Feature Subset Selection Based on Ant Colony Optimization and Support Vector Machine

Wan-liang WANG Yong JIANG S.Y. CHEN

(Collego of Software Engineering, Zhejiang University of Techonology, Hangzhou 310014, China)

Abstract: One of the significant research problems in pattern recognition is the feature subset selection. It is applied to select a subset of features, from a much larger set, through the elimination of variables that produce noise or strictly correlated with other already selected features, such that the selected subset is sufficient to perform the classification task. A hybrid method using ant colony optimization and support vector machine is proposed. The ant colony optimization searches the feature space guided by the result of the SVM. The tests on datasets show the effectiveness of the method.

Keywords: Feature subset selection; Ant colony optimization; Support vector machine

1 Introduction

The problem of classification in machine learning consists of using labeled examples to induce a model that classifies objects into a set of known classes. The objects are described by a vector of features, some of which may be irrelevant or redundant and may have a negative effect on the accuracy of the classifier. There are two basic approaches to feature subset selection [1,2]:wrapper[3] and filter[4,5]methods. Wrappers treat the induction algorithm as a black box that is used by the search algorithm to evaluate each candidate feature subset. While giving good results in terms of the accuracy of the final classifier, wrapper approaches are computationally expensive and may be impractical for large data sets. Filter methods are independent of the classifier and select features based on properties that good feature sets are presumed to have, such as separability or high correlation with the target. Although filter methods are much faster that wrappers, filters may produce disappointing results, because they completely ignore the induction algorithm.

Blum and Langley[6] argued that most existing feature selection algorithms consist of the following four components:

(1) Starting point in the feature space. The search for feature subsets could start with (i) no features, (ii) all features, or (iii) random subset of features. In the first case, the search proceeds by adding features successively, while in the second case, features are successively removed. When starting with a random subset, features could be successively added/removed, or reproduced by a certain procedure. In this paper, we try all these alternatives and compare under simulations.

(2) Search procedure. Ideally, the best subset of features can be found by evaluating all the possible subsets, which is know as exhaustive search. However, this becomes prohibitive as the number of features increases.

(3) Evaluation function. An important component of any feature selection method is the evaluation of feature subset. Evaluation functions measure how good a specific subset can be in discriminating between classes. Performance of classification algorithms is used to select features for wrapper methods. In this paper, the support vector machine is used as the classifier.

(4) Criterion for stopping the search. Feature

....

selection methods must decide when to stop searching through the space of feature subsets. In this paper, the ant colony algorithm stops when there is not improvement on the solution after several iterations or when $n\max$ number of iterations is reached.

A number of search algorithms have been proposed in the literature. Some of the most famous ones are the stepwise, branch-and-bound, and Genetic Algorithms (GA). This paper presents a hybrid approach using ant colony algorithm and support vector machine for feature subset selection problems. The rest of this paper is organized as follow. In the next section, an introduction on ACO and SVM is discussed. In sections 3 and 4, the proposed hybrid methodology is discussed, followed by a discussion on the experimental tests and the results.

2 Ant colony optimization and support vector machine

2.1 Ant colony optimization

The nature has always been fascinating to human being and it has inspired many theories to be applied to various areas. The ant colony algorithm emulates the behavior of real ants. Ants deposit a substance called pheromone on the path that they have traversed from the source to the destination nest ant the ants coming at a later stage apply a probabilistic approach in selecting the node with the highest pheromone trail on the paths. Thus the ants move in an autocatalytic process (positive feedback), favoring the path along which more ants have traveled and by and by traverse all the nodes.

Derived from the behavior of real ants, Dorigo et al. [7] defined a model, which used artificial ant colonies as an optimization tool. In the proposed ant systems, ants are defined as simple computational agents having some memory, they are not completely “blind” like real ants and live in an environment where time is discrete. The underlying idea was to parallelize search over several constructive computational threads, based on a dynamic memory structure incorporating information on the effectiveness of previously obtained results.

The characteristics of ACO include:

(1) a method to construct solutions which balances pheromone trails with a problem-specific heuristic,

(2) a method to both reinforce & evaporate pheromone, and

(3) local search to improve solutions.

ACO methods have been successfully applied to diverse combinatorial optimization problems including traveling salesman, quadratic assignment, vehicle routing problems, telecommunication networks, graph coloring, constraint satisfaction, Hamiltonian graphs and scheduling [8].

1.2 Support vector machines(SVMs)

SVMs [9] are a kind of learning machine based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated as follows: first, map the input vectors into one feature space, possible in higher space, either linearly or nonlinearly, which is relevant with the kernel function. Then, within the feature space from the first step, seek and optimized linear division, that is, construct a hyper-plane which separates two classes. It can be extended to multi-class. SVMs training always seek a global optimized solution and avoid over-fitting, so it has ability to deal with a large number of feature. A complete description about SVMs is available in .

In the linear separable case, there exists a separating hyper-plane whose function is

$$\mathrm{w}\cdot\mathrm{x}+b=0$$

which implies

$$y_i(\mathrm{w}\cdot\mathrm{x}+b=0)\geq 1, \quad i=1,\ldots,N$$

By minimizing $\|\mathbf{w}\|$subject to this constraint, the SVM approach tries to find a unique separating hyper-plane. Here $\|\mathbf{w}\|$ is the Euclidean norm of $\mathbf{w}$, and the distance between the hyper-plane and the nearest data points of each class is $2/\|\mathbf{w}\|$. By introducing Lagrange multipliers α_i, the training procedure amounts to solving a convex quadratic problem. The solution is a unique globally optimized result, which has the following properties

$$w=\sum_{i}^{N}\alpha_i y_i x_i$$

Only if corresponding $a_i > 0$, these x_i are called support vectors.

When SVMs are trained, the decision function can be written as

$$f(x) = sign(\sum \alpha_i y_i (\mathrm{x} \cdot \mathrm{x_i}) + b)$$

For a linear non-separable case, SVMs perform a nonlinear mapping of the input vector x from the input space $\Re^d$ into determined by Kernel function. Two typical kernel functions are listed as follows:

Polynomial: $k(x, y) = ((x \cdot y) + coef)^d, d \in N.$

RBF: $k(x, y) = \exp\{-\frac{\| x - y \|^2}{\sigma^2}\}.$

According to the different classification problems, the different kernel function can be selected to obtain the optimal classification results.

The discussion above deals with binary classification where the class labels can take only two values: 1 and -1. In the real world problem, however, multi-class classification strategy is needed. The earliest used implementation for SVM multi-class classification is one-against-all (OAA) methods. It constructs k SVM models where k is the number of classes. The ith SVM is trained with all of examples in the ith class with positive labels, and all the other examples with negative labels. Another major method is called one-against-one (OAO) method. This method constructs $k(k-1)/2$ classifiers where each one is trained on data from two classes.

3 Feature subset selection based on ACO and SVM

3.1Overview of the approach

The original data set S containing N number of features is reduced to different subsets $s_1, s_2, s_3, ...$ each having $n_1, n_2, n_3, ...$ number of features respectively using ant colony optimization. These subsets will be then fed to a designed multi-class SVM, such that the classification accuracy is maximized. The feature subset selection representation exploited by artificial ants includes the following:

(1) Each ant initializes in the first step will select a subset of n features from the original set of N features. We try three alternatives for the value of n. (i)the value of n starts from a minimum value, and the features are added successively till n reaches the maximum at a constant rate. The starting and the stopping value of n, and the increment rate are all defined by the user depending on the specific classification problem. (ii)the value of n starts from the minimum value, and then features are removed, till n reaches the maximum at a constant rate. (iii) Each ant randomly chooses a feature subset of m features where m is between the upper and lower limit of value of n. Specifically, for instance a problem having 30 features shall have a minimum and maximum value of n as 5 and 28. We call these methods described above as (i) from maximum value to minimum value (MaxMin), (ii) from minimum value to maximum value (MinMax), (iii) random value (Random). The flow chart of the method changes a little with different feature number initializing alternatives as described in Fig.1.

(2) A number of artificial ants and to search through the feature space. The number of ants initialized depends upon the number of features of the given problem. This will be further explained by means of simulations on specific dataset.

(3) $\tau(i, j)$ the intensity of pheromone trail associate with feature f_i and f_j which reflect the previous knowledge about the importance of f_j while the subset has contained the feature f_i.

The steps of each iteration of ACO are described as follows:

Step 1: the ants develop solution consisting of n number of features each based on a probability distribution rule.

Step 2: the r ants construct r different solutions, each containing a subset of n different features. SVM evaluates each subset by determining the error in predication of some data using that of n features.

Step 3: Once all ants have completed constructing their subsets, a global updating rule is applied to the solution set which produces the least classification error.

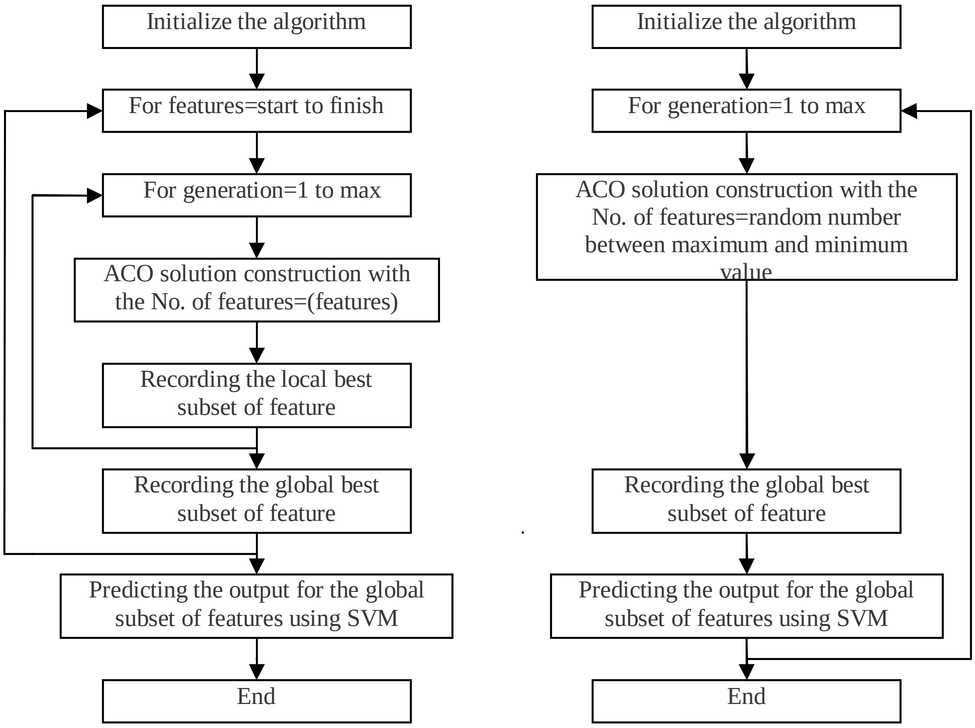
Step 4: a local pheromone updating rule is applied to the rest of the ants.

The above steps are repeated differently according to different feature number initialized method described above. The whole flow chart of the approach is showed in Fig.1.

3.2 Solution construction of ACO

For the ACO, every feature is treated as a location to be visited by the ants. For every ant, it starts in a virtual feature location (such that the ant can select all features equally), and then it builds solution by applying a probabilistic decision policy to move through adjacent states. In this case each subset of feature represents a state. An ant k located in feature f_i selects the next feature location f_j as follows:

$$j = \begin{cases} \arg\max_{u \in J_k(i)} \{\tau_{iu} [\eta_u]^\beta\} & if \quad q \le q_0 \\ \dfrac{\tau_{ij} [\eta_j]^\beta}{\sum_{u \in J_k(i)} \tau_{iu} [\eta_u]^\beta} & otherwise \end{cases} \qquad (1)$$


Initialize the algorithm
For features=start to finish
For generation=1 to max
ACO solution construction with the No. of features=(features)
Recording the local best subset of feature
Recording the global best subset of feature
Predicting the output for the global subset of features using SVM
End
Initialize the algorithm
For generation=1 to max
ACO solution construction with the No. of features=random number between maximum and minimum value
Recording the global best subset of feature
Predicting the output for the global subset of features using SVM
End

a. Flow chart of the method using MaxMin/MinMax b. Flow chart of the method using Random

Fig.1. Flow chart of the proposed method

For a particular ant k , η_u (η_u =1 in this paper) is a heuristic associated with the feature f_u and J_k is the set of features, which are not a part of the solution set, developed by ant k . β is a parameter, which determines the relative importance of pheromone versus heuristic. q is a random number uniformly distributed in between [0…1]. The parameter q_0 , exploitation probability factor, determines the relative importance of exploitation versus exploration. In exploitation, ants select those features which have a maximum of the product described in Equ.1, where as in biased exploration the probability of each feature to be selected by ants corresponds to the value of the above-mentioned product. This helps the ants to keep exploring new states which are close to the optimal solution.

1.3 Pheromone update rule of ACO

As described above, when all ants construct their solutions, and every subset is associated with a classification error. Then, the pheromone trails of the best ant tour (the best feature subset) is updated as

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \rho\sigma, \quad \forall \mathrm{f}_i, \mathrm{f}_\mathrm{j} \in S^{bs}, i \neq j$$

Where S^{bs} is the best feature subset, $\sigma = Q / Err_{s^{bs}}$, Err_s is the classification error associated with the subset s fed into the SVM, Q and ρ are parameters. The purpose of the global updating rule is to encourage the ants to produce subset with least classification error. Global updating rule is only applied to that subset of feature, which has produced the lease error in the current iteration. Besides, the pheromone levels of all features pairs of other subsets are adjusted according to the following formula:

$$\tau_{ij} \leftarrow (1-\kappa)\tau_{ij} + \kappa\tau_0, \quad \forall \mathrm{f}_i, \mathrm{f}_\mathrm{j} \in S^k, i \neq j$$

Where $0 < \kappa < 1$ is a parameter and τ_0 is the initial pheromone level at the beginning of the problem.

4 Experimental tests and results

4.1 Datasets used

In order to evaluate the approach discussed in the previous sections, datasets (WDBC, Image Segment, Dermatology, Hypothyroid and Wine) from UCI [10] were tested. Detail information about the datasets is available at its website. The SVM classifier was realized based on the SVM-Lib [11].

4.2 Parameters setting

(1) Number of ants

The selection of the “right” number of ants is a very critical issue affecting the performance of the algorithm. The number of ants must be sufficient to explore all potential states, while expending the least possible time. Each ant initialized the features number randomly as described above (comparison of the three alternatives for initial feature number will be discussed below) and the performance of the algorithm was tested using 5, 10, 15, 20, 25 ants. Other basic parameters of ACO are: $q_0 = 0.4, Q = 0.001, \rho = 0.8$, $\kappa = 0.9$ .The ACO stopped when a user-defined solution number $s\,\mathrm{max}$ =300 was reached. For the given data sets, it was observed that the algorithm gave best results for 15 ants within a specified time.

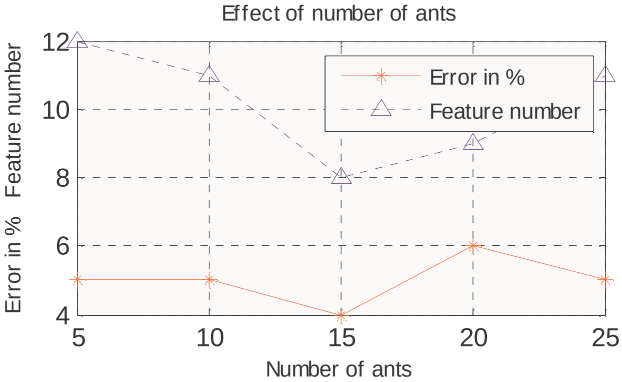


Fig.2. Effect of the number of ants on performance for WDBA

(2) Exploitation probability factor q_0

The parameter q_0 determines the relative importance of exploitation versus exploration. By setting $0.8 < q_0 < 1$, it can speed up the convergence of the algorithm, but also makes the algorithm fall into local optimal point easily. Fig. 3 shows that the given datasets, the algorithm performance is better when $q_0 < 0.6$. Other parameters of ACO were set as follows: 15 ants, $Q = 0.001, \rho = 0.8, \kappa = 0.9$.

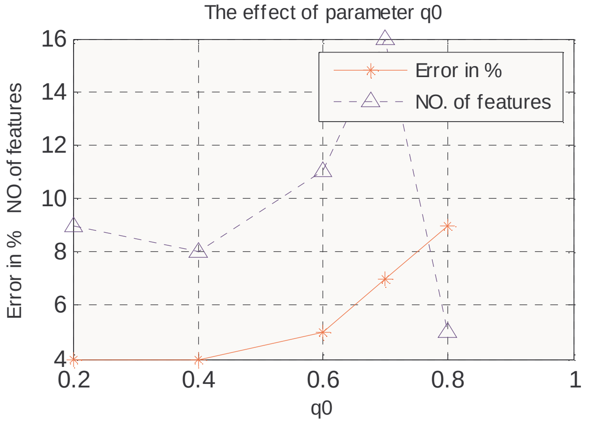


Fig.3. Effect of parameter q_0 on the performance for WDBA

4.3 Number of features initializing methods

Three methods for initializing the number of features are presented above. For the MaxMin and MinMax methods, it's hard to determine the number of generations and the adding/removing rate. What's more, two other interesting topics should be focused: (1) search efficiency. Suppose the number of features of the best subset is n for a given problem, then the algorithm can have the chance to get the optimal subset when the initial number of the features is set to n. As discussed above, the MaxMin/MinMax method adds/removes the feature at a user-defined rate, thus when the number of features initialized is smaller/bigger than n, the left search seems to be useless cause the search space doesn't contain the optimal solution. (ii) The Random method has stronger robustness. The state is transited according to the pheromone level between the feature locations in ACO, so to great extent the current best solution is determined by the knowledge exploited by the ants (pheromone levels). For a specific problem, the number of features of the best subset, n may be close to the maximum/minimum value, which makes the exploited knowledge limited when the initial number of features is n. Table 1 shows the comparison of the performance using the MinMax and Random methods with different maximum and minimum value. Fig.4 shows that the algorithm with Random method performed better than the algorithms with the other two methods.

Table 1 Comparison of the performance for WDBA using the MinMax and Random

	No. of features /Classification accuracy in %	
Minimum-Maximum	MinMax	Random
4-20	12/95%	9/97%
8-20	16/95%	8/96%

4.4 Experimental results

The results obtained are presented in Table 2. Feature subset selection using the proposed approach can improve the performance of the patter classifier since feature selection is not only concerned with reducing the number of features but also eliminating the variables that produce noise or, are correlated with other already selected variables. The basic parameters of the ACO are set as follows: 15 ants, $q_0 = 0.4$, $Q = 0.001, \rho = 0.8, \kappa = 0.9$.

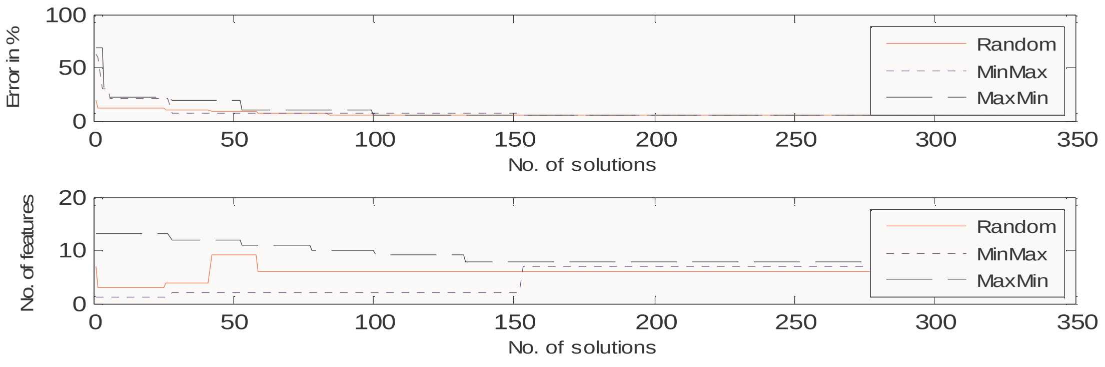


Fig.4. Comparison of performance for WDBA using different No. of features initializing methods

Table 2 Results of the SVM prediction using the reduced subset and the set of complete features

Dataset	Set size Training/Test	No. of features	Reduced subset	Accuracy(all features)	Accuracy(reduced features)	Subset
WDBC	300/200	30	8	75.50%	96.00%	1,2,6,8,11,18,21,27
Image Seg	210/2100	19	8	34.86%	79.57%	3-5,8,12,16,18,19
Dermatology	266/100	34	8	84.00%	95.00%	4,5,8,13,15,20,31,32
Hypothyroid	300/100	22	11	77.00%	96.00%	1,3,8,9,11,12,14,15,17,21,22
Wine	120/58	13	6	31.03%	94.83%	1,2,3,7,9,12

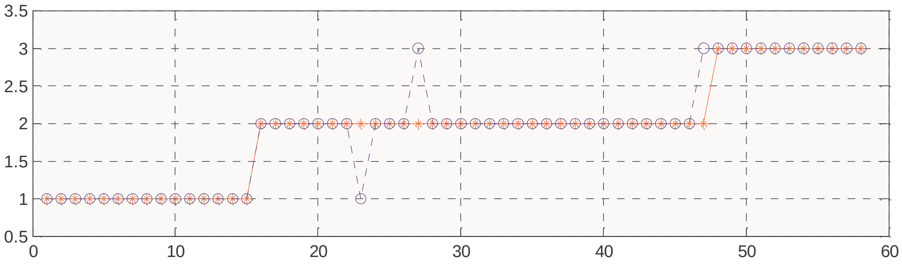


Fig.5. Graph of the actual output against the SVM predicted output using the reduces subset for Wine dataset

5 Conclusions

In this paper, we presented a novel feature selection method based on ant colony optimization and support vector machine. The simulation results show that the method offers an attractive approach to solve the feature subset selection problem. In addition, comparison of number of features initializing methods is discussed by means of experimental simulations. The potential future work includes developing and testing performance of the method for very large state spaces.

Acknowledgments

This project is supported by the National Natural Science Foundation of China (NSFC No.60374056, 60573123).

References

[1] John G., Kohavi, R., Phleger K.: Irrelevant features and the feature subset problem[C]. In: Proceedings of the 11th International Conference on Machine Learning, Morgan Kaufmann(1994): 121-129. .

[2] Dash M., Liu H. Feature selection for classification[J]. Intelligent Data Analysis, An International Journal, Elsevier, 1997, 1(3).

[3] QIAO Liyan, PENG Xiyuan, PENG Yu. BPSO-SVM wrapper for feature subset selection[J]. ACTA ELECTRONICA SINICA, 2006, 34(3): 496-499.

[4] Koller D., Sahami M. Toward optimal feature selection[C]. In: Proceedings of International Conference on Machine Learning, 1996.

[5] WU Zhifeng, CHEN Dongxia.An algorithm of feature subset selection based on genetic algorithm[J]. Journal of the Hebei Academy of Sciences, 2006,23(2):48-50.

[6] Blum A.L.,P. Langley . Selection of relevant features and examples in machine learning[J]. Artificial Intelligence, 1997, 97:245-271.

[7] Dorigo M., Maniezzo V.,and Colorni A. Ant system: Optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man and Cybernetics-Part B, 1996, 26: 29-41.

[8] WU Qidi, WANG Lei. Intelligent Ant colony algorithm and its application[M]. Shanghai: Shanghai scientific and Technological Education Publishing, 2004.

[9] Cotes C , Vapnik V. Support vector networks [J] . Machine Learning ,1995, 20 (3) :273-295.

[10]Blake C. L., Merz C.J. UCI repository of machine learning database, http://www.ics.uci.edu/-mlearn.

[11]Chang Chih-Chung, Lin Chih-Jen, LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/~djlin/libsvm.