

# Parameter determination and feature selection for C4.5 algorithm using scatter search approach

Shih-Wei Lin · Shih-Chieh Chen

Published online: 13 May 2011  
© Springer-Verlag 2011

**Abstract** The C4.5 decision tree (DT) can be applied in various fields and discovers knowledge for human understanding. However, different problems typically require different parameter settings. Rule of thumb or trial-and-error methods are generally utilized to determine parameter settings. However, these methods may result in poor parameter settings and unsatisfactory results. On the other hand, although a dataset can contain numerous features, not all features are beneficial for classification in C4.5 algorithm. Therefore, a novel scatter search-based approach (SS + DT) is proposed to acquire optimal parameter settings and to select the beneficial subset of features that result in better classification results. To evaluate the efficiency of the proposed SS + DT approach, datasets in the UCI (University of California, Irvine) Machine Learning Repository are utilized to assess the performance of the proposed approach. Experimental results demonstrate that the parameter settings for the C4.5 algorithm obtained by the SS + DT approach are better than those obtained by other approaches. When feature selection is considered, classification accuracy rates on most datasets are increased. Therefore, the proposed approach can be utilized to identify effectively the best parameter settings for C4.5 algorithm and useful features.

**Keywords** C4.5 · Decision tree · Scatter search · Optimization · Feature selection

---

S.-W. Lin (✉) · S.-C. Chen  
Department of Information Management,  
Chang Gung University, Taoyuan, Taiwan  
e-mail: swlin@mail.cgu.edu.tw

S.-C. Chen  
Department of Industrial Management, National Taiwan  
University of Science and Technology, Taipei, Taiwan

## 1 Introduction

Machine learning algorithms such as decision tree (DT), back-propagation network (BPN), and support vector machine (SVM) are very popular and can be applied to various areas. However, most of the machine learning algorithms will suffer parameter setting and feature-selection problems (Han and Kamber 2006). Before applying these methods to solve the problems, the parameter values must be set in advance to avoid constructing an over-fitting or under-fitting model. There are no clear rules for the “best” parameter settings and feature selection. In general, it is a trial-and-error process and may affect the classification accuracy of the model. A number of automated techniques have been proposed that search for “good” parameters and selected features. These techniques typically use a hill-climbing approach that starts with an initial value or feature subset; however, these techniques can quickly get a result, and it may easily fall into a sub-optimal situation.

DT can be used easily in numerous domains as it does not impose restrictions (e.g., variables should be independent or variables should follow a normal distribution) that are imposed by other techniques such as discriminant analysis and regression (Berry and Linoff 2001). Moreover, a DT has the following benefits: (1) a DT is a simple method for presenting knowledge, (2) it can handle nominal and categorical data and perform well and the DT has relatively faster learning speed than other classification methods (Han and Kamber 2006), (3) a DT provides information about relevance of features for prediction purposes. As a feature moves close the tree root, its relevance for predicting decisions for a class of data increases (Freitas 1998).

There are various DT algorithms such as iterative dichotomiser 3 (ID3) (Quinlan 1986), classification and

regression tree (CART) (Quinlan 1987), supervised learning in quest (SLIQ) (Han and Kamber 2006), scalable parallelizable induction of decision tree (SPRINT) (Sun et al. 2007), and C4.5 algorithm (Sun et al. 2007). Four criteria—predictive accuracy, speed, robustness, and interpretability—are used to analyze the model created by a DT. The C4.5 algorithm, which satisfies these criteria, is the most popular DT algorithm (Han and Kamber 2006). However, before applying the C4.5 algorithm to solve problems, parameters such as minimum case and pruning confidence level must be set in advance. Parameter settings for C4.5 algorithm must be determined carefully to avoid over- or under-fitting. The minimum case controls the tree whether it grows or not in the construct tree phase, and the pruning confidence level influences whether the node of the tree will be deleted or not in the pruning phase. For example, if the minimum case is set to a small value, the tree may be very large and may have too many branches, and some may reflect anomalies due to noise or outliers. That is, the classification accuracy rate will be very good in the training data, but have poor classification accuracy rate in the testing (unseen) data. This situation is called the over-fitting problem. On the other hand, if the minimum case is set to be a large value, then the tree may be very small and the classification accuracy rate of the training data may be worse. Furthermore, the classification accuracy rate of testing data may be much larger than that of the training data.

Selecting the right set of features for classification is a difficult problem when designing a good classifier. Typically, one does not know a priori which features are relevant for a particular classification task. One common approach is to collect as many features as possible prior to the learning and data-modeling phase. However, in most pattern classification problems, given a large set of potential features, identifying a small subset to classify data object is generally necessary. Data without feature selection might be redundant or noisy, and decrease classification efficiency. The primary benefits of feature selection are as follows: (1) computational cost and storage requirements are reduced; (2) degradation of classification efficiency due to the irrelevant or redundant features used in the training samples is overcome; (3) training and prediction time are reduced, and (4) data understanding and visualization are facilitated (Abe and Kudo 2005). In feature selection, whether each feature is useful must be determined; the task of finding an optimal subset of features is inherently combinatorial. Therefore, feature selection becomes an optimization problem. An optimal approach is then needed to examine all possible subsets. This study presents a novel scatter search-based approach that provides the best parameter settings for C4.5 algorithm, and identifies the beneficial subset of features for

different problems such that the classification accuracy rate of C4.5 algorithm is maximized.

The remainder of this paper is organized as follows. Section 2 reviews previous studies of DTs, feature selection, and scatter search. Section 3 describes the proposed SS + DT approach for determining optimal parameter settings for C4.5 algorithm, and identifies the most beneficial feature subset. Section 4 presents experimental results. Conclusions and future research directions are given in Sect. 5.

## 2 Literature review

### 2.1 Decision tree

Most DTs employ a top-down strategy that recursively partitions a dataset into small subdivisions. These procedures form the basis of a set of tests applied to each tree branch. The tree-like structure is composed of a root node (formed from all data), a set of internal nodes (splits), and a set of terminal nodes (leaves). Each interior node corresponds to a variable; an arc to a child represents a possible value for that variable. A leaf represents the predicted value of a target variable given the values of variables represented by the path from the root node.

The DT constructing process has two principal phases: the growth phase and pruning phase (Kim and Koehler 1995). During the growth phase, for a set of samples in partition  $S$ , a test feature  $X$  is selected for further partitioning the set into  $S_1, S_2, \dots, S_L$ , which are added to the decision tree as children of node  $S$ . Additionally, the node for  $S$  is labeled with test  $X$ , and partitions  $S_1, S_2, \dots, S_L$  are then recursively partitioned.

The interactive dichotomizer 3 (ID3) algorithms (Quinlan 1986, 1987) and their successor C4.5 algorithm (Quinlan 1993) are the primary focus of research in the field of DT learning. During the growth phase, the central choice by the ID3 algorithm is selection, during which features are tested at each node in the most useful way for classifying examples. The C4.5 algorithm uses an information entropy evaluation function as selection criteria (Quinlan 1993). The entropy evaluation function is calculated as follows.

Step 1: Calculate  $Info(S)$  to identify the class in the training set  $S$ .

$$Info(S) = - \sum_{i=1}^k \{ [freq(C_i, S) / |S|] \log_2 [freq(C_i, S) / |S|] \}, \quad (1)$$

where  $|S|$  is the number of cases in the training set,  $C_i$  is a class,  $i = 1, 2, \dots, k$ ,  $k$  is the number of classes, and  $freq(C_i, S)$  is the number of cases in  $C_i$ .

Step 2: Calculate the expected information value,  $\text{Info}_x(S)$ , for feature  $X$  to partition  $S$ .

$$\text{Info}_x(S) = - \sum_{i=1}^L [(|S_i|/|S|)\text{Info}(S_i)], \quad (2)$$

where  $L$  is the number of outputs for feature  $X$ ,  $S_i$  is a subset of  $S$  corresponding to the  $i$ th output, and  $|S|$  is the number of cases in subset  $S_i$ .

Step 3: Calculate the information gained after partitioning according to feature  $X$ .

$$\text{Gain}(X) = \text{Info}(S) - \text{Info}_x(S) \quad (3)$$

Step 4: Calculate the partition information value,  $\text{SplitInfo}(X)$ , acquired for  $S$  partitioned into  $L$  subsets.

$$\text{SplitInfo}(X) = - \sum_{i=1}^L \left[ \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \right] \quad (4)$$

Step 5: Calculate the gain ratio of  $\text{Gain}(X)$  over  $\text{SplitInfo}(X)$ .

$$\text{GainRatio}(X) = \text{Gain}(X)/\text{SplitInfo}(X) \quad (5)$$

The  $\text{GainRatio}(X)$  compensates for the weak point of  $\text{Gain}(X)$ , which is the quantity of information provided by  $X$  in the training set. Thus, the feature with the highest  $\text{GainRatio}(X)$  is adopted as a decision tree root. The gain ratio criterion is robust and results in small trees (Quinlan 1993). In order to avoid the over-fitting, splits can be stopped if a certain specified threshold (e.g., the minimum number of cases for a split search) is met (Osei-Bryson 2007). This is the so-called minimum case.

The aim of the pruning phase is to generalize the DT generated during the growth phase by generating a sub-tree that avoids over-fitting the training data. The actions in the pruning phase are often called post-pruning. The approach taken in C4.5 is called the confidence level, which uses the estimated error to determine whether the tree built in growth phase requires pruning or not at certain nodes. The probability of error cannot be determined exactly; however, there exists a probability distribution that is generally summarized as a pair of confidence limits. For a given confidence level, the upper limit of this probability can be determined from the confidence limits for the binomial distribution. Then, C4.5 simply equates the estimated error rate at a leaf with this upper limit, based on the argument that the tree has been constructed to minimize observed error rate (Quinlan 1993).

For constructing the DT model, the most difficult task is to obtain a good balance between accuracy and simplicity. Unfortunately, the minimum cases ( $M$ ) for the leaf and pruning confidence level ( $CF$ ) are varied for different problems. Generally, the  $M$  is preferred to be a high value when data is noisy; on the other hand, the  $CF$  should be a

lower value when the test error rate of pruned tree exceeds the estimated error rate. Determining these two parameters is an optimization problem (Quinlan 1993).

John (1994) observed that determining how to set parameters is an important issue associated with C4.5 algorithms. This study investigates several cross-validation-based approaches (C4.5\*, CVC4.5, and small CVC) to identify the best parameter values for the C4.5 algorithm. Moreover, Kohavi and John (1995) developed a best-first search algorithm to determine the parameter values of C4.5 algorithm using the minimum estimated error. Many studies extended the DT. Carvalho and Freitas (2002) utilized a genetic algorithm to discover small-disjunct rules and compared the results obtained by three versions of the C4.5 algorithm alone, and in eight public domain datasets. Gray and Fan (2008) proposed a genetic algorithm approach to construct DTs called randomly generated evolved tree (TARGET) that performs a better search of the tree model space than the greedy search algorithm. Aitkenhead (2008) created an evolutionary approach to increase DT flexibility using co-evolving competition between the tree and training dataset. Orsenigo and Vercellis (2004) developed an algorithm for creating a DT in which multivariate splitting rules were based on a new concept of discrete support vector machines (LDSDT<sub>TS</sub>). These studies focused on tree construction and rule generation. However, they did not consider parameter settings and feature selection simultaneously.

## 2.2 Feature selection

The DT requires a dataset for model construction. A dataset can have many features; however, not all features are useful for classification. When a dataset has considerable noise and complex dimensionality, a DT may have limitations associated with learning the classification patterns. Although the C4.5 algorithm has a feature-selection strategy that encompasses its learning performance, this strategy is not optimal. Correlated and irrelevant features may reduce the performance of the induced classifier (Perner and Apte 2004).

Feature selection can be defined as selecting the smallest subset of an original set of features that are necessary and sufficient for describing a target concept. The approaches for feature selection can be categorized into two models: a filter model and a wrapper model (Liu and Motoda 1998). Filter models utilize statistical approaches, such as factor analysis (FA), independent component analysis (ICA), principal component analysis (PCA), and discriminant analysis (DA), to investigate indirect performance measures, primarily based upon distance and information measures in feature selection. Sun et al. (2007) developed a PCA method on the C4.5 algorithm. The PCA is utilized to

reduce the number of features and C4.5 algorithm is trained to generate a DT model for diagnosis of rotating machinery. Last et al. (2001) presented an information-theoretical algorithm for feature selection to enhance C4.5 algorithm; this algorithm finds a set of features by removing irrelevant and redundant features. Perner and Apte (2004) created C4.5 algorithm and a contextual merit (CM) algorithm to select features. They showed that accuracy of the C4.5 classifier can be improved with an appropriate feature pre-selection phase for the learning algorithm. Although this model is fast, the resulting feature subset may not be optimum (Liu and Motoda 1998).

In the wrapper approach, feature subset selection is performed by an induction algorithm as a black box. The feature subset selection algorithm conducts a search for a good subset using the induction algorithm as part of the evaluation function. Some studies have proposed that when the objective is to minimize classifier classification error rate, and measurement cost for all features is equal, then the classification accuracy rate of the classifier is most appealing. That is, a classifier should be constructed with the goal of achieving the highest classification accuracy rate possible, and selecting the features used by the classifier as optimal features. This model is the so-called wrapper model, which uses selection methods to choose feature subsets and then evaluates the selection result after the classification algorithm calculates the accuracy rate. When the relevant features can be selected or noise removed, the classification accuracy rate of the classifier can be improved.

Smith and Bull (2005) utilized genetic programming to preprocess data then applied the C4.5 algorithm to ten well-known datasets from the UCI (University of California, Irvine) repository. López et al. (2006) developed three scatter search-based algorithms, the sequential scatter search with a greedy combination (SSS-GC), sequential scatter search with a reduced greedy combination (SSS-RGC), and parallel scatter search (PSS), to solve the feature-selection problem using three algorithms, the instance-based algorithm, Naive Bayes algorithm, and C4.5 algorithm. However, these algorithms do not consider parameter settings for the C4.5 algorithm. Thus, the optimal solution may be excluded. Few studies have considered parameter settings and feature selection simultaneously for the C4.5 algorithm. As irrelevant and redundant features exist in classification problems, when parameter settings and feature selection are not considered simultaneously, the optimal model may be excluded. Su and Shiu (2003) proposed the GA/DT approach to determine the optimization parameter values and a feature subset for production control systems. The GA/DT approach was only adopted for a specific problem; thus, further comparisons cannot be made.

In order to illustrate the feature-selection problem in the DT algorithm, an example shown in Table 1 was used.

**Table 1** Data values of example

No.	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	Class label	No.	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	Class label
1	80	88	63	0	1	3	17	80	88	63	1	1	3
2	63	53	64	1	0	1	18	63	53	34	1	1	2
3	72	58	62	0	0	2	19	80	38	34	1	1	3
4	61	70	64	1	0	1	20	61	70	64	0	1	1
5	76	78	64	0	0	1	21	76	35	64	0	1	2
6	78	47	70	0	0	2	22	80	47	70	0	1	3
7	67	62	65	0	0	1	23	67	62	62	0	1	2
8	71	70	64	1	0	1	24	71	70	64	1	0	1
9	82	75	65	0	1	3	25	82	75	65	0	0	1
10	80	57	61	0	1	3	26	66	57	61	0	1	2
11	52	75	75	1	1	1	27	52	75	75	1	1	1
12	63	32	66	1	0	1	28	63	72	66	0	0	2
13	63	30	56	0	0	2	29	63	30	66	1	0	1
14	78	41	70	0	0	2	30	78	41	70	0	1	2
15	80	53	64	1	1	3	31	63	53	64	1	0	1
16	80	72	73	0	1	1	32	80	72	63	0	0	1

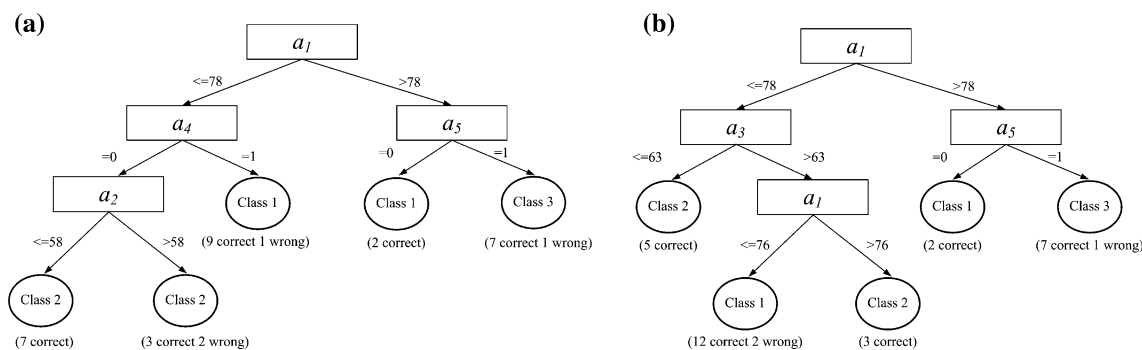
This example has 32 instances, and 5 variables,  $a_1$ ,  $a_2$ ,  $a_3$ ,  $a_4$  and  $a_5$ , can be used to classify its class. There are three classes, labeled 1, 2, and 3.

If default parameter setting in C4.5 ( $M = 2$  and  $CF = 25\%$ ) is used and all of five variables (feature selection is not applied) are fed to C4.5, the C4.5 algorithm will use four variables ( $a_1$ ,  $a_2$ ,  $a_4$  and  $a_5$ ) to build the classification model. The classification accuracy rate is 87.5% (28/32) and the tree structure is shown in Fig. 1a. If a certain feature selection method is performed, only three variables,  $a_1$ ,  $a_3$  and  $a_5$ , are necessary to construct the classification model. The classification accuracy rate is 90.63% (29/32), and the tree structure is shown in Fig. 1b.

Figure 1 points out that the C4.5 algorithm may easily fall into a local optimal (lower classification accuracy rate) due to the greedy search that is used. If feature selection is not used before constructing the DT model, the root node is  $a_1$ . The left nodes are  $a_4$  and  $a_2$  and the right node is  $a_5$ . If feature selection is used before constructing the DT model, only three variables are needed to construct the C4.5 model with a higher classification accuracy rate. The most important variable (root node) is the same, but the second important variable changes to  $a_3$  in the left node. Moreover, the feature selection can use fewer variables to construct the classification model, which has higher classification accuracy and may have simplified tree structure.

### 2.3 Scatter search

Introduced by Glover (1977), the scatter search (SS) is a population-based approach that starts with a collection of



**Fig. 1** Tree structure of the DT example ( $M = 2$  and  $CF = 25\%$ ). **a** Without feature selection. **b** With feature selection

reference solutions obtained by applying preliminary heuristic processes. In 1998, Glover published the SS template (Glover 1998) which presents an algorithmic description of the SS method and is considered a milestone in SS literature; many different applications were subsequently developed that have shown potential for solving various complicated optimization problems (Martí 2006). The scatter search is a powerful meta-heuristic approach and has been applied to many various applications successfully. A sample list of these applications can be found in Laguna and Martí (2003). To the best of our knowledge, some studies applied the scatter search to machine learning algorithm. For examples, Su et al. (2005) proposes a hybrid procedure combining neural networks and scatter searches to optimize the continuous parameter design of back-propagation neural network. López et al. (2006) developed three scatter search-based algorithms, to solve the feature-selection problem. Rasha et al. (2006) proposed a scatter search-based automatic clustering problem to discover cluster number and cluster centers without prior knowledge of a possible number of class, and without any initial partition. However, they did not apply SS to parameter determination and feature selection simultaneously.

Briefly, unlike genetic algorithms, an SS operates on a small set of solutions and makes only limited use of randomization as a proxy for diversification when searching for a globally optimal solution. Based on formulations initially proposed for combining decision rules and constraints, an SS uses strategies to combine solutions and create a balance between quality and diversification in the reference set.

Generally, the principal components in an SS can be described as follows.

- (1) *A diversification generation method* generates a population of solutions that satisfy a critical level of diversity.
- (2) *An improvement method* transforms a trial solution into an enhanced feasible trial solution.

- (3) *A reference set update method* builds and maintains a reference set that is a collection of high-quality solutions and diverse solutions. The reference set is the basis for creating new combined solutions.
- (4) *A subset generation method* is applied to the reference set and produces a subset of solutions as a basis for creating combined solutions.
- (5) *A solution combination method* transforms a given subset of solutions produced by the subset generation method into one or more combined new solutions.

Since each of these methods in italics can be implemented in a variety of ways and with different degrees of complexity, the SS procedure is very adaptable to different problems. Only four of the five components are strictly required in an SS. The improvement method, the only exception, is applied to generate high-quality solutions when they are not provided by other components.

### 3 The proposed approach

This study presents a novel SS-based approach that provides the best parameter settings for the C4.5 algorithm and finds the beneficial subset of features for different problems, such that the classification accuracy rate of the C4.5 algorithm is maximized. The way in which to apply scatter search to parameter determination and feature selection of DTs, solution representation and objective function calculation, and the procedure of the SS + DT is discussed as follows.

#### 3.1 Solution representation and objective function value

This study adopted an SS-based approach, called the SS + DT, for parameter determination and feature selection in the C4.5 algorithm. For the C4.5 algorithm without feature selection two parameter values,  $M$  and  $CF$ , were necessary. For the C4.5 algorithm with feature selection, if



$n$  features were needed to determine which features were selected, then additional  $n$  indicative variables had to be identified. The values of  $n$  variables range from 0 to 1. If a variable value is  $\leq 0.5$ , then its corresponding feature was not chosen. Conversely, if a variable value  $> 0.5$ , then its corresponding feature was chosen. For example, if the dataset had four features and the C4.5 algorithm requires two parameters, there were six variables used as shown in Fig. 2. This solution can be decoded as follows. The  $M$  is 5, the CF is 32%, and the selected features are 1, 2, and 4. The range of  $M$  and CF in the solution representation was 0–1, and the real value of  $M$  and CF was scaled to a specific range related to the input dataset.

Although four criteria exist, (predictive accuracy, speed, robustness, and interpretability) for evaluating the model created by the DT, the classification accuracy rate was employed most. Therefore, the classification accuracy rate was adopted as the objective function in this study.

### 3.2 Applying SS for C4.5 DT

The proposed SS + DT approach follows the steps of SS template (Laguna and Martí 2003) and is described as follows.

- (1) Diversification generation method: Population  $P$  with  $P_{\text{size}}$  solutions was generated randomly. Because all values ranged from 0 to 1 in the solution presentation, each value was uniformly generated from 0 to 1.
- (2) Improvement method: The improvement method was optional in the SS template, and therefore, was not applied in this study.
- (3) Reference set update method: The reference set, RefSet, collects both high-quality solutions and diverse solutions that are used to generate new solutions by applying the Combination method. The size of the reference set was  $b = b_1 + b_2 = |\text{RefSet}|$ , where  $b_1$  was the number of high-quality solutions and  $b_2$  was the number of diverse solutions. Construction of the initial reference set starts with selecting  $b_1$  best solutions (solutions with the highest classification rates) from  $P$ . These solutions were added to RefSet and deleted from  $P$ . For each solution in the  $P$ -RefSet, the minimum Euclidean distance to the solutions in RefSet was calculated. The solution with the maximum of the minimum distances was selected. This solution was then added to RefSet and deleted from  $P$ , and the minimum distances were updated accordingly. This process was repeated  $b_2$

5	32%	0.64	0.75	0.33	0.91
---	-----	------	------	------	------

Fig. 2 Solution representation of SS + DT

times, where  $b_2 = b - b_1$ . Thus, the resulting reference set had  $b_1$  high-quality solutions and  $b_2$  diverse solutions.

- (4) Subset generation method: The size of subsets was set to 2; that is, only subsets consisting of all pair-wise combinations of solutions in RefSet were considered. Therefore, at maximum,  $b(b - 1)/2$  subsets exist.
- (5) Solution combination method: The method employed consisted of finding linear combinations of reference solutions. Each combination of two reference solutions, denoted as  $Y'$  and  $Y''$ , were employed to create three trial solutions. These three trial solutions were (1)  $Y = Y' - d$ , (2)  $Y = Y' + d$ , and (3)  $Y = Y'' + d$ , where  $d = u(Y'' - Y')/2$  and  $u$  was a random number with values of 0–1.

### 3.3 Demonstration of SS + DT procedure

For demonstration purpose, the dataset in Sect. 2.2 was used and the parameter values for SS + DT were set as follows:  $P_{\text{size}} = 8$ ,  $b = 4$ ,  $b_1 = 2$ ,  $b_2 = 2$ ,  $M$  ranges from 2 to 5, and CF ranges from 1 to 30%.

#### 3.3.1 Diversification generation method

Because all variables were ranged from 0 to 1 in the solution,  $P_{\text{size}}$  solutions could be generated by setting each variable uniformly generated from 0 to 1, and the results are shown in Table 2.

In this table, solution 1 represents that  $M = 3$  ( $0.295 \times (5 - 2) + 2$ , round up to integer), CF = 28% ( $0.921 \times (30 - 1\%) + 1\%$ , round up to integer), the features 1, 2, 3, and 5 ( $x_3 > 0.5$ ,  $x_4 > 0.5$ ,  $x_5 > 0.5$ ,  $x_7 > 0.5$ ) were used for creating a DT model. Meanwhile, solution 2 represents that  $M = 3$  ( $0.258 \times (5 - 2) + 2$ ), CF = 8% ( $0.233 \times (30 - 1\%) + 1\%$ ), the features 1, 3, and 4 ( $x_3 > 0.5$ ,  $x_5 > 0.5$  and  $x_6 > 0.5$ ) were used for creating a DT model. Other solutions could be described in the same way. After the DT model was created, the classification

Table 2 Diverse solutions

Solution	$x_1$ (M)	$x_2$ (CF)	$x_3$ (F1)	$x_4$ (F2)	$x_5$ (F3)	$x_6$ (F4)	$x_7$ (F5)	obj(X) (%)
$X^1$	0.295	0.921	0.751	0.783	0.534	0.234	0.840	84.4
$X^2$	0.258	0.233	0.522	0.294	0.689	0.893	0.007	78.1
$X^3$	0.948	0.923	0.603	0.837	0.346	0.926	0.387	81.2
$X^4$	0.512	0.594	0.248	0.459	0.475	0.678	0.887	56.2
$X^5$	0.419	0.005	0.906	0.094	0.125	0.800	0.487	78.1
$X^6$	0.051	0.070	0.464	0.710	0.846	0.518	0.004	62.5
$X^7$	0.967	0.846	0.417	0.366	0.292	0.607	0.016	56.2
$X^8$	0.035	0.527	0.856	0.687	0.416	0.424	0.859	87.5

accuracy rate for each solution could be calculated and is shown in the last column of Table 2.

### 3.3.2 Reference set update method

Table 3 shows the best  $b_1$  solutions in  $P$ , which were immediately added to the RefSet. The first column in this table shows the solution number in  $P$ , followed by the variable values and the objective function value. Therefore, solutions  $X^8$  and solution  $X^1$  in  $P$  had the highest objective function value and became the first solution and the second solution in RefSet, respectively.

We then calculated the minimum distance  $d_{\min}(X)$  between each solution  $X$  in  $P$ -RefSet and the solution  $Y$  currently in RefSet. That is,  $d_{\min}(X) = \text{Min}_{Y \in \text{RefSet}} \{d(X, Y)\}$ , where  $d(X, Y)$  is the Euclidean distance between  $X$  and  $Y$ .

$$\text{Mathematically, } d(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}.$$

For example, the minimum distance between solution (i.e.,  $X^2$ ) in Table 2 and the RefSet solution in Table 3 (i.e.,  $X^8$  and  $X^1$ ) was calculated as follows:

$$\begin{aligned} d_{\min}(X^2) &= \text{Min}\{d(X^2, X^8), d(X^2, X^1)\} \\ &= \text{Min}(1.193, 1.289) = 1.193. \end{aligned}$$

The maximum  $d_{\min}$  value for the solution in  $P$ -RefSet corresponds to Solution  $X^7$  in  $P$  ( $d_{\min}(X^7) = 1.300$ ). We added this solution to RefSet, deleted it from  $P$ , and updated the  $d_{\min}$  values. The new maximum  $d_{\min}$  value of 0.1135 corresponded to solutions  $X^6$  in  $P$ , the diverse solutions added to RefSet are shown in Table 4.

### 3.3.3 Subset generation method

This method consisted of a generated subset of reference solutions to be subjected to the combination method. Due to the size of subset being set to 2 in this study, there was a maximum of  $b \times (b - 1)/2 = 4 \times (4 - 1)/2 = 6$  subsets.

**Table 3** High-quality solution in RefSet

Solution	$x_1$ (M)	$x_2$ (CF)	$x_3$ (F1)	$x_4$ (F2)	$x_5$ (F3)	$x_6$ (F4)	$x_7$ (F5)	obj(X) (%)
$X^8$	0.035	0.527	0.856	0.687	0.416	0.424	0.859	87.5
$X^1$	0.295	0.921	0.751	0.783	0.534	0.234	0.840	84.4

**Table 4** Diverse solutions in RefSet

Solution	$x_1$ (M)	$x_2$ (CF)	$x_3$ (F1)	$x_4$ (F2)	$x_5$ (F3)	$x_6$ (F4)	$x_7$ (F5)	obj(X) (%)
$X^7$	0.967	0.846	0.417	0.366	0.292	0.607	0.016	56.2
$X^6$	0.051	0.070	0.464	0.710	0.846	0.518	0.004	62.5

### 3.3.4 Combination method

Suppose solutions  $X^1$  and  $X^8$  were selected for the use in the combination method. Three trial solutions were (1)  $Y = Y' - d$ , (2)  $Y = Y' + d$ , and (3)  $Y = Y'' + d$ , where  $d = u(Y'' - Y')/2$  and  $u$  was a random number with values of 0–1. It should be noted that if the value of a variable was lower than 0, the value was set to 0; if the value of a variable was large than 1, the value was set to 1.

Suppose  $u$  was 0.543  $Y'' = X^1$  and  $Y' = X^8$ , three trial solutions were then obtained and are shown in Table 5. The best solution in Table 4 is  $X^9$  with the objective function value of 90.6%.

Using the subsets that are generated from the subset generation method, more combinations were obtained from the subset generation method and could be used to create additional trial solutions. The search continued in a loop that consisted of applying the combination method followed by the reference update method. This loop terminated when termination conditions were met.

## 3.4 System architecture of SS + DT

The SS-based approach for parameter determination and feature selection of a DT was constructed following the steps and detailed explanation as follows.

- (1) Input dataset and data pre-processing: After the dataset was input, the  $k$ -fold approach developed by Salzberg was applied with  $k = 10$  (Salzberg 1997). Thus, the dataset was segmented into 10 portions, with each portion of the data sharing the same proportion of each class of data. Nine data portions were applied in the training process, whereas the remaining portion was utilized in the test process. Since the number of data in each class was not a multiple of 10, the dataset could not be partitioned equitably. However, the ratio of the number of data in the training set to the number of data in the testing set was maintained as closely as possible to 9:1.
- (2) Feature subset selection and determination of parameter values: Each solution generated by the SS was the selected subset of features and parameter values. The selected features, parameter values, and training

**Table 5** New trial solution generated form the combination of solution  $X^1$  and  $X^8$

Solution	$x_1$ (M)	$x_2$ (CF)	$x_3$ (F1)	$x_4$ (F2)	$x_5$ (F3)	$x_6$ (F4)	$x_7$ (F5)	obj(X) (%)
$X^9$	0.000	0.420	0.885	0.810	0.303	0.476	0.864	80.4
$X^{10}$	0.106	0.634	0.804	0.461	0.529	0.372	0.854	87.5
$X^{11}$	0.366	1.000	0.722	0.111	0.947	0.182	0.835	90.6

dataset were then used for building the DT classifier model. Each DT classifier model was created by calling the C4.5 algorithm provided by Quinlan (1993).

- (3) Objective function value calculation: After the classification model was constructed, the objective function value can be calculated. The higher the classification accuracy rate, the better is the objective function value.
- (4) Termination criteria: When termination criteria were satisfied, the process ended; otherwise, the next iteration was run. The termination criterions utilized in this study were the maximal solutions evaluated,  $S_{\max}$ , and the allowable number of successive non-improving solutions evaluated,  $N_{\text{non-improving}}$ . That is, the number of solutions evaluated exceeds  $S_{\max}$  or the best objective function value obtained was not improved in  $N_{\text{non-improving}}$  successive solutions and the SS procedure was terminated.
- (5) The SS process: In this step, the system generated other solutions using SS as described in Sect. 3.2.

#### 4 Experimental results

The proposed SS + DT approach was implemented using C language and a Windows XP operating system on a personal computer with a Pentium IV 3.0 GHz CPU and 512 MB of RAM. To verify the effectiveness of the proposed SS + DT approach, 23 datasets from the UCI Machine Learning Repository (Hettich et al. 1998) were implemented. These dataset included several high dimensional datasets (Anneal, Breast cancer new, Ionosphere structure, and Sonar), several large datasets (Adult, Segmentation, and Wave), and one high dimensional and large dataset (Connect). Table 6 shows the number of features, instances, and classes for these datasets. As the C4.5 algorithm can handle missing values, the missing value was replaced by a “?” and the instance with missing values was reserved during the experiments. The range of  $M$  is 2–20, whereas the range of CF is 0.01–0.35 (1–35%).

After running a few datasets with several combinations of parameter settings under the situation in which feature selection was not considered; that is, only  $M$  and CF values in the C4.5 algorithm were necessary to be searched, the parameter values for  $P_{\text{size}}$ ,  $b_1$ ,  $b_2$ ,  $N_{\text{non-improving}}$  and  $S_{\max}$  for a SS were 30, 5, 5, 300, and 1,500, respectively. If the number of solutions evaluated exceeded 1,500, the proposed approach was terminated. If the best solution obtained was not improved in 300 successive solutions, the proposed approach was also terminated.

Because tenfold cross-validation was utilized, the minimum cases,  $M$ , confidence level, CF, and classification

**Table 6** Datasets from the UCI repository

Dataset	Number of features	Number of instances	Number of Classes
Adult	14	45,222	2
Anneal	38	798	5
Australian (credit scoring)	15	690	2
Breast cancer (new)	30	569	2
Breast cancer (original)	9	699	2
Bupa liver	6	345	2
Car evaluation	6	1,728	4
Connect	42	67,557	3
CRX	15	690	2
Glass	9	214	6
Heart disease	13	303	2
Hepatitis	19	155	2
House	13	506	2
Ionosphere structure	34	351	2
Monk2	7	601	2
New thyroid	5	215	3
Pima Indians diabetes	8	768	2
Segmentation	19	2,310	7
Sonar	60	208	2
Vehicle	18	846	4
Voting	16	348	2
Wave	21	5,000	3
Wine	13	178	3

accuracy rate were obtained by executing the proposed SS + DT approach once for each fold. As the proposed SS + DT approach was non-deterministic, the solutions obtained may not be equal for the same data. Thus, the proposed SS + DT approach was executed three times for each fold in the dataset to calculate average classification accuracy rate. That is, the SS + DT approach executed 30 ( $10 \times 3$ ) times for each dataset. The classification results obtained were then compared with those obtained by the C4.5\*, CVC 4.5, Small CVC 4.5 (John 1994), and the C4.5 algorithm (using default value) (Table 7). Classification accuracy rates were cited from their original studies. Notably, only the classification accuracy rate for Monk2 (66.95% in SS + DT, 67.1% in the C4.5\*, and 67.1% in the small CVC 4.5) obtained by SS + DT was not the best among the other approaches. However, other classification accuracy rates obtained by the proposed SS + DT approach were the best.

For feature selection, because the solution space was increased by  $n$  indicative variables,  $S_{\max}$  was increased to 3,000 and the other parameters for SS + DT remained unchanged. The classification results obtained were compared with those obtained by the Sequential Scatter Search with Greedy Combination (C4.5 + SSS-GC) (López et al.



**Table 7** Comparison between various approaches without feature selection (%)

Dataset	SS + DT	C4.5 <sup>a</sup>	CVC 4.5	Small CVC 4.5	C4.5
Australian (credit scoring)	90.98 ± 4.17 <sup>a</sup>	85.0	81.29 ± 0.04	81.5	81.0
Breast cancer (new)	97.50 ± 1.82 <sup>a</sup>	95.7	94.34 ± 1.10	94.4	94.4
Monk2	66.95 ± 1.70	67.1 <sup>a</sup>	65.20 ± 1.04	67.1 <sup>a</sup>	65.0
Pima Indians diabetes	77.78 ± 4.01 <sup>a</sup>	74.6	72.74 ± 1.23	72.3	68.4

Plus or minus denotes the standard deviation

<sup>a</sup> The higher classification accuracy rate among approaches

**Table 8** Comparison of other approaches with feature selection (%)

Dataset	SS + DT	C4.5 + SSS–GC	C4.5 + GA	C4.5 + GAP
Anneal	95.14 ± 1.96 <sup>a</sup>	91.45 ± 2.19	92.43 ± 2.03	N/A
Australian	92.65 ± 3.15 <sup>a</sup>	83.94 ± 2.17	83.51 ± 1.26	N/A
Breast cancer (new)	98.63 ± 1.10 <sup>a</sup>	N/A	N/A	95.71
Breast cancer (original)	98.77 ± 1.11 <sup>a</sup>	N/A	N/A	95.56
Bupa liver	77.31 ± 3.10 <sup>a</sup>	N/A	N/A	66.55
Glass	84.66 ± 3.90 <sup>a</sup>	N/A	N/A	71.84
Ionosphere	98.10 ± 2.39 <sup>a</sup>	87.07 ± 3.81	88.43 ± 2.19	90.69
New thyroid	99.44 ± 1.67 <sup>a</sup>	N/A	N/A	96.49
Pima Indians diabetes	80.27 ± 3.66 <sup>a</sup>	N/A	N/A	73.64
Sonar	96.37 ± 3.97 <sup>a</sup>	N/A	N/A	75.89
Vehicle	82.67 ± 2.63 <sup>a</sup>	N/A	N/A	72.11
Wine	100.00 ± 0.00 <sup>a</sup>	N/A	N/A	96.10

Plus or minus denotes the standard deviation

N/A approach did not use this dataset for test

<sup>a</sup> The higher classification accuracy rate among approaches

2006), Genetic algorithm (C4.5 + GA), and genetic programming (C4.5 + GAP) (Smith and Bull 2005) (Table 8).

Classification accuracy rates on these three methods were from their original studies. Based on the classification accuracy rates obtained, the proposed SS + DT approach was the best for all datasets.

Furthermore, the classification results obtained by SS + DT with feature selection were compared with those obtained by other DT-based methods, including Hybrid C4.5/GA (Carvalho and Freitas 2002), TARGET (Gray and Fan 2008), co-evolutionary DT (Aitkenhead 2008), and LDSDT<sub>TS</sub> (Orsenigo and Vercellis 2004) (Table 9). Their classification accuracy rates were also obtained from their original studies. Only classification accuracy rates for Connect (73.30% in SS + DT and 75.93% in Hybrid C4.5/GA) and Hepatitis (83.42% in SS + DT and 84.97% in Hybrid C4.5/GA) obtained by the SS + DT were not the best among the other approaches. The other classification accuracy rates obtained by the proposed SS + DT approach were best, demonstrating that the proposed SS + DT approach performs well in various problems. To sum up, only 3 out of 23 of the classification accuracy rates

obtained by the proposed approach for dataset were worse than those of other approaches.

Finally, to determine whether a significant difference existed between the proposed SS + DT approach with feature selection and that without feature selection, the classification results obtained by the proposed SS + DT approach with and without feature selection were compared (Table 10). Although classification accuracy rate for Monk2 was reduced, the classification accuracy rates for all other datasets were increased. For datasets where the classification accuracy rate increased, only one dataset (Anneal) was not significantly different; all other datasets have values of  $P < 0.05$ , meaning that significant difference existed.

Table 10 shows that all the average computations in the acceptable range, and in some datasets the computation cost between without/with feature selection, were quite small (breast-cancer new, breast-cancer original, bupa liver, car evaluation, CRX, monk2, new thyroid, pima Indians diabetes, sonar, voting, and wine). The  $N_{\text{non-improving}}$  stop criterion worked well, it balanced the SS + DT algorithm between the model computation cost

**Table 9** Comparison between the SS + DT approach with feature selection and other approaches (%)

Dataset	SS + DT	Hybrid C4.5/GA ( $S = 10$ )	TARGET	Co-evolutionary DT	LSDST <sub>TS</sub>
Adult	83.95 <sup>a</sup>	79.83	N/A	N/A	N/A
Breast cancer (new)	98.63 <sup>a</sup>	N/A	N/A	N/A	97.8
Breast cancer (original)	98.77 <sup>a</sup>	N/A	95.5	N/A	N/A
Bupa liver	77.31 <sup>a</sup>	N/A	N/A	N/A	75.3
Car evaluation	90.64 <sup>a</sup>	N/A	N/A	89.20	N/A
Connect	75.30	75.93 <sup>a</sup>	N/A	N/A	N/A
CRX	90.21 <sup>a</sup>	86.12	N/A	N/A	N/A
Glass	84.66 <sup>a</sup>	N/A	N/A	82.40	N/A
Heart disease	89.18 <sup>a</sup>	N/A	N/A	N/A	85.2
Hepatitis	83.42	84.97 <sup>a</sup>	N/A	N/A	N/A
House	97.56 <sup>a</sup>	N/A	N/A	N/A	96.5
Ionosphere	98.10 <sup>a</sup>	N/A	N/A	N/A	94.6
New thyroid	99.44 <sup>a</sup>	N/A	93.6	N/A	N/A
Pima Indians diabetes	80.27 <sup>a</sup>	N/A	77.0	N/A	80.2 <sup>a</sup>
Segmentation	97.55 <sup>a</sup>	93.62	N/A	N/A	N/A
Voting	97.31 <sup>a</sup>	92.30	N/A	N/A	N/A
Wave	85.22 <sup>a</sup>	79.60	N/A	N/A	N/A

N/A approach did not use this dataset for test

<sup>a</sup> The higher classification accuracy rate among approaches

**Table 10** Experimental results of SS + DT with/without feature selection on datasets

Dataset	With feature selection			Without feature selection		P-value
	Accuracy rate (%)	No. selected features	Avg. time (s)	Accuracy rate (%)	Avg. time (s)	
Adult	83.95 ± 3.24	9.80 ± 0.92	3,199.41	83.75 ± 3.24	3,694.34	0.001079
Anneal	95.14 ± 1.96	20.63 ± 1.75	395.57	95.12 ± 1.90	196.01	0.944067
Australian (credit scoring)	92.65 ± 3.15	7.83 ± 1.37	262.33	90.98 ± 4.17	160.80	<0.000001
Breast cancer (new)	98.63 ± 1.10	5.57 ± 1.33	9.33	97.50 ± 1.82	6.98	<0.000001
Breast cancer (original)	98.77 ± 1.11	13.37 ± 1.92	35.10	97.18 ± 1.97	34.53	<0.000001
Bupa liver	77.31 ± 3.10	4.60 ± 0.77	8.17	72.86 ± 5.57	8.26	<0.000001
Car evaluation	90.64 ± 1.86	5.30 ± 0.47	8.51	90.30 ± 1.74	6.62	0.020409
Connect	75.30 ± 5.56	26.03 ± 2.80	5,458.56	74.03 ± 5.61	2,701.84	<0.000001
CRX	90.21 ± 2.91	7.57 ± 1.00	12.86	88.94 ± 3.30	10.77	0.000145
Glass	84.66 ± 3.90	5.20 ± 1.54	15.72	78.77 ± 4.05	11.77	<0.000001
Heart disease	89.18 ± 2.71	8.00 ± 1.14	10.03	83.06 ± 5.94	7.24	0.000003
Hepatitis	83.42 ± 3.69	10.67 ± 3.65	10.39	79.72 ± 7.51	6.30	0.007067
House	97.56 ± 2.07	8.77 ± 1.10	61.36	96.64 ± 2.01	45.24	0.001194
Ionosphere structure	98.10 ± 2.39	15.77 ± 2.13	28.75	94.64 ± 2.88	23.63	<0.000001
Monk2	66.39 ± 1.32	1.67 ± 1.73	4.76	66.95 ± 1.70	4.94	–
New thyroid	99.44 ± 1.67	3.17 ± 0.38	6.00	95.98 ± 2.64	5.05	<0.000001
Pima Indians diabetes	80.27 ± 3.66	6.07 ± 1.03	17.47	77.78 ± 4.01	16.93	0.000043
Segmentation	97.55 ± 0.70	9.33 ± 1.67	143.93	96.48 ± 1.21	130.92	0.000003
Sonar	96.37 ± 3.97	28.77 ± 3.38	42.37	86.83 ± 3.49	42.24	<0.000001
Vehicle	82.67 ± 2.63	11.10 ± 2.07	54.30	78.88 ± 3.19	45.44	0.000010
Voting	97.31 ± 2.67	6.47 ± 1.96	6.83	96.83 ± 2.71	4.93	0.022623
Wave	85.22 ± 5.09	12.63 ± 1.40	1,526.42	84.04 ± 5.25	1,518.99	0.000282
Wine	100.00 ± 0.00	6.20 ± 1.24	7.82	95.90 ± 6.47	5.45	0.001918

Confidence level  $\alpha = 0.05$ 

Plus or minus denotes the standard deviation

– The classification accuracy rate is reduced with feature selection

and the classification accuracy rate. The proposed approach can cope with high dimensional datasets (Anneal, Breast cancer new, Ionosphere structure, and Sonar), large datasets (Adult, Segmentation, and Wave), and high dimensional and large dataset (Connect). The result showed the proposed approach performed well and the feature selection could enhance the classification accuracy rate and remove irrelevant or redundant features.

In order to verify the problem of the over-fitting, and under-fitting, both the classification accuracy rates on the training data and the testing data of SS + DT are shown in Table 11. Because there was no large difference between the training data and testing data, the use of the proposed approach seems not to have suffered the problem of over-fitting and under-fitting. Moreover, in Table 11 we provide the average tree size (average number of nodes) of each dataset. It can be noted that even feature selections did not have significant improvement of classification accuracy rates in several datasets; the SS + DT can remove some irrelevant features and produce smaller tree structure for these datasets.

Table 11 shows that the proposed approach can provide high-quality classification accuracy rate of the training data

and the testing data, and the feature selection could help the C4.5 algorithm reduce or keep the same tree size while classification accuracy rate is increased or keeps the similar result.

This study applied scatter search for C4.5 algorithm to determine the parameter and beneficial feature subset simultaneously for different problems to obtain higher classification accuracy rate. The experimental results showed that the scatter search is indeed beneficial for C4.5 to determine the parameter values and feature selection. Therefore, the scatter search has the potential to be applied to other machine learning algorithms in future research.

## 5 Conclusion and future research

Machine learning algorithms such as DT, BPN, and SVM are very popular and can be applied to various areas. However, most machine learning algorithms will suffer the parameter setting and feature selection problems. This study applied the SS-based approach to search for the best parameter settings for the C4.5 algorithm. This study used high dimensional datasets (Anneal, Breast cancer new,

**Table 11** Validation of the proposed scatter-based meta-heuristics for the training and testing result (%)

Dataset	With feature selection			Without feature selection		
	Training accuracy	Testing accuracy	Tree size	Training accuracy	Testing accuracy	Tree size
Adult	84.22	83.95	335.30	84.40	83.75	379.90
Anneal	95.37	95.14	65.20	95.66	95.12	75.60
Australian (credit scoring)	92.72	92.65	21.90	91.74	90.98	24.00
Breast cancer (new)	98.63	98.63	8.20	97.71	97.50	9.60
Breast cancer (original)	98.87	98.77	14.80	98.07	97.18	13.60
Bupa liver	78.83	77.31	11.80	76.65	72.86	14.60
Car evaluation	95.58	90.64	137.90	95.31	90.30	145.80
Connect	79.47	75.30	2,032.90	82.51	74.03	2,559.70
CRX	91.35	90.21	17.60	89.21	88.94	22.00
Glass	88.11	84.66	29.00	86.54	78.77	29.80
Heart disease	91.23	89.18	16.20	86.74	83.06	16.20
Hepatitis	83.44	83.42	14.70	80.38	79.72	16.10
House	98.52	97.56	69.20	98.31	96.64	72.40
Ionosphere structure	98.49	98.10	11.00	94.98	94.64	11.00
Monk2	68.00	66.39	7.60	72.14	66.95	23.00
New thyroid	99.50	99.44	7.40	98.14	95.98	7.20
Pima Indians diabetes	81.96	80.27	11.20	80.51	77.78	18.80
Segmentation	98.05	97.55	57.20	97.18	96.48	81.60
Sonar	96.41	96.37	21.00	94.28	86.83	20.90
Vehicle	84.26	82.67	55.60	84.15	78.88	63.40
Voting	97.80	97.31	7.90	96.87	96.83	8.50
Wave	94.82	85.22	311.00	94.60	84.04	332.80
Wine	99.38	100.00	8.80	98.54	95.90	8.40

Ionosphere structure, and Sonar), large datasets (Adult, Segmentation, and Wave), and high dimensional and large dataset (Connect), the proposed approach performed well. On the other hand, Table 10 shows that computation costs in the acceptable time means the proposed approach could solve high dimensional, large, and high dimensional and large dataset problems, and did not require large computation costs. Compared with previous studies, the proposed SS + DT approach shows good performance by obtaining higher classification accuracy rates. With feature selection, the proposed SS + DT approach effectively deletes some moderating or non-affecting features while maintaining the same or superior classification accuracy rate. Furthermore, the effects of the remaining features on classification can be examined in the future. The main contributions of the proposed approach include:

- (1) The trial-and-error method traditionally used for C4.5 algorithm in determining the parameter is time-consuming and cannot guarantee the better result. The proposed approach can be used for automatic parameter determination for C4.5 algorithm.
- (2) The feature selection could help the C4.5 algorithm reduce or keep the same tree size while classification accuracy rate is increased, or keep the similar result compared with the method.
- (3) The experimental result showed that the scatter search is indeed beneficial for C4.5 to determine the parameter values and feature selection. Therefore, the scatter search has the potential to be applied to other machine learning algorithms.

More studies can be done in the future. First, as the proposed SS-based meta-heuristic is versatile; exploring the potential application of this approach to other data-mining techniques, (such as BPN, SVM, probabilistic graphical models, and probabilistic graphical models) could improve the classification result. Second, the proposed SS + DT approach can be applied to other real-world problems to determine whether it can effectively solve such problems. Finally, ensemble architecture can be used. The original concept of ensemble is based on a committee machine. The purpose of a committee machine is to integrate options of many experts rather than only one expert to obtain a good classification result. Therefore, a multi-decision tree model can be utilized in ensemble architecture to enhance further classification accuracy rate; this is currently being investigated by the authors of this study.

**Acknowledgments** The authors would like to thank the National Science Council, Republic of China (Taiwan), for financially supporting this research under Contract No. NSC 97-2410-H-182-020-MY2 and the Chang Gung University, Taiwan, for financially supporting this research under Contract No. UARPD390081. We are

grateful to anonymous referees for their constructive suggestions for improving the paper.

## References

- Abe N, Kudo M (2005) Non-parametric classifier-independent feature selection. *Pattern Recognit* 39:737–746. doi:10.1016/j.patcog.2005.11.007
- Aitkenhead MJ (2008) A co-evolving decision tree classification method. *Expert Syst Appl* 34:18–25. doi:10.1016/j.eswa.2006.08.008
- Berry MJA, Linoff G (2001) *Data mining techniques: for marketing, sales and customer support*. Wiley, London
- Carvalho DR, Freitas AA (2002) A genetic-algorithm for discovering small-disjunct rules in data mining. *Appl Soft Comput* 2:75–88. doi:10.1016/S1568-4946(02)00031-5
- Freitas AA (1998) *Data mining: and knowledge discovery with evolutionary algorithm*. Springer, Berlin
- Glover F (1977) Heuristics for integer programming using surrogate constraints. *Decis Sci* 8:156–166. doi:10.1111/j.1540-5915.1977.tb01074.x
- Glover F (1998) A template for scatter search and path relinking. In: Hao JK, Lutton E, Ronald E, Schoenauer M, Snyers D (eds) *Artificial evolution, Lecture notes in computer science*, vol 1363, Springer, Berlin, pp 13–54
- Gray JB, Fan G (2008) Classification tree analysis using TARGET. *Comput Stat Data Anal* 52:1362–1372. doi:10.1016/j.csda.2007.03.014
- Han J, Kamber M (2006) *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco
- Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. Available via DIALOG. <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- John GH (1994) Cross-validated C4.5: using error estimation for automatic parameter selection, Technical Report, Computer Science Department, Stanford University. Available via DIALOG. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.5518>
- Kim H, Koehler GJ (1995) Theory and practice of decision tree induction. *Omega* 23:637–652. doi:10.1016/0305-0483(95)00036-4
- Kohavi R, John G (1995) Automatic parameter selection by minimizing estimated error. In: Prieditis A, Russell S (eds) *Machine learning: Proceedings of the twelfth international conference*, Morgan Kaufmann
- Laguna M, Martí R (2003) *Scatter search: methodology and implementations in C*. Kluwer Academic Publishers, Boston
- Last M, Kandel A, Maimon O (2001) Information-theoretic algorithm for feature selection. *Pattern Recognit Lett* 22:799–811. doi:10.1016/S0167-8655(01)00019-8
- Liu H, Motoda H (1998) *Feature selection for knowledge discovery and data mining*. Kluwer Academic, Boston
- López FG, Torres GM, Batista BM (2006) Solving feature subset selection problem by parallel scatter search. *Eur J Oper Res* 169:477–489. doi:10.1016/j.ejor.2004.08.010
- Martí R (2006) Scatter search-wellsprings and challenges. *Eur J Oper Res* 169:351–358. doi:10.1016/j.ejor.2004.08.003
- Orsenigo C, Vercellis C (2004) Discrete support vector decision trees via tabu search. *Comput Stat Data Anal* 47:311–322. doi:10.1016/j.csda.2003.11.005
- Osei-Bryson KM (2007) Post-pruning in decision tree induction using multiple performance measures. *Comput Oper Res* 34:3331–3345. doi:10.1016/j.cor.2005.12.009

- Perner P, Apte C (2004) Empirical evaluation of feature subset selection based on a real-world data set. *Eng Appl Artif Intell* 17:285–288. doi:[10.1016/j.engappai.2004.03.005](https://doi.org/10.1016/j.engappai.2004.03.005)
- Quinlan JR (1986) Introduction of decision trees. *Mach Learn* 1:81–106
- Quinlan JR (1987) Simplifying decision trees. *Int J Man Mach Stud* 27:221–234. doi:[10.1016/S0020-7373\(87\)80053-6](https://doi.org/10.1016/S0020-7373(87)80053-6)
- Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, Menlo Park
- Rasha SAW, Monmarché N, Slimane M, Moaid AF, Saleh HH (2006) A scatter search algorithm for the automatic clustering problem. *Lect Notes Comput Sci* 4065:350–364. doi:[10.1007/11790853\\_28](https://doi.org/10.1007/11790853_28)
- Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min Knowl Discov* 1:317–327. doi:[10.1023/A:1009752403260](https://doi.org/10.1023/A:1009752403260)
- Smith M, Bull L (2005) GAP: constructing and selection features with evolutionary computing. In: Jain LC, Ghosh A (eds) *Evolutionary computation in data mining*. Springer, Berlin
- Su CT, Shiue YR (2003) Intelligent scheduling controller for shop floor control systems: a hybrid genetic algorithm/decision tree learning approach. *Int J Prod Res* 41(12):2619–2641. doi:[10.1080/0020754031000090612](https://doi.org/10.1080/0020754031000090612)
- Su C-T, Chen M-C, Chan H-L (2005) Applying neural network and scatter search to optimize parameter design with dynamic characteristics. *J Oper Res Soc* 56:1132–1140. doi:[10.1057/palgrave.jors.2601888](https://doi.org/10.1057/palgrave.jors.2601888)
- Sun W, Chen J, Li J (2007) Decision tree and PCA-based fault diagnosis of rotating machinery. *Mech Syst Signal Process* 21:1300–1317. doi:[10.1016/j.ymsp.2006.06.010](https://doi.org/10.1016/j.ymsp.2006.06.010)