Bio-Medical Materials and Engineering 26 (2015) S2011–S2017
DOI 10.3233/BME-151505
IOS Press

Principal component-based feature selection for tumor classification

Lin Sun[a,b,*], Jiucheng Xu[a,b] and Ying Yin[a]
[a]*College of Computer and Information Engineering, Henan Normal University, Xinxiang, China*
[b]*Engineering Technology Research Center for Computing Intelligence and Data Mining, Henan Province, China*

Abstract. One of the important problems in microarray gene expression data is tumor classification. This paper proposes a new feature selection method for tumor classification using gene expression data. In this method, three dimensionality reduction methods, including principal component analysis (PCA), factor analysis (FA) and independent component analysis (ICA), are first introduced to extract and select features for tumor classification, and their corresponding specific steps are given respectively. Then, the superiority of three algorithms is demonstrated by performing experimental comparisons on acute leukemia data sets. It is concluded that PCA compared with FA and ICA is the best under feature load ratio. However, PCA cannot make full use of the category information. To overcome the weak point, Fisher linear discriminant (FLD) is employed as those components of PCA, and a new approach to principal component discriminant analysis (PCDA) is proposed to retain all assets and work better than both PCA and FLD for classification. The further experimental results show that the classification ability of selected feature subsets by means of PCDA is higher than that of the other related dimensionality reduction methods, and the proposed algorithm is efficient and feasible for tumor classification.

Keywords: Feature selection, principal component, discriminant analysis, classification

1. Introduction

Tumor classification is one of the conventional problems in microarray gene expression data and includes tumor detection and prediction of some rare diseases [1]. Its goal is to build an efficient and effective model that can differentiate the gene expressions of samples [2]. In fact, high-dimensional and heterogeneous tumor profiles challenge current machine learning methodologies for its small number of samples and large or even huge number of variables (genes) [3]. In recent years, gene expression profiles based on molecular diagnosis of tumor has attracted many researchers for realizing precise and early tumor diagnosis, however, the curse of dimensionality of tumor dataset seriously challenges the tumor classification [4]. The proposed algorithms for selecting gene subset are categorized into three types: statistical test-based, wrapper-based and transform-based gene selections. Here, the transform-based gene selection, including principal component analysis, independent component analysis, nonnegative matrix factorization, may be mostly used data reduction techniques for popularity and efficiency.

*Address for correspondence: Lin Sun, College of Computer and Information Engineering, Henan Normal University, Xinxiang, China. Tel.: 03733329075; E-mail: slinok@126.com.

Principal component analysis (PCA) is helpful in greatly reducing the high-dimensional microarray data space and identifying outliers while retaining implicit correlations with smaller number of variables. Usually, the results of this statistical procedure are presented in PCA plots representing measured variables in principal components that include as much as possible the variability of input data [5]. Bicciato et al. [6] described a computational procedure for classification of multiclass gene expression data through the application of disjoint principal component models. Whipple et al. [7] used PCA to identify a predictor vector between two mutually exclusive and collectively exhaustive classes. Yasumune et al. [8] developed metabolic distance estimation based on PCA of metabolic turnover. Independent component analysis (ICA) is a useful extension of PCA that has been developed in context with blind separation of independent sources from their linear mixtures [9]. Chen et al. [10] combined gene ranking with ICA to further improve the classification performance based on support vector machine (SVM). Zheng et al. [11] applied the sequential floating forward selection technique and SVM to find the most discriminating ICA features for classification. Factor analysis (FA) can reduce the dimensionality of dataset while retaining as much as possible the variation in datasets, therefore Wang et al. [12] integrated a feature score criterion with FA to further improve the SVM-based classification performance of gene expression data.

One drawback of PCA analysis is, however, that class information is not utilized for class prediction [12]. What's more, a result of PCA is visualized as a 2D or 3D scatter plot and it shows us both relations and degree of relations among the elements spatially [13]. In PCA, all of the observed variance is analyzed, while in FA it is only the shared variances which are analyzed. Compared with PCA, ICA is not always reproducible when it is used to analyze gene expression data [9]. While linear discriminant analysis (LDA) was found to perform the best, in order to utilize the method, the number of genes selected had to be drastically reduced from thousands to tens using a univariate filtering criterion [13]. The discriminant standard of LDA mainly includes distance, Fisher and Bayes. In fact, many of the selection and reduction methods can be combined, and combination of the methods may give us better results [12]. This paper focuses on creating such a solution. In this paper, Fisher linear discriminant (FLD) is applied to those components of PCA, then a new approach that is called principal component discriminant analysis (PCDA) retains all assets of FLD without being burdened by its limitations. The proposed method improves the classification ability and works better than both PCA and LDA. Experiments show that the hybrid method performs well in reducing dimension and improving the performance.

2. Materials and methods

2.1. Specific steps of PCA method

Step 1: Input original data and denote its standardization method by $x_{ij}^{*} = \frac{x_{ij}-\bar{x}_j}{s_j}$, where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$ are the average value and the standard deviation of feature values respectively, x is the feature value of original variable, $i = 1, 2, \cdots, n$, and $j = 1, 2, \cdots, p$.

Step 2: Calculate the covariance matrix of sample data, which is expressed as $M = (\mathrm{s}_{ij})_{p\times p} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)_{p\times p}$, where $i = 1, 2, \cdots, n$, and $j = 1, 2, \cdots, p$.

Step 3: Obtain the principal component of original variables $F_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$, where $i = 1, 2, \cdots, m$, $a_{i1}, a_{i2}, \cdots, a_{ip}$ are the coefficient, and $X_1, X_2, \cdots, X_p$ are the feature values of original variables. Mathematically, all variables of the input data are involved in the linear combinations to compute principal components in PCA [3]. The eigenvalue of covariance matrix to original variable is the variance of main component, so that the front m larger eigenvalues λ_i can represent the first m

larger variance values of principal component well. Then, the eigenvectors corresponding to the front covariance matrix are the coefficient a_i of F_i, where $i = 1, 2, \cdots, m$. Moreover, this selection ensures that the principal component variance increases in turn.

Step 4: If F_1 is used as the principal component indicator and has the most information, its variance in all of the linear combinations of $X_1, X_2, \cdots, X_p$ is largest and should be the first, otherwise F_2 is selected, where F_1 and F_2 remain independent and unrelated to, and $Cov(F_1, F_2) = 0$. Likewise, $F_1, F_2, \cdots, F_m (m \leq p)$ are the new variable indicators, and the variance accumulative contribution ratio calculates m value of $F_1, F_2, \cdots, F_m$ by $G(m) = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{k=1}^{p} \lambda_k}$. When the variance accumulative contribution ratio is more than 85%, it can reflect the information of the original feature.

Step 5: Calculate principal component loads to reflect the correlation degree between F_i and X_j, where the load of X_j with respect to F_i is denoted by $l_{ij} = \sqrt{\lambda_i} a_{ij}, i = 1, 2, \cdots, m$, and $j = 1, 2, \cdots, p$.

To validate the performance of PCA method, the public acute leukemia datasets are firstly downloaded at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, and the original datasets are preprocessed by using Bhattacharyya distance. Then the size of selected gene is 34*34, in which there are 2 kinds of acute lymphoblastic leukemia (T_cell), 18 kinds of acute lymphoblastic leukemia (B_cell), and 14 kinds of acute myeloid leukemia. For datasets there are 11 kinds of feature values whose contribution rates are more than 85%. Figure 1 shows a scatter plot with 11 principal components selected by PCA, where each color describes a kind of principal components. It can be seen from Figure 1 that the distance among the sample points of same tumor subtype is larger, even better than that among the sample points of different ones. Hence, to realize personalized medicine, one can put all the sample point areas into grids, and sample points in different grids can adopt different treatment modes.

2.2. Specific steps of FA method

Step 1: Input the original data X_{n*p}, and calculate the correlation coefficient matrixes.

Step 2: Calculate the eigenvalue λ_i ($\lambda_i \geq 0, i = 1, 2, \cdots, p$) of correlation coefficient matrix and the corresponding feature vector L_i of standard orthogonal.

Step 3: Setup the number of public factor, and rotate load matrix to explain common factor better.

Step 4: Design the FA model $x_i = \mu_i + a_{i1} f_1 + a_{i2} f_2 + \ldots + a_{im} f_m + \varepsilon_i$, where $i = 1, 2, \cdots, p$, $f_1, f_2, \cdots, f_m$ are common factor, ε_i is a specific factor of x_i, $a_{i1}, a_{i2}, \cdots, a_{im}$ are the loads of x_i on common factor. Then, calculate the contribution rate and the accumulate contribution rate of each factor.

Step 5: Give a professional interpretation for public factor.

Here, the acute leukemia data sets are used. If the public factor number is 2, one can get 2×2 rotation matrix, and if the number is 4, a 4×4 rotation matrix is obtained. The results are shown in Table 1.

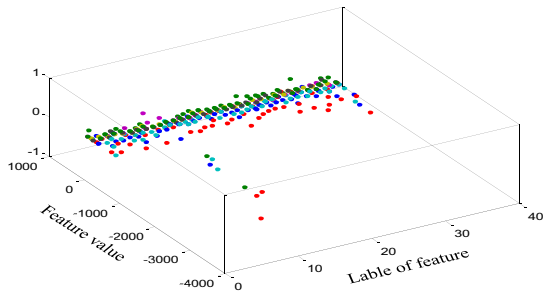

Fig. 1. Scatter plot with 11 principal components from acute leukemia datasets.

Table 1
Comparison of factor classification to acute leukemia datasets

Number of public factor	Contribution rate (%)		Accumulate contribution rate (%)	
2	74.1132	62.1731	74.1132	136.2863
4	54.6000	32.3778	54.6000	86.9778
	29.6459	27.9934	116.6237	144.6171

2.3. Specific steps of ICA method

Step 1: Make observation data X centralized, and let its mean $E\{S_iS_j\} = E\{S_i\}E\{S_j\} = 0$, where S is the feature value of original variable, and $i, j = 1, 2, \cdots, p$.

Step 2: Whiten data by $Z(t) = W_0X(t)$, where W_0 is a whitening matrix, and Z is a whitening vector.

Step 3: Choose the estimated component number m, and let the iterative number $p = 1$.

Step 4: Choose an initial weight vector W_p randomly.

Step 5: Let $W_p = E\{Zg(W_p^TZ)\} - E\{g'(W_p^TZ)\}W$, where g is a nonlinear function with $g_1(y) = tanh(y)$, $g_2(y) = yex(p - \frac{y^2}{2})$, or $g_3(y) = y^3$, and so on.

Step 6: Let $W_p = W_p - \sum_{j=1}^{p-1}(W_p^TW_j)W_j$, and make $Y = W^TX$ the biggest Non-Gaussianity from FastICA learning rule. Negentropy of Non-Gaussianity equation $N_g(W^TX)$ measures the approximate value, where $N_g(Y) = H(Y_{Gauss}) - H(Y)$, Y_{Gauss} is a Gaussianity random variable which has the same variance as Y, and $H(Y) = -\int p_Y(\xi)\log p_Y(\xi)d\xi$. The variance constraint of W^TX is 1, and equal to that the norm of constraint W for the whitening data.

Step 7: According to the conditions of Kuhn-Tucker, under the constraint $E\{(W^TX)^2\} = ||W||^2 = 1$, the optimum value of $E\{G(W^TX)\}$ satisfies $E\{Xg(W^TX)\} + \beta W = 0$, where β is a steady state value, and Jacobian matrix $JF(W) = E\{XX^Tg'(W^TX)\} - \beta I$ is obtained from Newton iteration method, where I is an unit matrix. The data can be turned into diagonal matrix by spheroidizing, followed by the approximate Newton iteration equation $W^* = W - \frac{E\{Xg(W^TX)\}-\beta W}{E\{g'(W^TX)\}-\beta}$ and $W = \frac{W^*}{||W^*||}$ after inverse processing. So one can assume that $W_p = \frac{W_p}{||W_p||}$.

Step 8: If W_P is non-convergent, then return to Step 5.

Step 9: Let $p = p + 1$, and if $p \leq m$, return to Step 4.

The data obeys Gaussian distribution in factor analysis (FA), which will restrict the application. Independent component analysis (ICA) has the potential for non-Gaussianity and becomes a strong contender for FA. For acute leukemia datasets, one chooses genes related to tumors by filtering method, and extracts principal components among them. The 11 kinds of characteristics load results of feature category extracted from the above datasets are obtained under feature load ratio shown in Figure 2, from which the extraction results of principal component analysis (PCA) are more accurate than FA and ICA.

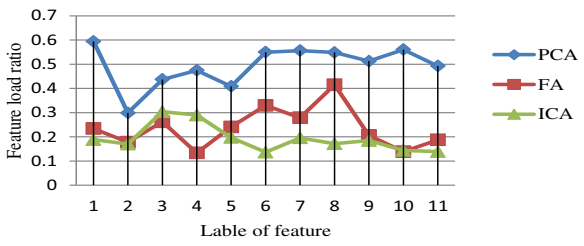

Fig. 2. Comparison of feature load ratio for PCA, FA and ICA for acute leukemia datasets.

2.4. Specific steps of PCDA method

Principal component analysis (PCA) can get a set of uncorrelated principal components through a set of linear transformations, and then the maximum energy of principal components is selected to achieve the purpose of the feature selection. The traditional PCA can choose a new set of linearly independent feature sets, but it does not make full use of the category information. To overcome the weak point, FLD is introduced to those components of PCA. Using FLD, one can always find a best direction to make the sample projection to the straight line which is distinguished. Then, PCDA method is proposed to retain all assets, and the original category information on classification and clustering problems can be fully used. The specific steps of PCDA algorithm can be expressed as follows:

Step 1: Select the features of original variable, choose principal components whose accumulating contribution rates are more than 85%, and form a test sample w_3 with PCA algorithm.

Step 2: Separate the training sample set X from the tumor subtypes into two subsets w_1 and w_2.

Step 3: Calculate sample mean vector $M_i = \frac{1}{n_i}\sum_{x_k \in X_i} x_k$ of all kinds of sample sets in d-dimensional characteristic space, where x_k is the feature value of test sample, and $i = 1, 2$.

Step 4: Choose projection direction $a = (a_1, a_2, \cdots, a_p)'$, and make x_{ij} project in a direction. Then one can get $y_{ij} = a' x_{ij}$, where $i = 1, 2, \cdots, k$, and $j = 1, 2, \cdots, n_i$, so generally one restrains a as a unit vector. Discrete degree matrix $\overrightarrow{S_i} = \sum_{x_k = X_i} (x_k - M_i)(x_k - M_i)^T$ can be calculated in all kinds of samples, where $i = 1, 2$.

Step 5: Calculate discrete degree matrix $\overrightarrow{S_w} = \overrightarrow{S_1} + \overrightarrow{S_2}$ among all kinds.

Step 6: Calculate inverse matrix $\overrightarrow{S_w}^{-1}$ of $\overrightarrow{S_w}$.

Step 7: According to Fisher criterion function $J_F(w) = \frac{|m_1 - m_2|}{s_1^2 + s_2^2}$, one can make J_F the biggest solution w^*, which is the best solution vectors. It follows that the results derived above are put into $J_F(w)$, and $w^* = \frac{\gamma}{\lambda}\overrightarrow{S_w}^{-1}(M_1 - M_2)$ can be obtained, where $\frac{\gamma}{\lambda}$ is a proportion factor. Then, $w^* = \overrightarrow{S_w}^{-1}(M_1 - M_2)$ can be obtained after simplification, followed by w^*.

3. Experimental results

Three subtypes of publicly microarray datasets are used to study the tumor classification problem. They are ALL-T (acute lymphoblastic leukemia, T_cell), ALL-B (acute lymphoblastic leukemia, B_cell) and AML (acute myeloid leukemia), which can be downloaded at http://www.broad. mit.edu/cgi-bin/cancer/datasets.cgi. The data set contains 72 case samples, and each sample includes 7129 gene expression data sets. Then, acute lymphoblastic leukemia (ALL) data sets after adjustment are shown in Table 2. For ALL and AML with PCDA, 10 gene samples are used as class ALL (w_1) and class AML (w_2) for training samples. Using PCA, one selects 11 groups of features with maximum contribution to form the test samples (w_3). They are classified from FLD analysis, and the results are shown in Figure 3.

Table 2
Acute lymphoblastic leukemia data sets after adjustment

Data set and type	Training data (38)			Test Data (34)		
	ALL-T	ALL-B	AML	ALL-T	ALL-B	AML
Original sample classification	8	19	11	2	18	14
Adjusted sample classification	6	21	11	4	16	14
Sample proportion	27		11	20		14

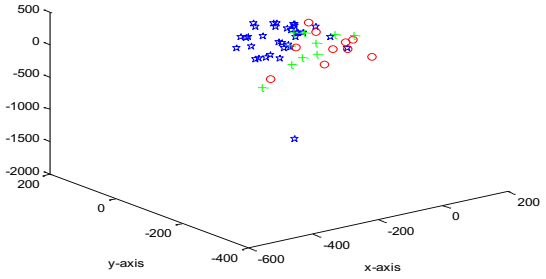

Fig. 3. Distribution diagram of training data w_1, w_2 and original sample data w_3.

Figure 3 illustrates a spacial distribution diagram of the training datasets ALL (w_1), AML (w_2) and original sample data (w_3), in which the green plus is class ALL (w_1), the red round is class AML (w_2), and the blue star is 34 sets of test samples (w_3). Then, 34 groups genetic data of test samples are divided into two categories, and 13 gene samples are divided into class AML (w_2), and the other 21 components for class ALL (w_1). Thus, the classification results of PCDA are obtained as follows : the test sample gene numbers of ALL are 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,27, and those of AML are 21,22,23,24,25,26,28,29,30,31,32,33,34.

In the following, 11 groups of features are selected from 34 cases of test samples, and to illustrate the performances of FLD analysis in PCDA, the 34 cases are divided into class ALL and class AML. Then, the FLD analysis results are shown in Figure 4, where the ordinate is the gene samples, and the abscissa is the features of the gene samples. Experimental results can be more precise than those of PCA.

To further evaluate the classification performances of PCDA method, our proposed PCDA (PCA+FLD) algorithm is compared with the other two dimensionality reduction methods with respect to FA and ICA, which include that PCA is replaced by FA and ICA for construct FA + FLD and ICA + FLD algorithms, respectively. The experimental results for tumor classification are outlined in Table 3, where the error rate represents the percentage of the wrong number versus the total number, and it can reflect the quality of classification results intuitively. In Table 3, the middle two columns describe the sample numbers in every class after FLD analysis. It can be seen that the results of ICA + FLD are the worst, because all of the sample points are sorted into AML. Principal component discriminant analysis (PCDA) performs markedly better than the other two methods on classifying effect. However, only the 27th sample is incorrectly classified into ALL. Therefore, in the current experiments, the classification ability of feature subsets by using PCDA is higher than that of the others, especially of direct feature selection. Principal component discriminant analysis is an effective method to extract feature subsets, and has obvious advantages compared with the other related dimensionality reduction methods.

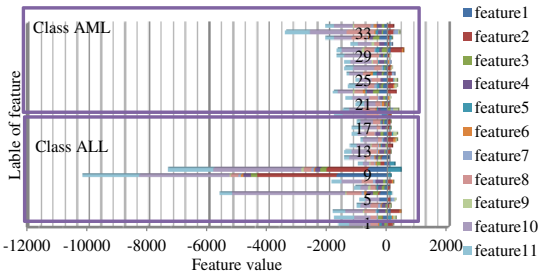

Fig. 4. Fisher linear discriminant (FLD) analysis results in PCDA for acute leukemia data sets.

Table 3
Classification results of three dimensionality reduction methods for tumor classification

Dimensionality reduction methods	Class ALL	Class AML	Error rate (%)
FA + FLD	8	26	35.294
ICA + FLD	0	34	58.824
PCA + FLD	21	13	2.941

4. Conclusion

One important application of gene data is classification of samples into categories. Gene selection plays a key role in diagnosing tumors, so a new PCDA-based feature selection method for tumor classification was designed. The method involves regularizing gene expression data using PCA, followed by the classification applying FLD analysis. The superiority of our algorithms is demonstrated by performing experimental comparisons with other related dimensionality reduction algorithms on acute leukemia data sets. The experiments show that the proposed hybrid method performs well in reducing dimension and improving the performance, and it is effective and efficient in predicting normal and tumor samples.

Acknowledgements

This work is supported by National Natural Science Foundation of China (61370169, 61402153), Project of Henan Science and Technology Department (142102210056), Project of Henan Educational Department (12A520027, 13A520529), and Education Fund for Youth Teachers of Henan Normal Univ.

References

[1] J. Cao, L. Zhang, B.J. Wang, et al., A fast gene selection method for multi-cancer classification using multiple support vector data description, Journal of Biomedical Informatics **53** (2015), 381–389.
[2] J.C. Xu, L. Sun, Y.P. Gao and T.H. Xu, An ensemble feature selection technique for cancer recognition, Bio-Medical Materials and Engineering **24** (2014), 1001–1008.
[3] H. Han and X.L. Li. Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery, BMC Bioinformatics **12** (2011), S7.
[4] S.L. Wang, X.L. Li, S.W. Zhang, et al., Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction, Computers in Biology and Medicine **40** (2010), 179–189.
[5] K. Jakubowska, W. Skrzeczanowski, M. Paduch, et al., Principle component analysis of the spectroscopic and neutron parameters characterizing PF-1000 device plasmas, Applied Physics B **117** (2014), 389–394.
[6] S. Bicciato, A. Luchini and C.D. Bello, PCA disjoint models for multiclass cancer analysis using gene expression data, Bioinformatics **19** (2003), 571–578.
[7] M.E. Whipple, E. Mendez, D.G. Farwell, et al., A log likelihood predictor for genomic classification of oral cancer using principle component analysis for feature selection, Studies in Health Technology and Informatics **107** (2004), 823–826.
[8] N. Yasumune, P.P. Sastia, B. Takeshi, et al., Metabolic distance estimation based on principle component analysis of metabolic turnover, Journal of Bioscience and Bioengineering **118** (2014), 350–355.
[9] D.S. Huang and C.H. Zheng, Independent component analysis-based penalized discriminant method for tumor classification using gene expression data, Bioinformatics **22** (2006), 1855–1862.
[10] H.W. Chen, J. Wang, D.X. Zhang, et al., Molecular diagnosis of tumor based on independent component analysis and support vector machines, Lecture Notes in Computer Science **4456** (2007), 46–56.
[11] C.H. Zheng, D.S. Huang and L. Shang. Feature selection in independent component subspace for microarray data classification, Neurocomputing **69** (2006), 2407–2410.
[12] S.L. Wang, J. Wang, et al., The classification of tumor using gene expression profile based on support vector machines and factor analysis, Sixth International Conference on Intelligent Systems Design and Applications **2** (2006), 471–476.
[13] N. Kunihiro, A. Koji, I. Shumpei, et al., A PCA based method of gene expression visual analysis, Genome Informatics **14** (2003), 346–347.