Article:
Wei, H.L. and Billings, S.A. (2007) Feature subset selection and ranking for data dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29 (1). pp. 162-166. ISSN: 0162-8828

https://doi.org/10.1109/TPAMI.2007.250607

Feature Subset Selection and Ranking for Data Dimensionality Reduction

Hua-Liang Wei and Stephen A. Billings

Abstract—A new unsupervised forward orthogonal search (FOS) algorithm is introduced for feature selection and ranking. In the new algorithm, features are selected in a stepwise way, one at a time, by estimating the capability of each specified candidate feature subset to represent the overall features in the measurement space. A squared correlation function is employed as the criterion to measure the dependency between features and this makes the new algorithm easy to implement. The forward orthogonalization strategy, which combines good effectiveness with high efficiency, enables the new algorithm to produce efficient feature subsets with a clear physical interpretation.

Index Terms—Dimensionality reduction, feature selection, high-dimensional data.

1 Introduction

In the literature many approaches have been proposed for dimensionality reduction [1], [2], [3]. The existing dimensionality reduction methods can roughly be categorized into two classes: feature extraction and feature selection. In feature extraction problems [3], [4], the original features in the measurement space are initially transformed into a new dimension-reduced space via some specified transformation. Significant features are then determined in the new space. Although the significant variables determined in the new space are related to the original variables, the physical interpretation in terms of the original variables may be lost. In addition, although the dimensionality may be greatly reduced using some feature extraction methods, such as principal component analysis (PCA) [5], the transformed variables usually involve all the original variables. Often, the original variables may be redundant when forming the transformed variables. In many cases, it is desirable to reduce not only the dimensionality in the transformed space, but also the number of variables that need to be considered or measured [6], [7].

Unlike feature extraction, feature selection aims to seek optimal or suboptimal subsets of the original features [7], [8], [9], [10], [11], [12], [13], [14], [15], by preserving the main information carried by the collected complete data, to facilitate future analysis for high-dimensional problems. In fact, in many cases, the inclusion of insignificant variables will inevitably complicate data inspection and modeling without providing any extra information, because the insignificant variables are, in a sense, irrelative or redundant and, thus, can be ignored [16]. Detailed discussions on various feature selection algorithms can be found in [3], [8], [11].

It is worth mentioning that dimensionality reduction is not necessarily always the best solution to all high-dimensional problems [17]. Consider the following scenario: Assume that there are hundreds or even thousands of features and each feature potentially carries a small amount of information. The problem is how to extract and integrate these little pieces of information. Instead of reducing the dimensionality, Breiman [17] suggested an attractive and almost opposite approach to handle this problem: increase the dimensionality by adding many functions of the predictor variables. Two outstanding examples of work in this direction are the Amit-Geman method [18], [19] and support vector machines [20].

• *The authors are with the Department of Automatic Control and Systems Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD UK. E-mail: {w.hualiang, s.billings}@sheffield.ac.uk.*

Manuscript received 30 Sept. 2005; revised 15 June 2006; accepted 19 June 2006; published online 13 Nov. 2006.
Recommended for acceptance by M.A.T. Figueiredo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0528-0905.

This study introduces a new unsupervised feature selection and ranking method, which belongs to the second class aforementioned. This is a *forward orthogonal search* (FOS) algorithm by *maximizing* the *overall dependency* (MOD), to detect significant variables and select a subset from a library consisting of all the original variables. The main idea behind the new method is that the overall features in the original measurement space should be sufficiently represented, using the selected subset. The new feature selection method, which will be referred to as the FOS-MOD algorithm, provides a ranked list of selected features ordered according to the percentage contribution (the capability for representing the overall features). The new unsupervised learning algorithm is different from other selection methods in that it subtly combines the forward orthogonalization scheme with the maximization of the overall dependency. The mechanism of the FOS-MOD algorithm is simple and quite easy to implement and can produce efficient subsets with a direct link back to the underlying system.

2 The New Unsupervised Learning Algorithm

2.1 The Basic Idea

Let $S = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be the collected full data set formed by a total of N observations (instances) and n attributes in the measurement space, where the kth instance vector is $[\mathbf{x}_1(k), \mathbf{x}_2(k), \cdots, \mathbf{x}_n(k)]$ and the observation vector for the jth attribute is $\mathbf{x}_j = [x_j(1), x_j(2), \cdots, x_j(N)]^T$. The objective of feature selection is to find a subset $S_d = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_d\} = \{\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_d}\}$, which can be used to represent the original features, where $\mathbf{z}_m = \mathbf{x}_{i_m}$, $i_m \in \{1, 2, \cdots, n\}$, $m = 1, 2, \ldots, d$ with $d \leq n$ (generally $d << n$ if the measurement space is of large dimension). The basic requirement is that the overall features in the measurement space should be sufficiently represented using S_d by ensuring that the variation in the overall features can be explained by the elements of S_d with an acceptable degree of accuracy. This means that any data vector $\mathbf{x}_i$ in the measurement space should be well approximated using S_d in the sense that

$$\mathbf{x}_i = f_i(\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_d) + \mathbf{e}_i, \tag{1}$$

where f_i is an unknown function describing the relationship between the ith variable and the selected variables, and $\mathbf{e}_i$ is an unobservable error representing the discrepancy in the approximation. In the present study, the commonly used linear model will be considered

$$\mathbf{x}_i = \sum_{m=1}^{d} \theta_{i,m} \mathbf{z}_m + \mathbf{e}_i. \tag{2}$$

The performance of the selected subset S_d can be evaluated by inspecting the approximation capability of S_d in reproducing individual features $\mathbf{x}_i (i = 1, 2, \ldots, n)$ in the measurement space, for example, what percentage of the variation in $\mathbf{x}_i$ can be accounted for by the elements in S_d. Assume that the percentage that the variation in $\mathbf{x}_i$ can be accounted for by the elements in S_d is $p_i(d)$, the average percentage that the variation in the overall features $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ can be accounted for by S_d can then be defined as $p(d) = (1/n) \sum_{i=1}^{n} p_i(d)$. If the percentage $p(d)$ is larger than a given threshold, S_d can then be determined as the final subset; otherwise, new significant variables need to be added into S_d.

2.2 Feature Detection and Ranking

The objective of feature selection is to seek a number of significant features to form a feature subset, which is representative and can characterize the main property of all the original features. Feature selection starts from a given full data set $S = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, and significant features are selected in a stepwise way, one feature at a time. Many criteria [8] can be employed to measure the similarity between features. In the present study, the squared-correlation coefficient [21], [22] will be used to interfere with the selection

procedure. The squared-correlation coefficient between two random vectors $\mathbf{x}$ and $\mathbf{y}$ of size $N \times 1$ is given below

$$sc(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x}^T\mathbf{y})^2}{(\mathbf{x}^T\mathbf{x})(\mathbf{y}^T\mathbf{y})} = \frac{(\sum_{i=1}^N x_i y_i)^2}{\sum_{i=1}^N x_i^2 \sum_{i=1}^N y_i^2}. \quad (3)$$

At the first step, let

$$C[i, j; 1] = sc(\mathbf{x}_i, \mathbf{x}_j)\ , i, j = 1, 2, \ldots, n, \quad (4)$$

$$\overline{C}[j; 1] = \frac{1}{n}\sum_{i=1}^n C[i, j; 1], \quad (5)$$

$$\ell_1 = \arg \max_{1 \le j \le n} \{\overline{C}[j; 1]\}. \quad (6)$$

The first significant variable can then be selected as $\mathbf{z}_1 = \mathbf{x}_{\ell_1}$, and the associated orthogonal variable can be chosen as $\mathbf{q}_1 = \mathbf{z}_1$. Notice that the first selected feature $\mathbf{z}_1 = \mathbf{x}_{\ell_1}$ explains the variation in the overall features with the highest percentage, compared with any other single feature in the candidate set S. In other words, $\mathbf{z}_1 = \mathbf{x}_{\ell_1}$ is the most relevant feature in S to represent all the other features.

Assume that a feature subset S_{m-1}, consisting of $(m-1)$ significant variables, $\mathbf{z}_1, \cdots, \mathbf{z}_{m-1}$, has been determined at step $(m-1)$, and the $(m-1)$ selected variables have been transformed into a new group of orthogonalized variables $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$ via some orthogonal transformation. The mth significant feature $\mathbf{z}_m$ will be chosen in such a manner: The subset $S_{m-1} + \{z_m\}$ should be the most "representative" and, thus, the most "informative" subset compared with any other subsets formed by adding a candidate feature to S_{m-1}. To select the mth significant variable $\mathbf{z}_m$, let $\alpha_j \in S - S_{m-1}$. Orthogonalize α_j with $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$ as below

$$\mathbf{q}_j^{(m)} = \alpha_j - \frac{\alpha_j^T\mathbf{q}_1}{\mathbf{q}_1^T\mathbf{q}_1}\mathbf{q}_1 - \cdots - \frac{\alpha_j^T\mathbf{q}_{m-1}}{\mathbf{q}_{m-1}^T\mathbf{q}_{m-1}}\mathbf{q}_{m-1}. \quad (7)$$

The squared-correlation coefficient between $\mathbf{x}_i$ and $\mathbf{q}_j^{(m)}$ is

$$C[i, j; m] = sc\Big(\mathbf{x}_i, \mathbf{q}_j^{(m)}\Big). \quad (8)$$

Let

$$\overline{C}[j; m] = \frac{1}{n}\sum_{i=1}^n C[i, j; m], \quad (9)$$

$$\ell_m = \arg \max_{1 \le j \le n} \{\overline{C}[j; m]\}. \quad (10)$$

The mth significant variable can then be chosen as $\mathbf{z}_r = \mathbf{x}_{\ell_m}$, and the associated orthogonal variable can be chosen as $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$. The $(m-1)$ features $\mathbf{z}_1, \cdots, \mathbf{z}_{m-1}$ (respectively, the associated orthogonalized variables $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_{m-1}$), by including the mth feature $\mathbf{z}_m = \mathbf{x}_{\ell_m}$ (respectively, the $\mathbf{q}_m = \mathbf{q}_{\ell_m}^{(m)}$), can explain the variation in the overall features with a higher percentage than by including any other candidate feature.

Subsequent significant variables can be selected in the same way step by step. At each step, the "best" variable that accounts for the variation of the overall features with the highest percentage is selected. The FOS-MOD algorithm is thus quite easy to implement and can often produce sparse feature subsets for general selection problems. This algorithm, as a greedy nonexhaustive search method, however, may not always produce an optimal feature subset. In fact, for any nonexhaustive search algorithm, there is no guarantee that the algorithm can produce an optimal feature subset [23].

2.3 Monitoring the Search Procedure

Assume that a subset $S_m = \{\mathbf{z}_1, \cdots, \mathbf{z}_m\} = \{\mathbf{x}_{i_1}, \cdots, \mathbf{x}_{i_m}\} \subseteq S$ has been obtained, where each element of S_m is considered to be "important" for representing the overall features. In the linear case, each data vector $\mathbf{x}_j (j = 1, 2, \ldots, n)$ in the measurement space can be approximated using a linear combination of $\mathbf{z}_1, \cdots, \mathbf{z}_m$ as below

$$\mathbf{x}_j = \sum_{k=1}^m \theta_{j,k}\mathbf{z}_k + \mathbf{e}_j, \quad (11)$$

or in a compact matrix form

$$\mathbf{x}_j = \mathbf{P}\boldsymbol{\theta}_j + \mathrm{e}_j, \quad (12)$$

where the matrix $\mathbf{P} = [\mathbf{z}_1, \cdots, \mathbf{z}_m]$ is of full column rank, $\boldsymbol{\theta}_j = [\theta_{j,1}, \cdots, \theta_{j,m}]^T$ is a parameter vector, and $\mathbf{e}_j$ is an approximation error. From the above feature selection procedure, the full rank matrix $\mathbf{P}$ can be orthogonally decomposed as

$$\mathbf{P} = \mathbf{Q}\mathbf{R}, \quad (13)$$

where $\mathbf{R}$ is an $m \times m$ unit upper triangular matrix and $\mathbf{Q}$ is an $N \times m$ matrix with orthogonal columns $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_m$. Substituting (13) into (12) yields

$$\mathbf{x}_j = (\mathbf{P}\mathbf{R}^{-1})(\mathbf{R}\boldsymbol{\theta}_j) + \mathbf{e}_j = \mathbf{Q}\mathbf{g}_j + \mathbf{e}_j, \quad (14)$$

where $\mathbf{g}_j = [g_{j,1}, \cdots, g_{j,m}]^T = \mathbf{R}\boldsymbol{\theta}_j$ is an auxiliary parameter vector. Using the orthogonal property of $\mathbf{Q}$, $g_{j,k}$ can be directly calculated from $\mathbf{x}_j$ and $\mathbf{Q}$ using $g_{j,k} = (\mathbf{x}_j^T\mathbf{q}_k)/(\mathbf{q}_k^T\mathbf{q}_k)$ for $k = 1, 2, \ldots, m$. The unknown parameter vector $\boldsymbol{\theta}_j$ can then be easily calculated from $\mathbf{g}_j$ and $\mathbf{R}$ by substitution using the special structure of $\mathbf{R}$.

From (14), the total sum of squares of the independent variable $\mathbf{x}_j$, with respect to $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_m$ (or, equivalently, with respect to $\mathbf{z}_1, \cdots, \mathbf{z}_m$), can be expressed as

$$\mathbf{x}_j^T\mathbf{x}_j = \sum_{k=1}^m g_{j,k}^2\mathbf{q}_k^T\mathbf{q}_k + \mathbf{e}_j^T\mathbf{e}_j. \quad (15)$$

Following [21], [22], the kth error reduction ratio (ERR) introduced by including $\mathbf{q}_k$ (or, equally by including $\mathbf{z}_k$) in to the subset, is defined as

$$\mathrm{ERR}[j; k] = \frac{g_{j,k}^2(\mathbf{q}_k^T\mathbf{q}_k)}{\mathbf{x}_j^T\mathbf{x}_j} \times 100\% = \frac{(\mathbf{x}_j^T\mathbf{q}_k)^2}{(\mathbf{x}_j^T\mathbf{x}_j)(\mathbf{q}_k^T\mathbf{q}_k)} \times 100\%, \quad k = 1, 2, \ldots, m. \quad (16)$$

The *sum of error reduction ratio* (SERR) due to $\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_m$ (or equally due to $\mathbf{z}_1, \cdots, \mathbf{z}_m$) are defined as [24]

$$\mathrm{SERR}[j; m] = \sum_{k=1}^m \mathrm{ERR}[j; k]. \quad (17)$$

The percentage of the variation in the overall features that can be accounted for by the subset S_m can then be calculated as

$$\overline{SERR}[m] = \frac{1}{n}\sum_{j=1}^n SERR[j; m]. \quad (18)$$

The criterion $\overline{SERR}$ can be used to measure the performance of the selected subset S_m and to monitor the search procedure. If $\overline{SERR}$ is larger than a given threshold, the associated subset S_m can then be considered to be sufficient to represent the overall features, otherwise, more significant variables need to be included.

The time required to implement the FOS-MOD algorithm is mainly determined by two parts: the orthogonalization procedure (7) and the calculation of the correlation matrix (8). The orthogonalization procedure (7) is of the complexity $O((m-1)N)$ with $m << n$, and the calculation of the correlation matrix (8) is of the complexity $O(n^2N)$, where n is the number of candidate features and N is the number of observations. The overall computational complexity of the FOS-MOD algorithm for each search step is thus of the order $O(n^2N)$.

TABLE 1
Feature Detection and Ranking Results for the Alate Adelges Data

Step	Feature No.	$\overline{\text{ERR}}$ (%)	$\overline{\text{SERR}}$ (%)
1	13	69.4245	69.4245
2	17	11.2188	80.6433
3	11	4.4604	85.1037
4	5	3.5045	88.6082
5	19	2.4312	91.0394
6	18	1.6673	92.7067
7	9	1.1296	93.8363
8	6	1.0512	94.8875
9	10	0.9699	95.8574
10	15	0.7766	96.6340
11	1	0.7318	97.3658
12	14	0.7041	98.0699
13	16	0.5112	98.5811
14	8	0.3926	98.9737
15	2	0.2947	99.2684
16	4	0.2802	99.5487
17	3	0.2115	99.7602
18	12	0.1233	99.8835
19	7	0.1165	100.000

3 Experiments

3.1 Example 1—The Alate Adelges Data

The Alate Adelges data set comprises 19 variables measured on each of 40 winged aphids (alate adelges) that had been caught in a light trap. This data set was studied in [25] using principal component analysis. The full 40×19 data matrix is available in [7], where a very efficient procrustes analysis method has been proposed to select variables that preserve multivariate data structure.

The original data were standardized and the following analysis was based on the normalized data. Denote the 19 variables (attributes) by $x_1, x_2, \cdots, x_{19}$. By applying the new FOS-MOD algorithm to the data set, the significance of the 19 variables has been detected and the detection results are shown in Table 1, where variables are ranked according to the percentage contribution to the underlying overall characteristics. Note that the first three features, x_{13}, x_{17}, x_{11}, selected by the FOS-MOD algorithm are identical to those selected by the B4 method in [26]. The B4 method is a PCA based approach, which involves the use of the first p components themselves. Candidate variables are associated with each of the first p components in some specified manner and p variables are retained and the remaining variables are rejected (see [16] and the references therein for details about the B4 method).

If the threshold for $\overline{SERR}$ is set to be 0.95, a subset of nine features should then be considered. To evaluate how well the 9-feature subset captures the structure of the complete data, a further principal component analysis was done on both the complete data and the data formed by the selected nine features. Fig. 1a presents the two-dimensional graph of the complete data matrix while Fig. 1b presents the two-dimensional representation of the 9-feature subset. Clearly, the 9-feature subset provides a satisfactory representation for the complete data providing that capturing the data structure is the prime objective. In Fig. 1a, both of the first two principal components (PCs) are functions of all the 19 variables, while in Fig. 1b, the first two PCs only involve the nine selected variables. Table 1 clearly shows which of these individual variables contribute most and provides a ranked list of these. This aids interpretation because PCs in general cases are functions of all the original variables but FOS-MOD shows individual contributions.

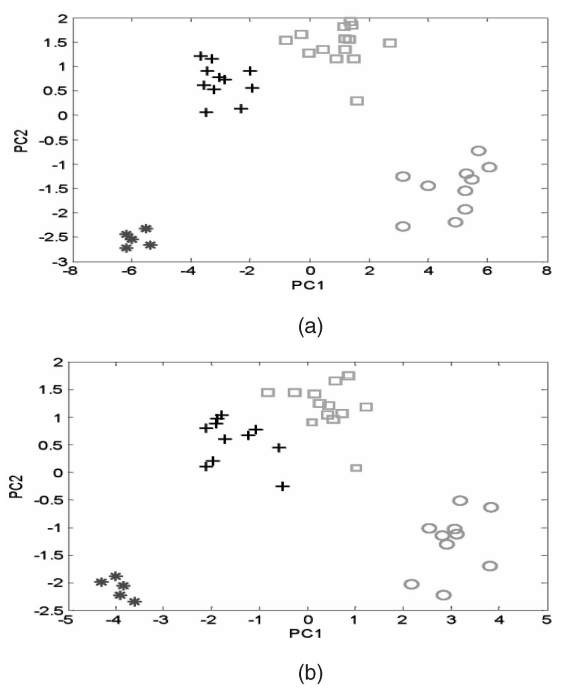

Fig. 1. Alate Adelges data plotted against the first two principal components. (a) Computed from all the 19 variables and (b) computed from the first nine selected variables listed in Table 1.

Notice that Fig. 1 only graphically presents the performance of the FOS-MOD algorithm by qualitatively comparing the structure formed by the first two associated PCs. From this visual illustration, however, it is difficult to obtain a quantitive measure about how efficient the subsets selected by the FOS-MOD algorithm are. In the following, the FOS-MOD algorithm was thus applied to pattern classification by analysing several real data sets, to quantitively inspect the efficiency of the new algorithm.

3.2 Example 2—Data Sets from UCI Machine Learning Repository

Five real data sets, taken from the UCI machine learning repository [27], are considered. The objective is to select a subset for each data set using the FOS-MOD algorithm and the selected subset is then used to replace the associated complete data for designed pattern classification. The threshold for $\overline{SERR}$ in the FOS-MOD algorithm was set to be 0.95 for all the five data sets. Details about the five data sets and associate experiments are given below:

- *Wisconsin Breast Cancer* (WBC). The Wisconsin breast cancer data contains 699 samples, where 458 are benign samples (65.52 percent) and 241 are malignant samples (34.48 percent). Each instance is characterized by nine attributes. The objective is to predict diagnosis results that are either benign or malignant.
- *Wisconsin Diagnostic Breast Cancer* (WDBC). This data set contains 569 samples, where 357 are benign samples (62.74 percent) and 212 are malignant samples (37.26 percent). Each instance is characterized by 30 real-valued attributes. The objective is as in the WBC data.
- *Johns Hopkins University Ionosphere*. This data set contains 351 samples and 34 real-valued attributes. This data set involves a binary classification task.

TABLE 2
A Comparison of the Classification Accuracy over the Original Complete Data and the Associated Selected Subsets, Using the k-NN Algorithm

Dataset	No. Attributes		Accuracy (%)		CPU Time for Subset Search (sec)
	Complete Data	Selected Subset	Complete Dataset	Selected Subset	
WBC	9	4	98.16 ± 2.03 (5-NN)	97.42 ± 2.16 (15-NN)	0.06
WDBC[N]	30	13	97.94 ± 1.67 (5-NN)	97.04 ± 1.65 (7-NN)	0.64
Forest [N]	10	5	66.07 ± 2.48 (1-NN) 64.39 ± 1.81 {5-NN}	64.45 ± 2.23 (9-NN) 63.18 ± 2.38 {5-NN}	0.78
Ionosphere	34	19	87.55 ± 3.20 (1-NN) 84.22 ± 4.38 {11-NN}	86.39 ± 5.51 (3-NN) 81.08 ± 5.63 {11-NN}	0.93
Arrhythmia	261	96	65.38 ± 7.20 (7-NN) 56.92 ± 7.70 {95-NN}	63.65 ± 4.39 (5-NN) 56.92 ± 7.70 {95-NN}	179

[N]: the original data were normalized in the experiments. { }: the value of k in k-NN rule was suggested by [8].

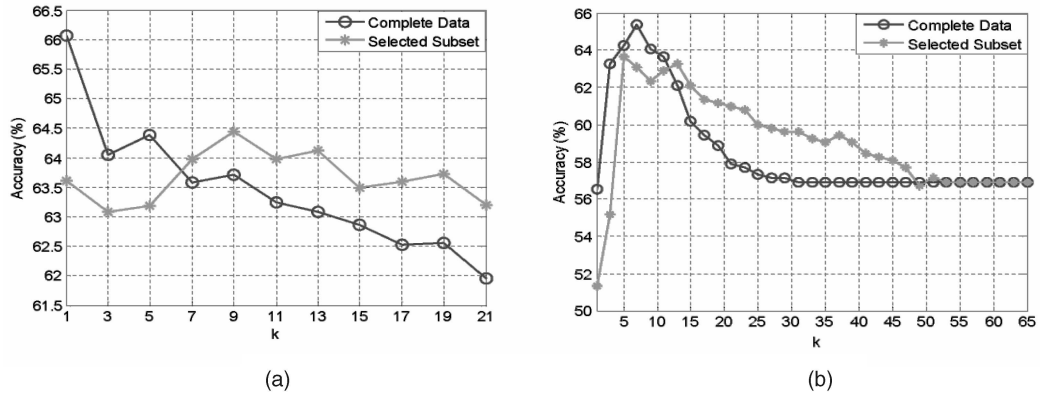

Fig. 2. A compassion of the classification accuracy based on the complete data and the associated subset, using the k-NN algorithm with different values of k. (a) For the Forest data and (b) for the Arrhythmia data.

- *Cardiac Arrhythmia*. This data set contains 452 instances and 279 attributes. The task is to classify a patient into one of the 16 classes of cardiac arrhythmia. This data set was preprocessed as below. Some values are missing for the attributes numbered by 11, 12, 13, and 15, and the missing values were filled with some values chosen randomly according to the distribution of the known values for the three attributes. Most of the values for the 14th attribute are missing and this attribute was not included in our experiment. Among the 279 attributes, 17 are trivial because all the observations for these attributes are zero. The 17 zero-valued attributes were not used in our experiment.
- *Forest Cover Type*. This data set represents the forest cover types in a region. There are 54 attributes, 581,012 instances and seven classes of cover types. The first 11,340 instances were used as the training data and the next 3,780 instances were used as the test data. Following [8], only the first 10 numerical-valued attributes were considered.

To inspect the performance of the new FOS-MOD algorithm, the k-nearest-neighbor (k-NN) algorithm was applied to evaluate the classification accuracy calculated by performing the following random cross-validation procedure. The k-NN algorithm was performed 20 times over the training and validation data defined as below: at each time, about 10 percent of the samples were randomly selected and left out, and these were used as the test data; the remaining 90 percent samples were used as the training data. The average classification accuracy of the 20 runs of the k-NN algorithm, over the test data, was then calculated. The value of k, in the k-NN rule, was chosen by performing many experiments for different values of k, where $1 \leq k \leq \sqrt{N_{tr}}$ and N_{tr} is the number of the samples in the training set, and k was chosen as the one that gives the best classification performance.

A feature subset for each of the five data sets, WBC, WDBC, Forest, Ionosphere, and Arrhythmia, was selected. The number of features in the selected subsets for the five data sets was 4, 13, 5, 19, and 96, respectively. The k-NN algorithm was applied to both the original complete data and the associated feature subset for each of the five data sets. A comparison between the classification accuracy based on the complete data and the associated subset for the five data sets is reported in Table 2, where the associated algorithms are implemented using Matlab (R14) on a Sun-Blade-2500 workstation (1.28GHz).

It can be seen from Table 2 that the classification accuracy based on the selected subsets is comparable with those based on the complete data. This means that the selected feature subsets are representative and informative and, thus, can be used to replace the complete data for pattern classification. Table 2 only presents the classification accuracy at some specific value of k, where the k-NN rule provides the best classification performance. It may be informative to compare the overall classification accuracy for different values of k, with respect to both the selected subset and the associated complete data. As a benchmark, Fig. 2 depicts such a comparison for the two data sets Forest and Arrhythmia.

For the data set WBC, the classification accuracy based on the selected subset is 97.42 percent, which is very near to the best result (97.5 percent) given in [28], where many classifiers were compared. For the data set WDBC, the classification accuracy based on the selected subset here is near to the result in [15], where the number of features involved in selected subsets is much more than the 13 used here. In this sense, the subset produced by the proposed algorithm for the data set WDBC is more compact. While for the data set Forest, the result produced by the FOS-MOD algorithm is comparable with those in [8], where several feature selection algorithms were compared, for the data sets Ionosphere and Arrhythmia, the results here are slightly better than those in

[8]. The mechanism of the FOS-MOD algorithm, however, is quite easy and the implementation of this algorithm only involves the calculation of the squared-correlation matrix and the maximization of the overall dependency. The results of the analysis of these data sets using several methods are already given in [8]. Comparing the results of the FOS-MOD algorithm with those in [8] therefore provides a full comparison of the various methods.

4 Conclusions

A new unsupervised learning algorithm has been proposed for feature selection and dimensionality reduction. The main advantage of the new algorithm is that the implementation only involves the calculation of the designed correlation matrix and the forward orthogonalization procedure. The new algorithm, which combines good effectiveness with high efficiency, often produces efficient feature subsets and, thus, provides an effective solution to the dimensionality reduction problem. The algorithm assumes that a linear relationship exists between sample features. In many cases, where features are linked by some nonlinear relationship, this assumption may become unreasonable. In such cases, more variables may need to be included in the final subset to achieve a satisfactory recognition result. This is a disadvantage of this type of approach. Future work will involve adapting the present method to accommodate nonlinear relationships and to seek more powerful dependence measurement criteria.

References

[1] M.A. Carreira-Perpinan, "Continuous Latent Variable Models for Dimensionality Reduction and Sequential Data Reconstruction," PhD dissertation, Dept. of Computer Science, Univ. of Sheffield, Sheffield, U.K., 2001.

[2] I.K. Fodor, "A Survey of Dimension Reduction Techniques," Technical Report UCRL-ID-148494, Lawrence Livermore Nat'l Laboratory, Center for Applied Scientific Computing, June 2002.

[3] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37, Jan. 2000.

[4] A.R. Webb, *Statistical Pattern Recognition,* second ed. Wiley, 2002.

[5] I.T. Jolliffe, *Principal Component Analysis,* second ed. Springer, 2002.

[6] G.P. McCabe, "Principal Variables," *Technometrics,* vol. 26, pp. 137-144, May 1984.

[7] W.J. Krzanowski, "Selection of Variables to Preserve Multivariate Data Structure Using Principal Components," *Applied Statististics,* vol. 36, no. 1, pp. 22-33, 1987.

[8] P. Mitra, C.A. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 3, pp. 301-312, Mar. 2002.

[9] B. Krishnapuram, A.J. Hartemink, L. Carin, and M.A.T. Figueiredo, "A Bayesian Approach to Joint Feature Selection and Classifier Design," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1105-1111, Sept. 2004.

[10] M.H.C. Law, M.A.T. Figueiredo, and A.K. Jain, "Simultaneous Feature Selection and Clustering Using Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1154-1166, Sept. 2004.

[11] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence,* vol. 97, nos. 1-2, pp. 273-324, Dec. 1997.

[12] A.J. Miller, *Subset Selection in Regression.* Chapman and Hall, 1990.

[13] P. Pudil, J. Novovicova, and J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters,* vol. 15, no. 11, pp. 1119-1125, Nov. 1994.

[14] S.K. Pal, R.K. De, and J. Basak, "Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach," *IEEE Trans. Neural Networks,* vol. 11, no. 2, pp. 366-376, Mar. 2000.

[15] K.Z. Mao, "Identifying Critical Variables of Principal Components for Unsupervised Feature Selection," *IEEE Trans. Systems, Man, and Cybernetics, Part B,* vol. 35, pp. 339-344, 2005.

[16] I.T. Jolliffe, "Discarding Variables in a Principal Component Analysis-I: Artificial Data," *Applied Statistics,* vol. 21, no. 2, pp. 160-173, 1972.

[17] L. Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science,* vol. 16, no. 3, pp. 199-215, Aug. 2001.

[18] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," *Neural Computation,* vol. 9, no. 7, pp. 1545-1588, Oct. 1997.

[19] Y. Amit, D. Geman, and K. Wilder, "Joint Induction of Shape Features and Tree Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 11, pp. 1300-1305, Nov. 1997.

[20] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[21] M. Korenberg, S.A. Billings, Y.P. Liu, and P.J. McIlroy, "Orthogonal Parameter Estimation Algorithm for Non-Linear Stochastic Systems," *Int'l J. Control,* vol. 48, pp. 193-210, 1988.

[22] S.A. Billings, S. Chen, and M.J. Korenberg, "Identification of MIMO Non-Linear Systems Suing a Forward Regression Orthogonal Estimator," *Int'l J. Control,* vol. 49, pp. 2157-2189, June 1989.

[23] T.M. Cover and J.M. Van Campenhout, "On the Possible Orderings in the Measurement Selection Problem," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 7, no. 9, pp. 657-661, Sept. 1977.

[24] H.L. Wei, S.A. Billings, and J. Liu, "Term and Variable Selection for Nonlinear System Identification," *Int'l J. Control,* vol. 77, no. 1, pp. 86-110, Jan. 2004.

[25] J.N.R. Jeffers, "Two Case Studies in the Application of Principal Component Analysis," *Applied Statistics,* vol. 16, no. 3, pp. 225-236, 1967.

[26] I.T. Jolliffe, "Discarding Variables in a Principal Component Analysis. II: Real Data," *Applied Statistics,* vol. 22, no. 1, pp. 21-31, 1973.

[27] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, 2006.

[28] Faculty of Physics, Dept. of Informatics, Nicolaus Copernicus Univ., Torun, Poland,http://www.phys.uni.torun.pl/kmk/projects/datasets.html, 2006.