TotalPLS: Local Dimension Reduction for Multicategory Microarray Data

Wenjie You, Zijiang Yang, Mingshun Yuan, and Guoli Ji

***Abstract*—Dimension reduction is an important topic in data mining, which is widely used in the areas of genetics, medicine, and bioinformatics. We propose a new local dimension reduction algorithm TotalPLS that operates in a unified partial least squares (PLS) framework and implement an information fusion of PLS-based feature selection and feature extraction. This paper focuses on extracting the potential structure hidden in high-dimensional multicategory microarray data, and interpreting and understanding the results provided by the potential structure information. First, we propose using PLS-based recursive feature elimination (PLSRFE) in multicategory problems. Then, we perform feature importance analysis based on PLSRFE for high-dimensional microarray data to determine the information feature (biomarkers) subset, which relates to the studied tumor subtypes problem. Finally, PLS-based supervised feature extraction is conducted on the selected specific genes subset to extract comprehensive features that best reflect the nature of classification to have a discriminating ability. The proposed algorithm is compared with several state-of-the-art methods using multiple high-dimensional multicategory microarray datasets. Our comparison is performed in terms of recognition accuracy, relevance, and redundancy. Experimental results show that the algorithm proposed by us can improve the recognition rate and computational efficiency. Furthermore, mining potential structure information improves the interpretability and understandability of recognition results. The proposed algorithm can be effectively applied to microarray data analysis for the discovery of gene coexpression and coregulation.**

***Index Terms*—Dimension reduction, feature extraction, feature selection, microarray data analysis, partial least squares (PLS).**

I. Introduction

In recent years a large number of high-dimensional and ultrahigh-dimensional datasets [1], [2] have appeared in various areas of application, such as microarray, image recognition, and text categorization due to technological advances in data acquisition and storage. More detailed information about objective phenomenon can be provided with increasing data dimensions, thus making the acquired dataset increasingly large. Fukunaga [3] provided the variability analysis in performance estimates of classifiers trained and tested using a finite number of samples. They proved that for the first-order classification (only first-order statistics), the training sample size is proportional to the number of the dimensions of the dataset. For second-order classifiers (only second-order statistics), the training sample size is proportional to the square of dimension of the data. For nonparametric classifiers, the relationship between training sample size and data dimension is exponential. However, the training sample size is always limited for typical microarray data analysis. As the dimension increases, the size of training samples becomes relatively small. These reasons imply that many data mining classification algorithms lack efficiency or even fail in high-dimensional microarray data analysis [4].

Constructing effective dimension reduction methods is crucial. Dimension reduction uses a small amount of features or synthesis features (component variables) to substitute a feature subset containing strong correlations in the original data [5]. The intrinsic dimension of high-dimensional feature space showed that there is often a low-dimensional structure embedded in the feature space of high-dimensional data [4]. Low-dimensional representations, revealing internal structure and ultimately understandable patterns are the key issues in high-dimensional data mining research. In the context of limited loss of information, the accuracy of classification can be improved if high-dimensional data are projected into a low-dimensional subspace and classified in this low-dimensional subspace. Feature extraction is a common method for dimension reduction of high-dimensional data. The original high-dimensional feature space is projected on to low-dimensional feature space. Potential features obtained after the projection are either a linear or nonlinear combination of the original features. By this we mean that feature extraction rotates the original coordinate system and then selects a number of important potential features. It is obvious that feature extraction is a global dimension reduction method; thus, it yields better results when the dataset is globally relevant.

Feature selection is another dimension reduction method that only uses a subset of features from the original data. The advantage of feature selection is that the selected data table does not rotate. Therefore, it is easy to interpret the results. The drawback is that this method may cause information loss. Information loss does not occur, however, when there are redundant or irrelevant features. This is because the redundant information is repeatedly captured in one or more features and the presence

Manuscript received May 8, 2013; revised August 17, 2013 and October 18, 2013; accepted October 27, 2013. Date of publication November 21, 2013; date of current version January 16, 2014. This work was supported by funds from the Natural Sciences and Engineering Research Council of Canada, the National Natural Science Foundation of China (No. 61174161, 61201358, and 61203176), the Natural Science Foundation of Fujian Province of China (No. 2012J01154), the specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130121130004 and 20120121120038), the Key Research Project of Xiamen City of China (No. 3502Z20123014), and the Fundamental Research Funds for the Central Universities in China (Xiamen University: No. 2011121047, 2013121025 and CBX2013015). This paper was recommended by Associate Editor J. K. Lee.

W. You and M. Yuan are with the Department of Automation, Xiamen University, Xiamen, Fujian 361005, China (e-mail: wenjie.you@hotmail.com; yzk78@sina.com).

Z. Yang is with the School of Information Technology, York University, Toronto, ON M3 J 1 P3, Canada (e-mail: zyang@yorku.ca).

G. Ji is with the Department of Automation, Xiamen University, Xiamen, Fujian 361005, China. He is and also with the Innovation Center for Cell Biology, Xiamen University, 361102 Xiamen, China (e-mail: glji@xmu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2013.2288777

of redundant and irrelevant features decreases the classification accuracy. Conceptually, feature subset selection is the process that searches all possible feature subsets. There are many different search strategies. Nevertheless, computational cost should be relatively low and optimal or near optimal feature subsets should be found. In fact, it is often impossible to satisfy both of these requirements. Therefore, a tradeoff between computational cost and an optimal feature subset should be considered.

In data mining and machine learning, the problem of multicategory classification is another challenge. Multicategory classification is often regarded as the expansion of two-category classification. Integration of support vector machine (SVM), recursive feature elimination (RFE) feature selection (see [6]), and SVM-supervised learning are some of the best methods so far for dimension reduction and classification in high-dimensional microarray datasets. Because SVM is originally designed for two-category classification, SVM-based multicategory classification is currently an important part of the study [7]. There are usually two solving strategies. One is to implement a decomposition algorithm in which a multicategory classification is decomposed into several two-category classifications and multiple SVM-based two-category classifiers are constructed. In the next step of this strategy, the multiple classifiers are combined to achieve multicategory classification such as "one-versus-one (OVO)", "one-versus-all (OVA)", and "directed acyclic graph SVM" [8]–[10]. The other strategy is to implement an overall approach that takes into account all categories in an optimization formula [11]. This strategy, however, is slow and the increased classification accuracy obtained is not at a significant advantage. Lee *et al.* [12] designed multicategory SVMs using the hinge loss function, which has a promising theoretical property. Its solution for the multiclass problem resembles Bayes rule asymptotically. However, in real microarray analysis for ultrahigh-dimensional data, OVO and OVA are still considered a better fit. Nevertheless, the OVO strategy makes the size of the training samples even smaller, whereas the OVA strategy causes class imbalance in the training samples [13]. Furthermore, as the number of categories increases, the recognition rate becomes very poor.

In view of this high-dimensional and multicategory microarray data analysis, this paper presents a new local dimension reduction technique based on partial least squares (PLS), which can be directly applied to supervised multicategory classification problems. PLS-based recursive feature elimination (PLSRFE) selects a group of information feature subsets that relate to the studied pattern classification. Subsequently, PLS-based supervised feature extraction is applied on the selected information feature subsets to extract comprehensive features that best reflect classification nature and have a discriminating ability. This paper also focuses on mining the potential structure for multicategory microarray data analysis. Compared with other dimension reduction techniques, our algorithm can achieve a higher recognition rate with better intelligibility. Moreover, the introduced method achieves both feature selection and feature extraction in a unified PLS framework using our algorithm TotalPLS.

The remainder of this paper is organized as follows. Related dimension reduction methods are provided in Section II. Section III proposes PLSRFE and TotalPLS. In Section IV, TotalPLS is compared with the state-of-the-art methods in multicategory tumor microarray datasets. Further analysis of interpretability, understandability, and visual representation of the mining results are provided in Section V. Finally, we present our conclusions and discuss future work in Section VI.

II. Related Work

Here, we use the terms "feature selection" and "gene selection" interchangeably. In addition, we focus on analyzing high-dimensional multicategory problems where "high-dimensional datasets" refer to datasets with more than 10 000 dimensions. Numerous real datasets are high-dimensional, multicategory, and nonlinear. Effective analysis of such data is a research priority in machine learning, multivariate data analysis, and data processing in cognitive science. The purpose of dimension reduction is to find some form of structure hiding in the original high-dimensional data. Traditional dimension reduction methods can be generally classified into feature extraction (or feature transformation) and feature selection (or variables selection) [14], [15].

Feature selection algorithms use some statistical indicators or separability criteria to select the most representative features from the original feature sets. The goal is to get an informative feature subset that has a small number of features but contains strong identification information. Usually, two types of methods can be used in feature selection: filter and wrapper methods. A filter method is a single feature-scoring method based on certain criteria, which usually selects features with high scores for further analysis. The most popular filter methods involve using a t-test statistic or F-test statistic (see [16] and [17]), analyzing signal-to-noise ratio (see [18]–[20]), performing a nonparametric test such as the Wilcoxon rank sum test or Kruskal-Wallis rank sum test (see [21], [22]), studying mutual information (see [23]–[25]), implementing a Relief algorithm (see [26]), or a variation of a Relief algorithm in which the evaluation function is not related to the classifier (see [27]). A wrapper method is a feature-selection associated with a classifier. The output of the classifier is treated as a feature-selection criterion. Wrapper methods include Genetic Algorithms (see [28]), SVM (see [6], [29], [30]), the RFE method, and Boosting. A wrapper method usually uses the error probability of a classifier as the evaluation function. In filter method, each feature is evaluated depending only on its inherent self-information and the feature evaluation is not related with the information of other features. The advantage of a filter method is that it is fast and universal (independent of other features and classifier). Its disadvantage is that it ignores the relationship among features. A more accurate approach needs to consider the joint distribution among features. It must take into account all the features and should allow for the detection of features that have a relatively small main effect but has at least one strong interaction effect [31]. Those features with a smaller main effect may contain more important information for the studied issues and the analysis of these features may provide more comprehensive understanding of expression patterns for the issues studied.

Good feature selection methods should perform the following functions [32]:

1) take full account of the interaction among features;
2) performance should be based on the feature subset rather than individual features associated with classification;
3) the feature selection algorithm should be reasonable and efficient and the selected subset should contain fewer features than in the initial set;
4) detect features that have a relatively small main effect but have a strong interaction effect [31].

Therefore, feature selection can be divided into single-feature selection (univariate method) and multifeature selection (multivariate method) based on whether the approach considers interaction among features or not. SVM-based SVMRFE algorithms [6] and Relief algorithms [26] based on an iteratively adjusted margin are currently known to be the best methods for multifeature selection. However, the original two methods apply to two-category classification problems. ReliefF, which is an improved version of the Relief algorithm (see [27]), can solve multicategory classification and regression problems. OVO- and OVA-based multiclass SVM-RFE (see [33]) and the extensions of SVM-RFE are used in multiclass gene selection [10]. These algorithms consider the correlation among features to some extent. Feature selection based on mutual information in information theoretic learning has been extensively studied [25]. There are many approaches to offset the impact of interaction among features [34]–[36]. The limitation of these methods is the high-computational cost. In addition, most of the algorithms based on information measurements can be unified to a general criterion function about mutual information in feature selection [37].

Feature extraction is a mapping process of the original feature space. It projects the original feature space into low-dimensional space that better reflects the structure information of the original sample data [15], [38]. There are several commonly used feature extraction methods including principal component analysis (PCA) (see [39]), independent component analysis (ICA) (see [40]), linear discriminant analysis (LDA) (see [39]), and between-group analysis (BGA) (see [41]). They achieve very good results for most datasets with a linear structure. Using the idea of largest variance and minimum deviation, PCA discovers the main direction of the dataset to achieve dimension reduction. However, PCA has limitations when datasets are highly nonlinear. ICA assumes the existence of inherent variables in the dataset and uses blind source analysis of information theory to obtain the synthesis features. It does not consider the global and local nature of the observation space of the data. LDA and BGA are supervised feature extraction methods and take into account the *a priori* information of categories. In addition, the methods based on space-frequency transformation including Fourier transform (FT), discrete cosine transform, Hadamard transform, and wavelet analysis such as Gabor wavelet are generally used in image processing [42]. However, wavelet transforms, FTs, and other orthogonal transforms essentially transform the spatial datasets into the frequency domain to achieve dimension reduction. This process does not have an intuitive geometric interpretation. Yan *et al.* proposed the incremental orthogonal centroid algorithm, which is based on the orthogonal centroid algorithm [15], and Li *et al.* proposed the maximum margin criterion method [43]. These algorithms have good performance.

There are two major research directions regarding model selection on high-dimensional datasets. One is based on a regularization method in which shrinkage estimates are obtained using a penalty function. Examples include ridge estimation (see [44], [45]); Lasso (see [46]), Least Angle Regression (LARS, see [1]), and Elastic net (see [47]). The other direction is based on components such as PCA and PLS [48]. The PLS-based supervised feature extraction method can achieve good results with many public databases and benchmark tests. PLS is a nonparametric method based on the idea of high-dimensional projection, and it has been extensively used in high-dimensional microarray data analysis. Nguyen and Rocke [49] and Dai *et al.* [18] applied PLS in a dimension reduction method called PLSDR for high-dimensional microarray data analysis. Boulesteix *et al.* systematically compared several of the most popular PLS approaches and their applications in the fields of tumor classification, identification of relevant genes [50], [51]. Ji *et al.* proposed PLS-based gene selection to identify tumor-specific genes in two-category and multicategory tumor subtypes classifications [31]. Yang *et al.* proposed information fusion of PLS-based feature selection and SVM-based classification to predict business failure [52]. Gutkin *et al.* presented a PLS-based feature selection that dealt with two-category classification [53]. Cao *et al.* presented a computational methodology called sparse PLS to perform variable selection in a multiclass classification framework [54]. Chakraborty and Dutta proposed SVA-PLS method to excavate the hidden sources of expression heterogeneity in gene expression studies [55]. Zhao *et al.* developed higher-order partial least squares aiming to predict a tensor (multiway array) Y from a tensor X through projecting the data onto the latent space and performing regression on the corresponding latent variables [56]. All indicate that PLS has good performance on dimension reduction and classification problems.

III. TotalPLS—Local Dimension Reduction Algorithm

Traditional dimension reduction techniques studied global dimension reduction more for high-dimensional data [57]. Each dimension of the synthesis features extracted by the global dimension reduction technique contains information of all dimensions in the original data. Therefore, a global dimension reduction technique reflects all features in the dataset. The difference between a global dimension reduction technique and local dimension reduction technique is that each dimension of the synthesis features extracted by a local dimension reduction technique contains only a local information feature subset of the original dataset.

The extracted synthesis features better reflect the discrimination classification information in raw data. A local dimension reduction method is a two-stage design process. First, feature selection is completed on the high-dimensional dataset to exclude redundant and irrelevant features, which are irrelevant to the classification information. This is done so that a group of important information feature subsets can be gained. Subsequently, feature extraction is performed on the information feature subsets. A certain mapping is used to project the feature subsets into low-dimensional space that better reflects the information in the original sample. The objective is to find some form of hidden structure (i.e., potential feature subset) in the original high-dimensional data. The latent feature subset obtained from a good dimension reduction method can make the classification process that follows easier. This paper presents a new local dimension reduction technique that involves a PLS-based feature selection (PLSRFE) that can be directly applied to supervised multicategory classification problems. PLSRFE selects a group of information feature subset that relates to the studied pattern classification. Subsequently, PLS-based supervised feature extraction is applied in the selected information feature subset. As the proposed algorithm achieves both PLS-based feature selection and feature extraction in a unified PLS framework, we call it TotalPLS dimension reduction.

A. Partial Least Squares

Let $X = (x_1, x_2, \ldots, x_p)$ be a $n \times p$ matrix that has been normalized to have a mean of zero and $Y = (y_1, y_2, \ldots, y_q)$ be a $n \times q$ matrix that has been normalized to have a mean of zero. The goal of PLS is to find a pair of projection directions (weight vectors) w and v so that the projections (i.e., PLS components) $t = Xw$ and $u = Yv$ can meet the requirement that t and u carry as much information on variation as possible in X and Y, respectively, and the correlation coefficient between t and u is a maximum. PLS requires maximizing the covariance of t and u. The projection directions w and v of PLS can be obtained by solving the following criterion function [58]:

$$J(w, v) = \frac{(w^T \Sigma_{XY} v)^2}{w^T w \cdot v^T v} \tag{1}$$

where $w^T w = v^T v = 1$, and Σ_{XY} is the covariance matrix for the random vectors X and Y. The unit vector w and v obtained by solving the optimization criterion function in (1) are referred to as the PLS projection directions. When the samples are projected to the projection vector, t and u have the maximum covariance.

Under the orthogonal constraint $w_k^T w_i = 0, v_k^T v_i = 0 (1 \leq i < k)$ and using the Lagrangian multiplier method, the problem can be transformed into solving the following Eigen equations:

$$\begin{aligned} \Sigma_{XY} \Sigma_{YX} w &= \lambda^2 w \\ \Sigma_{YX} \Sigma_{XY} v &= \lambda^2 v. \end{aligned} \tag{2}$$

Let r be the maximum number of the effective relevant projection vectors (i.e., is the number of nonzero eigenvalues λ^2 of the matrix $\Sigma_{XY}\Sigma_{YX}$). The $d (d \leq r)$ pairs of relevant projection vectors are obtained by d eigenvectors corresponding to the largest d eigenvalues in (2). As $\Sigma_{XY}\Sigma_{YX}$ and $\Sigma_{YX}\Sigma_{XY}$ are both symmetric matrices and $\mathrm{rank}(\Sigma_{XY}\Sigma_{YX}) = \mathrm{rank}(\Sigma_{YX}\Sigma_{XY}) \leq \mathrm{rank}(\Sigma_{XY})$, the two Eigen equations have the same nonzero eigenvalues $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2$ and the number of nonzero eigenvalues $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2$ are at most $\mathrm{rank}(\Sigma_{XY})$. If the nonzero eigenvalues $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2$ meet the requirement of $\lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_r^2 > 0$, the corresponding r pairs of eigenvectors are orthogonal

$$w_i^T w_j = v_i^T v_j = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \tag{3}$$

To obtain the uncorrelated projection vectors, the first pair of weight unit vectors is obtained as the first pair of the optimal projection directions $\{w_1, v_1\}$. The $k + 1 (k \geq 1)$ pairs of the optimal projection directions meet the following constraints:

$$w_{k+1}^T \Sigma_X w_i = v_{k+1}^T \Sigma_Y v_i = 0 \quad (i = 1, 2, \ldots, k). \tag{4}$$

When the criterion function $J(w, v)$ is maximized, the vector pairs $\{w_{k+1}, v_{k+1}\}$ are obtained.

Let $W = (w_1, w_2, \ldots, w_k)$ and let $V = (v_1, v_2, \ldots, v_k)$. Then, the constraints aforementioned can be rewritten as $w_{k+1}^T \Sigma_X W = v_{k+1}^T \Sigma_Y V = 0$. The optimal projection direction $\{w_{k+1}, v_{k+1}\}$ that satisfies the statistical irrelevance constraint is the eigenvector corresponding to the largest eigenvalue of the Eigen equation

$$\begin{aligned} P\Sigma_{XY}\Sigma_{YX} w_{k+1} &= \lambda w_{k+1} \\ Q\Sigma_{YX}\Sigma_{XY} v_{k+1} &= \lambda v_{k+1} \end{aligned} \quad [58],\ [59] \tag{5}$$

where

$$\begin{aligned} P &= I - (\Sigma_X W)[(\Sigma_X W)^T (\Sigma_X W)]^{-1} (\Sigma_X W) \\ Q &= I - (\Sigma_Y V)[(\Sigma_Y V)^T (\Sigma_Y V)]^{-1} (\Sigma_Y V). \end{aligned} \tag{6}$$

B. PLS-based Recursive Feature Elimination for Multicategory Data

Given a dataset with known categories $\{(X_i, y_i) | X_i \in R^p, y_i \in Y_C, i = 1, \ldots, n\}$ where $Y_C = \{c_1, c_2, \ldots, c_k\}$ is the class label set and k denotes the number of categories, the original class labels $(y)_{n\times 1}$ are encoded. The dependent variable of our PLS model is defined as $Y = (y_{ij})_{n\times k}$ (n is the number of observed samples and k is the number of categories) where

$$y_{ij} = I(y_i = c_j) = \begin{cases} 1 & y_i = c_j \\ 0 & \text{others} \end{cases}, i = 1, \ldots, n,\ j = 1, \ldots, k. \tag{7}$$

Therefore, the dependent variable matrix is encoded as $Y = (y_{ij})_{n\times k}$.

The dependent variable (class) after coding can be deemed as a collection of multiple response variables. Therefore, this PLS model can be directly applied in multiple-class pattern classification problems. In fact, it is a PLS-based model with multiresponse variables. The most common solution to the PLS-based model with multiple response variables is SIMPLS [60]. Let *nfac* be the number of factors. SIMPLS can calculate $w_1, w_2, \ldots, w_{\mathrm{nfac}}$ by solving the following optimization problem:

$$\begin{cases} \max\ J(w) = cov^2(Xw_i, Y) \\ s.t. \qquad \|w_i\| = 1; \\ \qquad w_i^T (X^T X) w_j = 0; \\ \qquad j = 1, \ldots, i-1. \end{cases} \tag{8}$$

The component t_h extracted from the PLS calculation represents as much variation information of X as possible. It is also associated with Y as much as possible to explain the information

```
Algorithm 1  PLSRFE
Input:   TrainX_{n×p},   ClassY_{n×g},   nfac
Output:  idx
Initialization: Set idx = [ ] and S = [1,2,…,p]
Repeat
    Update TrainX, whose features only include the features in S
    For i = 1 to p do
        Calculate each feature vip(i) in terms of Eq. (9)
    Sort vip in descending order and record the sorted array weight and the index rank
    Find the feature e so that weight(e) has the minmum magnitude among weight
    Update idx by adding the feature e on top of idx,
    Update S by removing feature e from S
Until  |S| <= nfac
Return  idx = [S, idx]
```

Fig. 1. PLS-based recursive feature elimination.

of Y. To analyze the explanation of variation in the variables with X to Y, we further quantitatively denote the impact of each x_j to Y. The variable importance in projection (VIP) [61] is introduced.

Definition 1 (Variant Importance in Projection): Let $r(x_i, x_j)$ be the correlation coefficient between two variables x_i and x_j. Given the explanation of variation of component t_h to $Y, Rd(Y; t_h) = \frac{1}{q}\sum_{k=1}^{q} r^2(y_k, t_h)$, and the accumulation of variation explanation of $t_1, t_2, \ldots, t_m$ to $Y, Rd(Y; t_1, t_2, \ldots, t_{nfac}) = \sum_{h=1}^{m} Rd(Y; t_h)$, we define the following:

$$\text{VIP}(x_j, m) = \sqrt{\frac{p}{Rd(Y; t_1, \ldots, t_m)} \sum_{h=1}^{m} Rd(Y; t_h) w_{jh}^2} \quad (9)$$

as VIP of x_j to Y where m is the number of latent variables (factors) and w_{jh} is the jth weight of axis w_h indicating the marginal contribution of x_j to construct components t_h.

Therefore, the indicator of VIP can be used in feature selection to realize PLS-based feature ranking. Its computational complexity is $O(mnp)$ [31]. And when $m > 1$, it is a multifeature ranking method (multivariate method).

Now, we present a new multifeature selection method based on an RFE strategy, the PLSRFE (see Algorithm 1 in Fig. 1). PLSRFE calculates VIP indicators based on the projection vector t and projection direction vector w generated in PLS modeling. The features with smaller sorting coefficients are removed in each iteration and eventually the features are listed in the descending order. In each recursive loop, PLSRFE removes the features with the existing smallest sorting coefficient and then the PLS model is remodeled using the remaining features to obtain new ranking coefficients. PLSRFE implements this process recursively and finally the feature ranking list is obtained. We can get a good compact feature subset using the ranking list. In fact PLSRFE is a Filter method and has highly efficient computing power.

C. TotalPLS Algorithm and Parameter Settings

In machine learning, an easy-to-learn system should have as few adjustable parameters as possible to reduce the uncertainties of the learning system. Similar to the kernel parameter selection in SVM, PLS parameter selection (i.e., the selection of the *nfac* variable) is a difficult task. Cross-validation can be used in PLS parameter selection. However, the drawback to cross-validation is that it significantly increases the computation cost and the problem to a certain extent becomes even more difficult to handle. Therefore, it is very important to select the parameter heuristically to reduce the computational cost and make defined parameters have a clear physical meaning. If *nfac* is too small, the model will underfit and if it is too large, the model complexity increases, which causes overfitting and weakens the generalization of the model. To simplify the model, we assume that *nfac* of TotalPLS satisfies the following parameter settings for multicategory classification problems: if k is the number of categories, then *nfac* is equal to k for feature selection (FS($nfac = k$)), and *nfac* is inferior or equal to k for feature extraction (FE($nfac \leq k$)). This assumption also reflects our purpose to mine the potential structural information in the multicategory problem.

PLS is applied to find the basic structure (path analysis) to simplify the observation system. PLS can synthesize the complex dimensions to a few potential features (unobservable random variables) to guide the classification process. For example, the k-category classification problem can be intuitively understood as *k,* unobservable synthesizable features representing the expression patterns for k categories of classification from a statistical point of view. Or there may be *k*–1 unobservable synthesizable features that represent the expression patterns for the classification of k-categories. As the prior knowledge of category is known, these *k*–1 synthesizable features represent the expression patterns of the *k*–1 categories, whereas the remaining one is the kth category pattern, which is similar to the concept of degrees of freedom in statistics.

Our new local dimension reduction algorithm (TotalPLS) is presented below in a unified PLS framework, which implements an information fusion of PLS-based feature selection and feature extraction (see Fig. 2).

IV. Experiments

We experimentally verify that our proposed algorithm not only improves the recognition accuracy, but also focuses on the intelligibility of mining results. Multiple datasets from cancer microarray database are selected to validate the proposed algorithm. The current literature shows that most researchers use a large number of datasets for comparison of their work to validate the recognition accuracy of their methods. However, they ignore the intelligibility of the recognition results and any practical significance they may offer. Our experiments focus more on supervised PLS-based feature extraction. We use the proposed method to remove irrelevant and redundant features and implement supervised feature extraction on the selected feature subset. Another goal is mining the potential structural information hidden in the high-dimension multicategory cancer microarray data (for details see Section V).

```
Algorithm 2   TotalPLS
Input:  $TrainX_{n×p}$,  $ClsY_{n×1}$,  $Dim$
Output: $XScore$   // $XScore$ is the score on PLS-based
                      latent factor
(1) Initialization
      —Encode class label $ClsY_{n×1}$ and generate $ClassY_{n×g}$
        using Eq. (7)
      —Set $nfac$ = $unique(ClsY)$   //$unique(ClsY)$ denotes
        the number of categories
(2) Feature Selection
      —Obtain $idx$ by calling Algorithm 1
        PLSRFE($TrainX$, $ClassY$, $nfac$)
      —Update $TrainX$, whose features only include top $Dim$
        features in $idx$
(3) Feature Extraction
      —For $j$=1 to $nfac$ do
      ——Calculate score matrix $T_j = <TrainX, w_j>$
          using Eq. (8)
      ——Update $XScore$ so that $XScore = [XScore, T_j]$
Return  $XScore$
```

Fig. 2. TotalPLS-based feature reduction algorithm.

TABLE I
DATASETS

Data Set	# Instances (Train:Test)	# Feat.	# Clas s	# Instances per class
(Bio) MLL	72 (57:15)	12582	3	24/20/28
(Bio) Stjude	327 (215:112)	12558	7	15/27/64/20/43/79/79
(Bio) GCM	198 (144:54)	16063	14	12(*4)/14/22/11(*4)/10(*2)/20/30
(Bio)CLL-SUB-111	111	11340	3	11/49/51
(Bio) Brain	42	5597	5	10/10/10/4/8
(Bio) Lung	203	12600	5	139/17/21/20/6
(Bio) NCI60	61	5244	8	9/5/7/8/8/9/6/9
(Bio) Tumors-11	174	12533	11	27/8/26/23/12/11/7/26/6/14(*2)

In the second column of the table, the numbers in parenthesis are the number of training and test samples. The training set and test set were separated in the original literature.

A. Multicategory Microarray Datasets

We chose eight cancer microarray datasets, where the high-dimensional datasets are standard. All microarray datasets in bioinformatics are from the Kent Ridge bio-medical dataset and the Arizona State University feature selection repository. Table I summarizes the datasets that we used. The selected microarray datasets have high dimensions and many categories. In addition, the datasets are not balanced. In particular, the GCM and NCI60 datasets are recognized as difficult datasets by many researchers in the literature.

B. Experimental Design

All experiments are implemented in MATLAB (2010a) on a desktop with Intel Core i3 CPU 2.4 GHz and 2-GB RAM. The specific steps in our experimental implementation are provided below: 1) Training and testing datasets are generated by an independent verification method [holdout cross-validation (HOCV)] and k-fold cross-validation (k-fold CV). A variety of dimension reduction methods are used to reduce the dimension on the space of the high-dimensional training set. 2) Classifiers are trained based on the resulting low-dimensional space and tested on the low-dimensional testing sets. 3) The recognition rate of different dimension reduction methods in low-dimensional space are calculated as evaluation criteria and the differences in terms of statistical significance are analyzed.

F-tests, the ReliefF algorithm, and the MSVMRFE algorithm are three state-of-the-art feature selection methods. We also used the ReliefF and MSVMRFE algorithms as both are suitable to process multicategory data. While the F-test is a typical single-feature selection method (see [16]), ReliefF is a typical multifeature selection method (see [17]), and MSVM-RFE is a typical multifeature selection method based on the RFE strategy (see [10], [33]). To compare the proposed algorithm with F-test, ReliefF, and MSVMRFE, our experiment applied all four of these different algorithms to the training set. The information feature subsets that contained the top 30, 50, 100, 500, and 1000 features were selected by each algorithm, respectively. Subsequently, supervised PLS-based feature extraction and LDA-based feature extraction are implemented in the selected information feature subsets, respectively. The different dimension reduction methods are denoted as:

PLSDR1 (F-test & PLS) [48] LDADR1(F-test & LDA)
PLSDR2 (ReliefF & PLS) LDADR2 (ReliefF & LDA)
PLSDR3 (MSVMRFE & PLS) LDADR3 (MSVMRFE & LDA)
TotalPLS (PLSRFE & PLS) LDADR4 (PLSRFE & LDA).

We first use HOCV to assess a model's generalization ability on three microarray datasets where the training set and test set were separated in the original literature, and then further assess the model's generalization ability using stratified k-fold CV on all the datasets. All of the dimension reduction methods are implemented on the same training set and testing set, which was randomly generated by k-fold CV: we applied different dimension reduction methods 1000 times on different subsets of the original dataset (tenfold CV was repeated 20 times, fivefold CV was repeated 40 times,and each time the feature subsets contained the top 30, 50, 100, 500, and 1000 features that were selected).

C. Result of Experiments

1) Experiment 1 (Analysis of Redundancy and Relevancy): We implemented HOCV in the microarray datasets (training set and test set) using eight local dimension reduction methods. For the k-category classification problems, LDA can get at most k–1 optimal discriminant vectors because of the LDA rank limitation. Thus, in this experiment, we also set parameter $nfac$ in feature extraction as k–1 for a fair comparison. The relationship between the number of selected feature and recognition accuracy on test sets are identified in Fig. 3. The recognition rate from TotalPLS is much better than the other methods.

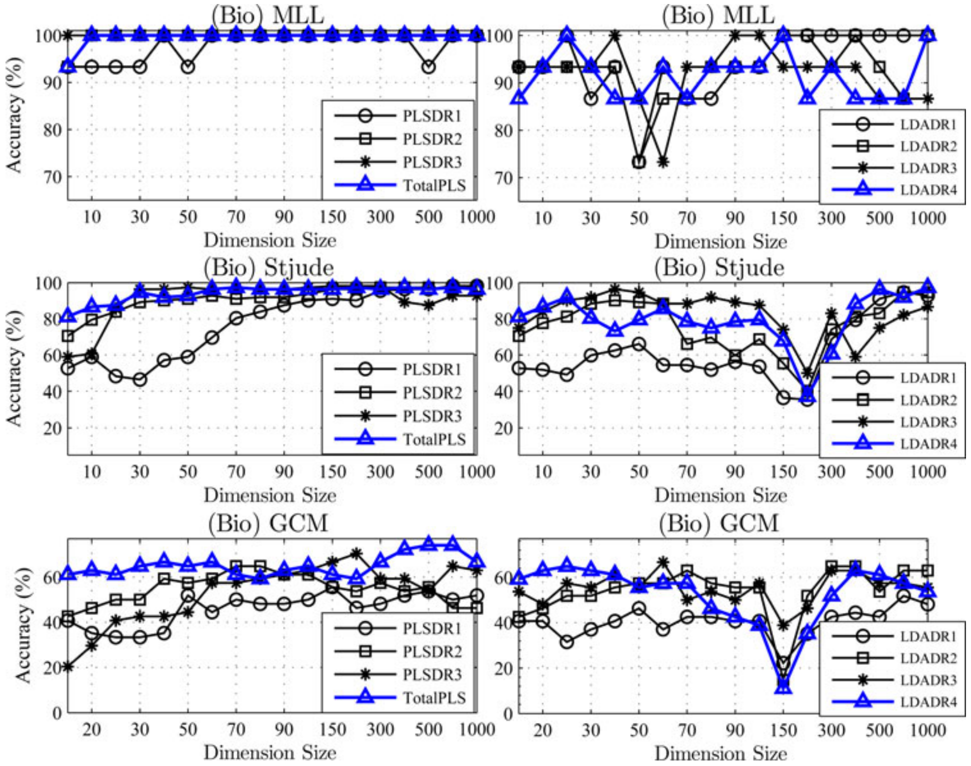

Fig. 3. Relationship between the dimension size and the recognition accuracy using different dimension reduction methods where the number of factors in feature extraction is fixed at k–1 (HOCV, Classifier: LDA).

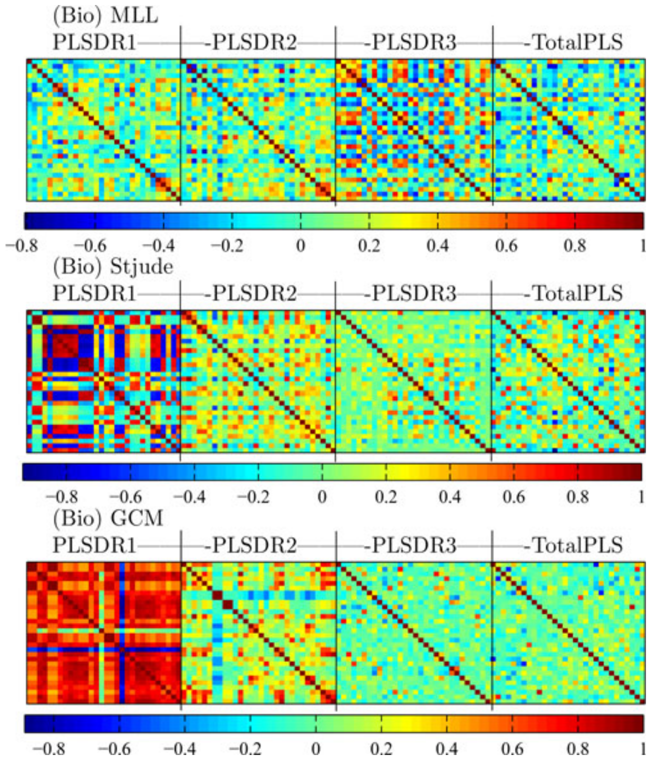

Fig. 4. Correlation heat map based on the top 30 selected features using different dimension reduction methods.

The results from PLS-based feature extraction are better than the results from LDA-based feature extraction, which is also consistent with the conclusions in the literature [18], [49], [50]. Therefore, we only focus on four PLS-based feature extraction methods in the following experiments and analysis.

Redundancy and relevancy are important criteria to evaluate dimension reduction methods. Excellent feature selection algorithms are able to reduce the correlation among features to improve classification accuracy. In Fig. 4, we see the heat map of the correlation matrix with the top 30 information features selected on three biological datasets. The lighter the color, the weaker the correlation among the selected features. Conversely, the darker the color, the stronger the correlation among the selected features. The darkness of the color can, to some extent, measure the degree of redundancy among the selected features.

In the MLL dataset set, the correlation heat map for the top 30 features selected by the four algorithms is consistent. This is not surprising as this dataset is easy to classify (see Table II). In the other two biological datasets (Stjude and GCM), PLSDR1 causes significant redundancy because positive correlation dominates the figure. The PLSDR2 and PLSDR3 algorithms have better performance as there is only some negative correlation. Our proposed TotalPLS method can significantly reduce redundancy to some extent as PLSDR1 (F-test) is a single feature selection method, whereas PLSDR2 (ReliefF) is a multifeature selection method, and TotalPLS (PLSRFE) and

TABLE II
CLASSIFICATION ACCURACY (%) OF FOUR DIMENSION REDUCTION METHODS PERFORMED ON DIFFERENT CLASSIFIERS (HOCV)

Classifier	#Dim.	PLSDR1	PLSDR2	PLSDR3	TotalPLS	PLSDR1	PLSDR2	PLSDR3	TotalPLS	PLSDR1	PLSDR2	PLSDR3	TotalPLS
		DataSet: (Bio) MLL				DataSet: (Bio) Stjude				DataSet: (Bio) GCM			
LDA	30	93.33(2)	100(2)	100(2)	100(2)	69.64(3)	90.18(6)	96.43(7)	94.64(6)	40.74(3)	53.70(11)	48.15(5)	64.81(10)
	50	93.33(2)	100(2)	100(2)	100(2)	76.79(4)	91.07(7)	98.21(6)	96.43(6)	51.85(14)	57.41(11)	48.15(12)	64.81(8)
	100	100(3)	100(1)	100(2)	100(2)	90.18(6)	93.75(7)	98.21(6)	97.32(6)	51.85(11)	61.11(13)	64.81(13)	64.81(14)
	500	93.33(3)	100(2)	100(2)	100(2)	96.43(6)	97.32(7)	90.18(4)	97.32(6)	53.70(12)	55.56(9)	55.56(13)	74.07(14)
	1000	100(3)	100(2)	100(2)	100(2)	98.21(7)	97.32(7)	93.75(6)	96.43(6)	51.85(13)	53.70(10)	62.96(14)	66.67(14)
	Mean	**96.00(2.6)**	**100(1.8)**	**100(2)**	**100(2)**	**86.25(5.2)**	**93.93(6.8)**	**95.36(5.8)**	**96.43(6)**	**50.00(10.6)**	**56.30(10.8)**	**55.93(11.4)**	**67.04(12)**
	Full	100				5.36				46.30			
SVM (Linear)	30	93.33(2)	100(2)	100(2)	100(2)	50.00(3)	79.46(6)	66.07(6)	83.04(6)	29.63(5)	22.22(10)	31.48(7)	38.89(14)
	50	100(3)	100(2)	100(2)	100(2)	63.39(4)	82.14(7)	88.39(7)	88.39(7)	38.89(8)	27.78(12)	31.48(11)	33.33(12)
	100	100(3)	100(2)	100(2)	100(2)	71.43(7)	79.46(6)	91.07(6)	90.18(5)	38.89(8)	25.93(10)	37.04(12)	35.19(14)
	500	100(3)	100(3)	100(2)	100(2)	86.61(6)	86.61(7)	90.18(6)	86.61(6)	27.78(14)	27.78(11)	33.33(12)	35.19(13)
	1000	100(3)	100(3)	100(2)	100(2)	84.82(6)	84.82(7)	90.18(7)	86.61(6)	27.78(14)	25.93(11)	29.63(12)	27.78(12)
	Mean	**98.67(2.8)**	**100(2.4)**	**100(2)**	**100(2)**	**71.25(5.2)**	**82.50(6.6)**	**85.18(6.4)**	**86.96(6)**	**32.59(9.8)**	**25.96(10.8)**	**31.85(10.8)**	**34.07(13)**
	Full	100				24.11				38.89			
NBC	30	93.33(2)	100(2)	100(2)	100(2)	24.11(2)	72.32(7)	92.86(6)	73.21(6)	42.59(6)	50.00(14)	51.85(13)	64.81(5)
	50	93.33(2)	100(2)	100(2)	100(2)	38.39(4)	82.14(6)	97.32(6)	83.04(6)	48.15(14)	53.70(12)	55.56(13)	70.37(13)
	100	100(2)	100(1)	100(2)	100(2)	70.54(5)	72.32(5)	91.96(6)	84.82(6)	46.30(4)	55.56(9)	66.67(13)	64.81(13)
	500	93.33(3)	100(2)	100(2)	100(2)	96.43(6)	94.64(7)	92.86(6)	89.29(6)	57.41(10)	66.67(14)	66.67(12)	68.52(11)
	1000	93.33(3)	100(2)	100(2)	100(2)	94.64(6)	94.64(7)	94.64(7)	96.43(7)	61.11(14)	61.11(14)	68.52(10)	68.52(10)
	Mean	**94.67(2.4)**	**100(1.8)**	**100(2)**	**100(2)**	**64.82(4.6)**	**83.21(6.4)**	**93.93(6.2)**	**85.36(6.2)**	**51.11(9.6)**	**57.41(12.6)**	**61.85(12.2)**	**67.41(10.4)**
	Full	86.67				0.00				59.26			
KNN (K=10)	30	93.33(2)	100(2)	100(2)	100(2)	57.14(3)	92.86(7)	94.64(6)	95.54(6)	46.30(6)	46.30(2)	46.30(9)	64.81(9)
	50	100(3)	100(2)	100(2)	100(2)	70.54(2)	92.86(6)	95.54(6)	96.43(6)	48.15(6)	53.70(10)	48.15(11)	64.81(12)
	100	100(2)	100(1)	100(2)	100(2)	86.61(4)	96.43(6)	98.21(6)	96.43(6)	50.00(7)	53.70(11)	66.67(13)	61.11(9)
	500	100(3)	100(2)	100(2)	100(2)	99.11(6)	97.32(7)	92.86(7)	97.32(7)	42.59(12)	62.96(9)	57.41(13)	62.96(14)
	1000	100(3)	100(2)	100(2)	100(2)	98.21(7)	97.32(7)	94.64(6)	94.64(5)	48.15(13)	62.96(13)	64.81(10)	66.67(13)
	Mean	**98.67(2.6)**	**100(1.8)**	**100(2)**	**100(2)**	**82.32(4.4)**	**95.36(6.6)**	**95.18(6.2)**	**96.07(6)**	**47.04(8.8)**	**55.93(9)**	**56.67(11.2)**	**64.07(11.4)**
	Full	93.33				5.36				53.70			
1NN	30	93.33(3)	100(2)	100(1)	100(2)	50.89(2)	86.61(6)	90.18(6)	90.18(7)	37.04(3)	50.00(11)	51.85(7)	64.81(6)
	50	100(3)	100(2)	100(2)	100(2)	70.54(2)	90.18(6)	96.43(6)	90.18(6)	48.15(13)	51.85(12)	53.70(8)	70.37(12)
	100	100(2)	100(1)	100(1)	100(2)	83.93(4)	95.54(6)	97.32(7)	94.64(7)	50.00(11)	61.11(9)	68.52(13)	66.67(12)
	500	100(3)	100(2)	100(2)	100(2)	98.21(6)	95.54(6)	91.07(7)	95.54(7)	48.15(10)	62.96(11)	64.81(13)	72.22(14)
	1000	100(3)	100(2)	100(2)	100(2)	93.75(6)	96.43(7)	93.75(7)	95.54(7)	48.15(10)	68.52(14)	70.37(14)	72.22(13)
	Mean	**98.67(2.8)**	**100(1.8)**	**100(1.6)**	**100(2)**	**79.46(4)**	**92.86(6.2)**	**93.75(6.6)**	**93.21(6.8)**	**46.30(9.4)**	**58.89(11.4)**	**61.85(11)**	**69.26(11.4)**
	Full	93.33				5.36				61.11			

Full means the original full dimensional space with no feature selection. The number in parenthesis is the number of components at which the maximal classification accuracy is attained, and the number of factors in feature extraction is inferior or equal to *k*. The underlined results are the best over all four dimension reduction methods.

PLSDR3 (MSVMRFE) are multifeature selection methods based on the RFE strategy, which can significantly reduce redundancy and relevance. In other words, multifeature selection methods can identify small effect features with strong interaction effects, and RFE strategies significantly reduce redundancy.

2) Experiment 2 (Comparative Analysis on Multiple Classifiers): The proposed dimension reduction algorithm is independent of a classifier and can be used in conjunction with any classifier. Here, we combine different dimension reduction methods with different classifiers and use the recognition accuracy of classifiers to evaluate dimension reduction methods. In the evaluation process, we only focus on the impact of the identification results for different dimension reduction methods. We can adjust the parameters of the classifier (complex classifiers) to obtain higher recognition accuracy. The five different classifiers used are: 1) Fisher Linear Discriminant Analysis (LDA), 2) Extended multiclass Support Vector Machine (MSVM linear kernel), 3) Naive Bayesian Classifier (NBC), 4) K-Nearest Neighbor Classifier ($K = 10$), 5) Nearest Neighbor Classifier (1NN).

The recognition results on the four different dimension reduction methods in the five different classifiers listed above are presented in Table II. We can see from Table II that dimension reduction methods can always improve the recognition rate of classifiers. Compared with other dimension reduction methods, our proposed algorithm is able to obtain higher recognition accuracy. The performance of the SVM classifier on the GCM dataset is not ideal because the GCM dataset is a multicategory dataset (up to 14 categories) and the learning ability of the expanded multicategory SVM classifier becomes poor as the number of categories increases (consistent with the conclusion in Section I).

3) Experiment 3 (Analysis of the Results Based on k-fold Cross-Validation): To avoid any "selection bias," stratified k-fold cross-validation is adopted in the following experiments on eight datasets. Because the number of minority class samples is less than ten in three biological datasets (Brain, Lung, and NCI60), fivefold cross-validation is adopted. For all other datasets, tenfold cross-validation is used. All samples of the datasets are randomly divided into k subsets of which k–1 of these subsets is treated as a training set and the other subset is treated as a testing set. We repeat this process k times. Thus, each subset will be used as a test set. The recognition rate with k-fold CV is recorded. All the dimension reduction methods

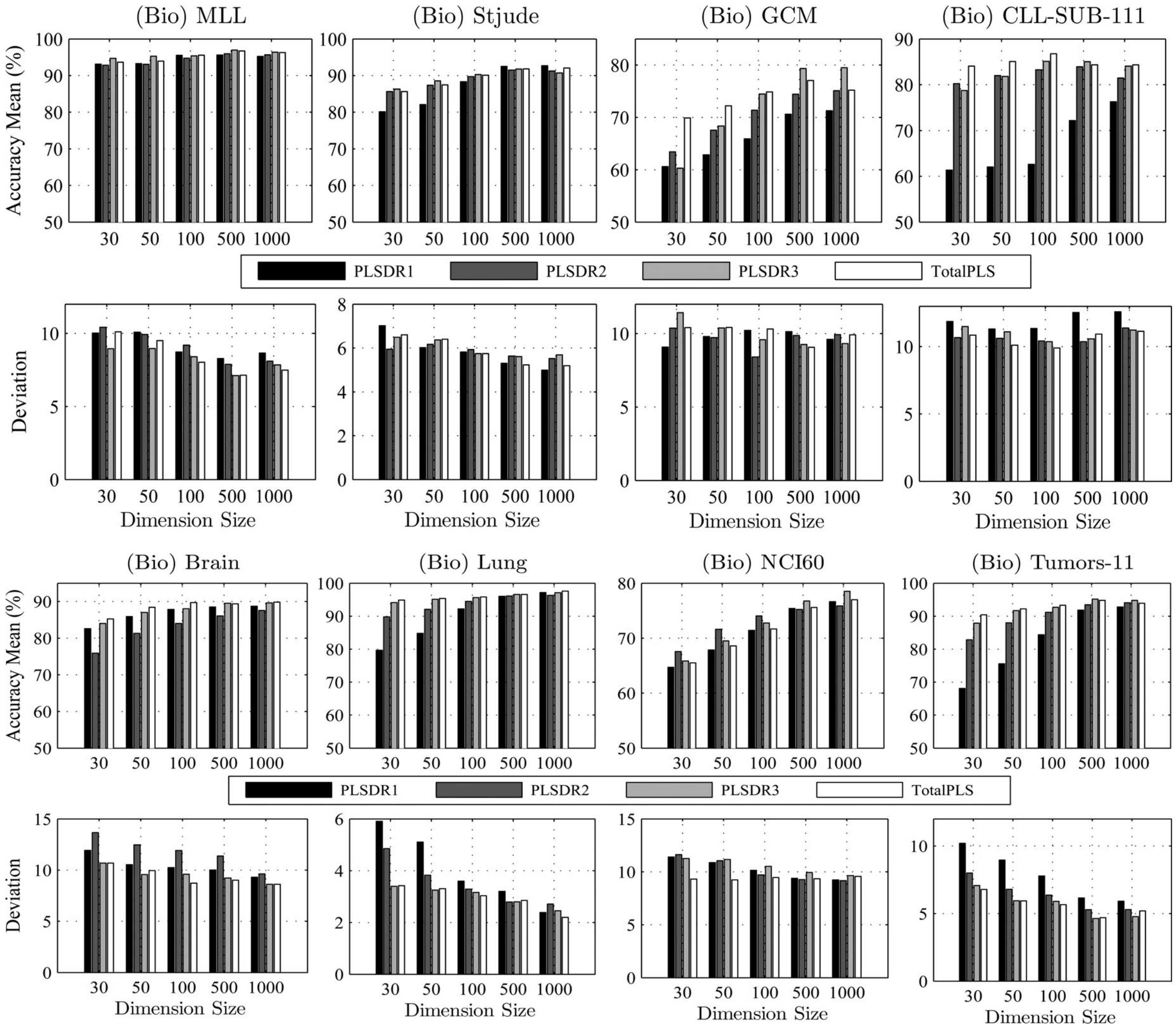

Fig. 5. Top: Recognition rate mean and Bottom: standard deviation based on different dimension reduction methods (PLSDR1, PLSDR2, PLSDR3, and TotalPLS) in eight biological datasets where the number of factors in feature extraction is fixed at k. (40 random results of fivefold CV for Brain, Lung, and NCI60, 20 random results of tenfold CV for all other datasets, Classifier: LDA).

are implemented on the same training set and testing set, which are randomly generated by k-fold CV. We only select the LDA classifier for comparison; however, the same logic applies to the other classifiers. Fig. 5 shows the average recognition rate (accuracy mean in%) and standard error (deviation) with 20 random results of tenfold CV (40 random results of fivefold CV) based on the four different dimension reduction methods.

Fig. 5 shows that the recognition accuracy of PLSDR2 is higher than the recognition accuracy of PLSDR1. For most of the datasets (except for *NCI60*), TotalPLS and PLSDR3 yielded outstanding performances. Their recognition accuracy rates are higher than both PLSDR1 and PLSDR2. In terms of the standard deviation, the results from TotalPLS and PLSDR3 have smaller standard error (STD) except for the GCM dataset (only when # Dim is equal to 30 and 50). This indicates that the TotalPLS method depends less on the training sets and is more robust. In summary, our proposed algorithm can improve the prediction accuracy and is more stable.

4) Experiment 4 (Evaluation of Computational Efficiency): The computing speed is another important indicator used to evaluate algorithms. The time complexity of the algorithm is the main factor affecting the computing speed, which is especially important for high or ultrahigh-dimensional data. From Table III, we can observe that TotalPLS has very good computing speed. Other methods (such as PLSDR3) also produce good results on some datasets, but they are still more time-consuming than TotalPLS. In fact there are many state-of-the-art algorithms, such as minimum-redundancy maximum-relevancy, random forest, and LASSO. However, they perform poorly with high-dimensional and ultrahigh-dimensional multicategory

TABLE III
CPU TIME PER RUN (IN SECONDS) OF THE FOUR DIMENSION REDUCTION METHODS

Data Set	PLSDR1	PLSDR2	PLSDR3	TotalPLS	Task
(Bio)MLL	7.64	9.73	3.12	**0.33**	HOCV (TABLE II)
(Bio)Stjude	8.92	35.82	66.33	**1.19**	HOCV (TABLE II)
(Bio)GCM	13.91	34.21	44.46	**2.36**	HOCV (TABLE II)
(Bio)CLL-SUB-111	7.25	16.24	9.47	**0.32**	10-fold CV (Fig. 5)
(Bio)Brain	18.70	12.81	6.62	**0.91**	5-fold CV (Fig. 5)
(Bio)Lung	87.13	159.13	63.24	**3.31**	5-fold CV (Fig. 5)
(Bio)NCI60	19.32	18.10	13.61	**1.72**	5-fold CV (Fig. 5)
(Bio)Tumors-11	101.52	303.44	348.25	**14.16**	10-fold CV (Fig. 5)

TABLE IV
RESULTS OF NONPARAMETRIC STATISTICAL TESTS PERFORMED IN EXPERIMENT 3

DataSet	TotalPLS vs. PLSDR1	TotalPLS vs. PLSDR2	TotalPLS vs. PLSDR3
	Paired signed rank	Paired signed rank	Paired signed rank
(Bio)MLL	(+) 0.0097	(=) 0.0613	(=) 0.0555
(Bio)Stjude	(+) 1.92×10^{-25}	(=) 0.0492	(=) 0.4927
(Bio)GCM	(+) 1.91×10^{-81}	(+) 2.60×10^{-25}	(+) 2.06×10^{-4}
(Bio)CLL-SUB-111	(+) 1.53×10^{-133}	(+) 1.01×10^{-18}	(+) 2.48×10^{-11}
(Bio)Brain	(+) 2.49×10^{-8}	(+) 5.13×10^{-39}	(=) 0.0101
(Bio)Lung	(+) 6.35×10^{-106}	(+) 1.96×10^{-63}	(+) 7.26×10^{-7}
(Bio)NCI60	(=) 0.0620	(-) 2.33×10^{-4}	(-) 5.16×10^{-4}
(Bio)Tumors-11	(+) 5.40×10^{-111}	(+) 3.74×10^{-34}	(+) 0.0023
Summary	**7(+) 1(=) 0(-)**	**5(+) 2(=) 1(-)**	**4(+) 3(=) 1(-)**

Here, (+) implies that the first algorithm is statistically better than the confronting one, (–) implies the contrary, and (=) means that the two algorithms have no significant differences between them. The *p*-values are given.

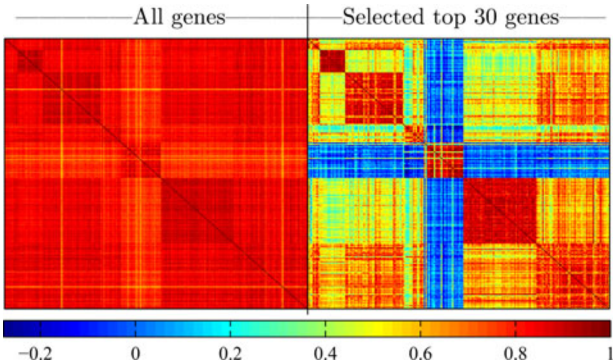

Fig. 6. Heat map of correlation coefficient matrix before and after TotalPLS-based dimension reduction (30 specific genes) in the sample set (Bio Stjude).

microarray data [62]. In addition, they are computationally very time consuming and cannot be applied to practical applications.

5) Experiment 5 (Statistical Tests on the Differences Between Algorithms): We use the previous 20 random results of tenfold CV (40 random results of fivefold CV) to do paired testing to compare the prediction accuracy by different algorithms (see Table IV). We use nonparametric statistical significance tests. As the comparison is implemented on the same training set and testing set, which is randomly generated by tenfold CV, any difference is caused by the difference of algorithms. There is no difference in the random composition of the sample set.

Table IV shows the comparison of our results. For TotalPLS and PLSDR1, TotalPLS is significantly better than PLSDR1 on all datasets except on the *NCI60* dataset. For TotalPLS and PLSDR2, TotalPLS is significantly better than the algorithm PLSDR2 on five datasets, the two compared algorithms have no significant differences on the two datasets; on the *NCI60* dataset, PLSDR2 is significantly better than the TotalPLS. For TotalPLS and PLSDR3, TotalPLS is significantly better than PLSDR3 on four datasets, on the *NCI60* dataset, PLSDR3 is significantly better than the TotalPLS. The two compared algorithms have no significant differences on the remaining three datasets using signed rank test. This result may be because both algorithms are multifeature selection based on an RFE strategy, and they show similar performance. In short, compared with the state-of-the-art algorithms, our proposed TotalPLS is highly effective in high-dimensional multicategory problem.

V. INTELLIGIBILITY AND INTERPRETABILITY OF THE POTENTIAL STRUCTURE

An important objective of machine learning and data mining is to effectively access the useful knowledge that we need. Therefore, the intelligibility and interpretability of the mining results is another important indicator to evaluate learning algorithms. Intelligibility and interpretability measure how much the learning outcomes are as close to our understanding as possible.

Algorithm evaluation also depends on whether the results have a biological interpretation. First, we screened a small number of specific genes (biomarkers) that are known to be closely related to the emergence and development of tumors and we further searched the phenomena of overexpression and underexpression for the related genes on tumor subtypes. Next, supervised feature extraction is used to construct the expression patterns for different tumor subtypes on the specifically selected genes to analyze how the related specific gene coregulates the expression of the tumor subtypes. These genes are likely to provide a new perspective that can be used to explore the mechanism of tumor genesis to find molecular targets for tumor therapy, provide information on the prediction of tumors, as well as provide insight on reliable diagnosis and treatment options.

Our study of the Stjude dataset was aimed at classifying subtypes of pediatric acute lymphoblastic leukemia. The dataset was divided into seven groups. Six of the groups were diagnostic groups (BCR-ABL, E2 A-PBX1, Hyperdiploid $>$ 50, MLL, T-ALL, and TEL-AML1), and the other group contained diagnostic samples that did not fit into any of the groups mentioned earlier (we labeled this as "Others"). There are 12 558 genes in total. According to the original publication (http://www.stjuderesearch.org), each group of samples has been randomized into training and testing parts.

Fig. 6 shows the heat map of the correlation coefficient matrix before and after dimension reduction. The matrix on the left in Fig. 6 is the heat map of the correlation coefficient matrix on all 12 558 genes and the matrix on the right in Fig. 6 is the heat map of the correlation coefficient matrix of the top 30 specific genes selected by TotalPLS. The comparison shows that the sample set has a distinct category-outline after TotalPLS based dimension reduction was used. Our proposed TotalPLS dimension reduction can remove a large number of redundant genes. Therefore, it enhances the similarity of within-category samples on the selected genes (i.e., the between-category information is clearer). It further approves the effectiveness of our methodology on removing redundant genes. It also shows that our feature selection method is very effective.

Fig. 7 shows the loading of the top 30 selected specific genes on the six latent factors (synthesis features) and the score of the whole training sample on the top six loading vectors. The weight of the loading vector denotes the relative impact on the latent variable to the corresponding predictor variables (30

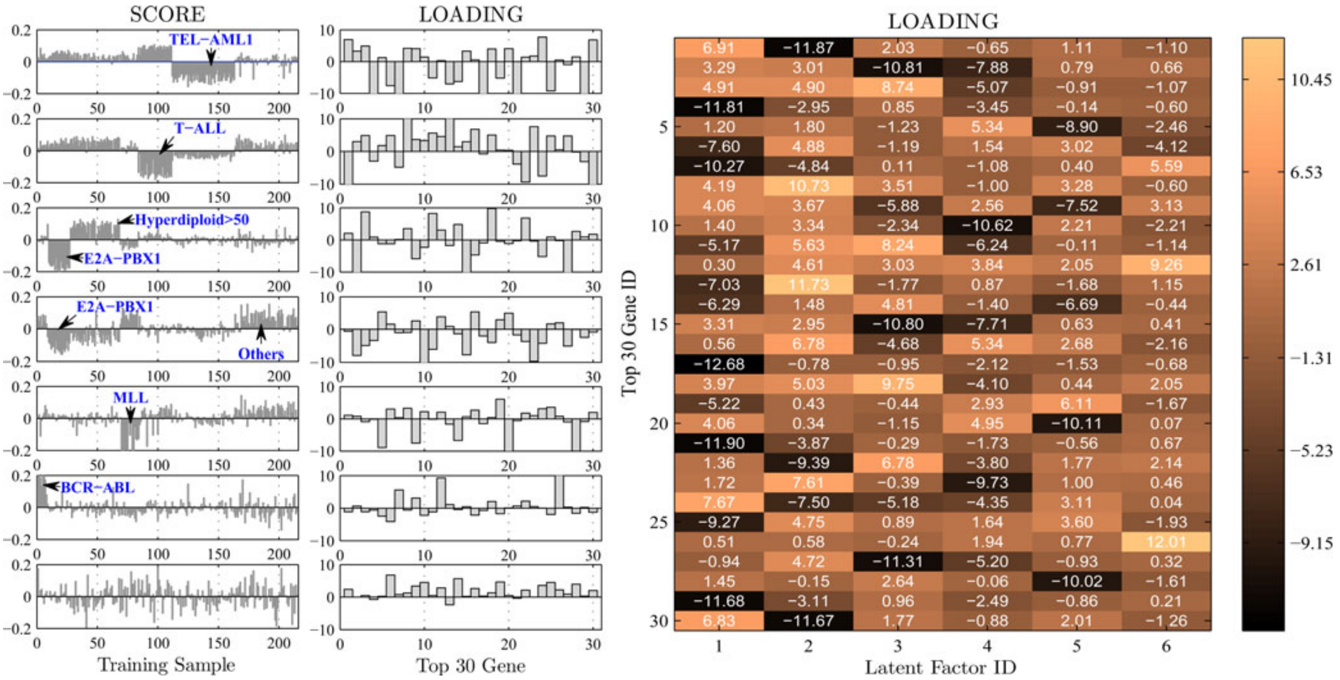

Fig. 7. Scores on the six synthesis features and loadings of top 30 selected genes on the six synthesis features for the (Bio) Stjude dataset.

TABLE V

DESCRIPTION OF THE TOP 30 COEXPRESSED GENES BASED ON TOTALPLS (HOLDOUT)

Rank	Reference number	Affymetrix number	Gene symbol	Gene name (gene description)	PLS loading over(+),under(-)	
17	X17025	36985_at	IDI1	isopentenyl-diphosphate delta isomerase	1(-)	
21	U69883	38203_at	KCNN1	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 1	1(-)	
4	AB012124	35614_at	TCFL5	transcription factor-like 5 (basic helix-loop-helix)	1(-)	**TEL-AML1**
29	AF070644	38652_at	FLJ20154	hypothetical protein FLJ20154	1(-)	
7	M63928	38578_at	TNFRSF7	tumor necrosis factor receptor superfamily member 7	1(-)	
25	AB020674	35260_at	MONDOA	KIAA0867 protein	1(-)	
1	AA919102	38319_at	CD3D	CD3D□antigen delta polypeptide TiT3 complex	2(-)	
30	X76220	38051_at	MAL	mal, T-cell differentiation protein	2(-)	**T-ALL**
24	AL034374	33821_at	HELO1	Human DNA sequence from clone RP3-483K16 on chromosome 6p12.1-21.1	2(-)	
13	M13560	35016_at		Human Ia-associated invariant gammachain gene, exon 8, clones lambda-y (1,2,3).	2(+)	
8	U90878	36937_s_at	PDLIM1	PDZ and LIM domain 1 elfin	2(+)	**T-ALL**
16		36945_at	ERP29	endoplasmic reticulum protein 29	2(+)	
27	J03473	41146_at	ADPRT	ADP-ribosyltransferase NAD poly ADPribose polymerase	3(-)	
2	AL049381	33355_at	PBX1	Homo sapiens cDNA FLJ12900 fis clone NT2RP2004321 (by CELERA search of target sequence = PBX1)	3(-)	**E2A-PBX1**
15	M86546	32063_at	PBX1	pre-B-cell leukemia transcription factor 1	3(-)	
18	AB005047	38968_at	SH3BP5	SH3-domain binding protein 5 BTKassociated	3(+)	
3	X02317	36620_at	SOD1	superoxide dismutase 1 soluble amyotrophic lateral sclerosis 1 adult	3(+)	
11	Y18004	38518_at	SCML2	sex comb on midleg Drosophila like 2	3(+)	**Hyperdiploid >50**
22	L10373	38408_at	TM4SF2	transmembrane 4 superfamily member 2	3(+)	
14	AL049940	37732_at	RYBP	RING1 and YY1 binding protein	3(+)	
10	W26633	41139_at	MAGED1	melanoma antigen family D 1	4(-)	**E2A-PBX1**
23	X04500	1520_s_at	IL1B	interleukin 1 beta	4(-)	
20	AI535946	33412_at	LGALS1	LGALS1 Lectin, galactoside-binding, soluble, 1 (galectin 1)	5(-)	
28	AF027208	41470_at	PROML1	prominin mouse like 1	5(-)	
5	AJ001687	36777_at	D12S2489E	DNA segment on chromosome 12 unique 2489 expressed sequence	5(-)	**MLL**
9	L19182	2062_at	IGFBP7	insulin-like growth factor binding protein 7	5(-)	
19	L12711	38789_at	TKT	transketolase Wernicke-Korsakoff syndrome	5(+)	
26	D88153	40196_at	HYA22	HYA22 protein	6(+)	
12	M55531	34362_at	SLC2A5	solute carrier family 2 facilitated glucose transporter member 5	6(+)	**BCR-ABL**
6	L11373	1373_at	PCDHGC3	protocadherin gamma subfamily C 3	6(-)	

The sixth column indicates the expression information on PLS loading of the selected genes: number, corresponding component factors and the tumor subtypes (reference to the original literature), symbol, and corresponding expression level (+: overexpression, -: underexpression).

specific genes). The sign reflects positive or negative correlation between the specific gene and the latent factors, whereas the size of the loading reflects the degree of the specific gene associated with the latent factors. Therefore, long bars on the figure (Loading subgraph) indicate that the loading weight is high and the correlation with the predicted response is large. Its corresponding visualization heat map is provided in Fig. 7 (right subgraph). In addition, Fig. 7 (left subgraph) also shows the corresponding category that each latent factor expresses from the scores. For example, the first row of the left subgraph in Fig. 7 shows that the first latent factor that expresses the information of leukemia subtype TEL-AML1, the corresponding distribution of loading is shown in the middle subgraph, and the visualization heat map is shown in the right subgraph. In detail, the absolute weights of the genes whose rank IDs are 4, 7, 17, 21, 25, and 29 are relatively large. Thus, we can trust that those genes coexpress the leukemia subtype TEL-AML1. Similarly, we can continue to obtain the relationship between other latent factors and their corresponding genes. Table V shows a simple corresponding relationship between latent factors and coexpressed genes.

In detail, the first row in the right subgraph of Fig. 7 shows that CD3D gene (rank id 1) has a specific expression in the Latent Factor 2. As the second row in the left subgraph of Fig. 7 indicates that the Latent Factor 2 mainly expresses the leukemia subtype T-ALL, the CD3D gene has a specific expression in the leukemia subtype T-ALL. The gene selection methods based on the chi-square and t-statistic in the literature [62] also screened this gene in leukemia subtype T-ALL. The PBX1 gene (rank id 2) has a specific expression phenomenon in the leukemia subtype E2 A-PBX1. The gene selection methods based on the Wilkins' and t-statistic in the literature [62] also screened this gene in leukemia subtype E2 A-PBX1. The SOD1 gene (rank id 3) has a specific expression phenomenon in the leukemia subtype Hyperdiploid $>$ 50 and the gene selection methods based on the chi-square, CFS, and t-statistic in the literature [62] screened this gene in leukemia subtype Hyperdiploid $>$ 50 as well. Similarly, other identification results are also surprisingly similar to those in the original literature [62], which applied five different gene selection methods to select specific genes in different leukemia subtypes. Therefore, the specific genes selected by our algorithm are credible to some extent.

After the resulting latent structure is further analyzed, the coexpression gene of the different leukemia subtypes can be found from the longitudinal direction of the right subgraph in Fig. 7. Moreover, we can also observe the phenomenon that some specific genes simultaneously express different subtypes from the transverse direction. For example, the TNFRSF7 gene (rank id 7) has specific expression in Latent Factors 1 and 6. Thus, the TNFRSF7 gene has specific expression phenomenon in leukemia subtypes TEL-AML1 and BCR-ABL. Our proposed algorithm can capture those specific genes that have relatively small main effect, but have strong interaction effect. For example, the ERP29 gene (rank id 16) is not captured in the original literature [62], which applied five different gene selection methods. However, our result in Table V shows that this gene may be involved in the expression of the leukemia subtype T-ALL, which was confirmed by the literature [63], [64].

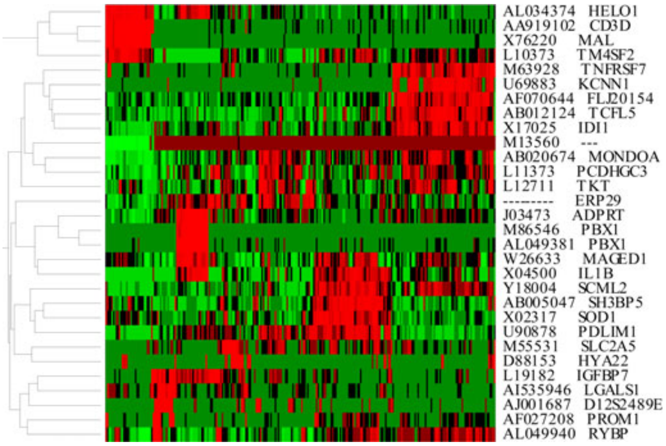

Fig. 8. Hierarchical clustering and heat map on top 30 specific genes. (Bio Stjude).

Furthermore, the hierarchical clustering and heat map of the top 30 identified specific genes are very consistent with the clustering results and the conclusions in Table V (see Fig. 8). All of these show that the proposed algorithm has a better biological explanation for recognition results. It can be effectively applied to microarray data analysis for the interpretation of gene coexpression and coregulation. By this, we mean that our algorithm has good performance on recognition accuracy. Furthermore the results of data mining have better understandability and visual representation.

VI. Conclusion

High-dimensional data bring great challenges in terms of computational complexity and classification performance. Therefore, it is necessary to effectively compress a high-dimensional feature space into a low-dimensional feature space to design a learner with superior performance. Feature extraction has a stronger ability to extract structure information in variables. Feature selection preserves the original features so that the obtained feature subset has better explanatory ability. Therefore, dimension reduction is essential to study and understand the mechanism of practical problems.

Having reasonable and effective access to useful knowledge that we need is an important goal of data mining. The interpretability, understandability, and visualization of mining results are important indicators to evaluate data mining algorithms. This paper introduced the PLS-based feature selection algorithm (PLSRFE) for multicategory classification, subsequently, a new local dimension reduction algorithm called TotalPLS is given that implements an information fusion of PLS-based feature selection and feature extraction in a unified PLS framework. The experimental results prove that the proposed algorithm improves not only the recognition accuracy, but also the interpretability and visualization to mine the potential structure for high-dimension multicategory data. Furthermore, the proposed method can be effectively applied to microarray data analysis to interpret gene coexpression and coregulation.

In general, dimension reduction methods can be classified into linear and nonlinear methods. We have analyzed the nonlinear feature extraction based on kernel PLS. Usually, it is able to improve the recognition rate through selecting the appropriate kernel function and its parameters. However, the resulting model and the results cannot be easily interpreted. Nonlinear methods will be further investigated in our future research. In addition, motivated by the idea of Reshef *et al.* [65], we plan to introduce maximal information-based nonparameteric exploration into this study to discover the nonlinear relationship among specific genes in tumor microarray data. How to introduce nonlinear method to mine the potential structural information hidden in high-dimensional multicategory microarray data to better interpret and understand coexpression and coregulation relationships is one of our future research directions.

Acknowledgment

The authors would like to thank Prof. E. Bass and three anonymous reviewers for their constructive comments that improved the quality and the readability of this paper. The authors would also like to thank their laboratory team members' assistance.

References

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–451, 2004.

[2] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic, 1990.

[4] A. K. Jain, R. P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[5] K. G. Anil, P. Chaudhuri, and C. A. Murthy, "Multiscale classification using nearest neighbor density estimates," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 5, pp. 1139–1148, Oct. 2006.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.

[7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[8] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[9] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000.

[10] X. Zhou and D. P. Tuck, "MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[11] E. J. Bredensteiner and K. P. Bennet, "Multicategory classification by support vector machines," *Comput. Optim. Appl.*, vol. 12, pp. 53–79, 1999.

[12] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 67–81, 2004.

[13] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognit.*, vol. 44, no. 8, pp. 1761–1776, 2011.

[14] I. Gheyas and L. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognit.*, vol. 43, no. 1, pp. 5–13, 2010.

[15] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, and Z. Chen, "Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 3, pp. 320–333, Mar. 2006.

[16] K. A. Lê Cao, A. Bonnet, and S. Gada, "Multiclass classification and gene selection with a stochastic algorithm," *Comput. Statist. Data Anal.*, vol. 53, pp. 3601–3615, 2009.

[17] J. Hua, W. D. Tembe, and E. R. Doughertya, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognit.*, vol. 42, no. 3, pp. 409–424, 2009.

[18] J. J. Dai, L. Lieu, and D. Rocke, "Dimension reduction for classification with gene expression microarray data," *Statist. Appl. Genet. Mol. Biol.*, vol. 5, no. 1, pp. 1–19, 2006.

[19] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[20] G. Z. Li and X. Q. Zeng, "Feature selection for partial least square based dimension reduction," *Stud. Comput. Intell.*, vol. 205, pp. 3–37 , 2009.

[21] L. Deng, J. W. Ma, and J. Pei, "Rank sum method for related gene selection and its application to tumor diagnosis," *Chin. Sci. Bull.*, vol. 49, no. 15, pp. 1652–1657, 2004.

[22] L. J. Wei, "Asymptotic conservativeness and efficiency of Kruskal-Wallis test for K dependent samples," *J. Amer. Statist. Assoc.*, vol. 76, no. 378, pp. 1006–1009, Dec. 1981.

[23] B. F. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," *IEEE Trans. Syst., Man, Cybern. A*, vol. 39, no. 1, pp. 36–46, Jan. 2009.

[24] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, Max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[25] H. Yan, X. T. Yuan, S. C. Yan, and J. Y. Yang, "Correntropy based feature selection using binary projection," *Pattern Recognit.*, vol. 44, no. 12, pp. 2834–2842, Dec. 2011.

[26] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 10th Nat. Conf. Artif. Intell.*, 1992, pp. 129–134.

[27] I. Kononenko, "Estimation attributes: Analysis and extensions of RELIEF," in *Proc. Eur. Conf. Mach. Learning*, 1994, pp. 171–182.

[28] C. H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, pp. 37–44, 2003.

[29] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Inf. Sci.*, vol. 179, no. 13, pp. 2208–2217, Jun. 2009.

[30] K. Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 60–67, Feb. 2004.

[31] G. Ji, Z. Yang, and W. You, "PLS-based gene selection and identification of tumor-specific genes," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 830–841, Nov. 2011.

[32] Y. Lu and J. W. Han, "Cancer classification using gene expression data," *Inf. Syst.*, vol. 28, no. 4, pp. 243–268, 2003.

[33] K. Duan, C. Jagath, Rajapakse, and M. N. Nguyen, "One-Versus-One and One-Versus-All multiclass SVM-RFE for Gene Selection in Cancer Classification," *Evol. Comput., Mach. Learning Data Mining Bioinformatics*, vol. 4447, pp. 47–56, 2007.

[34] G. Qu, S. Hariri, and M. Yousif, "A new dependency and correlation analysis for features," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 9, pp. 1199–1207, Sep. 2005.

[35] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[36] A. Al-Ani, M. Deriche, and J. Chebil, "A new mutual information based measure for feature selection," *Intell. Data Anal.*, vol. 7, no. 1, pp. 43–57, 2003.

[37] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognit.*, vol. 42, no. 7, pp. 1330–1339, 2009.

[38] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2001.

[39] "Dimension reduction," University College Dublin, Dublin, Ireland, Tech. Rep. UCD-CSI-2007–7, 2007.P. Cunningham

[40] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.

[41] W. H. Yang, D. Q. Dai, and H. Yan, “Feature extraction and uncorrelated discriminant analysis for high-dimensional data,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 601–614, May 2008.

[42] C. G. Rafael and E. W. Richard, *Digital Image Processing*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice Hall, 2002.

[43] H. Li, T. Jiang, and K. Zhang, “Efficient and robust feature extraction by maximum margin criterion,” *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.

[44] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY, USA: Springer-Verlag, 2001.

[45] F. Li and Y. M. Yang, “Analysis of recursive gene selection approaches from microarray data,” *Bioinformatics*, vol. 21, no. 19, pp. 3741–3747, 2005.

[46] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[47] H. Zou and T. Hastie, “Regularization and variable selection via the Elastic Net,” *J. Roy. Statist. Soc. B*, vol. 67, pp. 301–320, 2005.

[48] C. R. Rao and H. Toutenburg, *Linear Models: Least Squares and Alternatives*. New York, NY, USA: Springer-Verlag, 2001.

[49] D. V. Nguyen and D. M. Rocke, “Multi-class cancer classification via partial least squares with gene expression profiles,” *Bioinformatics*, vol. 18, no. 9, pp. 1216–1226, 2002.

[50] A. L. Boulesteix, “PLS dimension reduction for classification with microarray data,” *Stat. Appl. Genet. Mol. Biol.*, vol. 3, no. 1, pp. 1–30, 2004.

[51] A. L. Boulesteix and K. Strimmer, “Partial least squares: A versatile tool for the analysis of high-dimensional genomic data,” *Briefings Bioinformatics*, vol. 8, no. 1, pp. 32–44, 2006.

[52] Z. Yang, W. You, and G. Ji, “Using partial least squares and support vector machines for bankruptcy prediction,” *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8336–8342, 2011.

[53] M. Gutkin, G. Dror, and R. Shamir, “SlimPLS: A method for feature selection in gene expression-based disease classification,” *PLoS One*, vol. 4, no. 7, pp. 1–12, 2009.

[54] K.-A. Le Cao, S. Boitard, and P. Besse, “Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems,” *BMC Bioinformatics*, vol. 12, no. 253, pp. 1–16, 2011.

[55] S. Chakraborty and S. Datta, “Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies,” *Bioinformatics*, vol. 28, no. 6, pp. 799–806, 2012.

[56] Q. Zhao, C. F. Caiafa, D. P. Mandic, Z. C. Chao, Y. Nagasaka, N. Fujii, L. Zhang, and A. Cichocki, “Higher-order partial least squares (HOPLS): A generalized multilinear regression method,” *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 35, no. 7, pp. 1660–1673, Jul. 2013.

[57] C. Domeniconi, J. Peng, and D. Gunopulos, “Locally adaptive metric nearest neighbor classification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1281–1289, Sep. 2002.

[58] M. Barker and W. Rayens, “Partial least squares for discrimination,” *J. Chemometrics*, vol. 17, pp. 166–173, 2003.

[59] Y. S. Liu and W. Rayens, “PLS and dimension reduction for classification,” *Comput. Statist.*, vol. 22, pp. 189–208, 2007.

[60] S. de Jong, “SIMPLS: An alternative approach to partial least squares regression,” *Chemometrics Intell. Lab. Syst.*, vol. 18, pp. 251–263, 1993.

[61] S. Wold, W. Johansson, and M. Cocchi, “PLS - partial least-squares projections to latent structures,” in *3D QSAR in Drug Design, Theory Methods and Applications*. New York, NY, USA: Springer-Verlag, 1993.

[62] Y. Eng-Juh, E. R. Mary, A. S. Sheila, W. K. Williams, P. Divyen, M. Rami, G. B. Fred, C. R. Susana, V. R. Mary, P. Anami, C. Cheng, C. Dario, W. Dawn, X. D. Zhou, J. Y. Li, H. Q. Liu, C. H. Pui, E. E. William, N. Clayton, W. Limsoon, and R. D. James, “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling,” *Cancer Cell*, vol. 1, pp. 133–143, 2002.

[63] A. K. Anagnostopoulos, K. Vougas, A. Kolialexi, A. Mavrou, M. Fountoulakis, and G. T. Tsangaris, “The protein profile of the human immature T-cell line CCRF-CEM,” *Cancer Genomics Proteomics*, vol. 2, pp. 271–300, 2005.

[64] S. Geley, B. L. Hartmann, R. Hattmannstorfer, M. Loffler, M. J. Ausserlechner, D. Bernhard, R. Sgonc, E. M. Strasser-Wozak, M. Ebner, B. Auer, and R. Kofler, “p53-induce (2005),d apoptosis in the human T-ALL cell line CCRF-CEM,” *Oncogene*, vol. 20, pp. 2429–2437, 1997.

[65] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, “Detecting novel associations in large data sets,” *Science*, vol. 334, pp. 1518–1524, 2011.

Wenjie You received the B.Sc. degree in mathematics from Zhangzhou Normal University, Zhangzhou, China, in 1997 and the M.Eng. degree in control engineering from Xiamen University, Fujian, China, in 2009, where he is currently working toward the Ph.D. degree in systems engineering.

He is currently a Senior Member with China Computer Federation. His current research interests include statistical computing, data mining, and machine learning.

Zijiang Yang received the M.A.Sc. and Ph.D. degrees in industrial engineering from the University of Toronto, Toronto, ON, Canada, in 1999 and 2002, respectively.

She is currently an Associate Professor with the School of Information Technology, York University, Toronto, ON, Canada. Her current research interests include prediction, classification, performance analysis in the financial service industry, and data mining algorithms. She has published papers in IEEE Transactions on Neural Networks, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, IEEE Transactions on Engineering Management, *Communications in Nonlinear Science and Numerical Simulations*, *Knowledge-Based Systems*, *Expert System with Applications*, *Computers and Operations Research*, *Journal of the OR Society* (JORS), *Applied Mathematics and Computation*, *Mathematical and Computer Modeling*, *Annals of Operations Research*, *Engineering Computations*, *Journal of Applied mathematics*, and other peer-reviewed journals.

Mingshun Yuan received the B.Sc. degree in mathematics and application mathematics, in 2002 and the M.A.Sc. degree in computational mathematics, in 2005, both from the Chengdu University of Technology, Chengdu, China. He is currently working toward the Ph.D. degree in control theory and control engineering from the Xiamen University, Fujian, China.

From 2009 to 2011, he was a Lecturer with the Department of Mathematics and Computer Science, Fujian Normal University, China. His current research interests include evolutionary computation, data mining algorithms, and machine learning.

Guoli Ji received the B.Sc. degree in automation control and the M.A.Sc. degree in system engineering from Xi’an Jiaotong University, Xi’an, China, in 1982 and 1986, respectively.

He is currently a Full Professor with the Department of Automation, Xiamen University, Xiamen, China. He is also the President of Institute of Systems Engineering, and the President of Xiamen Association of Systems Engineering. He has led and participated in many research projects from the Natural Science Foundation of China, Natural Science Foundation of Fujian, and others. Recently, he has published papers in *Bioinformatics*, *Journal of Immunology*, *Journal of Process Control*, *Nucleic Acids Research*, *BMC Biotechnology*, *Genome Research*, *Proceedings of the National Academy of Sciences*, *BMC Bioinformatics*, *BMC Genomics*, *Cell Research*, *PLoS ONE*, *Knowledge-based Systems* and IEEE Transactions on Systems, Man, and Cybernetics, Part C-Applications and Review. His expertise and research areas include bioinformatics, computational biology and systems biology, advanced process control, model predictive control technology and software development, biological databases, data-mining technologies and platform development, modeling and simulation of complex systems, decision theory and decision support systems, management information systems and system integration for enterprises, etc.