Correspondence

Orthogonal Forward Selection and Backward Elimination Algorithms for Feature Subset Selection

K. Z. Mao

***Abstract*—Sequential forward selection (SFS) and sequential backward elimination (SBE) are two commonly used search methods in feature subset selection. In the present study, we derive an orthogonal forward selection (OFS) and an orthogonal backward elimination (OBE) algorithms for feature subset selection by incorporating Gram–Schmidt and Givens orthogonal transforms into forward selection and backward elimination procedures, respectively. The basic idea of the orthogonal feature subset selection algorithms is to find an orthogonal space in which to express features and to perform feature subset selection. After selection, the physically meaningless features in the orthogonal space are linked back to the same number of input variables in the original measurement space. The strength of employing orthogonal transforms is that features are decorrelated in the orthogonal space, hence individual features can be evaluated and selected independently. The effectiveness of our algorithms to deal with real world problems is finally demonstrated.**

***Index Terms*—Feature subset selection, orthogonal backward elimination (OBE), orthogonal forward selection (OFS).**

I. Introduction

Selecting a subset of features from a pool of many potential variables is a common problem in pattern classification. Quite often, data acquisition process collects samples on a large number of variables when it is unknown which specific ones are most important for class discrimination. The goal of feature subset selection is to identify and to select the most important and nonredundant variables from the large pool of potential variables. Generally, a feature subset selection algorithm involves a feature evaluation criterion and a search algorithm. The evaluation criterion evaluates the capacity of feature subsets to distinguish one class from another, while the search algorithm explores the potential solution space. Based on the evaluation criterion used, feature selection methods can be classified into filter and wrapper methods [1]. The wrapper method takes feature selection and pattern classification as a whole and evaluates feature subsets based on classification results directly, while the filter method employs intrinsic properties of data such as class separability measures as the criterion for feature subset evaluation. Because of its independence on classification algorithms, the feature subset selected by the filter method can be used by any classifier. In the present study, we evaluate feature subsets based on Mahalanobis class separability measure (see, for example, [5]). After selecting the evaluation criterion, we need to choose a suitable search algorithm. Exhaustive search method guarantees to find the optimal solution, but it has to evaluate all possible combinations of all variables. Exhaustive search is rarely attempted in practice when features are to be selected from a pool of many potential variables. Branch and bound [2] algorithm and genetic algorithms [3], [4] can provide optimal feature subset without exhaustive search. But the two methods are still computationally impractical if the pool of potential variables is large. In practice, suboptimal search methods such as sequential forward selection (SFS) algorithm and sequential backward elimination (SBE) algorithm (see, for example, [5]) are often employed. The drawback of SFS and SBE is that once a feature is selected/deleted, it cannot be deleted/re-selected at a later stage. As a consequence, redundant features might be selected. To alleviate this problem, the max–min algorithm [6], the plus-l–take-away-r $(l-r)$ algorithm (see, for example, [5]), and the floating search method [7] have been proposed. Owing to the incorporation of the deletion/re-selection procedure, the $(l-r)$ algorithm and the floating search method have been found to be powerful (see, for example, [8]).

The motivation of employing deletion/re-selection procedure in the $(l-r)$ and floating search algorithms is to reduce redundancy in the feature subset, which is caused mainly by correlations or interactions between candidate features. In the present study, we attempt to alleviate the redundancy problem by employing orthogonal decompositions. The strength of employing orthogonal decomposition is that features are decorrelated in the orthogonal space and they can be evaluated and selected independently. The orthogonal transforms used in our study are Gram–Schmidt transform and Givens transform (see, for example, [9] and [10]). The reason of employing Gram–Schmidt and Givens orthogonal transforms instead of the well known principal component analysis (PCA) is that features in the Gram–Schmidt and Givens orthogonal space can be made to associate with the same number of input variables of the measurement space, while the PCA features are linked with the full set instead of a subset of the input variables.

The present study is organized as follows. In Section II, Gram–Schmidt orthogonal transform is introduced, and an orthogonal forward feature subset selection algorithm is developed. In Section III, an orthogonal backward elimination (OBE) algorithm based on Givens rotation is derived. Experimental studies with real world problems are presented in Section IV. Concluding remarks are given in Section V.

II. Orthogonal Forward Selection Algorithm for Feature Subset Selection

A. Gram–Schmidt Orthogonal Transform

Suppose N samples $\mathbf{x}(1)$, $\mathbf{x}(2)$, ..., $\mathbf{x}(N)$ are available, and each sample is represented by an n-dimensional vector $\mathbf{x}(k) = [x_1(k), x_2(k), \ldots, x_n(k)]^T$. Feature vector $\mathbf{x}_i$ and feature matrix $\mathbf{X}$ are defined as

$$\mathbf{x}_i = [x_i(1), x_i(2), \ldots, x_i(N)]^T$$

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(N) & x_2(N) & \cdots & x_n(N) \end{bmatrix}. \quad (1)$$

The feature matrix $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \mathbf{QR}. \quad (2)$$

$\mathbf{R}$ is an upper triangular matrix, and $\mathbf{Q}$ is an orthogonal matrix

$$\mathbf{R} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ & \alpha_{22} & \cdots & \alpha_{2n} \\ & & \ddots & \vdots \\ & & & \alpha_{nn} \end{bmatrix}$$

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n] = \begin{bmatrix} q_1(1) & q_2(1) & \cdots & q_n(1) \\ q_1(2) & q_2(2) & \cdots & q_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ q_1(N) & q_2(N) & \cdots & q_n(N) \end{bmatrix}$$

Manuscript received March 10, 2001; revised March 15, 2002. This paper was recommended by Associate Editor P. K. Willett.

The author is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: ekzmao@ntu.edu.sg).

Digital Object Identifier 10.1109/TSMCB.2002.804363

where $\mathbf{q}_i$ is the new feature vector in the orthogonal space. In the Gram–Schmidt orthogonal decomposition, the orthogonal matrix $\mathbf{Q}$ is constructed using the following procedure (see, for example, [9]):

$$\mathbf{q}_1 = \mathbf{x}_1 \tag{3}$$

$$\mathbf{q}_i = \mathbf{x}_i - \sum_{j=1}^{i-1} \alpha_{ji}\mathbf{q}_j \tag{4}$$

where

$$\alpha_{ji} = \begin{cases} \dfrac{\mathbf{q}_j^T \mathbf{x}_i}{\mathbf{q}_j^T \mathbf{q}_j}, & \text{for } j = 1, 2, \ldots, i-1 \\ 1, & \text{for } j = i. \end{cases} \tag{5}$$

Equation (2) implements a mapping from space $\mathbf{X}$ to space $\mathbf{Q}$: $\mathbf{Q} = \mathbf{R}^{-1}\mathbf{X}$, and the feature vector $\mathbf{q}_i$ ($i = 1, 2, \ldots, n$) can be interpreted as sample distributions in the direction of feature q_k in the orthogonal space.

The quality of a feature subset can be evaluated based on its ability to provide large class separation. A few criteria for class separability measure are available such as the Mahalanobis distance measure. For a two-class classification problem, the Mahalanobis distance measure is defined as (see, for example, [5])

$$J(i, j) = [\mathbf{m}_i - \mathbf{m}_j]^T \mathbf{C}_{ij}^{-1} [\mathbf{m}_i - \mathbf{m}_j] \tag{6}$$

where $\mathbf{m}_i = [m_{1i}, m_{2i}, \ldots, m_{ni}]^T$ is the mean vector of samples in class i. $\mathbf{C}_{ij} = \mathbf{C}_i + \mathbf{C}_j$, and $\mathbf{C}_i$ and $\mathbf{C}_i$ are the covariance matrices of classes i and j, respectively. In the orthogonal space, the covariance matrix is diagonal

$$\mathbf{C}_{ij} = \operatorname{diag}\left[\sigma_{1ij}^2, \sigma_{2ij}^2, \ldots, \sigma_{nij}^2\right].$$

Hence, the class separability measure in the orthogonal space can be decomposed as

$$J(i, j) = \sum_{k=1}^{n} \frac{(m_{ki} - m_{kj})^2}{\sigma_{kij}^2}. \tag{7}$$

For multiclass problems, the average class separability measure can be used as feature subset evaluation criterion

$$\begin{aligned} J &= \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} J(i, j) \\ &= \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \sum_{k=1}^{n} \frac{(m_{ki} - m_{kj})^2}{\sigma_{kij}^2} \\ &= \frac{2}{L(L-1)} \sum_{k=1}^{n} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \frac{(m_{ki} - m_{kj})^2}{\sigma_{kij}^2} \end{aligned} \tag{8}$$

where L is the number of classes. By defining the average class separability measure in the direction of q_k as

$$J_k = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \frac{(m_{ki} - m_{kj})^2}{\sigma_{kij}^2} \tag{9}$$

we obtain

$$J = \sum_{k=1}^{n} J_k. \tag{10}$$

Equation (10) shows that the average class separability measure is the sum of average separability measure in individual directions in the orthogonal space. The advantage of using orthogonal transform is that features are decorrelated, consequently individual features can be evaluated and selected independently.

Next we investigate the link between features in the Gram–Schmidt orthogonal space and input variables in the measurement space. Equations (3)–(5) show that the first k features $q_1, q_2, \ldots, q_k$ in the orthogonal space are associated with the first k input variables $x_1, x_2, \ldots, x_k$ only. This property of the Gram–Schmidt forms the basis of our orthogonal feature subset selection method: features are first selected in the orthogonal space, the physically meaningless features are then linked back to the same number of input variables of the original measurement space. But care must be taken when implementing this idea. If the candidate features are simply orthogonalized in the order in which they happen to be put in the matrix $\mathbf{X}$, the important features in the orthogonal space are not necessarily sorted consecutively in the first few columns of $\mathbf{Q}$. As a result, the number of associated variables in the original measurement space would be larger than the number of features selected in the orthogonal space. For example, if feature q_n in the last column of $\mathbf{Q}$ is selected, it would be linked back to all variables in the original space. This problem can be solved by making important features first enter matrix $\mathbf{X}$ and $\mathbf{Q}$ using a sequential forward selection procedure. The combination of the orthogonal transform and the sequential forward selection leads to the following orthogonal forward feature selection algorithm.

B. Orthogonal Forward Feature Subset Selection Procedure

The orthogonal forward selection (OFS) algorithm that incorporates forward selection into Gram–Schmidt orthogonal transform is summarized as follows.

1) At the first step, consider all variables x_i ($i = 1, 2, \ldots, n$) as the candidate that first enters matrix $\mathbf{X}$

   $$\mathbf{q}_1^{(i)} = \mathbf{x}_i.$$

   Compute Mahalanobis distance measures provided by $\mathbf{q}_1^{(i)}$, $i = 1, 2, \ldots, n$. The variable that yields maximum class separability, say $\mathbf{x}_j$, is identified and is added to the feature subset. Let $\mathbf{q}_1 = \mathbf{x}_j$.
2) At the second step, consider all remaining $n-1$ variables as the candidate secondly entering matrix $\mathbf{X}$

   $$\mathbf{q}_2^{(i)} = \mathbf{x}_i - \alpha_{12}^{(i)}\mathbf{q}_1 \qquad 1 \le i \le n,\ i \ne j$$

   where

   $$\alpha_{12}^{(i)} = \mathbf{q}_1^T \mathbf{x}_i / \mathbf{q}_1^T \mathbf{q}_1$$

   and compute corresponding Mahalanobis distance measures. The feature that provides maximum class separation is identified and is added to the feature subsets.
3) The above procedure is continued until the class separability measure provided by the next best feature is less than a pre-specified threshold.

III. Orthogonal Backward Elimination (OBE) Algorithm for Feature Subset Selection

In contrast to the OFS algorithm, the OBE algorithm starts from orthogonal decomposition of the full feature set. If the feature to be deleted is in the last column of matrix $\mathbf{Q}$, the degradation of class separability measure when it is deleted is simply

$$J_n = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \frac{(m_{ni} - m_{nj})^2}{\sigma_{nij}^2}. \tag{11}$$

Hence, each feature can be moved to the last column to find which leads to the least deterioration in class separability measure after deletion. Gram–Schmidt procedure can be used to perform orthogonalization after each column exchange. But orthogonalization can be performed in a more efficient way based on Givens transform (see, for example, [9] and [10]).

A. Givens Transform

Consider the orthogonal decomposition

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]\,\mathbf{R}.$$

The degradation of class separability due to the deletion of the last column of $\mathbf{Q}$ can be measured using (11). If each column in $\mathbf{X}$ is in

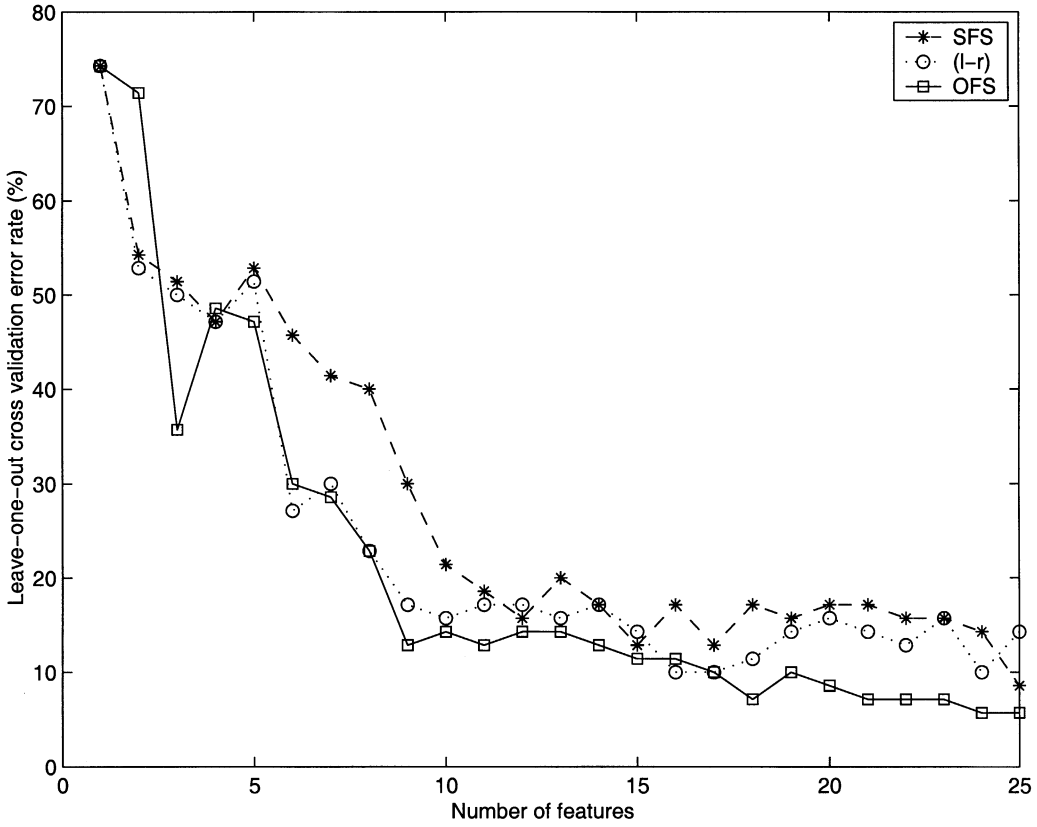

Fig. 1. Comparison of SFS, $(l-r)$ and OFS for Experiment 1.

turn moved to the last column, the corresponding degradation can be computed. For example, if $\mathbf{x}_i$ is exchanged with $\mathbf{x}_n$, the matrix $\mathbf{X}$ becomes

$$\begin{aligned}\mathbf{X}_{i\leftarrow n} &= [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_i] \\ &= [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n]\,\mathbf{I}_{i\leftarrow n} \\ &= \mathbf{X}\mathbf{I}_{i\leftarrow n} \end{aligned} \tag{12}$$

where $\mathbf{I}_{i\leftarrow n}$ is the permutation of an identity matrix whose ith and nth columns are exchanged. Substituting (12) into (2), we obtain

$$\mathbf{X}_{i\leftarrow n} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]\,\mathbf{R}\mathbf{I}_{i\leftarrow n}. \tag{13}$$

Obviously, $\mathbf{R}\mathbf{I}_{i\leftarrow n}$ is no longer triangular. To re-triangulate $\mathbf{R}\mathbf{I}_{i\leftarrow n}$, we start from exchanging columns i and $i+1$. After column exchange, we obtain

$$\mathbf{R}\mathbf{I}_{i\leftarrow i+1} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1i+1} & \alpha_{1i} & \cdots & \alpha_{1n} \\ & \alpha_{22} & \cdots & \alpha_{2i+1} & \alpha_{2i} & \cdots & \alpha_{2n} \\ & & \ddots & \vdots & \vdots & \vdots & \vdots \\ & & & \alpha_{ii+1} & \alpha_{ii} & \ldots & \alpha_{in} \\ & & & \alpha_{i+1i+1} & 0 & \cdots & \alpha_{i+1n} \\ & & & & & \ddots & \vdots \\ & & & & & & \alpha_{nn} \end{bmatrix}. \tag{14}$$

The element α_{i+1i+1} of $\mathbf{R}\mathbf{I}_{i\leftarrow i+1}$ can be reduced to zero by applying Givens rotation to the ith and the $(i+1)$th rows of $\mathbf{R}\mathbf{I}_{i\leftarrow i+1}$. Givens rotation is an orthonormal transform which is defined as (see, for example, [9] and [10])

$$\mathbf{g} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \tag{15}$$

where

$$\left.\begin{aligned} c &= \alpha_{ii+1} \Big/ \sqrt{\alpha_{ii+1}^2 + \alpha_{i+1i+1}^2} \\ s &= \alpha_{i+1i+1} \Big/ \sqrt{\alpha_{ii+1}^2 + \alpha_{i+1i+1}^2} \end{aligned}\right\}. \tag{16}$$

A full transformation matrix $\mathbf{G}$ is defined as

$$\mathbf{G} = \mathrm{diag}\,\{1, \ldots, 1, \mathbf{g}, 1, \ldots, 1\} \tag{17}$$

which is also orthonormal. Hence, we have

$$\mathbf{X}_{i\leftarrow i+1} = \mathbf{Q}\mathbf{R}\mathbf{I}_{i\leftarrow i+1} = \mathbf{Q}\mathbf{G}^T\mathbf{G}\mathbf{R}\mathbf{I}_{i\leftarrow i+1}. \tag{18}$$

Equation (18) shows that only two Givens rotations are needed in order to obtain the orthogonal decomposition of $\mathbf{X}_{i\leftarrow i+1}$: one rotation is applied to columns i and j of matrix $\mathbf{Q}$, and another rotation is applied to rows i and j of matrix $\mathbf{R}\mathbf{I}_{i\leftarrow i+1}$.

To obtain orthogonal decomposition of $\mathbf{X}_{i\leftarrow n}$, we can in turn exchange $\mathbf{x}_i$ with $\mathbf{x}_{i+1}$, $\mathbf{x}_{i+2}$, ..., $\mathbf{x}_n$, and perform Givens rotation after each exchange. Suppose we have obtained $\mathbf{Q}$ and $\mathbf{R}$ by applying Gram–Schmidt orthogonal decomposition to the full feature matrix $\mathbf{X}$, the procedure of orthogonalizing $\mathbf{X}_{i\leftarrow n}$ using Givens transform can be summarized as follows.

1) At the first step, exchange $\mathbf{x}_i$ with $\mathbf{x}_{i+1}$, and calculate parameters c and s of Givens rotation matrix using (16). Then apply Givens rotations to the ith and $(i+1)$th rows of $\mathbf{R}\mathbf{I}_{i\leftarrow i+1}$ and the ith and $(i+1)$th columns of $\mathbf{Q}$. Take matrices $\mathbf{X}_{i\leftarrow i+1}$, $\mathbf{Q}\mathbf{G}^T$ and $\mathbf{G}\mathbf{R}\mathbf{I}_{i\leftarrow i+1}$ as new $\mathbf{X}$, $\mathbf{Q}$, and $\mathbf{R}$, respectively.
2) At the kth step, exchange $\mathbf{x}_{i+k-1}$ with $\mathbf{x}_{i+k}$, calculate c and s of Givens rotation, apply the rotation to rows $i+k-1$ and $i+k$ of $\mathbf{R}\mathbf{I}_{i+k-1\leftarrow i+k}$ and columns $i+k-1$ and $i+k$ of $\mathbf{Q}$, and denote $\mathbf{X}_{i+k-1\leftarrow i+k}$, $\mathbf{Q}\mathbf{G}^T$ and $\mathbf{G}\mathbf{R}\mathbf{I}_{i+k-1\leftarrow i+k}$ as new $\mathbf{X}$, $\mathbf{Q}$, and $\mathbf{R}$, respectively.
3) The above procedure is continued until $k = n - i$.

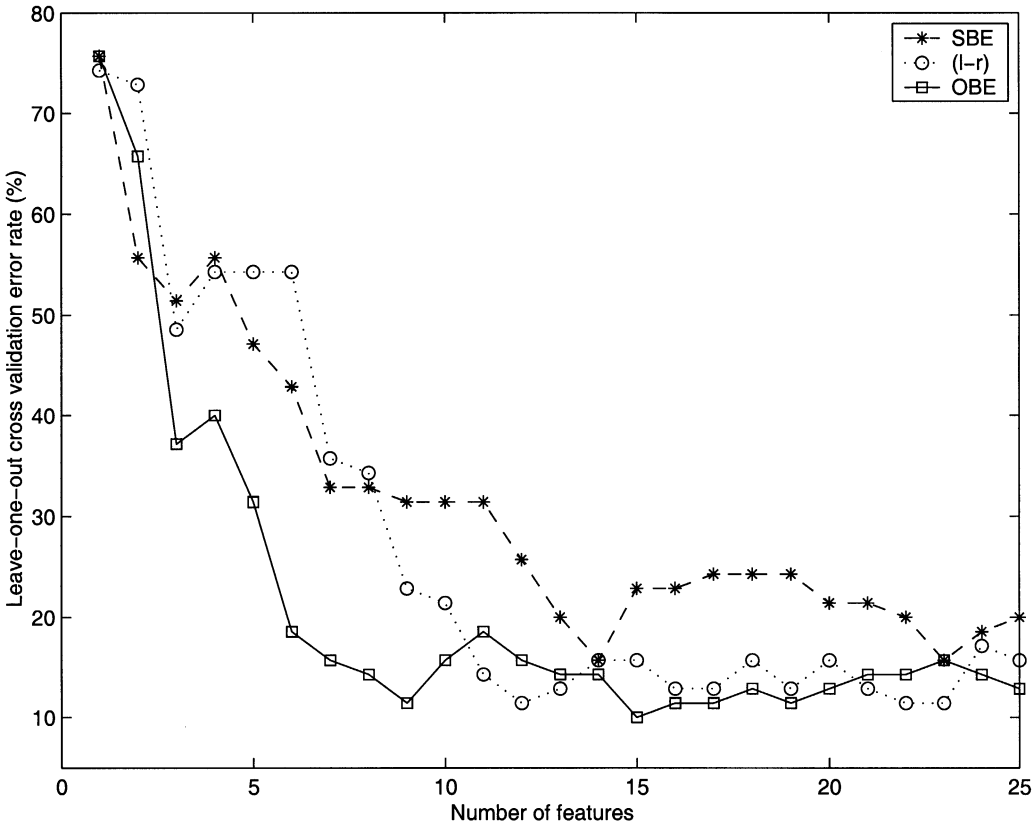

Fig. 2. Comparison of SBE, $(l - r)$, and OBE for Experiment 1.

B. Orthogonal Backward Elimination Feature Subset Selection Procedure

The OBE algorithm that combines Givens transform and sequential backward elimination algorithm can be summarized as follows.

i) Initially, all variables available are used to form the full feature matrix $\mathbf{X}$ and QR decomposition is performed using (3)–(5).
ii) At the first step, column $n - 1$ of matrix $\mathbf{X}$ is first exchanged with the last column. Givens rotation is then applied to obtain new $\mathbf{Q}$ and $\mathbf{R}$. Denote the degradation of class separability measure after removing the last column as J_1.
iii) At the kth step, column $n - k$ of matrix $\mathbf{X}$ is exchanged with the last column. Givens orthogonalization procedure Steps 1)–3) in Section III-A is then used to obtain new $\mathbf{Q}$ and $\mathbf{R}$. The class separability measure degradation after removing the last column is denoted as J_k.
iv) Step iii) is continued until $k = n-1$. The variable that results in minimum degradation to class separability measure after deletion is identified and is discarded. Set $n = n - 1$.
v) Repeat Steps ii)–iv) until the degradation of class separability measure resulted from the deletion of the next least important feature is larger than the pre-specified threshold.

IV. Experiments

The OFS algorithm and the OBE algorithm are developed to reduce redundancy in the selected feature subset. We have done a few experiments, and the results show that if correlations between candidate features are trivial, employing orthogonal transform does not make much difference; but orthogonal algorithms provide improvements if severe correlations exist.

A. Experiment 1

In the experiment, a hyperspectral data band selection problem was used to test the effectiveness of our algorithms.

Spectral imaging, which divides the ultraviolet, visible and infrared spectra into distinct bands for imaging, has become available in recent years. A typical hyperspectral sensor like the airborne visible/infrared imaging spectrometer (AVIRIS) is able to provide 224 contiguous spectral bands within the range of 0.8–2.4 μm, with a spectral resolution up to the order of 10 nm [11]. With this kind of fine spectral resolution, hyperspectral imaging provides necessary information for precise studies of objects and substances. Hyperspectral imaging has become an important tool in remote sensing and biomedical engineering [12] etc. Despite the advantage of having a fine spectral resolution, the high dimensionality of hyperspectral data can present problems to pattern classification. Dimensionality reduction is therefore an important issue in hyperspectral data classification. In [13], a projection pursuit method was proposed. In [14], the principal component analysis (PCA) for multispectral image classification was discussed. Both methods achieve massive reduction in dimensionality, but the features provided do not retain their original physical meanings.

In the present study, we use feature subset selection method to reduce dimensionality so that features selected retain their original physical interpretations. The problem under study is to classify five kinds of minerals which are Hematite, Montmorillonite, Muscovite, Olivine, and Topaz. The reflectance data of these minerals were downloaded from the USGS Spectral Lab [15], where reflectance at 480 bands were provided in the spectra range of 0.2–3 μm. The number of samples of each mineral are 12, 10, 13, 17, and 18, respectively. These are the only minerals that ten or more samples were provided by the USGS Spectral Lab.

The number of bands (features) available is 480, which is very large. Spectral band selection was performed first. For comparison reasons, we ranked features using SFS algorithm, $(l - r)$ algorithm and OFS

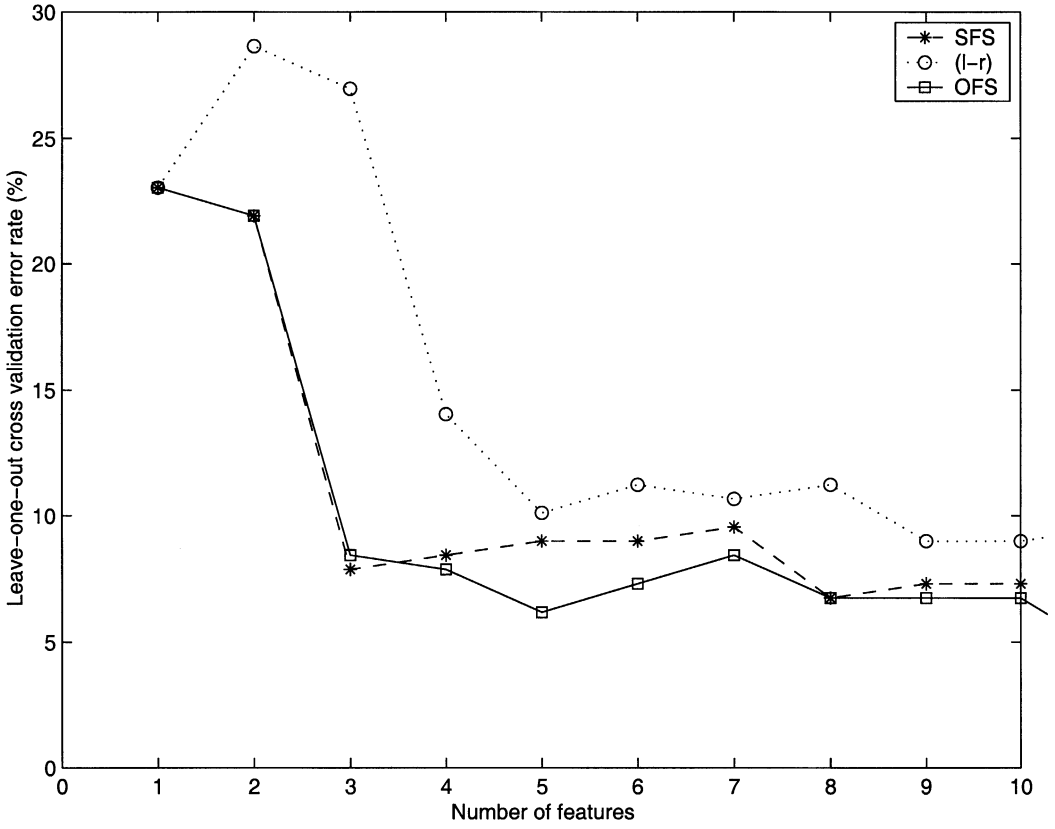

Fig. 3. Comparison of SFS, $(l - r)$ and OFS for Experiment 2.

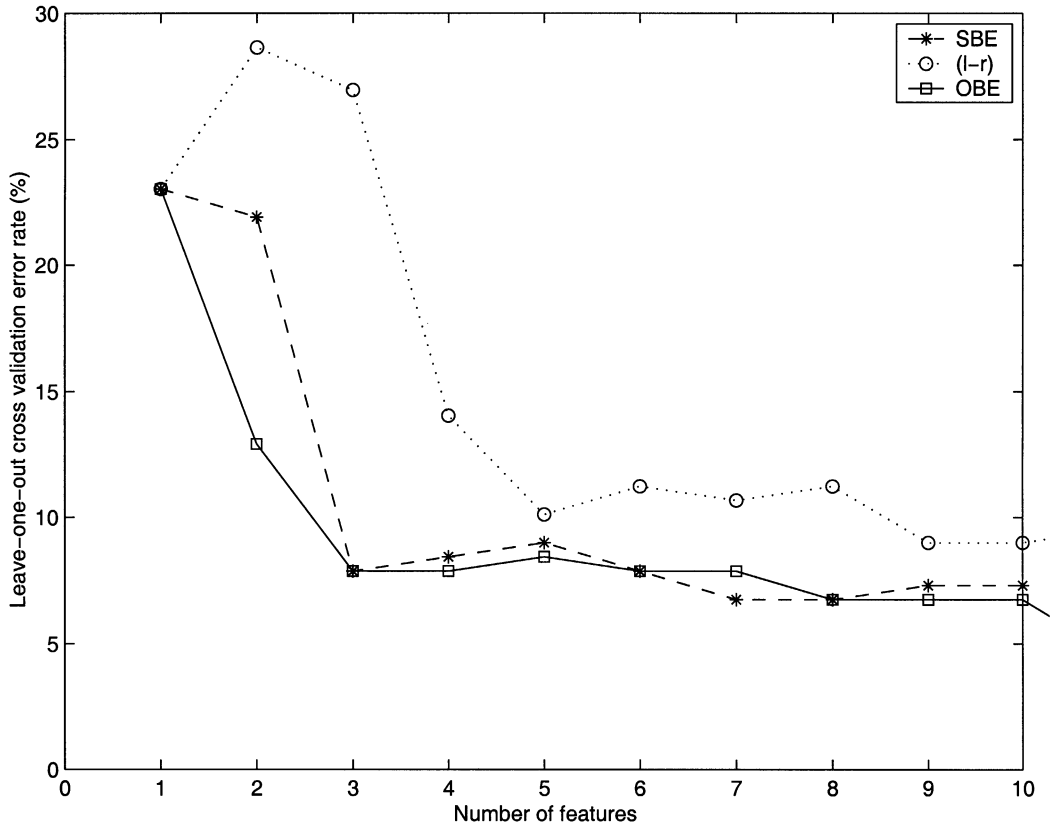

Fig. 4. Comparison of SBE, $(l - r)$ and OBE for Experiment 2.

algorithm respectively. For $(l - r)$ algorithm, l and r were set to 2 and 1, respectively. After feature ranking, a linear least square estimation algorithm was used to perform classification. Leave-one-out method was employed to evaluate classification error rate, and the results are

shown in Fig. 1. From Fig. 1 we can see that OFS achieves lower error rate than SFS and $(l - r)$ if the same number of features are used. On the other hand, OFS demands less features than SFS and $(l - r)$ to achieve the same classification error rate. The improvement is owing to the reduction of redundancy by the orthogonal transform. It should be pointed out, however, that OFS demands more computations than SFS and $(l-r)$ algorithms due to the orthogonal decomposition involved. In addition, when more features, say 50 features, are used, feature subsets provided by the three methods tend to provides the same classification results.

Due to the large number of candidate features, intensive computations would be demanded if backward elimination algorithms are applied to the dataset directly. In our experiment, the dimensionality of the dataset was first reduced from 480 to 50 using the sequential forward selection algorithm. OBE, SBE, and $(l-r)$ feature selection algorithms were then performed based on the reduced data. The leave-one-out cross validation classification error rates of the linear least square estimation classifier are shown in Fig. 2. Again, employing orthogonal transform is advantageous.

B. Example 2

In this example, wine dataset from UCI Machine Learning Repository [16] was used to test the effectiveness of our algorithms. The wine data are the results of a chemical analysis of wine grown in the same region in Italy but derived from three different cultivars. The dataset contains 178 samples that belong to three classes, respectively.

In the wine dataset, each sample is represented by 13 numerical variables, some of which are insignificant for class discrimination. Feature selection was therefore selected first. For comparison, we ranked features based on the SFS algorithm, $(l - r)$ algorithm, OFS algorithm, SBE algorithm, and OBE algorithm, respectively. For the $(l - r)$ algorithm, l and r was were set to 2 and 1, respectively. After feature subset selection, the linear least square estimation algorithm was used to classify the 178 samples, and the leave-one-out cross validation error rates corresponding to the five algorithms are shown in Figs. 3 and 4, respectively. Obviously, the orthogonal feature selection algorithms achieved good results compared with the sequential forward selection algorithm, the sequential back elimination algorithm and the $(l - r)$ algorithm as well. However, the improvement is limited due to minor correlations between candidate features in this example.

V. Conclusion

In the present study, we have proposed two feature subset selection algorithms by incorporating orthogonal transforms into the sequential forward selection and backward elimination procedures. The orthogonal feature subset selection algorithms are particularly powerful for problems with severe correlations among candidate features. The effectiveness of our algorithms has been tested using real world problems.

References

[1] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.
[2] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, pp. 917–922, Sept. 1977.
[3] W. Siedlecki and J. Skansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognit. Lett.*, vol. 10, no. 5, pp. 335–347, 1989.
[4] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, pp. 44–49, Feb. 1998.
[5] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.
[6] E. Backer and J. A. D. Schipper, "On the max–min approach for feature ordering and selection," in *The Seminar on Pattern Recognition*. Liege, Belgium: Liege Univ., 1977.
[7] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.
[8] A. Jain and D. Zongker, "Feature selection: Evaluation, application and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 153–158, Feb. 1997.
[9] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their applications to nonlinear system identification," *Int. J. Control*, vol. 50, pp. 1873–1896, 1989.
[10] G. A. F. Saber, *Linear Regression Analysis*. New York: Wiley, 1977.
[11] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The airborne visible infrared/imaging spectrometer (AVIRIS)," *Remote Sensing Environ.*, vol. 44, no. 1, pp. 12–143, 1993.
[12] E. Wilburn, J. O. G. Reddick, E. N. Cook, T. D. Elkin, and R. J. Deaton, "Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks," *IEEE Trans. Med. Imag.*, vol. 16, pp. 911–918, Dec. 1997.
[13] L. O. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, pp. 2653–2667, June 1999.
[14] C.-C. Hung, A. Fahsi, W. Tadesse, and T. L. Coleman, "A comparative study of remotely sensed data classification using principal component analysis and divergence," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 3, 1997, pp. 2444–2449.
[15] USGS Spectral Lab [Online]. Available: ftp://speclab.cr.usgs.gov
[16] C. L. Blake and C. J. Merz. (1998) UCI repository of machine learning databases. Tech. Rep., Dept. Inform. Comput. Sci., Univ. of California, Irvine, CA. [Online]. Available: http://www.ics.uci.edu/ mlearn/Machine-Learning.html

Analysis of the Weighting Exponent in the FCM

Jian Yu, Qiansheng Cheng, and Houkuan Huang

***Abstract*—The fuzzy c-means (FCM) algorithm is one of the most frequently used clustering algorithms. The weighting exponent m is a parameter that greatly influences the performance of the FCM. But there has been no theoretical basis for selecting the proper weighting exponent in the literature. In this paper, we develop a new theoretical approach to selecting the weighting exponent in the FCM. Based on this approach, we reveal the relation between the stability of the fixed points of the FCM and the data set itself. This relation provides the theoretical basis for selecting the weighting exponent in the FCM. The numerical experiments verify the effectiveness of our theoretical conclusion.**

***Index Terms*—Fixed point, fuzzy c-means, Hessian matrix, weighting exponent.**

I. Introduction

THE fuzzy c-means algorithm (FCM) is a popular fuzzy clustering method. Many of its applications are indicated in [1]. One of the most important parameters in the FCM is the weighting exponent m. When m is close to one, the FCM approaches the hard c-means algorithm.

Manuscript received April 22, 2002; revised September 6, 2002. This paper was recommended by Associate Editor N. Pal.

J. Yu and H. Huang are with the Department of Computer Science and Technology, Northern Jiaotong University, Beijing 100044, China (e-mail: jianyu@center.njtu.edu.cn).

Q. Cheng is with the Department of Information Science, School of Mathematics Science, Peking University, Beijing 100871, China.

Digital Object Identifier 10.1109/TSMCB.2003.810951