Modern Applied Science; Vol. 8, No. 1; 2014
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

A Wrapper-Based Combined Recursive Orthogonal Array and Support Vector Machine for Classification and Feature Selection

Wei-Chang Yeh[1,2], Yuan-Ming Yeh[3], Cheng-Wei Chiu[2] & Yuk Ying Chung[4]

[1] Integration and Collaboration Laboratory, Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, New South Wales, Australia

[2] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.

[3] Faculty of Science, University of Sydney, NSW, Australia

[4] School of Information Technologies, University of Sydney, NSW, Australia

Correspondence: Wei-Chang Yeh, Department of Industrial Engineering and Engineering Management, National Tsing Hua University, P.O. Box 24-60, Hsinchu, Taiwan 300, R.O.C. E-mail: yeh@ieee.org

Received: August 26, 2013 Accepted: November 24, 2013 Online Published: December 17, 2013

doi:10.5539/mas.v8n1p11 URL: http://dx.doi.org/10.5539/mas.v8n1p11

Abstract

In data mining, classification problems are among the most frequently discussed issues. Feature selection is a very important pre-processing function in the vast majority of classification cases. Its aim is to delete irrelevant or redundant features in order to reduce the feature dimension and computing complexity and increase the accuracy of classification. Current feature selection methods can be roughly divided into the filter method and the wrapper method. The former chooses the feature subset before classifying, whereas the latter chooses the feature subset during the classification procedure. In general, wrapper methods result in better performance than filter methods, but they are time-consuming. This paper therefore proposes a wrapper method called OA-SVM that uses an orthogonal array (OA) to make systemic rules of feature selection and uses support vector machine (SVM) as the classifier. The proposed OA-SVM is employed to test eight UCI databases for the classification problem. The results of these experiments verify that the proposed OA-SVM for feature selection can effectively delete irrelevant or redundant features, thereby increasing classification accuracy.

Keywords: classification, feature selection, orthogonal array, support vector machine

1. Introduction

With the rapid progress of technology development, access to huge databases and their management is an issue that many enterprises are likely to face. Data mining techniques have consequently become some of the most important applications in recent years for solving this issue. The main purpose of data mining is to discover and analyze the useful information from large databases, to provide a reference for managers or decision makers. In general, data mining's more commonly used capabilities are classification, clustering, affinity grouping, and prediction. Among those, classification problems are widely encountered in many fields. Classification, which is a type of supervised learning, uses a known training set to establish a prediction model for the categorization of data of an unknown class.

In practical applications, data is usually pre-processed before establishing a prediction model, and this process is often referred to as feature selection. Data usually contains a large amount of features, but not every feature is a useful classification target. The removal of irrelevant or redundant features while ensuring that classification does not affect the accuracy of the target concept and the desired information may significantly improve a complex operation and increase efficiency (John, Kohavi, & Pfleger, 1994). Thus, feature selection technique is our focus in this paper.

In order to increase accuracy and reduce the computing time, feature selection methods and data classification technology constitute the two major steps for classification problems. Many scholars have proposed different algorithms to improve the accuracy of classification in the feature selection methods, but the use of different methods on the same problem might produce different degrees of accuracy and efficiency. Thus, the choice of method is an important issue when determining how to address a particular problem. This study proposes a

wrapper method that uses an orthogonal array (OA, statistical methods) as a feature selection technique and support vector machine (SVM) for classification. The proposed method establishes a systematic rule for the selection of the feature subset to significantly reduce the computing time and increase classification accuracy.

This paper is organized as follows. Section 2 introduces the concept of feature selection and briefly reviews some feature selection methods. In Section 3, the basic concepts of SVM and the OA are presented. The ROA-SVA is proposed to solve the feature selection problem for classification in Section 4. In Section 5, the wine (recognition) dataset adapted from UCI is used to show how to implement the proposed ROA-SVM. Comparisons based on benchmark data listed in the UCI demonstrate the effectiveness of the proposed ROA-SVM in Section 6. Finally the conclusion and suggestions for future research are presented in Section 7.

2. Feature Selection Methods

The main purpose of feature selection is to delete irrelevant or redundant variables and reduce space dimensions. Although an exhaustive search method is able to find the best feature subset, it is usually unrealistic and costly. Many heuristic or random methods, called feature selection methods, have been proposed by scholars to solve the above issues. Dash and Liu (1997) summarized a typical feature selection method in four steps, as shown in Figure 1.

- Generation procedure: A procedure generates the feature subset which is evaluated in the next step.
- Evaluation function: Evaluate the feature subset and generate a goodness (such as accuracy) to determine the candidate feature.
- Stopping criterion: A criterion is used to decide when to stop the process to prevent an exhaustive search from taking place.
- Validation process: The stopping criterion is usually the last step of a feature selection process; however, a validation procedure is necessary to compare the result of other feature selection methods to prove that the proposed method is valid.

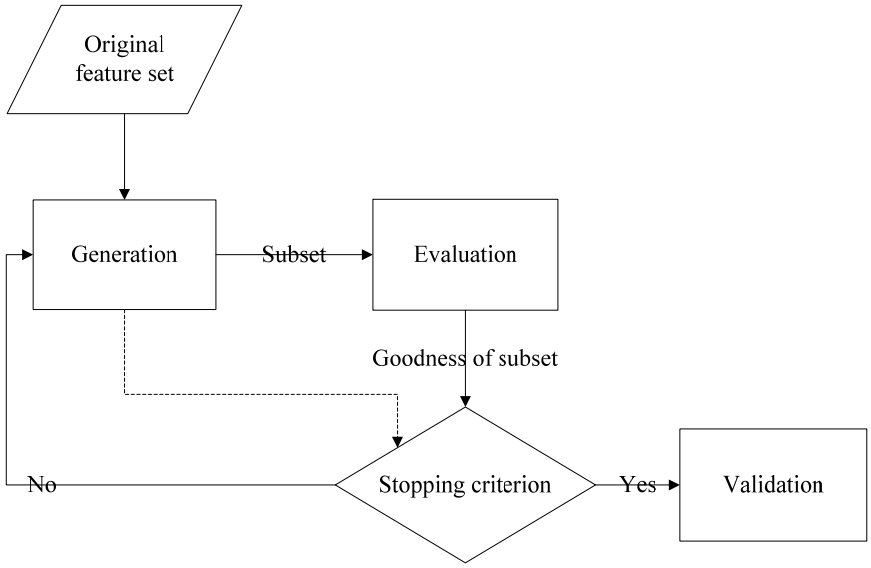

Figure 1. Feature selection process with validation (Dash & Liu, 1997)

In generally, there are two kinds of feature selection methods: filter and wrapper methods (Blum & Langley, 1997). Filter methods select the features subsets by analyzing the distance, information and other measures of the intrinsic data. Because filter methods do not rely on classification technology, the advantage of these methods is that calculation is simple and fast. The main disadvantage is that the mutual relations of the selecting subsets of features and classifier are ignored. Rokach et al. (2007) divided the filter method into ranker method and non-ranker method. The ranker method evaluates the features by a given measure and sorts the ranks; however, the non-ranker method only generates the feature subset and no ranks. The filter method is illustrated in Figure 2.

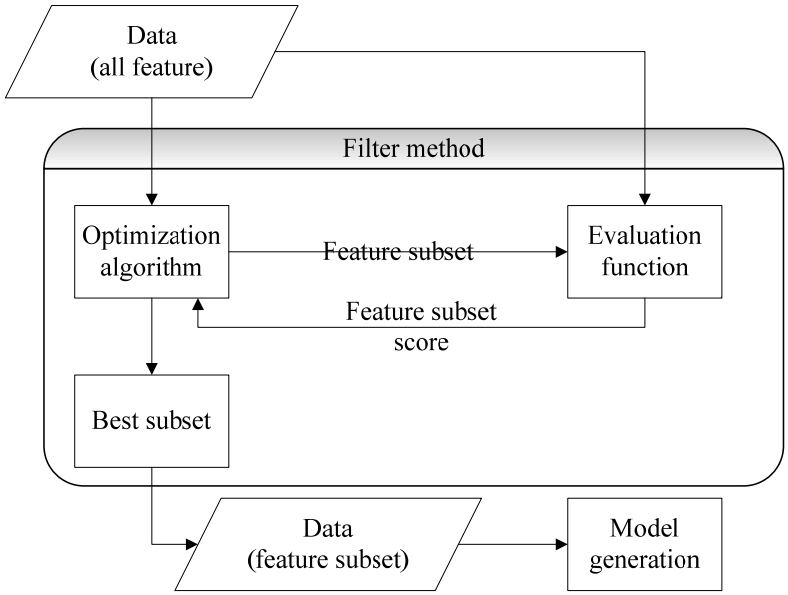

Figure 2. Filter method flow chart (Mladenić, 2006)

The wrapper method uses the classifier directly to select features. This method therefore combines the feature selection method and classification technology. The pros and cons of the wrapper methods are opposite to those of the filter methods. Wrapper methods are usually computationally expensive and costly, but they demonstrate better performance than the filter methods (Zhu, Ong, & Dash, 2007). The wrapper method is illustrated in Figure 3.

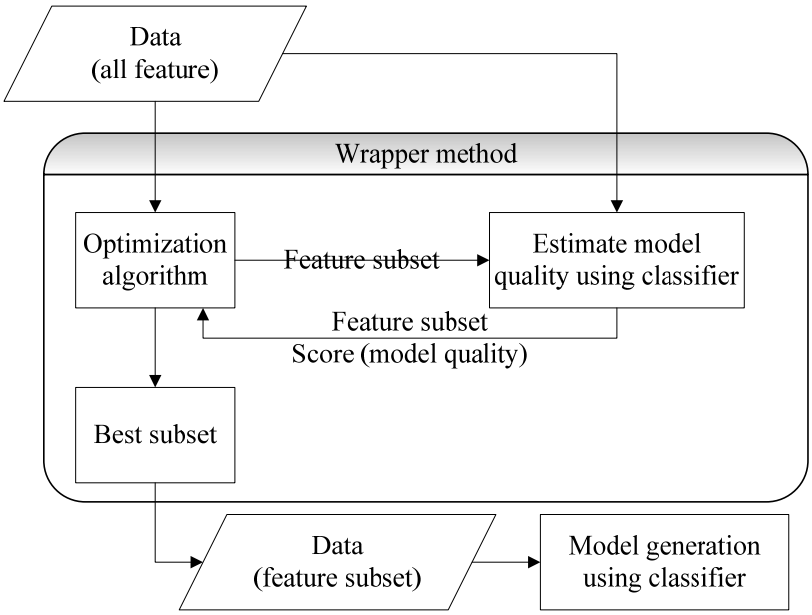

Figure 3. The flow chart of wrapper method (Mladenić, 2006)

3. Introduction of SVM and OA

The proposed ROA-SVM is based on the OA and SVM. Section 3.1 introduces the SVM for the classification method by illustrating the basic idea behind SVMs based on the linear model. The concept of OA will be introduced in Section 3.2.

3.1 SVM

SVMs (Vapnik, 1995, 1998) have been proven to give excellent performance in binary classification cases. Let $\mathbf{X}_i=(x_{i1}, x_{i2},\ldots, x_{id})\in R^d$ be the ith training data, and $y_i\in\{\tilde{1},1\}$ denote its class label for i=1,2,…,n. A hyper-plane can be written in the following form:

$$F(\mathbf{X})=\mathbf{W}^{\mathrm{T}}\mathbf{X}+b=0, \tag{1}$$

such that (as shown in Figure 4)

$$\mathbf{W}^{\mathrm{T}}\mathbf{X}_i+b\geq 1, \text{ for } y_i= 1 \tag{2}$$

$$\mathbf{W}^{\mathrm{T}}\mathbf{X}_i+b\leq -1, \text{ for } y_i=-1 \tag{3}$$

where $\mathbf{W}$ is normal to the hyper-plane, $|b|/\|\mathbf{W}\|$ is the perpendicular distance from the hyper-plane to the origin, and $\|\mathbf{W}\|$ is the Euclidean norm of $\mathbf{W}$.

The above two equations can be combined and rewritten as

$$y_i(\mathbf{W}^{\mathrm{T}}\mathbf{X}_i+b)\geq 1 \tag{4}$$

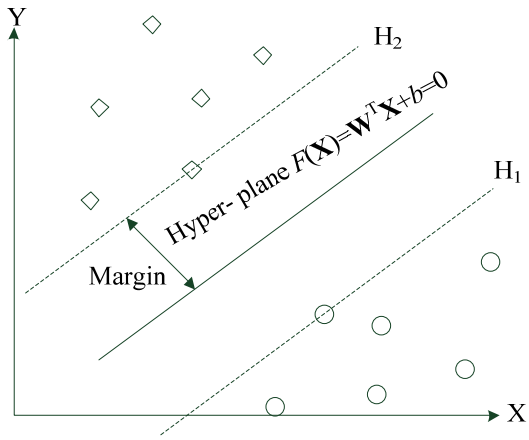

Figure4. Illustration of SVM

The purpose of SVM is to find $\mathbf{W}$ and b in Equation (1) to maximize the margin ρ between two support hyper-planes

$$\mathrm{H}_1\text{: } \mathbf{W}^{\mathrm{T}}\mathbf{X}_i+b=1 \tag{5}$$

$$\mathrm{H}_2\text{: } \mathbf{W}^{\mathrm{T}}\mathbf{X}_i+b=-1 \tag{6}$$

to separate two classes of data. Notice that $\rho=2d$, where d is the distance between the hyper-plane and any one of the support hyper-planes and defined as

$$d=\frac{(|b+1|-|b|)}{\|\mathbf{W}\|_2}=\frac{1}{\|\mathbf{W}\|_2} \tag{7}$$

By above Equations (4) and (7), the SVMs problem can be summarized as a quadratic programming problem:

$$Minimize\ \frac{\|\mathbf{W}\|^2}{2} \tag{8}$$

such that Equation (4) is held. The above quadratic programming problem is also a convex optimization problem which can be solved using the Lagrange multiplier method after translating the quadratic programming problem using the Lagrange multipliers $i\geq 0$, we have

$$L(\mathbf{W}, b,)=\frac{\|\mathbf{W}\|^2}{2}-\sum_{i=1}^{n}\alpha_i[y_i(\mathbf{W}^{\mathrm{T}}X_i-b)]+\sum_{i=1}^{n}\alpha_i \tag{9}$$

To find the extreme point to minimize Equation (9), the partial differentiations are taken to Equation (9) with

respect to **W** and b and set to zero:

$$\frac{\partial L(\mathbf{W},b,\alpha)}{\partial \mathbf{W}} = \mathbf{W} - \sum_{i=1}^{n} \alpha_i y_i X_i = 0 \tag{10}$$

$$\frac{\partial L(\mathbf{W},b,\alpha)}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{11}$$

The above two equations can be rewritten as follow:

$$\mathbf{W} = \sum_{i=1}^{n} \alpha_i y_i X_i \tag{12}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{13}$$

Substitute Equations (12) and (13) into Equation (9), we have

$$\text{Maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j X_i^{\mathrm{T}} X_j \tag{14}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{for all } i \geq 0. \tag{15}$$

For a convex problem, KKT conditions are necessary and sufficient to solve **W**, b and α_i. Therefore, solving the SVMs problem is equal to solving the KKT conditions. The related KKT conditions are included Equations (4), (12), and (13), and the rest are listed below:

$$\text{(Dual feasibility)} \ \alpha_i \geq 0 \tag{16}$$

$$\text{(Complementary slackness)} \quad {}_i[y_i(\mathbf{W}^{\mathrm{T}}\mathbf{X}_i + b) - 1] = 0. \tag{17}$$

Notice that α_i can be obtained by solving the quadratic programming problem listed in Equations (14) and (15). Next, Equation (12) is used to obtain **W**. Finally, b can be solved using Equation (17).

Even in high dimensional feature space or nonlinear classification problems, SVMs can translate these problems to linear separable problems by convert function. Therefore SVMs have been widely used in various fields for feature selection problems in recent years (Tong & Koller, 2001; Lodhi, Shawe-Taylor, Christianini, & Watkins, 2001; Burges, 1998; Papageorgiou, Evgeniou, & Poggio, 1998; Osuna, Freund, & Girosi, 1997; Viola & Jones, 2001; Byvatov & Schneider, 2003; Furey, Cristianini, Duffy, Bednarski, Schummer, & Haussler, 2000). In this paper, we will use SVM as our classification method.

3.2 OA

An OA is an array of positive integers (called levels) arranged in rows (denoted experiments) and columns (denoted factors). The ith column denotes the ith feature, and the 0 in any combination is set to select the feature, 1 as a waiver of the features. For example, only feature A is selected in Experiment 2 since A=0 and B=C=1 in Table 1. All columns exhibit the following properties of statistically independence in any OA:

- Self-balanced: The number of each level is the same in each column. For example, Table 1 is a 2-level 3-factor OA and level 0 appears the same number of times as level 1, i.e., twice in each column (factor).

Table 1. Two levels and three factors OA

Number of Experiment	Column (factor)		
	1	2	3
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0

- Mutual-balanced: The number of any level is the same in each column. For example, level 1 appears the same number of times, i.e., twice in any column of Table 1.

The above two properties are called the orthogonality. Algorithms for constructing OAs with various levels are found in (Rokach, Chizi, & Maimon, 2007). The details of OA are as follows. Let $L_n(s^m)$ be an OA for n experiments, m factors and s levels per factor, where L denotes a Latin square. Eighteen standard basic OAs are listed as in Table 2.

Table 2. The standard OA

OA	Row Number	Factor Number	Maximum column number at these levels			
			2	3	4	5
L_4	4	3	3			
L_8	8	7	7			
L_9	9	4		4		
L_{12}	12	11	11			
L_{16}	16	15	15			
L'_{16}	16	5			5	
L_{18}	18	8	1	7		
L_{25}	25	6				6
L_{27}	27	13		13		
L_{32}	32	31	31			
L'_{32}	32	10	1		9	
L_{36}	36	23	11	12		
L'_{36}	36	16	3	13		
L_{50}	50	12	1			11
L_{54}	54	26	1	25		
L_{64}	64	63	63			
L'_{64}	64	21			21	
L_{81}	81	40		40		

Note that an additional experiment will be tested by the factor weighted analysis (FWA) based on the self-balanced property. The FWA can evaluate the effects of respective factors (hereafter called 'features' in this research) and determine whether a feature is needed after the result (which is defined as and called the 'accuracy of classification' in this research hereafter) of each experiment is given. Let w_i denote the accuracy of experiment i, $x_{ij} \in \{0,1\}$ denote the level of experiment i of feature j, and the effect of feature j be defined as

$$e_j = \sum_{i=1}^{n} w_i x_{ij} \tag{18}$$

If

$$e_j \leq \frac{1}{2} \sum_{i=1}^{n} w_i \tag{19}$$

then the feature j is selected in the additional experiment. For convenience in interpreting the FWA, some assumed values are added to Table 3. For example, feature A is obtained in experiments 3 and 4, but it is neglected in experiments 1 and 2. The effect of feature A (feature 1) can be computed as followed:

$$e_j = \sum_{i=1}^{4} w_i x_{ij} = 0\times 75 + 0\times 80 + 1\times 85 + 1\times 60 = 145 \tag{20}$$

$$\leq \frac{1}{2}\sum_{i=1}^{4} w_i = \frac{75+80+85+60}{2} = 150 \tag{21}$$

Therefore, feature A is selected in the additional experiment. We determine whether features B and C are likewise selected. We set A=0, B=0, and C=1 in the fifth experiment which means that the features A and B are obtained in the fifth experiment. Finally SVM is used to compute the accuracy of the classification. The best feature subset for feature selection is selected by ranking the accuracy of each experiment and choosing the highest one.

Table 3. The additional experiment in Table 1

Number of Experiment	Features			Accuracy (%)
	A	B	C	
1	0	0	0	75
2	0	1	1	80
3	1	0	1	85
4	1	1	0	60
5	0	0	1	unknown

The OA is a special statistical design of experiments that studies the effects of several factors simultaneously to use the least number of experiments to explore the maximum number of factors and estimate the interaction between factors efficiently, rather than exploring all the possible combinations of assignments. Therefore, OA has the advantage of significantly reducing the number of experiments and simplifying the data analysis.

4. The Proposed ROA-SVM

This section discusses the details of how the proposed ROA-SVM combines recursive OA and SVM to conduct feature selection for classification problems. The proposed ROA-SVM is mainly based on the standard OA for two levels, such as $L_4(2^3)$, $L_8(2^7)$, $L_{12}(2^{11})$, $L_{16}(2^{15})$, $L_{32}(2^{31})$, and $L_{64}(2^{63})$. Let Z_k=0, 3, 7, 11, 15, 31, and 63, where k=0,1,2,…,6. When the number of features are m and $Z_k<m\leq Z_{k+1}$, the OA denoted by $L_{Z_{k+1}}(2^{Z_k})$ is proposed. The procedure is recursive until no better accuracy can be found in each experiment.

The proposed ROA-SVM is essentially the same for any number of features and experiments, but we will describe it in detail only for $L_4(2^3)$ shown in Table 1. SVM is used as the evaluation tools and classification method. We set 10-fold cross-validation in SVMs. In 10-fold cross-validation, the input data is randomly partitioned into 10 equal parts and a single part of the 10 parts is retained as the testing data for the model. The other 9 parts are used as training data. The cross-validation process is repeated 10 times, with each of the 10 parts being used exactly once as the testing data.

Finally, the 10 results can be averaged to produce a single accuracy. In experiment 1, the data with all features A, B and C are selected for SVM with 10-fold cross-validation to compute the accuracy of classification. In experiment 2, only feature B is obtained to compute the classification accuracy. Experiments 3 and 4 are proven likewise. In this way, we obtain the respective accuracy of each experiment. Note that, as mentioned in Section 3.2, an additional experiment will be conducted with the FWA, in addition to the original experiments. Those features in the experiment that have the best accuracy will be selected and the remainder will be discarded in the next run. This procedure is repeated until there is no further improvement in accuracy. Figure 5 illustrates the flow chart of ROA-SVM.

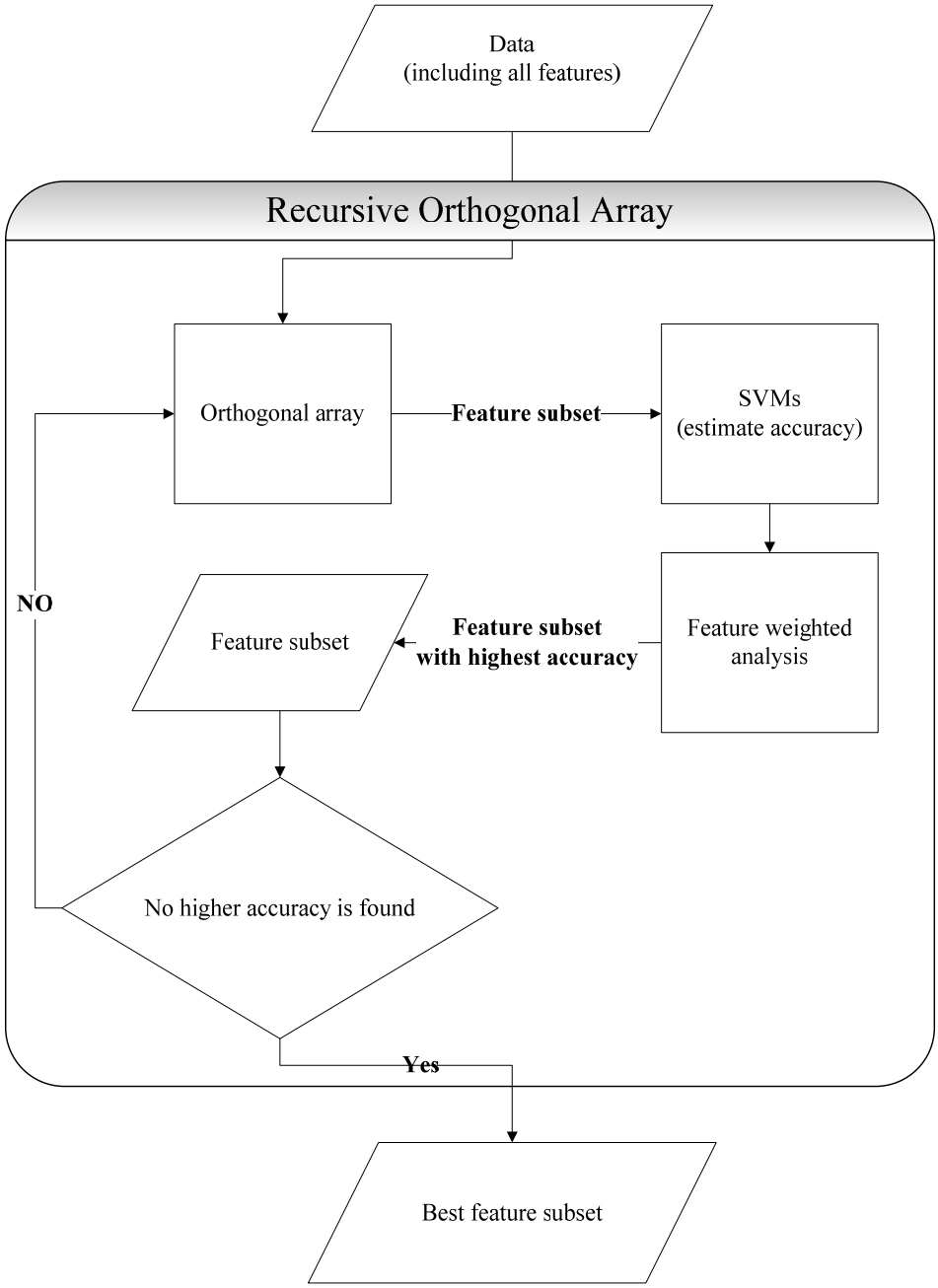

Figure 5. The flow chart of recursive orthogonal array

5. A Numerical Example: Wine Recognition Dataset

In this section, the wine (recognition) dataset adapted from UCI is used to show the procedure of ROA-SVM. Wine dataset has 178 data patterns and 13 features. For a complete test and comparisons, 10-fold cross-validation is used; therefore, there are always 90% data in the training set and 10% data as the testing data. Because 12≤13≤16, the $L_{16}(2^{15})$ OA is used. The result of the first run in feature selection using the proposed ROA-SVM is shown in Table 4.

Table 4. The first OA results of wine dataset

experiment	1	2	3	4	5	6	7	8	9	10	11	12	13	null	null	accuracy
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45.51%
2	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	83.15%
3	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	91.01%
4	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	65.17%
5	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	47.19%
6	0	1	1	0	0	1	1	1	1	0	0	1	1	0	0	83.15%
7	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	91.57%
8	0	1	1	1	1	0	0	1	1	0	0	0	0	1	1	66.29%
9	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	91.57%
10	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0	61.24%
11	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0	48.31%
12	1	0	1	1	0	1	0	1	0	1	0	0	1	0	1	80.34%
13	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	88.76%
14	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1	59.55%
15	1	1	0	1	0	0	1	0	1	1	0	1	0	0	1	47.19%
16	1	1	0	1	0	0	1	1	0	0	1	0	1	1	0	86.52%
17	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**93.82%**

In Table 4, the accuracy obtained by ROA-SVM is the percentage of correctly classified testing data. Table 4 shows that Experiment 17, which has zeros in rows 1, 6, 10, 12, has the highest accuracy. Hence, features 1, 6, 10 and 12 are selected for the next run.

Table 5. The second round results of wine database

experiment	1	6	10	12	null	null	null	accuracy
1	0	0	0	0	0	0	0	93.82%
2	0	0	0	1	1	1	1	90.45%
3	0	1	1	0	0	1	1	88.76%
4	0	1	1	1	1	0	0	69.10%
5	1	0	1	0	1	0	1	76.97%
6	1	0	1	1	0	1	0	65.73%
7	1	1	0	0	1	1	0	88.20%
8	1	1	0	1	0	0	1	73.60%
9	**0**	**0**	**0**	**0**	**1**	**1**	**1**	93.82%

The results of the second run are presented in Table 5. The best accuracy is still 93.82% as found in the first run. Therefore, the features 1, 6, 10 and 12 are the best subset features in our proposed ROA-SVM with the best accuracy being 93.82%.

6. Computational Experiments

To evaluate its quality and performance for data mining, the proposed ROA-SVM is applied to and compared with the original SVM in eight widely referenced real-world datasets (including the wine dataset discussed in Section 4) which are adopted from the UCI Machine Learning Repository (Asuncion & Newman, 2007).

These eight benchmark datasets are Balance scale weight & distance dataset (Balance), Iris plants dataset (Iris),

General description of thyroid disease dataset (Thyroid), Pima Indians diabetes dataset (Diabetes), Breast cancer dataset, Glass identification dataset (Glass), Wine recognition dataset (Wine), Australian credit approval dataset (Credit). The number of instances, classes, and features of these datasets are shown in Table 6.

Table 6. Summary of eight adapted UCI dataset

Dataset	class	feature	instances
Balance	3	4	625
Iris	3	4	150
Thyroid	3	5	215
Diabetes	2	8	768
Breast Cancer	2	10	684
Glass	7	10	214
Wine	3	13	178
Credit	2	14	690

For a fair comparison, all tests and methods are based on the 10-fold cross-validation method.

Table 7. Training and testing data number of dataset

Dataset	Number of Training Data	Number of Testing Data
Balance	562	63
Iris	135	15
Thyroid	193	22
Diabetes	691	77
Breast Cancer	615	69
Glass	192	22
Wine	160	18
Credit	621	69

To fully exploit the benefit and demonstrate the performance of the proposed ROA-SVM, two tests are used (Test1 and Test2). In Test1, the computational result provides a comparison between the proposed ROA-SVM and the conventional SVM. The datasets in Test1 for which the proposed ROA-SVM has failed to reduce the number of features are tested further in Test2. In Test2, the exhaustive method is implemented to remove all possible combinations of features to prove that all features are significant and none are removable in those datasets which were impossible to reduce in Test1.

6.1 Test1

Two SVM-based classifiers, ROA-SVM and traditional SVM, are implemented. The accuracy and the number of feature subsets on the eight UCI datasets based on SVM and ROA-SVM are summarized in Table 8.

Table 8. The result of feature selection on UCI data

Data	Features	Selected Number of Features after using ROA-SVM	SVM	ROA-SVM
Balance	4	4	**90.08%**	**90.08%**
Iris	4	4	**98.00%**	**98.00%**
Thyroid	5	2	75.81%	**95.81%**
Diabetes	8	2	65.10%	**72.40%**
Breast Cancer	10	4	65.79%	**96.49%**
Glass	10	2	98.13%	**99.07%**
Wine	13	4	45.51%	**93.82%**
Credit	14	1	55.65%	**85.51%**

The results of the above experiments (with the exception of the first two) show that the proposed ROA-SVM is superior to the conventional SVM in terms of both prediction accuracy and number of features.

6.2 Test2 Based on the Exhaustive Method

Excluding the balance and iris datasets, the accuracy of the other six datasets is increased with fewer selected features in the classification. To further test whether there are irrelevant or redundant features in the balance and iris datasets, the exhaustive method is used to test them. Since both datasets include only four features, a 15×4 OA is used that only has two possible values, that is, 0 and 1; 0 in any combination is set to select the feature, and 1 is set as a waiver of the features as shown in Section 3. All the possible combinations of feature subsets and accuracies estimated by SVM for the balance dataset and iris dataset are listed in Tables 9 and 10, respectively.

Table 9. The result of balance dataset using the exhaustive method

Experiment	1	2	3	4	Accuracy
1	0	0	0	0	**90.08%**
2	0	0	0	1	75.52%
3	0	0	1	0	75.52%
4	0	1	0	0	75.36%
5	1	0	0	0	75.68%
6	0	0	1	1	70.08%
7	0	1	0	1	66.56%
8	1	0	0	1	67.36%
9	0	1	1	0	67.04%
10	1	0	1	0	67.20%
11	1	1	0	0	71.04%
12	0	1	1	1	63.52%
13	1	0	1	1	63.52%
14	1	1	0	1	63.52%
15	1	1	1	0	63.52%

Table 10. The result of iris dataset using the exhaustive method

Experiment	1	2	3	4	Accuracy
1	0	0	0	0	**98.00%**
2	0	0	0	1	95.33%
3	0	0	1	0	95.33%
4	0	1	0	0	96.67%
5	1	0	0	0	96.67%
6	0	0	1	1	80.00%
7	0	1	0	1	95.33%
8	1	0	0	1	95.33%
9	0	1	1	0	96.00%
10	1	0	1	0	96.00%
11	1	1	0	0	95.33%
12	0	1	1	1	73.33%
13	1	0	1	1	54.00%
14	1	1	0	1	95.33%
15	1	1	1	0	96.00%

No better accuracy can be found in either Table 9 or 10, which means that there are no irrelevant or redundant features in either the balance dataset or the iris dataset. Thus, either the proposed ROA-SVM can effectively reduce the number of features and efficiently increase the accuracy, or all features are important and cannot be removed, such as in the balance and iris datasets.

7. Conclusions and Future Research

Classification is an important task in data mining. Feature selection is always an important issue in classification. This work describes a new classifier design method called ROA-SVM to provide a systematic method for the effective deletion of irrelevant or redundant features. According to the testing result from Table 8, the classification result in 5th column using the proposed ROA-SVM method is better than the 4th column using SVM to classify the eight UCI dataset which includes: Balance, Iris, Thyroid, Diabetes, Breast Cancer, Glass, Wine, Credit.

The comparisons based on eight common UCI benchmark datasets demonstrate the effectiveness of the proposed ROA-SVM method in deleting the irrelevant or redundant features and reducing the number of experiments, thereby increasing the accuracy of classification and computation time significantly.

Our experimental results had shown a good achievement with the default SVM parameter settings. However, the parameter settings have a deep impact on classification performance, so how to adjust the parameters to achieve better performance is still worth researching.

References

Asuncion, A., & Newman, D. (2007). *UCI Machine Learning Repository*. Retrieved from http://www.ics.uci.edu/~mlearn/MLRepository.html

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Res. Soc., 54*, 627-635. http://dx.doi.org/10.1057/palgrave.jors.2601545

Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell, 97*, 245-271. http://dx.doi.org/10.1016/S0004-3702(97)00063-5

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classifications and regression trees*. Pacific Grove, CA:Wadsworth.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc, 2*, 121-167.

Byvatov, E., & Schneider, G. (2003). Support vector machine applications in bioinformatics. *Appl. Bioinformat, 2*(2), 67-77.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Mach. Learn, 3*, 261-283.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intell. Data Anal*, 131-156.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn, 29*, 103-130. Retrieved from http://www.cc.gatech.edu/fac/Charles.Isbell/classes/reading/papers/bayes-opt.pdf

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Trans. Syst., Man, Cybern, 6*, 325-327. http://dx.doi.org/10.1109/TSMC.1976.5408784

Fix, E., & Hodges Jr, J. L. (1989). Discriminatory analysis, nonparametric discrimination: Consistency properties. *Int. Stat. Rev, 57*(3), 238-247.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*, 906-914. http://dx.doi.org/10.1093/bioinformatics/16.10.906

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset election problem. *Proc. Eleventh International Conference on Machine Learning* (pp. 121-129). Retrieved from http://www.cs.columbia.edu/~kathy/cs4701/documents/fs.pdf

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Appl. Statist, 29*(2), 119-127.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proc. Tenth National Conference on Artificial Intelligence* (pp. 223-228).

Lodhi, H., Shawe-Taylor, J., Christianini, N., & Watkins, C. (2001). Text classification using string kernels. *Adv. Neur. In, 13*. http://dx.doi.org/10.1162/153244302760200687

Mladenić, D. (2006). Feature selection for dimensionality reduction. *SLSFS Lecture Notes in Comp. Sci., 3940*, 84-102. Berlin Heidelberg: Springer. http://dx.doi.org/10.1007/0-387-25465-X_5

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data and a proposal. *J. Am. Stat. Assoc, 58*, 415-434.

Osuna, E., Freund, R., & Girosi, F. (1997). Training support vector machines: An application to face detection. *Comp. Vis. Patt. Recogn*. http://dx.doi.org/10.1109/CVPR.1997.609310

Papageorgiou, C., Evgeniou, T., & Poggio, T. (1998). A trainable pedestrian detection system. *IEEE Conference on Intelligent Vehicles*.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples; In D. Michie (Ed.), *Expert Systems in the Micro-electronic Age*. UK Edinburgh: Edinburgh University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn, 1*(1), 81-106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *Proc. IJCAI-01 Workshop on Empirical Methods in AI* (pp. 41-46).

Rokach, L., Chizi, B., & Maimon, O. (2007). A methodology for improving the performance of non-ranker feature selection filters. *International Journal of Recognition and Artificial Intelligence, 21*(5), 809-830.

Rumelhart, D. E., Hinton, D. E., & Williams, R. J. (1986). *Learning internal representations by error propagation in parallel distribution distributed processing* (pp. 318-362). Cambridge, MA: MIT Press. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a164453.pdf

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res*, 45-66. http://dx.doi.org/10.1162/153244302760185243

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Comp. Vis. Patt. Recogn*. http://dx.doi.org/10.1109/CVPR.2001.990517

Zhu, Z., Ong, Y. S., & Dash, M. (2007). Wrapper- filter feature selection algorithm using a memetic framework. *IEEE Trans. Syst., Man, Cybern. B, Cybern, 37*(1), 70-76. http://dx.doi.org/10.1109/TSMCB.2006.883267