that was used in Eq. 12.

V. SVM PREVIOUS STUDIES

Modified Particle Swarm Optimization (PSO) was utilized in [17] to solve SVM model selection problem. The modified PSO made through virtue of chaotic motion with sensitive count on initial conditions and ergodicity (CPSO) and the error of k-fold CV is utilized as the objective function of PSO. The authors used SVM with Gauss kernel (GSVM) and SVM with Wavelet kernel (WSVM). The performance of GSVM and WSVM is compared against BackPropagation Neural Network (BPNN) with four neurons in the input layer, ten neurons in the hidden layer, and one neuron in the output layer with TANSIG and PURELIN as activation function of hidden and output layer respectively and gradient descent as a training algorithm for BPNN. The experiments results show that WSVM has fast convergence speed and high generalization capability compared to GSVM and comparing against BPNN, SVM has simple construction, fast convergence speed with high generalization capability. Another study also used PSO to solve SVM model selection problem was proposed by [18]. In this work, the authors utilize standard type of PSO and each value of the particle was rounded to one decimal place where the parameter range was the same as in the grid search model selection. GA that works with real parameters to tune SVM parameters is also implemented. The authors compared their work with grid search algorithm. In order to evaluate their work, they examine their work on two datasets, which are, letter recognition database and speech dataset for speaker recognition process and utilized five-fold CV. Both PSO and GA seek for the optimal value for regularization parameter C

and RBF kernel SVM parameters in 20 iterations. The result showed that both PSO and GA were faster than grid search in model selection and both techniques present comparable results and their capability to optimize more than two parameters. The result for letter dataset show that PSO and GA obtain classification accuracy as good as grid search, while in case of speaker dataset, the result show that PSO can obtain good parameters and get classification accuracy five times better and faster than GA.

A novel productivity model selection strategy for SVM through kernel matrix approximation centre on productivity enhanced by SVM training was presented by [19]. This approach works as follow, first, kernel matrix approximation algorithm was introduced through utilized Monte Carlo to unsystematically sample the kernel matrix and performs the incomplete Cholesky factorization to get the low rank approximation of the sample matrix. After that, MOCIC was applied to calculate a low rank approximation of the kernel matrix, and utilize the approximate matrix to productively solve the convex quadratic programming of SVM and chose the optimal parameter via the Approximate Cross Validation Error (ACVE). The authors examined the feasibility and the efficiency of model selection on five University California, Irvin (UCI) datasets and the results show that the sampling size decrease within a certain range, the execution time falls sharply and the alternations of test set accuracy is omitted. As the future work, the authors suggest to build a complete approximate model selection theory which has close fitting approximation mistake range and can be directly or indirectly performed in the design of adaptive approximate model selection algorithms. The integration of GA with gradient descent method to build hybridized technique to select the optimal value for RBF and C parameters was proposed by [20]. This technique first selects the best chromosome as initial solution then utilizing gradient descent approach to seek for optimal kernel parameter for this best chromosome as its iterative solution. After examining a new generation, the new best chromosome was taken from three nominees: the best solution in the past generation, the best solution in the new generation, and the iterative enhanced version of the past best solution. The authors evaluated their approach on thirteen datasets and compared their approach with simple GA-SVM, 5-fold CV, radius margin bound (R-M bound), span bound and Adankon's approach and the results show that this approach produced better results compared with other approaches. The authors suggest examining their approach on more real applications and applying it with other kernel functions as well as use some heuristic mechanisms to speed up their approach.

Several studies have hybrid systems to enhance classification accuracy by using few and suitable feature subsets [2], [4], [7], [21], [22], [23], [24], [25], [26], [27], and [28]. In these studies, feature subset and SVM parameters (C and γ RBF kernel) variable are simultaneously optimized. SVM is then used to measure the quality of the solution for all the hybrid systems. However, each hybrid is different depending on the system which it is based on [4] and [21] proposed the use of a hybrid system which is based on GA and SVM. GA is employed to select suitable feature simultaneously with optimize SVM parameter which were represented in the encoded chromosomes. The authors in [23] and [24] on the other hand, chose to use a hybrid system which is based on Particle Swarm Optimization (PSO) and SVM. In [24], discrete and continuous PSO values are mixed to simultaneously select suitable feature and optimize SVM parameter while [25] used SA to simultaneously optimize model selection and features subset selection. Hide-and-Seek SA, which is a variant of SA is used in [25] to optimize the continuous values of SVM parameters and the features are represented as discrete values.

Utilization of Bees algorithm to simultaneously select the best combination of feature subset and SVM parameters values for the process of classifying faults in wood layer pieces was proposed in [22]. A hybrid system which is based on ACO and SVM was proposed by [7]. The classical ACO has been employed to simultaneously select suitable feature and optimize SVM parameter. A hybrid system which is based on Clonal Selection Algorithm (CSA) and SVM is proposed in [26]. CSA is then used to select suitable feature simultaneously and optimize SVM parameter. A hybrid system which is based on Cat Swarm Optimization (CSO) and SVM where CSO is employed to select suitable feature simultaneously with optimize SVM parameter is discussed in [27]. Two versions of Gravitational Search Algorithm (GSA) which are real value GSA (RGSA) to optimize the real value of SVM parameters and binary (discrete) value GSA (BGSA) to select features subset have been reported in [28]. GSA is considered as swarm based metaheuristic seek approach built on gravity's law and motion and it is derived from the Newtonian gravity. This concept has been applied on binary class classification problem and not on multi-class classification problems [28].

In conclusion, the above mentioned hybrid studies produced good results for classification accuracy with few numbers of selected features. Suggestions that were highlighted are as follows: the authors in [4] and [7] suggested applying their work on Support Vector Regression (SVR), because SVR accuracy counts mainly on SVR parameters and selected feature subset. While the authors in [4], [7], [23], and [24] suggested in using other types of kernel function besides RBF. Authors in [21], [23], [26], and [27] suggested applying their works on other real world problem. Finally, the authors in [7] suggested the use of continuous ACO to optimize the continuous value of SVM parameters.

VI. METHODOLOGY

The first proposed algorithm utilizes $IACO_R$ to optimize only SVM classifier parameters. The methodology for this proposed algorithm follows same methodology proposed in [32] and [35]. An ant's solution is used to represent a combination of the classifier parameters, C and γ, based on the Radial Basis Function (RBF) kernel of the SVM classifier. The classification accuracy of the built SVM classifier is utilized to direct the updating of solution archives. Based on the solution archive, the transition probability is computed to choose a solution path for an ant. In implementing the proposed scheme, this study utilizes the RBF kernel function for SVM classifier because of its capability to manage high dimensional data, good performance in major cases, and it

only needs to use one parameter, which is kernel parameter gamma (γ) [29-32]. The main steps are (1) selecting feature subset (2) initializing solution archive and algorithm parameters, (3) solution construction for C and γ, and (4) establishing SVM classifier model.

In features subset selection, F-score is used as a measurement to determine feature importance. This measurement is used to judge the favouritism capability of a feature. High value of F-score indicates favourable feature. The calculation of F-score is as follow [7]:

$$F - Score_i = \frac{\sum_{c=1}^{v}\left(\bar{x}_i^{(c)} - \bar{x}_i\right)^2}{\sum_{c=1}^{v}\left(\frac{1}{N_i^{(c)}-1}\sum_{j=1}^{N_i^{(c)}}\left(x_{i,j}^{(c)} - \bar{x}_i^{(c)}\right)^2\right)} \qquad (13)$$

where $i = 1, 2, ..., N_f$, v is the number of categories of target variable, N_f is the number of features, $N_i^{(c)}$ is the number of samples of the ith feature with categorical value c, $c \in \{1, 2, ..., v\}$, $\bar{x}_{i,j}^{(c)}$ is the jth training sample for the ith feature with categorical value c, $j \in \{1, 2, ..., N_i^{(c)}\}$, $\bar{x}_i$ is the ith feature, and $\bar{x}_i^{(c)}$ is the ith feature with categorical value c.

After computing F-score for each feature in the dataset, average F-score will be computed and it will be considered as threshold for choosing a feature in the feature subset. Feature with F-score equal to or greater than the threshold will be chosen and put in the feature subset. This subset will be presented to SVM.

In the initialization step, each ant established a solution path for parameter C and parameter γ. Two solution archives are needed to design the transition probabilities for C and γ. The range for C and γ values will be sampled according to random parameter k which is the initial archive size of solutions archives. The weight vector, w is then computed for each sample for C and γ as follow:

$$w_l = \frac{1}{qk\sqrt{2\pi}} e^{-\frac{(l-1)^2}{2q^2k^2}} \qquad (14)$$

where q is the algorithm's parameter to control diversification of search process. These values will be stored in solution archives. Once this step is completed, the sampling procedure will be constructed in two phases. Phase one involves choosing one of the weight vectors as following:

$$p_l = \frac{w_l}{\sum_{r=1}^{k} w_r} \qquad (15)$$

The second phase involves sampling selecting w via a random number generator according to a parameterized normal distribution.

This initialization will construct the transition probabilities. Like the solution archives, some important system parameters were set to such value as follows: the number of ants = 2, q = 0.1, *initial archive size* = 10, *Growth* = 5, *maximum archive size* = 15, *MaxStagIter* = 2, number of runs = 10, C range $\in [2^{-1}, 2^{12}]$ and $\gamma \in [2^{-12}, 2^2]$.

The third step is related to solution construction where each ant builds its own solution. This solution will be a combination of C and γ. In order to construct the solution, two transition probabilities with various solutions archives are needed. These transitions will be computed according to Eq. (14) and Eq. (15).

Classifier model will be constructed in step four. Solution is generated by each ant and will be evaluated based on classification accuracy obtained by SVM model utilizing k-fold CV with the training set. In k-fold CV, training data group is partitioned into k subgroups, and the holdout approach is repeated k times. One of the k sub-groups is utilized as the test set and the remaining $k-1$ subgroups are combined to construct the training group. The average errors along with all the k trails are calculated. CV accuracy is calculated as follows:

$$CV_{accuracy} = \frac{\sum_i test\_accuracy}{k}, i = 1, 2, ..., k \qquad (16)$$

Test accuracy is used to evaluate the percentage of samples that are classified in the right way to determine k-folds and it will be computed as follows:

$$Test\ Accuracy = \frac{no.of\ correctly\ predicted\ data}{total\ testing\ data} * 100\% \qquad (17)$$

The benefits of using CV are (1) each of the test groups is independent and (2) the dependent outcomes can be enhanced [7].

The second proposed algorithm has adopted $IACO_{MV}$ to optimize features subset selection and SVM classifier parameters. The methodology for this proposed algorithm flows same methodology proposed in [31] and [32]. An ant's solution is used to represent a combination of features subset and the classifier parameters, C and γ, and not only SVM parameters as in first proposed algorithm stated above. $IACO_{MV}$ is like $IACO_R$ in part of optimizing SVM model, but is different during the selection of feature subset. $IACO_{MV}$ will not use filter technique to select feature subset as in $IACO_R$, but the ant will select suitable feature subset. The ant solution therefore will include three parts; the first two parts are for C and γ while the third part is for feature subset. $IACO_{MV}$ will start with initializing three solution archives and these solution archives are needed to design the transition probabilities for first feature in the features subset, C and γ, and one pheromone table. The process for the ant to select a suitable feature subset begins with computing the F-score for each feature according to Eq. (1) and then computes the weight vector w for each feature as follows:

$$w_f = \frac{w_l}{u} + \frac{q}{\eta} \qquad (18)$$

where u is the number of *feature$_i$* is selected, η is the number of none selected features, and q in both equations is the algorithm's parameter to control diversification of search process. These values will be stored in the solution archive for features. In order to select other features that construct the features subset, the following probability transition is used:

$$Prob_{ij}^k = \begin{cases} \frac{(Prob_{ij})^\alpha (F-Score_j)^\beta}{\sum_{j\in I_i^k}(Prob_{ij})^\alpha (F-Score_j)^\beta} & if\ j \in I_i^k \\ 0 & otherwise \end{cases} \qquad (19)$$

Similarly $IACO_R$ parameters are used in $IACO_{MV}$ with the same values. Two additional parameters that appear in $IACO_{MV}$ are α which is set to 1 and β which is set to 2.

In order to construct the transition probability for the ant to construct feature subset, the solution archive for feature as well as the pheromone table is needed. The transition probability for feature will be computed according to Eq. (6).

The pheromone table will populated using the equations below::

$$T_{ij}(t + 1) = pT_{ij} + \sum_{k=1}^{m} \Delta T_{ij}^k(t) \qquad (20)$$

$$\Delta T_{ij}^k = \begin{cases} CVACC^k * Weight_i^k * Weight_j^k & if\ ant\ k\ use\ edge\ (i,j) \\ 0 & otherwise \end{cases} \qquad (21)$$

VII. RESULTS

Ten datasets were used in evaluating the proposed $IACO_R/IACO_{MV}$-SVM algorithms. The datasets are Australian, German, Heart, Image Segmentation, Ionosphere, Iris, Pima-Indian Diabetes, Sonar, Splice, and Vehicle datasets, available from UCI Repository of Machine Learning Databases [34]. The summary of these datasets are presented in Table 1.

Table 1 Summarization of UCI's Datasets

Datasets	No. of Instances	No. of Features	No. of Classes	Features' Type
Australian	690	14	2	Categorical, Integer, Real
German	1000	20	2	Categorical, Integer
Heart	270	13	2	Categorical, Real
Image	2310	19	7	Real
Ionospher	351	34	2	Integer, Real
Iris	150	4	3	Real
Diabetes	768	8	2	Integer, Real
Sonar	208	60	2	Real
Splice	3190	61	3	Categorical
Vehicle	846	18	4	Integer

All input variables were scaled during data pre-processing phase to avoid features with higher numerical ranges from dominating those in lower numerical ranges and also to reduce the computational effort. The following formula was used to linearly scale each feature to [0, 1] range.

$$\bar{x} = \frac{x - min_i}{max_i - min_i} \tag{22}$$

where x is the original value, $\bar{x}$ is the scaled value, and max_i and min_i are the maximum and minimum values of $feature_i$, as stated by [7].

Each dataset is randomly re-arranged and divided into ten approximately equal size subsets, one subset as testing set and the remaining as training sets and repeated ten times. The performance of the proposed $IACO_R/IACO_{MV}$-SVM was compared with GA $_{with\ feature\ chromosome}$-SVM [21], SA-SVM [25], and GA-SVM with feature selection [4]/ GA-SVM without feature selection [23].

C programming language was used to implement the $IACO_R/IACO_{MV}$-SVM. Experiments were performed on Intel(R) Core (TM) 2 Duo CPU T5750, running at 2.00 GH_Z with 4.00 GB RAM and 32-bit operating system.

Table 2 shows the average classification accuracy that was produced in all the ten runs. The classification accuracy of classify pattern of the proposed $IACO_R/IACO_{MV}$-SVM algorithms is compared with the classification accuracy by the hybrid systems mentioned above. The proposed algorithms classify patterns with higher accuracy compared to the GA $_{with\ feature\ chromosome}$-SVM, SA-SVM, and GA-SVM with feature selection / GA-SVM without feature selection in eight datasets. This is because the proposed algorithms handle directly the continuous value of SVM parameters without the need to discretize it. While in one dataset (Ionosphere dataset) GA $_{with\ feature\ chromosome}$-SVM was slightly better than the proposed $IACO_R$-SVM algorithm and in other dataset (Iris dataset) GA $_{with\ feature\ chromosome}$-SVM and SA-SVM with feature selection were also slightly better than the proposed algorithms of $IACO_R/IACO_{MV}$-SVM, because the proposed algorithm of $IACO_R$-SVM optimized only SVM model selection while GA $_{with\ feature\ chromosome}$-SVM and SA-SVM with feature selection optimized SVM model selection and feature subset selection. Also, the performance of the proposed algorithm $IACO_{MV}$-SVM was better than the proposed algorithm $IACO_R$-SVM. This is because the proposed algorithm $IACO_{MV}$-SVM simultaneously optimizes feature subset selection and model selection for SVM.

Table 3 shows the average selected features subset size that was produced in all the ten runs. The proposed $IACO_R$-SVM algorithm produced lower average number of selected features in nine datasets when compared with GA $_{without\ feature\ chromosome}$-SVM, SA-SVM without feature selection , and GA-SVM without feature selection , while in one dataset (Iris dataset) GA $_{without\ feature\ chromosome}$-SVM was slightly better than the proposed algorithm $IACO_R$-SVM. Also, the proposed algorithm $IACO_{MV}$-SVM produced lower average number of selected features in eight datasets when compared with GA $_{with\ feature\ chromosome}$-SVM, SA-SVM with feature selection, and GA-SVM with feature selection, while in two datasets (Ionosphere Iris datasets) the proposed algorithm $IACO_{MV}$-SVM produced same feature subset size comparing with GA -SVM with feature selection. However, the proposed $IACO_{MV}$-SVM algorithm produced lower feature subset size when compared with the proposed $IACO_R$-SVM algorithm. This is because the proposed $IACO_{MV}$-SVM algorithm is based on wrapped feature technique that depends on the inductive learning approach which has the ability to integrate the classification accuracy and necessary features introduced to the SVM classifier from $IACO_{MV}$ algorithm. The biggest reduction in number of features for $IACO_{MV}$-SVM was 87% for Australian dataset while the smallest number of feature reduction was 75% for Diabetes dataset. For $IACO_R$-SVM, the biggest reduction in number of features was 76% for Australian dataset while the smallest number of feature reduction was 54% for Heart dataset.

Table 2 Classification Accuracy %

Dataset	IACO_R-SVM	GA-SVM without Feature chromosome	SA-SVM without feature selection	GA-SVM without feature selection	IACO_MV-SVM	GA-SVM with Feature chromosome	SA-SVM with feature selection	GA-SVM with feature selection
Australian	94.62 ± 0.52	86.81 ± 3.64	88.34	88.09	96.96 ± 0.53	91.59 ± 2.14	92.19 ± 3.23	88.10 ± 2.25
German	91.56 ± 0.1	80.80 ± 2.10	-	84.24	97.23 ± 0.46	86.10 ± 1.97	-	85.60 ± 1.96
Heart	96.28 ± 0.21	91.11 ± 2.58	-	-	98.01 ± 0.35	95.56 ± 2.34	-	94.80 ± 3.32
Segmentation	97.29 ± 0.08	-	-	-	98.96 ±0.41	98.12	-	-
Ionosphere	98.90 ± 0.52	98.57 ± 2.02	97.50	96.61	99.99 ± 0.02	99.43 ± 1.21	99.07 ± 0.73	98.56 ± 2.03
Iris	99.86 ± 0.29	96.00 ± 3.44	-	97.56	99.98 ± 0.29	100 ± 0	-	100 ± 0
Diabetes	91.25 ± 0.13	81.97 ± 5.34	80.19	82.98	97.22 ± 0.81	83.84 ± 5.14	82.22 ± 3.55	81.50 ± 7.13
Sonar	98.92 ± 0.08	95.00 ± 2.36	91.85	95.22	99.99 ± 0.02	99.00 ± 2.11	95.99 ± 3.90	98.00 ± 3.5
Splice	96.76 ± 0.48	-	-	-	98.65 ±0.55	90.53	-	-
Vehicle	93.20 ± 0.18	84.74 ± 2.32	88.76	85.87	93.92 ± 0.29	88.24 ± 1.47	90.4 ± 2.21	84.06 ± 3.54

Table 3 Average Feature Subset Size

Dataset	IACO_R-SVM	GA-SVM without Feature chromosome	SA-SVM without feature selection	GA-SVM without feature selection	IACO_MV-SVM	GA-SVM with Feature chromosome	SA-SVM with feature selection	GA-SVM with feature selection
Australian	3.3 ± 0.46	6.7 ± 3.16	15	14	1.8 ± 0.4	5.2 ± 2.15	10.60 ± 2.46	3 ± 2.45
German	6.4 ± 0.63	11.80 ± 3.33	-	24	3.9 ± 0.3	10.3 ± 1.76	-	13 ± 1.83
Heart	6.0 ± 0.45	7 ± 1.05	-	-	2 ± 0	6.2 ± 1.12	-	5.4 ± 1.85
Segmentation	8.3 ± 0.46	-	-	-	3 ± 0	18	-	-
Ionosphere	10.6 ± 1.74	15.4 ± 3.32	34	34	6 ± 0	13.9 ± 3.45	22.36 ± 5.02	6 ± 0
Iris	2 ± 0	1.8 ± 0.38	-	4	1 ± 0	1.2 ± 0.28	-	1 ± 0
Diabetes	2.6 ± 0.49	5.1 ± 1.63	8	8	2 ± 0	3.7 ± 1.26	5.86 ± 1.17	3.7 ± 0.95
Sonar	21.7 ± 1.42	28.7 ± 4	60	60	12 ± 0	26.4 ± 3.20	48.18 ± 45.53	15.5 ± 1.1
Splice	11.7 ± 1	-	-	-	9.5 ± 4.18	61	-	-
Vehicle	9.3 ± 1	10.3 ± 2.72	18	18	2.8 ± 0.4	9.2 ± 1.71	15.56 ± 2.17	9.2 ± 1.4

VIII. CONCLUSION AND FUTURE WORK

This study has investigated the hybridization of $IACO_R/IACO_{MV}$ and SVM technique to obtain optimal model parameters as well as to simultaneously optimize SVM parameters and features subset. Experimental results on five public UCI datasets showed promising performance in terms of test accuracy and features subset size. Possible extensions can focus on the area where other kernel parameters besides RBF, application to other SVM variants and multiclass.

REFERENCES

[1] X. Wang, X. Liu, W. Pedrycz, X. Zhu, & G. Hu, "Mining axiomatic fuzzy association rules for classification problems," *European Journal of Operational Research,* vol. 218, pp. 202-210, 2012.

[2] W. Liu & D. Zhang, "Feature subset selection based on improved discrete particle swarm and support vector machine algorithm," *Paper presented at the IEEE International Conference of the Information Engineering and Computer Science, (2009),* (pp.1-4).

[3] V. Vapnik &A. Vashist, "A new learning paradigm: learning using privileged information," *Neural networks: the official journal of the International Neural Network Society,* vol. 22, no.5-6, pp.544–557, 2009.

[4] C. Huang & C. Wang, "A GA-based feature selection and parameters optimization for support vector machines," *Expert system with Applications,* vol.31, no.2, pp.231-240, 2006.

[5] N. Ayat, M. Cheriet, & C. Suen, "Automatic model selection for the optimization of SVM kernels," *Pattern Recognition*, vol.38, no.10, pp.1733-1745, 2005.

[6] R. Sivagaminathan and S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 49-60, 2007.

[7] C. Huang, "ACO-based hybrid classification system with feature subset selection and model parameters optimization," *Neurocomputing*, vol.73, no.1-3, pp.438-448, 2009.

[8] S. Vieira, J. Sousa, & T. Runkler, "Ant colony optimization applied to feature selection in fuzzy classifiers," *Paper Presented at IEEE International Conference on Fuzzy System 2008, (pp. 778-788).*

[9] M. Kabir, M. Shahjahan, & K. Murase K., "An efficient feature selection using ant colony optimization algorithm", In C. Leung & J. Chan (Eds.), *Neural Information Processing*, (pp. 242-252). Springer-Verlag: Berlin Heidelberg, 2009.

[10] M. Kabir, M. Shahjahan, K. Murase, "A new hybrid ant colony optimization algorithm for feature selection", *Expert Systems with Applications*, vol. 39, no. 3, pp. 3747-3763, 2012.

[11] N. Abd-Alsabour, "Feature selection for classification using an ant system approach," *Distributed, Parallel and Biologically Inspired Systems, IFIP Advances in Information and Communication Technology*, vol. 329, pp 233-241, 2010.

[12] S. Sanz, M. Cumplido, F. Cruz, & C. Calzon, "Feature selection via genetic optimization," In J. Dorronsoro (Ed.), *Artificial Neural Networks*, (pp. 547–552). Springer-Verlag: Berlin Heidelberg, 2002.

[13] K. Socha & M. Dorigo, "Ant colony optimization for continuous domain," *European Journal of Operational Research,* vol.185, no.3, pp.1155-1173, 2008.

[14] T. Liao, M. Dorigo, & T. Stutzle, "Improved ant colony optimization algorithms for continuous and mixed discrete-continuous optimization

problems," Retrieved from http:// http://www.swarmbots.org/~mdorigo/HomePageDorigo/thesis/dea/Liao MAS.pdf, 2011.

[15] R. Fan, P. Chen, C. Lin, "Working set selection using second order information for training support vector machines," *Machine Learning Research*, vol. 6, pp. 1889-1918, 2005.

[16] K. Socha, "Ant colony optimization for continuous and mixed-variables domain," Doctoral dissertation, Universite' Libre de Bruxelles, 2008, Retrieved from: *iridia.ulb.ac.be/~mdorigo/HomePageDorigo/thesis/SochaPhD.pdf*, 2010.

[17] S. Wang & B. Meng, "Parameter selection algorithm for support vector machine," *Procedia Environmental Sciences*, vol. 11, pp. 538-544, 2011.

[18] M. Hric, M. Chmulik, & R. Jarina, "Model parameters selection for SVM classification using particle swarm optimization," *Paper presented at 21st IEEE International Conference on Radioelektronika (RADIOELEKTRONIKA)*, 2011, pp. 1-4, Brno.

[19] S. Liao, C. Yang, & L. Ding, "Approximate parameter tuning of support vector machine," *Paper presented at 3rd IEEE International Workshop on Intelligent Systems and Applications (ISA)*, 2011, pp. 1-4, Wuhan.

[20] X. Zhou J. Xu, "A SVM model selection method based on hybrid genetic algorithm and empirical error minimization criterion," In H. Wang et al. (Eds.), *The Sixth ISNN 2009, AISC 56*, (pp. 245–253). Springer-Verlag: Berlin Heidelberg, 2002.

[21] M. Zhao, C. Fu, L. Ji, K. Tang, & M. Zhou, "Feature selection and parameter optimization for support vector machines: a new approach based on genetic algorithm with feature chromosomes," E*xpert System with Applications,* vol. 38, no. 5, pp. 5197-5204, 2011.

[22] D. Pham, Z. Muhamad, M. Mahmuddin, A. Ghanbarzadeh, E. Koc, & S. Otri, "Using the bees algorithm to optimise a support vector machine for wood defect classification,". *Paper presented at Innovative Production Machines and Systems Virtual Conference,* 2007.

[23] S. Lin, K. Ying, S. Chen, & Z. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert System with Applications*, vol. 35, no. 4, pp. 1817-1824, 2008.

[24] C. Huang & J. Dun, "A distributed PSO-SVM hybrid system with feature selection and parameter optimization," *Applied Soft Computing*, vol. 8, no. 4, pp. 1381-1391, 2008.

[25] S. Lin, Z. Lee, S. Chen, T. Tseng, "Parameter determination of support vector machine and feature selection using simulated annealing approach," *Applied Soft Computing*, vol. 8, no. 4, pp. 1505-1512, 2008.

[26] S. Ding & S. Li, "Clonal Selection Algorithm for Feature Selection and Parameters Optimization of Support Vector Machines," *Paper presented at 2nd International Symposium IEEE on knowledge acquisition and modeling 2009, (pp. 17-20).*Yichang.

[27] K. Lin & H. Chien, "CSO-based feature selection and parameter optimization for support vector machine," *Paper presented at IEEE pervasive computing (JCPC)* 2009*, (pp. 783-788).* Tamsui, Taipei.

[28] S. Sarafrazi & H. Nezamabadi-Pour, "Facing the classification of binary problems with a GSA-SVM hybrid system," *Mathematical and Computer Modeling*, in press.

[29] Moustakidis S. & Theocharis J., SVM-FuzCoC: A Novel SVM-Based Feature Selection Method Using a Fuzzy Complementary Criterion, *Pattern Recognition*, Vol. 43, No. 11, 2010, pp. 3712-3729.

[30] Zhang H., Xiang M., Ma C., Huang Q., Li W., Xie W., Wei Y., Yang S., Three-Class Classification Models of LogS and LogP Derived by Using GA-CG-SVM Approach, *Molecular Diversity*, Vol. 13, No. 2, 2009, pp. 261-268.

[31] Alwan, H.B & Ku-Mahamud, K.R., Incremental Mixed Variable Ant Colony Optimization for Feature Subset Selection and Model Selection. *Paper Presented at Proceedings the 12th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED '13)*, 2012, pp. 131-136. Cambridge, UK.

[32] Alwan, H.B & Ku-Mahamud, K.R., Hybrid Pattern Classification Algorithm based on ACO-SVM. *Paper Presented at Proceedings the 13th WSEAS International Conference on Applied Computer Science (ACS '13)*, 2013, pp. 140-145. Morioka City, Iwate, Japan.

[33] T. Barbu, "SVM-based human cell detection technique using histograms of oriented gradients," *Paper Presented at Proceedings the 17th WSEAS International Conference on Applied Mathematics (AMATH '12)*, 2012, pp. 156-160. Montreux, Switzerland.

[34] *UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California, Irvine, CA, <http://www.ics.uci.edu/mlearn/MLRepository>, 2012.*

[35] Alwan, H.B & Ku-Mahamud, K.R., Incremental Continuous Ant Colony Optimization Technique for Support Vector Machine Model Selection Problem, *Paper Presented at Proceedings the 17th WSEAS International Conference on Applied Mathematics (AMATH '12),* 2012, pp. 165-170. Montreux, Switzerland.

Hiba Basim Alwan holds a Bachelor in Computer Science from Al-Mansour University College, Baghdad-Iraq in 2004 and a Masters degree in Computer Science from University of Technology, Baghdad-Iraq, in 2006. As an academic, her research interests include Ant Colony Optimization, Genetic Algorithm, Machine Learning, Support Vector Machine, and Classification. She is currently a PhD. student in the School of Computing at the College of Arts and Science, UUM.

Prof. Dr. Ku Ruhana Ku-Mahamud holds a Bachelor in Mathematical Science and a Masters degree in Computing, both from Bradford University, United Kingdom in 1983 and 1986 respectively. Her PhD in Computer Science was obtained from University Pertanian Malaysia in 1994. As an academic, her research interests include computer system performance modelling and swarm intelligence. She is currently a professor in the School of Computing at the College of Arts and Science, UUM.