Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 49–60
Expert Systems with Applications
www.elsevier.com/locate/eswa

A hybrid approach for feature subset selection using neural networks and ant colony optimization

Rahul Karthik Sivagaminathan, Sreeram Ramakrishnan *

Department of Engineering Management and Systems Engineering, 1870 Miner Circle, University of Missouri – Rolla, Rolla, MO 65409, USA

Abstract

One of the significant research problems in multivariate analysis is the selection of a subset of input variables that can predict the desired output with an acceptable level of accuracy. This goal is attained through the elimination of the variables that produce noise or, are strictly correlated with other already selected variables. Feature subset selection (selection of the input variables) is important in correlation analysis and in the field of classification and modeling. This paper presents a hybrid method based on ant colony optimization and artificial neural networks (ANNs) to address feature selection. The proposed hybrid model is demonstrated using data sets from the domain of medical diagnosis, yielding promising results.

Keywords: Feature subset selection; Ant colony optimization; Neural networks

1. Introduction

1.1. Pattern classification

Pattern classification is the task of classifying any given input feature vector into pre-defined set of classes of patterns (Kulkarni & Vidyasagar, 1997) where as pattern recognition is the task of making important decisions based on complex patterns of information (Ripley, 1996). A detailed discussion on definition and various tools for pattern classification can be found in (Kulkarni, Lugosi, & Santosh, 1998). These methods include Artificial Neural Networks (ANNs), nearest neighbor, kernel and histogram methods, and support vector machines. Researchers in this area focus on characterizing problems to determine if a particular problem can be learned or not, the amount of data required for learning, and then developing the necessary algorithms for learning. Among the existing methods, ANNs have attracted many researchers and has emerged as the most popular tool for pattern recognition and classification. One domain where such applications have found significant utility is the analysis of medical data sets.

Certain problems in the medical diagnosis domain can be considered as a problem of pattern recognition and classification. The use of ANNs is not new in medical diagnosis. For example, in Lanzarini and Giusti (2000), ANNs were used to recognize patterns in medical images. In Zhou, Jiang, Yang, and Chen (2002), an automatic pathological diagnosis procedure named Neural Ensemble based Detection (NED) is proposed, which utilizes an ANN ensemble to identify lung cancer cell images. Unlike other researchers employing back propagation neural networks, in Desai, Lin, and Desai (2001) a neural network based on Kohonen's Linear Vector Quantization is used for the diagnosis of prostate cancer. In this paper, data sets from medical diagnosis are used to demonstrate the feature reduction method using ant optimization.

1.2. Importance of feature selection in classification methods

Many practical pattern classification tasks (Blum & Langley, 1997) (e.g., medical diagnosis) require learning

* Corresponding author. Tel.: +1 573 341 6787; fax: +1 573 341 6567.
E-mail address: Sreeram@umr.edu (S. Ramakrishnan).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.04.010

of an appropriate classification function that assigns a given input pattern (typically represented by using a vector of feature values) to one of a set of classes. The choice of features used for classification could have an impact on the accuracy of the classification function, the time required for classification, training data set requirements, and implementation costs associated with the classification.

The accuracy of the classification function that can be learned using an inductive learning algorithm such as ANNs depends on the set of input features. The attributes or features used to describe the pattern implicitly define a correlation. If the correlation is not accurate enough it would fail to capture the information that is necessary for classification and hence regardless of the learning algorithm used the accuracy of the classification function would be limited. In their paper (John, Kohavi, & Pfleger, 1994) explained the importance of identifying relevant and irrelevant features. It should also be noted that the time and the size of training data set(s) needed for learning a sufficiently accurate classification function increases for more complex patterns with many features (Punch et al., 1993).

The cost for measuring a feature is a critical issue to be considered while selecting a subset. In case of medical diagnosis the features may be observable symptoms or diagnostic tests. Each clinical test is associated with its own diagnostic value, cost and risk. The challenge is in selecting the subset of features with minimum risk, least cost yet which is significantly important in the determining its class/pattern. In Gorunescu, Gorunescu, Darzi, and Gorunescu (2005) a probabilistic neural network with heuristics is used for feature selection in cancer diagnosis.

As can be seen from the above discussion, the issue of feature subset selection in automated design of pattern classifiers is an important research issue. The feature subset selection problem refers to the task of identifying and selecting a useful subset of features to be used to represent patterns from a larger set of often mutually redundant, possibly irrelevant, features with different associated measurement costs and/or risks. Examples of feature subset selection problem include large-scale data mining, power system control, and medical diagnosis (Yang & Honavar, 1998).

1.3. Background

The existing literature in this domain is rich with different solution techniques. Initial methods included exhaustive search in which all combinations of subsets were evaluated. The method guarantees an optimal solution, but finding the optimal subset of features is NP hard. For large number of features, evaluating all states is computationally non-feasible (Boz, 2002) necessitating the need for heuristic search methods. As in Doak (1992), these methods can be classified as exponential, sequential or randomized methods.

The "exponential method" includes methods such as "branch and bound" (Narendra & Fukunaga, 1977) which starts from a full set and removes features using a first depth strategy. The method guarantees an optimal solution under the monotonic assumption that the children of the nodes whose objective function values are lesser than the current best will not contain a better solution and so these features will not be further explored. The other method in this category includes beam search (Doak, 1992). In this method, the features are arranged in a queue with the best states placed at the head of the queue. At each iteration, beam search evaluates all possible states that result from adding a feature subset.

Sequential search algorithms (SSA), also known as stepwise methods (Pudil, Novovicova, & Kittler, 1994), have a relatively lower complexity and use the "hill climbing" strategy to find an optimal solution. Depending upon the different starting points SSA is classified in to sequential forward selection (Devijver & Kittler, 1982) starting with an empty set, Sequential Backward Selection starting from the complete feature set. Meta-heuristic methods are generally considered as random search methods. Some popular meta-heuristic algorithms include genetic algorithm (Leardi, Boggia, & Terrile, 1992; Yang & Honavar, 1998) and simulated annealing (Debuse & Smith, 1997).

This paper presents as ant colony optimization (ACO) approach for feature selection problems using data sets from the field of medial diagnosis. This paper presents a novel approach for heuristic value calculation, which will reduce the set of available features. The rest of this paper is organized as follows. In the next section, an introduction on ACO applications in feature selection problems is discussed. The different methods for feature selection problems (based on the existence of classification function) are presented in Section 3. In Sections 4 and 5, the proposed hybrid methodology is discussed, followed by a discussion on the experimental setup, datasets used and the results.

2. Ant colony optimization

Ant algorithm was first proposed by Dorigo and Gambardella (1997) as a multi-agent approach for difficult combinatorial optimization problems such as traveling sales man problem (TSP) and the quadratic assignment problem (QAP). From then, researchers have applied ACO to many discrete optimization problems (Bonabeau, Dorigo, & Theraulaz, 1999; Corne, Dorigo, & Glover, 1999).

ACO is a meta-heuristic approach which has been applied to various NP hard problems such as static/dynamic combinatorial optimization. ACO applications in static combinatorial optimization problems include job shop scheduling (Blum & Sampels, 2002; Colorine, Dorigo, & Maniezzo, 1994), flow shop (Stützle, 1998), open shop (Blum, 2003), group shop (Sampels, Blum, Mastrolilli, & Rossi-Doria, 2002), vehicle routing (Bullnheimer, Hartl, & Strauss, 1998), sequential ordering (Gambardilla & Dorigo, 1997), graph coloring (Costa & Hentz, 1997) and shortest common super sequences (Micheal & Middendorf, 1999). ACO application to dynamic combinatorial optimi-

zation problems includes connection oriented network routing (Schoonderwoerd, Holland, Bruten, & Rothkrantz, 1996) and connection less network routing (Sim & Sun, 2001).

In Ani (in press, 2005) an ACO approach was presented for feature selection problems. In this paper, the author calculates a term called "updated selection measure (USM)" which is used for selecting features, a function of the pheromone trail and the so called "local importance" which has replaced the heuristic function. A major application of the algorithm developed in this paper is in the field of texture classification and classification of speech segments. Similarly, another application of ACO can be found in Jensen and Shen (2003) where an entropy-based modification of the original rough set-based approach for feature selection problems was presented. Other applications include Schreyer and Raidl (2002) where an ACO approach is used for labeling point features, a pre-processing step which reduces the search space.

This paper presents a relatively simpler model of ACO. The major difference from previous works is in the calculation of the heuristic values. Heuristic value calculations are application specific and help the algorithm reach the optimal solution quickly by reducing the search domain. In medical diagnosis applications heuristic value can be a function of diagnostic value, cost or risk. Generally, the value of these parameters, except cost, is fuzzy and the function cannot be generalized for different applications. In this paper, the heuristic value is treated as a simple function of cost. Clearly, the features associated with lesser costs will be preferred by the algorithm. The algorithm uses ANNs as a classification function to evaluate the "goodness" of the subset developed at each stage, instead of the nearest neighborhood algorithm used otherwise.

3. Different approaches for feature subset selection problems

A classification function is essentially the tool used for classifying patterns or the tool to evaluate the efficiency of each subset to predict the class output or pattern. Depending on whether a classification function is used or not, feature subset selection algorithms can be divided into two (John et al., 1994) – filter approach and the wrapper approach.

3.1. Filter approach

In the filter approach, no classification function is used – feature subsets are evaluated by other means. In "focus algorithm" (Almuallim & Dietterich, 1991), a type of filter approach, an exhaustive search is utilized to examine all the subsets of features. The method then identifies the subset with minimum number of features which classifies the training set instances with acceptable level of accuracy. Relief method (Kira & Rendell, 1992) is random search method based on filter approach model. Here, a weight is assigned to each feature based on the relevance to the target concept, and instances are selected randomly to find the relevance of features. Another filter approach model (Cardie, 1993) uses a nearest neighborhood algorithm.

3.2. Wrapper approach

In a wrapper approach, a classification function is used to evaluate the "goodness" of the feature subsets developed. The feature subset selection algorithm is wrapped around the classification function, thus the name. In Caruana and Freitag (1994) tree caching is used for "greedy" attribute selection. Caching can be used with deterministic decision trees and do not usually use all of the available features. If decision trees use n of the N total features, all feature subsets which have all of these n features will create the same tree with the same accuracy.

3.3. Comparison of the Wrapper versus Filter approach

Most meta-heuristic feature subset selection algorithms use a wrapper approach model because of some inherent advantages (Boz, 2002). In the filter approach the feature selection is performed as a pre-processing step. The disadvantage is that it ignores the effect of the selected feature subset on the performance of the induction algorithm. In John et al. (1994) it is claimed that to determine a useful subset of features, the subset selection algorithm must take into account the biases of the induction algorithm in order to select a subset. The current paper builds a wrapper approach model using ANN as the classification function.

3.4. Artificial neural networks

In a number of examples of practical interest, where mathematical models are unavailable but real-life data relating inputs to outputs exist, ANNs can be used to construct an empirical model. These models then may be used to predict the outputs for a set of new inputs not employed in the construction of the model. But one of the major drawbacks of such methods is that the structure of the model must be specified a priori- it requires a set of data for training and developing the model, which may not be necessarily available.

4. Methodology

4.1. A hybrid approach with artificial neural networks and ant colony optimization

It should be recalled that this method is based on observations by earlier researchers that ants in real-life, while walking from their food source to nest are able to optimize their path, without making use of any apparent visual

clues. This is possible because of an indirect communication mode called "stigmergy" using pheromone – an odorous chemical. The quantity of the pheromone depends on the distances, quantity and quality of the food source. Each ant follows a direction rich in pheromone smell, thus making it a loop of positive feedback. The pheromone decays over time and evaporates, resulting in less pheromone on the less popular paths. Due to this "evaporation", ants explore other paths as well and eventually end up with the most optimal path. Here, ANNs are used as the classification function, where as the ACO serves as the evaluation algorithm (Fig. 1).

The original data set $\boldsymbol{S}$ containing $\boldsymbol{N}$ number of features is reduced to different subsets $s_1, s_2, s_3, \ldots$ each having $n_1, n_2, n_3, \ldots$ number of features respectively using ant algorithm. These subsets will be then fed to a pre-designed ANN trained by Levenberg–Marquard Back Propagation, and having a fixed number of neurons in the hidden layer. Generally, the number of neurons in the hidden layer depends on the dimensions of the input feature vector. However, it should be noted that in the proposed methodology, the dimension of the input feature vector changes as the algorithm proceeds. Moreover, the number of neurons in the hidden layer depends on the search domain. This methodology defines uses the maximum and the minimum value of $\boldsymbol{n}$ (the number of features in the chosen subset) to set limits for the number of neurons (further discussed in Section 4.2). In this architecture, training was achieved using the Levenberg–Marquard algorithm (due to its inherent advantages associated with speed of training and accuracy of learning).

Each of the data sets in this method is divided to training and testing data points. Once training is accomplished the network will be tested for some unseen data points, and the number of class mismatch is treated as the error corresponding to each subset. Thus, every subset of feature will be associated with some prediction error, which will help to determine the best feature subset. Thus, ANN gives direction to ant algorithm to find the optimal solution set, and the final subset developed by ant algorithm (as the most optimal subset) is again evaluated by the ANN for a larger number of epochs.

4.2. Hybrid artificial neural network – ant algorithm

In this section, an overview of the proposed methodology is presented (Fig. 2). Informally, the ant system works as follows. First, a set of ants is initialized. The number of ants initialized depends upon the number of features the given problem. This will be further explained in a set of examples in the next section. Each ant initialized in the first step will select a subset of $\boldsymbol{n}$ features from the original set of $\boldsymbol{N}$ features. The value of $\boldsymbol{n}$ increases at a constant rate. This rate is a user-defined function. In this paper, considering the relatively smaller state space, an increment rate of 1 is used. It will be interesting to further explore the method with different values of this rate. The initial and final values of $\boldsymbol{n}$ are problem-specific and hence a user defined function. For instance a problem having 30 features shall have a starting and finishing value of $\boldsymbol{n}$ as 5 and 28. If the user knows the minimum requirement of features for the classification is 5, and reducing to a subset of 28 features will have no significant reduction. In short, by setting the upper and lower limit on the value of $\boldsymbol{n}$, the user is specifying the search domain.

Step 1: Initially, when the pheromone level or the desirability measure for all the features are the same, the ants develop solution consisting of $\boldsymbol{n}$ number of features each, using an initialization rule (for

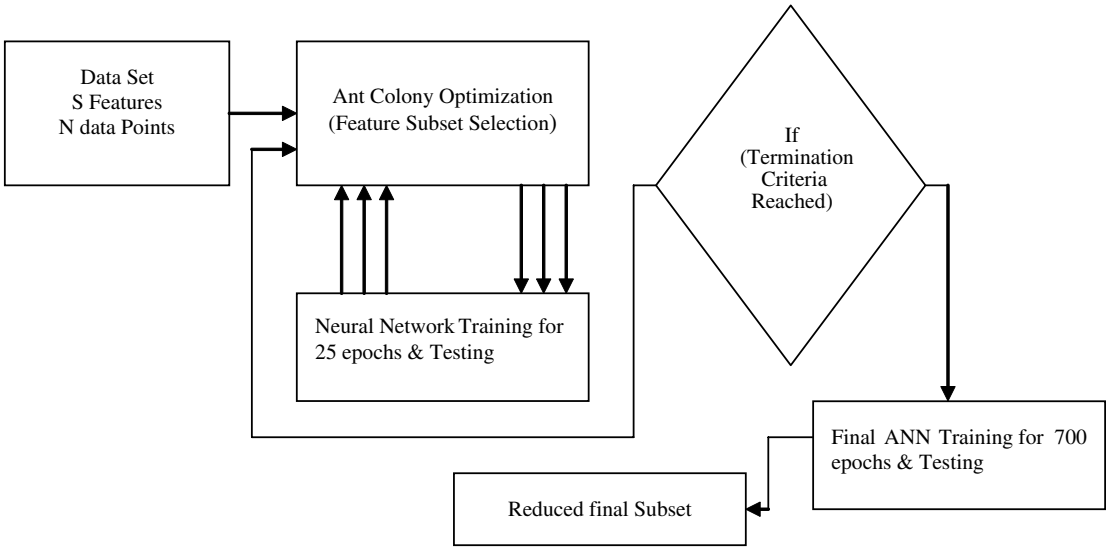

Fig. 1. Hybridizing ant algorithm with ANN.

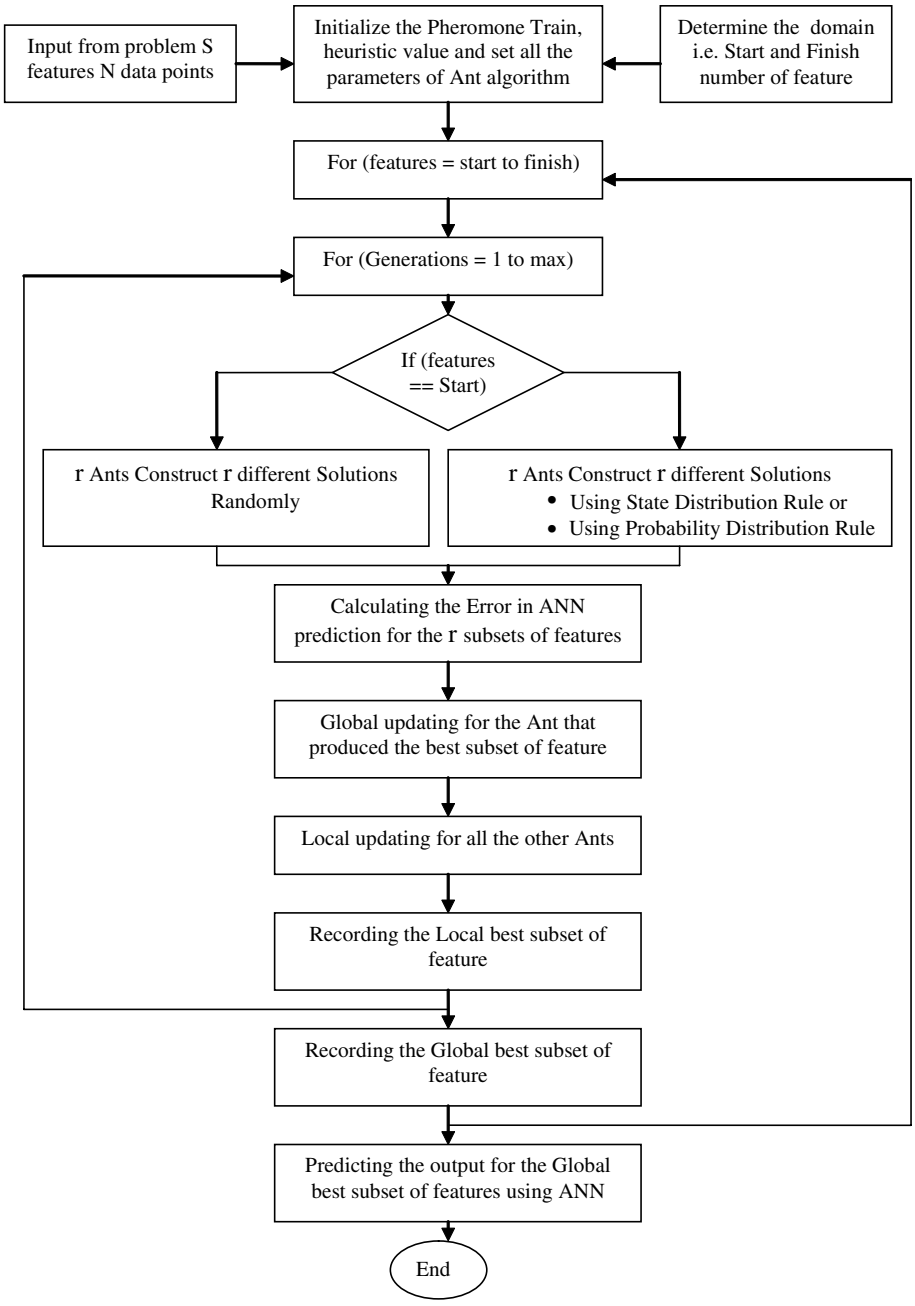

Fig. 2. Flow chart for the proposed hybrid algorithm.

example, randomly). Later, ants select features based on *state distribution rule* or *probability distribution rule.*

Step 2: Each of the ***r*** ants construct ***r*** different solutions, each containing a subset of ***n*** different features. ANN (after sufficient training) evaluates each subset by determining the error in prediction for unseen data points using that subset 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. Each time the ant that has produced the solution with least error, it is "rewarded" by increasing the desirability of all the features which are part of its (ant's) solution.

Step 4: Similarly, a *local "pheromone"-updating rule* is applied to the rest of the ants. That is, those features which were selected the ants (except the winning ant) is subject to this rule, in which their desirability is decreased by a minimal amount.

The above steps are repeated for all the values of ***n*** between its starting and finishing value. During each iteration, the best subset and its corresponding error is recorded.

4.3. Stepwise algorithm: implementing ant algorithm for feature selection problem

Here, the functioning of the AA for the proposed hybrid approach is discussed. Let $S = \{a, b, c, d \ldots z\}$ be the set of given N features, and $s = \{p, q, r \ldots t\}$ where (s c S). Furthermore, let $\delta(f)$ represent the cost parameter associated with the feature f in its measurement. For instance, in medical diagnosis it cane be considered as the cost for taking each test or the cost associated for measuring or assigning values to the visible symptoms. Further, let $\tau(f)$ be the desirability measure (pheromone level) of feature f to be in the selected subset of features s. Initially, the desirability of each feature will be the same, but as the algorithm proceeds, with global and local updating steps, those features which are more important to determine the class/output will see their "desirability" increase compared to that of the other features.

The state transition rules enable enables ants to select features using the pheromone trail and the heuristic value (the inverse of cost parameter). Each ant chooses a particular feature by maximizing a product of these two parameters. This is further explained in Section 4.3.1. Once a particular set of features are selected, the next step is global updating to increment the features which were selected by the winner ant (Section 4.3.2). The last step is the local updating with the objective of decreasing the pheromone trail of the other features which were selected by ants but did not produce a good solution (Section 4.3.3).

4.3.1. State transition rule

The objective function of this optimization algorithm is to minimize the classification error in predicting the output. In this hybrid approach, the role of each ant is to build a solution subset. The "ants" build solutions applying a probabilistic decision policy to move through adjacent states. In this case each subset of feature represents a state. The state transition rules are discussed here.

In the proposed method, an ant chooses a feature as follows:

$$s = \begin{cases} \arg\max\{\tau(u)^* \eta(u)^\beta\} & \text{if } (q < q_0) \quad \text{(exploitation)} \\ \dfrac{\tau(s)^* \eta(s)^\beta}{\sum_{u \in j_k(r)} \tau(u)^* [\eta(u)^\beta]} & \text{otherwise (if } s \in j_k(r)) \quad \text{(biased exploration)} \end{cases} \tag{1}$$

For a particular ant, r, η represents the inverse of the cost parameter and j_k is the set of features, which are not a part of the solution set, developed by ant r. β is a parameter, which determines the relative importance of pheromone versus heuristic. The value of β is application and user specific (represents how much importance has to be given to cost while selecting the subset of features). Setting the value of β at zero will give equal priority to all features irrespective of their costs, where as $\beta = 1$, will give equal importance to cost minimization while selecting features; q is a random number uniformly distributed in between [0...1]. Thus, Eq. (1) favors the choice of features which are associated with low costs and high amount of pheromone level. The pheromone deposited acts as the memory, while heuristic information is simply the inverse of cost parameter. The ants search for a good solution and cooperate through pheromone mediated indirect and global communication. Informally, each ant adds new features to a partial solution by exploiting both information gained from past experience and a heuristic. Thus, in exploitation the feature which has highest pheromone trail and low cost is selected, while in exploration any feature is randomly selected by probability.

The equation discussed earlier consists of two components – exploitation and biased exploration corresponding to the state transition rule (stochastic greedy rule) and the random proportional rule respectively. 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 above product, where as in biased exploration the probability of each feature to be selected by ants corresponds to the value of the above-mentioned product (the feature with the highest product has the highest probability of selection). This helps the ants to keep exploring new states which are close to the optimal solution. Since the probability is a function of the previous information and heuristics, it is referred to as biased exploration.

In this methodology, if the cost parameter associated with the features is unknown, it is assumed to be unity ($q_0 = 1$) to allow all features to be selected with equal probability. Alternatively, to scale the cost parameter, the cost of the most expensive feature can be divided by the individual cost of the features. In certain medical diagnosis applications, practitioners know that certain feature(s) is absolutely important to be included in any model. In such scenarios, the cost of that feature could be taken as a negligible amount. This would make its inverse a very large number, which in turn, increases the value of the product of pheromone trail and heuristics, forcing the ants to select that particular feature.

4.3.2. Global updating rule

As discussed earlier, the ants would have, by this stage, accomplished the task of constructing a solution subset, and each subset would be associated with a classification error – the number of instances where ANN produced wrong results, for unseen data points using the given subset of features. Logically, the next step is to appreciate the ant which has produced the subset which has produced the least classification error.

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 least error in the current iteration. By this rule, the pheromone level of all the individual features, which were a part of the best feature subset will be

incremented. Thus, the ant that develops the best solution is allowed to deposit pheromone on the set of features that it has selected as solution. This choice together with the use of random proportional rule is intended to make the search more directed. Ants search in the neighborhood of the best state found up to the current iteration of the algorithm. Global updating is performed only after all the ants have developed their respective solutions. The pheromone level is incremented by applying the global updating rule:

$$\tau(s+1) = (1-\kappa) * \tau(s) + \kappa * \sigma \quad (2)$$

σ is the inverse of Δx, where Δx is the least classification error of the globally best solution and κ is the pheromone decay parameter. Global updating is intended in providing greater amount of pheromone to the solution set that produces less classification error. Thus, those features which are repeatedly a part of the best solution subset will be incremented frequently, which would make them more attractive for the future generation ants to select them. That is, these features have a more probability of being selected in future by the ants while constructing a solution subset.

4.3.3. Local updating rule

The local updating rule not only makes the irrelevant features less desirable, but also helps ants select those features which have never been explored previously. This updating rule will either decrease the pheromone trail or maintain the same level depending on whether a particular feature has been selected or not. By employing this updating rule, the pheromone level of the features that have been a part of the best feature subset in the previous iterations, will decrease by a very minimal amount, where as the pheromone level of the features that have never been a part of the best feature subset will remain the same. Thus, the pheromone level of the features will never be less than the pheromone level to which they are initialized. The change in the pheromone level is obtained as

$$\tau(s+1) = (1-\alpha) * \tau(s) + \alpha * \tau_0 \quad (3)$$

where $0 < \alpha < 1$ is a parameter called local pheromone update strength parameter and τ_0 is the initial pheromone level at the beginning of the problem.

Thus, by using the local and global updating rules, the pheromone level of some features (which were a part of the best feature subset in the previous iterations) will diminish by a minimal amount; for features which are a part of the best feature subset in the current iteration, the value will increase and for the rest of the features the value will not change. This prevents the ants from converging to a common path. This characteristic, which was observed experimentally in real-life ants (Dorigo, Caro, & Gambardella, 1999) is a desirable property. This is because, if the ants explore different paths, then there is a higher probability that one of them will find an improving solution as opposed to the possibility of convergence to the same tour.

5. Experimental setup

5.1. Data sets used

In order to evaluate the methodology discussed in the previous sections, real-world data sets from the UC-Irvine repository (Merz & Murphy, 1996), shown in Table 1, were tested. This library has data sets and domain theories that can be used to evaluate learning algorithms.

5.2. Setting ACO parameters

Tuning the parameters for any optimization algorithm is at least as important as designing the algorithm itself. The controllable parameters which affect the performance of ACO include the number of ants, generations, $\boldsymbol{q}_0$ (the Exploitation probability factor), Pheromone decay parameter ($\boldsymbol{\kappa}$) and local pheromone update strength parameter ($\boldsymbol{\alpha}$). In order to tune the parameters, different values of the parameters were tested on the thyroid disease (Australia) dataset.

5.2.1. 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. A range of 3–12 ants were considered for exploration in this study. It should be noted that the "optimum" number of ants is specific to the data set(s) considered. The discussion here is limited to the data sets considered here, and the reader is urged to focus on the approach used for identifying the "optimal" number of ants for the data sets considered in this experiment. In this implementation, the performance of the algorithm was tested using 3, 5, 8, 10, 12 ants. For the given data sets, it was observed that the algorithm gave best results for five ants. Increasing the number of ants not only resulted in higher time requirements to reach a solution but also increased the testing error of ANN using the subset of features developed as solution. This effect is shown for the above-cited data set in Fig. 3.

Table 1
Details of datasets used

Data sets used	Number of attributes	Attribute type	Number of classes	Size of data set
Thyroid disease (Australia)	28	Numeric, nominal	2	206
Thyroid disease (discordant)	28	Numeric, nominal	2	206
Thyroid disease (hypothyroid)	21	Numeric	3	400
Dermatology	34	Numeric	6	366
Breast cancer (Wisconsin diagnostic)	32	Numeric, nominal	2	569
Breast cancer (Wisconsin prognostic)	34	Numeric, nominal	2	198

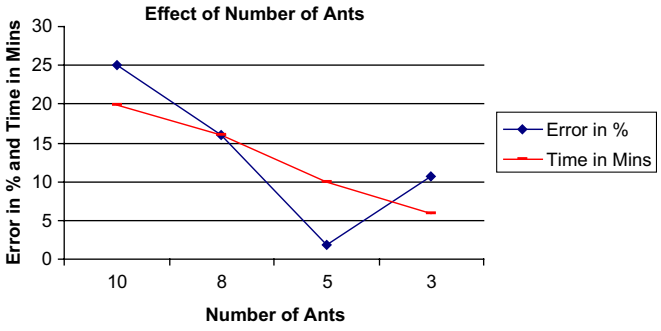

Fig. 3. Effect of number of ants on performance for thyroid disease (Australia) data set.

5.2.2. Exploitation probability factor

In Dorigo and Gambardella (1997), it was argued that the entire search space (the domain of traveling salesman problem was employed by the authors) can be divided into three categories – best edges, testable edges and unused edges. Similarly, the entire set of features can be divided into three sets: (i) best features (BF) – features which have repeatedly been in the best subset; (ii) testable features (TF) – features which have been in the best subset in previous iterations, and (iii) unused features (UF) – features that have never been in the best subset. Recalling Section 4.3.1, in stochastic greedy rule, ants exploit features which fall in the category of BF, where as in random proportional rule ants explore the subsets falling in the edges of BF and TF. The value of exploitation probability factor, q_0, determines how much ants should exploit BF and explore TF. By setting q_0 at 0.8, ACO favors features falling on the edges of TF and BF.

Ideally in ACO, the features in BF which are not consistently performing well will be downgraded to TF, and the features belonging to TF shall be downgraded to UF, unless they happen to belong to the new best subset. If the value of q_0 is set to lesser than 0.8, ants may favor features falling in the category of TF and UF, exploring new states but misled to poor results. Similarly, if q_0 is set to 1, ants may not explore features falling in the edges of BF and TF, selecting only features falling in BF – resulting in all the ants follow the same path. The graph, Fig. 4, shows that the given data sets, the algorithm performance is best when q_0 is 0.8.

5.2.3. Pheromone decay parameter and local pheromone update strength parameter

Pheromone decay parameter (κ) and local pheromone update strength parameter (α) help ants maintain a well coordinated pheromone mediated cooperation. The values of κ and α should be close to 0.9 or 0.8 so that in each iteration just the right amount of pheromone is deposited so as to influence the decision of future generation ants in the right direction. Depositing or degrading more amount of pheromone in each step may lead to pheromone accumulation or depletion on certain features, clouding the understanding of which features are more important. By decreasing the value of κ, the amount of pheromone deposited in each iteration increases, making them more desirable for the future generation ants. This may not let the ants select those features which have the capability of producing a good subset but which have not been selected before. Similarly, if κ is set at 1, it will lead to no pheromone deposition on the features which are producing good subsets, making the ants non-cooperative and thus leading to poor performance. (Fig. 5 shows the performance of ACO for different values of pheromone decay parameter.)

5.2.4. Number of generations

Similarly, number of generations is an important parameter. Increasing the number of generations increases runtime of the algorithm tremendously while fewer generations make ants explore less possible states for each value of n, leading to poor/pre-mature convergence. The graph (Fig. 6) shows the algorithm performance and time consumed varying number of generations. As we can see that the algorithm performs its best when five generations of ants are produced in each step.

5.3. Setting the ANN parameters

ANNs are used to evaluate the "goodness" of the subsets (ability to correctly classify the class/pattern) developed by ants as solution in each iteration and to test the

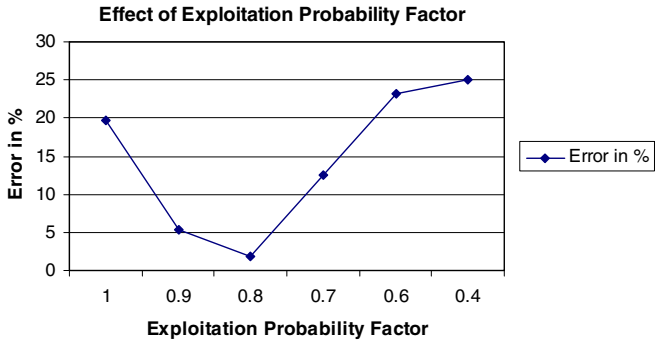

Fig. 4. Effect of exploitation probability factor on performance for thyroid disease (Australia) data set.

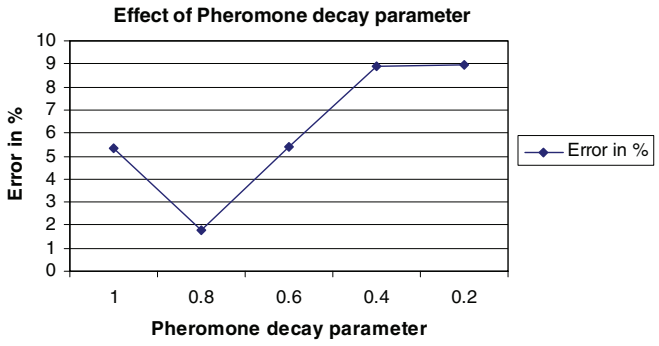

Fig. 5. Effect of pheromone decay parameter on performance for thyroid disease (Australia) data set.

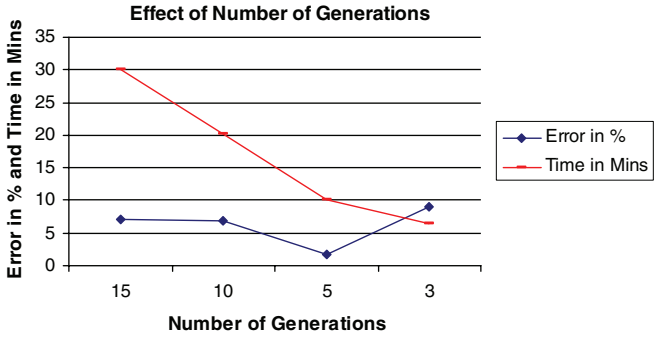

Fig. 6. Effect of number of generations on the performance for thyroid disease (Australia) data set.

final subset produced. As stated previously, the networks are trained using Levenberg–Marquard's back propagation algorithm. Two different ANN models were used for these two purposes. The only difference between the two was the training epochs – for evaluating the subsets the ANN model developed is trained for a mere 25 epochs, due to limited time. For instance, consider a problem consisting of 20 features to be reduced. An initial step involves selecting a subset of four features; even though the solution developed at this stage may not be a global optimum, the step is needed for the algorithm to check which of the ***r*** ants have produced the best subset of four features. For this purpose, an initial number of 25 epochs is sufficient to obtain a "good generalization". It should be further noted that the final training for the global "best subset" is performed for 700 epochs.

Selecting the number of neurons in the hidden layer for the ANN designed to evaluate the subsets depends on the search horizon, i.e., the maximum and minimum value of ***n***. This in-fact limits the application of this algorithm from being used for very high dimensional (in thousands) feature selection problems. In such scenarios, a possible approach is to divide the entire state space into pre-defined ranges, and then apply the algorithm to each segment. It should be noted that in that approach, the number of neurons in the hidden layer will be different for each step. In each of the applications approximately 80% of the entire data set was used for training and the remaining 20% was used for testing. Table 2 provides the details of the ANN models.

6. Results and conclusions

The results obtained are presented in Table 3. As stated earlier, feature subset selection may in some cases improve

Table 2
Details of the artificial neural network models used

Sr. no.	Data sets	Training set size	Testing set size	ANN model for evaluating subsets			ANN model for final subset evaluation		
				Hidden layer neurons	Hidden layer transfer function	Output layer transfer function	Hidden layer neurons	Hidden layer transfer function	Output layer transfer function
1	Thyroid disease (Australia)	150	56	7	Tansig	Logsig	9	Tansig	Logsig
2	Thyroid disease (discordant)	150	56	7	Tansig	Logsig	9	Tansig	Logsog
3	Thyroid disease (hypothyroid)	300	100	8	Tansig	Purelin	8	Tansig	Purelin
4	Dermatology	266	100	10	Tansig	Purelin	12	Tansig	Purelin
5	Breast cancer (Wisconsin diagnostic)	469	100	10	Tansig	Purelin	12	Tansig	Purelin
6	Breast cancer (Wisconsin prognostic)	148	50	12	Tansig	Purelin	14	Tansig	Purelin

Table 3
Results of the ANN prediction using the reduced subset and the set of complete features

Sr. no.	Data sets	No. of attributes	Reduced subset	% Reduction	ANN prediction using all features		ANN prediction for reduced subset	
					Training error	Testing accuracy (%)	Training error	Testing accuracy (%)
1	Thyroid disease (Australia)	28	12	57.14	0.00034	91.08	0.00024	98.22
2	Thyroid disease (discordant)	28	4	85.71	0.00035	92.86	0.00085	96.42
3	Thyroid disease (Hypothyroid)	21	14	58.82	0.00041	86.00	0.00041	94.50
4	Dermatology	34	7	66.66	0.000051	68.00	0.000021	95.00
5	Breast cancer (Wisconsin diagnostic)	32	12	62.5	0.000025	69.00	0.000045	95.57
6	Breast cancer (Wisconsin prognostic)	30	14	58.82	0.000055	64.29	0.000025	77.50

the performance of the pattern 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. To demonstrate the method, first, the entire sets of features were used in predicting the output. The results obtained are discussed in 6th and 7th column of Table 3. Then the reduced subsets were used to predict the output using the same neural network model for the same number of epochs, results shown in 8th and 9th columns of Table 3.

From Table 3, it can be seen that the performance of the classifier improves in all the test cases considered in the implementation. For the data set 6, where the testing accuracy is "lower" than that observed in other data sets, we hypothesize that it is not suitable for feature selection application. The algorithm discussed in this paper attempts to determine inter-variable relationship amongst a reduced subset which can predict the output accurately. This relation may or may not exist in datasets. For certain datasets using all features may be necessary to predict the output. In such scenarios, feature selection algorithms such as these cannot add significant value. But since the accuracy using reduced subset for the dataset 6 is still more than the accuracy using the set of complete features, we conclude the algorithm has removed the noisy features to some extent. Thus, from the above presented results it could be seen that the method proposed in this paper shows promising results. The graphs (Figs. 7 and 8) show the prediction of ANN in graphical form for the thyroid disease (hypothyroid) data set and the thyroid disease (Australia) data set. The graph compares the actual output using the entire set of features and the output obtained from the ANN using the reduced subset of features.

This paper shows that ant algorithm offers an attractive approach to solve the feature subset selection problem (under a different cost and performance constraints) in inductive learning of neural network pattern classifier. The algorithm considers both the individual performance and performance in a subset to predict the output, while selecting each feature. The potential future work in this area includes developing a heuristic model specifically for medical diagnosis applications as a function of diagnostic value, cost and risk associated with each test. This will help selecting those features which are associated with high diagnostic value, low risk and low cost, thereby reducing the overall cost. In this paper, the performance of the method for very large state spaces which may

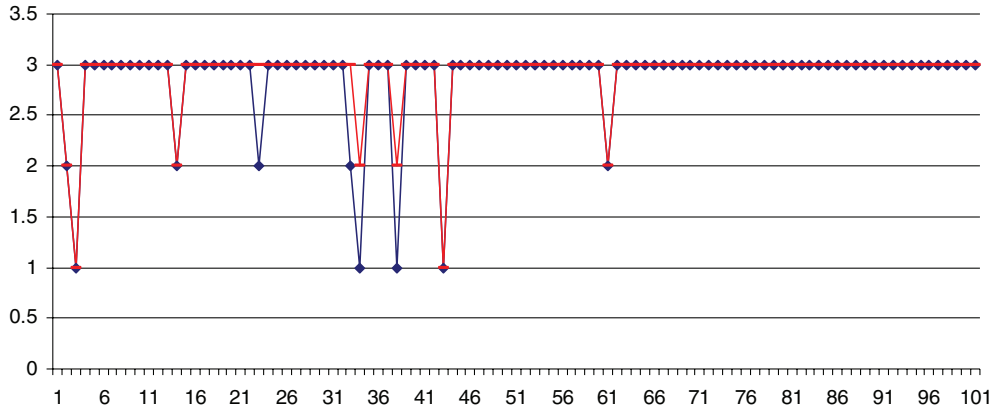

Fig. 7. Graph of the actual output against the ANN predicted output using the reduced subset for thyroid disease dataset (hypothyroid).

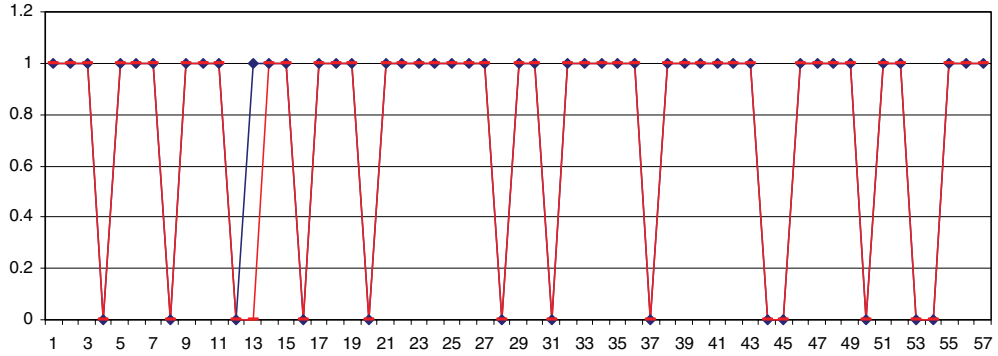

Fig. 8. Graph of the actual output against the ANN predicted output using the reduced subset for thyroid disease dataset (Australia).

require segmenting was not explored and is worth studying further. In addition, comparison of the method discussed in this paper with other learning methods, impact of pheromone decay parameter and local pheromone update strength parameter on efficiency of the hybrid method, and quantifying the impact of exploitation probability factor are potential directions for further studies.

References

Almuallim, H., & Dietterich, T. G. (1991). Learning with many irrelevant features. *In Proceedings of the ninth national conference on artificial intelligence (AAAI-91)* (Vol. 2, pp. 547–552). Anaheim, CA: AAAI Press.

Ani, A. Al. (in press). An ant algorithm based approach for feature subset selection. In *International Conference on Artificial Intelligence and Machine Learning*.

Ani, A. Al. (2005). Feature subset selection using ant colony optimization. *International Journal of Computational Intelligence, 2*(1), 53–58.

Blum, C. (2003). An ant-colony optimization algorithm to tackle shop scheduling problems. Tech. report, TR/IRDIA/ 2003-01, IRDIA, Universite' Libre de Bruxelles, Belgium.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 245–271.

Blum, C., & Sampels, M. (2002). Ant colony optimization for fop shop scheduling: A case study on different pheromone representations. In *Proceedings of the 2002 congress on evolutionary computing (CEC'02)* (pp. 1558). New York: IEEE Press.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). From nature to artificial swarm intelligence. New York: Oxford University Press.

Boz, O. (2002). Feature subset selection using sorted feature relevance. In *The proceedings of ICMLA, international conference of machine learning and applications*, Los Angeles, USA (pp. 147–153).

Bullnheimer, B., Hartl, R. F., & Strauss, G. (1998). Applying the ant system for the vehicle routing problem. In *Meta-heuristics: Advances and trends in local search paradigms for optimizations* (pp. 109–120).

Cardie, C. (1993). Using decision trees to improve case-based learning. In *Proceedings of the tenth international conference on machine learning* (pp. 25–32). Los Altos, CA: Morgan Kaufmann Publishers.

Caruana, R., & Freitag, D. (1994). Greedy attribute selection. In *Proceedings of the eleventh international conference on machine learning* (pp. 180–189). Los Altos, CA: Morgan Kaufmann Publishers.

Colorine, A., Dorigo, M., & Maniezzo, V. (1994). Ant system for job shop scheduling. *Belgium Journal of Operations Research, Statistics and Computer Science (JORBEL), 34*, 39–53.

Corne, D., Dorigo, M., & Glover, F. (1999). New ideas in optimization. Maidenhead: McGraw Hill.

Costa, D., & Hentz, A. (1997). Ants can color graph. *Journal of the Operational Research Society, 48*, 295–305.

Debuse, J. C. W., & Smith, V. J. R. (1997). Feature subset selection within a simulated annealing data mining algorithm. *Journal of Intelligent Information Systems, 9*, 57–81.

Desai, R., Lin, F. C., & Desai, G. R. (2001). Medical diagnosis with a Kohonen LVQ2 neural network. In *Proceedings of the 8th international conference on neural information processing*, Cd-ROM, Shanghai.

Devijver, P. A., & Kittler, J. (1982). Pattern recognition: A statistical approach. Englewood Cliffs, NJ: Prentice Hall International.

Doak, J. (1992). Intrusion detection: The application of feature selection – A comparison of algorithms, and the application of a wide area network analyzer. Master's thesis, Department of Computer Science, University of California, Davis.

Dorigo, M., Caro, G. D., & Gambardella, L. M. (1999). Ant algorithm for discrete optimization. *Artificial Life, 5*(2), 137–172.

Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transaction on Evolutionary Computation, 1*(1), 53–66.

Gambardilla, L. M., & Dorigo, M. (1997). HAS-SOP: An hybrid ant system for the sequential ordering problem. Tech. report 11-97, Lugano, Switzerland: IDSIA.

Gorunescu, F., Gorunescu, M., Darzi, E. El., & Gorunescu, S. (2005). An evolutionary computational approach to probabilistic neural network with application to hepatic cancer diagnosis. In *18th IEEE symposium on computer-based medical systems (CBMS-05)* (pp. 461–466).

Jensen, R., & Shen, Q. (2003). Finding rough set reducts with ant colony optimization. In *Proceedings of the 2003 UK workshop on computational intelligence* (pp. 15–22).

John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine learning: Proceedings of the eleventh international conference* (pp. 121–129). Los Altos, CA: Morgan Kaufmann Publishers.

Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the 10th national conference on artificial intelligence* (pp. 129–134). San Jose, CA: MIT Press.

Kulkarni, R. S., Lugosi, G., & Santosh, V. S. (1998). Learning pattern classification – A survey. *IEEE Transaction on Information Theory, 44*(6), 2178–2206.

Kulkarni, R. S., & Vidyasagar, M. (1997). Learning decision rules for pattern classification under a family of probability measures. *IEEE Transactions on Information Theory, 43*(1), 154–166.

Lanzarini, L., & Giusti, D. A. (2000). Pattern recognition in medical images using neural networks. *IEEE Transaction on Image and Signal Processing Analysis*.

Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemo-metrics, 6*, 267–281.

Merz, C. J., & Murphy, P. M. (1996). UCI Repository of machine learning databases. Irvine, CA: Department of Information and Computer Science, University of California. Available from http://www.ics.uci.edu/~mlearn/MLRepository.html.

Micheal, R., & Middendorf, M. (1999). An ACO algorithm for the shortest common super sequence problem. New ideas in optimization. Maidenhead: McGraw Hill.

Narendra, P., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset se lection. *IEEE Transactions on Computing, 77*(26), 917–922.

Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters Archive, 15*(11), 1119–1125.

Punch, W. F., Goodman, E. D., Pei, M., Chia-shun, L., Hovland, P., & Enbody, R. (1993). Further research on feature selection and classification using genetic algorithms. In *The proceedings of 5th international conference on genetic algorithm* (pp. 557–564).

Ripley, B. D., & Hjort, N. L. (1996). Pattern recognition and neural networks. New York: Cambridge University Press.

Sampels, M., Blum, C., Mastrolilli, M., & Rossi-Doria, O. (2002). Metaheuristics for group shop scheduling. In *The proceedings of seventh international conference on parallel problem solving from nature, PPSN-VII. Lecture notes in computer science*, Berlin, Germany (Vol. 2439, pp. 631–640).

Schoonderwoerd, R., Holland, O., Bruten, J., & Rothkrantz, L. (1996). Ant-based load balancing in telecommunications networks. *Adaptive Behavior, 5*(2), 169–207.

Schreyer, M., & Raidl, G. R. (2002). Letting ants labeling point features. In *Proceedings of the 2002 IEEE congress on evolutionary computation at the IEEE world congress on computational intelligence* (pp. 1564–1569).

Sim, K. M., & Sun, W. H. (2001). A comparative study of ant-based optimization for dynamic routing. In *The Proceedings of conference on active media technology*. *Lecture notes computer science*, Hong Kong (pp. 153–164).

Stützle, T (1998). An ant approach for the flow shop problem. In *Proceedings of the 6th European congress on intelligent techniques and soft computing* (Vol. 3, pp. 1560–1564). Germany: Aachen.

Yang, J., & Honavar, V. (1998). *Feature subset selection using a genetic algorithm*. *Feature extraction, construction, and subset selection: A data mining perspective*. New York: Kluwer.

Zhou, Z. H., Jiang, Y., Yang, Y. B., & Chen, S. F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine, 24*(1), 25–36.