



Approximate LDA Technique for Dimensionality Reduction in the Small Sample Size Case

Kuldip K. Paliwal

*School of Engineering, Griffith University,
Brisbane, Australia*

k.paliwal@griffith.edu.au

Alok Sharma

*University of Tokyo, Japan
Signal Processing Lab, Griffith University, Australia
University of the South Pacific, Fiji*

aloks@ims.u-tokyo.ac.jp

Abstract

The regularized linear discriminant analysis (LDA) technique overcomes the small sample size (SSS) problem by adding a regularization parameter to the eigenvalues of within-class scatter matrix. However, it has some drawbacks. In this paper we address its drawbacks and propose an improvement. The proposed technique is experimented on several datasets and promising results have been obtained.

Keywords: Approximate linear discriminant analysis (ALDA), dimensionality reduction, small sample size problem, classification accuracy, regularized LDA.

1. Introduction

In many classification problems, the dimensionality of the sample space (or the feature space) is very large and number of data samples is small. It is therefore necessary to reduce the dimensionality of the space for improving the robustness (or generalisation capability) and computational complexity of the pattern classifier.

One of the most widely studied techniques for dimensionality reduction is the linear discriminant analysis (LDA) technique [1]. The LDA technique finds an orientation W that transforms high dimensional feature vectors belonging to different classes to a lower dimensional feature space such that the projected feature vectors of a class on this lower dimensional space are well separated from the feature vectors of other classes. If the dimensionality reduction is from d -dimensional (R^d) space to h -dimensional (R^h) space (where $h < d$), then the size of the orientation matrix W is $R^{d \times h}$; i.e., it has h column vectors. The orientation matrix W is obtained by maximising the Fisher's criterion function, in other words by the eigenvalue decomposition (EVD) of $S_W^{-1}S_B$, where $S_W \in R^{d \times d}$ is within-class scatter matrix and $S_B \in R^{d \times d}$ is between-class scatter matrix. For a c -class problem, the value of h will be $c - 1$ or less, a constraint due to S_B . If the dimensionality d is very large compared to the number of training samples n then S_W becomes singular and the evaluation of eigenvalues and eigenvectors of $S_W^{-1}S_B$ becomes impossible. This drawback is considered to be the main problem of LDA and is commonly known as the small sample size (SSS) problem [2].

In order to overcome this SSS problem, several techniques have been proposed in the literature. For example, PCA+LDA [3],[4], also known as the Fisherface technique, avoids this singularity problem in two stages. The problem with this two-stage Fisherface technique is that though it avoids singularity of S_W matrix, the initial dimensionality reduction by PCA reduces the information content and as a result decreases the classification accuracy. Another technique that overcomes the singularity problem in two steps is the null LDA

technique [5],[6]. It has been shown in the literature [7] that both the range space and null space of S_W are meaningful for classification purposes. Since the null LDA technique discards the range space of S_W , it throws away some important information useful for classification. The other technique is the pseudoinverse technique [8] which avoids the singularity problem but discards the null space of S_W . Another way to overcome the singularity problem is through the regularized LDA technique [9],[10] which adds a small positive constant to the eigenvalues of S_W to make it non-singular. Though this technique uses the null space as well as the range space of the S_W matrix, it suffers from the following two problems: 1) S_W matrix is of very large size and it is not very easy to compute its inverse¹, and 2) The value of the small perturbation term added to each diagonal element of S_W is not known a priori. The perturbation term has to be determined heuristically through the cross-validation method which is an expensive procedure and a poor choice of the term can degrade the generalization performance of the method. The perturbation term has been added just to make the inverse operation feasible and it has no physical meaning.

Like the regularized LDA technique, the proposed technique finds orientation matrix W in a way that the null space and range space of S_W are retained, and at the same time the singularity problem in the computation of the inverse of S_W is avoided. Also, the technique does not suffer from the two above-mentioned problems associated with the regularized LDA technique. Instead of adding the small amount of perturbation term and not knowing its value a priori, we approximate S_W^{-1} by a very well defined matrix [11]. We use this approximation with S_B to compute the EVD of $S_\alpha^{-1}S_B$, where S_α^{-1} is the approximation of S_W^{-1} . This gives us the orientation matrix W . The proposed technique does not use any heuristic approach to evaluate the perturbation term through the cross-validation procedure. The advantages of the proposed technique are demonstrated experimentally in this paper on several datasets. In addition, we demonstrate empirically the relative importance of the range and null spaces of S_W in classification.

2. Regularized LDA

When the dimensionality of feature space is very large compared to the number of training samples available, then the S_W matrix becomes singular. To overcome this singularity problem in the regularized LDA method, a small perturbation to the S_W matrix has been added. This makes the S_W matrix non-singular. However, the size of S_W matrix is very large due to very high dimensional feature space and its inversion becomes computationally infeasible. To overcome this computational problem, one can first project the original feature space onto the range space of S_T and then can apply regularization to the within-class scatter matrix in the reduced feature space. Let the within-class scatter matrix and between-class scatter matrix in the reduced dimensional space be \hat{S}_W and \hat{S}_B . The regularization can be applied as follows:

$$(\hat{S}_W + \delta I)^{-1} \hat{S}_B w_i = \lambda_i w_i,$$

where $\delta > 0$ is a perturbation term or regularization parameter. The addition of δ in the regularized method helps to incorporate both the null space and range space of \hat{S}_W . However, the drawback is that there is no direct way of evaluating the parameter as it requires heuristic approaches to evaluate it and a poor choice of δ can degrade the generalization

¹ However, it is possible to avoid this problem by reducing the dimensionality first by throwing the null space of S_T and keeping the range space of S_T , and then applying regularized method. Therefore this large size of S_W can be controlled by this procedure.

performance of the method. The parameter δ has been added just to perform the inverse operation feasible and it has no physical meaning.

3. Null LDA technique

In the null LDA technique [6], the h column vectors of the orientation $W = [w_1, w_2, \dots, w_h]$ are taken to be in the null space of the within-class scatter matrix S_W ; i.e., $w_i^T S_W w_i = 0$ for $i = 1 \dots h$. In addition, these column vectors have to satisfy the condition $w_i^T S_B w_i \neq 0$ for $i = 1 \dots h$.

Since the dimensionality of the null space of S_W is $d - (n - c)$, we will have $d - (n - c)$ linearly independent vectors satisfying the two above mentioned conditions. Since $d - (n - c)$ is greater than h , Chen et al. [6] have used eigen analysis of S_B matrix to select h leading eigenvectors from these $d - (n - c)$ vectors to form the orientation matrix W . Thus, in the null space method, W is found by maximizing $|W^T S_B W|$ subject to the constraint $|W^T S_W W| = 0$, i.e.,

$$W = \arg \max_{|W^T S_W W|=0} |W^T S_B W|.$$

The null LDA technique finds the orientation W in two stages. In the first stage, it computes W such that $S_W W = 0$: i.e., data is projected on the null space of S_W and throws the range space of S_W . Then in the second stage it finds W that satisfies $S_B W \neq 0$ and maximizes $|W^T S_B W|$. The second stage is commonly implemented through the PCA method applied on S_B . When the dimensionality d of the original feature space is very large in comparison to sample size n , the evaluation of null space becomes nearly impossible as the eigenvalue decomposition of such a large $d \times d$ matrix will lead to serious computational problems. This is a major problem. There are two main techniques in this respect suggested in the literature for computing the orientation W . In the first technique, a pre-processing step is introduced where the PCA technique is applied to reduce the dimensionality from d to $n - 1$ by removing the null space of S_T . In the reduced $n - 1$ dimensional space it is possible to compute the null space of S_W . This pre-processing step is then followed by the two steps of the null space LDA method [5]. In the second technique, no pre-processing is necessary but the required null space of S_W is computed in the first stage by first finding the range space of S_W , then projecting the data onto this range space followed by subtracting it from the original data. After this step, the PCA method is applied to carry out the second stage [12]. It can be seen that in both the techniques range space of S_W was thrown which could have some discriminant information for classification. As a result the null LDA method is also suboptimal.

4. Pseudoinverse method

In the pseudoinverse method [8] the inverse of within-class scatter matrix S_W is estimated by its pseudoinverse and then the conventional eigenvalue problem is solved to compute the orientation matrix W . The pseudoinverse method implemented in this paper is described as follows:

In this method, a pre-processing step is used where feature vectors are projected on the range space of S_T to reduce the computational complexity. After the pre-processing step, the reduced dimensional within-class scatter matrix \hat{S}_W is decomposed as

$$\hat{S}_W = U_w D_w^2 U_w^T, \text{ where } U_w \in R^{t \times t}, D_w \in R^{t \times t}, t = \text{rank}(S_T),$$

$$D_w = \begin{bmatrix} \Lambda_w & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \Lambda_w \in R^{w \times w} \text{ (} w \text{ is the rank of } S_W \text{ such that } w < t \text{)}.$$

Let the eigenvectors corresponding to the range space of \hat{S}_W is U_{wr} and the eigenvectors corresponding to the null space of \hat{S}_W is U_{wn} , i.e., $U_w = [U_{wr}, U_{wn}]$, then the pseudoinverse of S_W can be expressed as

$$\hat{S}_W^+ = U_{wr} \Lambda_w^{-2} U_{wr}^T,$$

The orientation matrix W can now be computed by solving the following conventional eigenvalue problem $\hat{S}_W^+ \hat{S}_B w_i = \lambda_i w_i$, where \hat{S}_B is the between-class scatter matrix in the reduced space. It can be observed that the null space of within-class scatter matrix is discarded which would sacrifice some discriminant information. In the next section, we empirically demonstrate the importance of null space and range space of within-class scatter matrix.

5. Relative importance of the range space and null space of within-class scatter matrix in classification

In order to develop the concept of the proposed technique it is helpful to exhibit the relative importance of the range space and null space of S_W in classification. Several techniques that overcome SSS problem discard either null space or range space of S_W . However, it has been mentioned in the literature [7] that the null space as well as the range space of S_W contain information useful for classification, but has not been demonstrated explicitly. Here we demonstrate it experimentally on a number of datasets. In order to illustrate this, a comparison between the classification accuracy of the null space of S_W and the range space S_W has been given in Table 1. The description of datasets has been depicted in Section 7. In these experiments, the data samples have been transformed to the lower dimensional space by using the range space of total-scatter matrix S_T as a pre-processing step. Then the within-class scatter matrix is computed in the reduced dimensional space. Let $\hat{S}_W \in R^{t \times t}$ (where $t = rank(S_T)$) be the within-class scatter in t -dimensional space. Then we can decompose \hat{S}_W into its eigenvalues and eigenvectors as $\hat{S}_W = U_w D_w U_w^T$, where $U_w \in R^{t \times t}$ and $D_w \in R^{t \times t}$ are its eigenvectors and eigenvalues, respectively. We can further decompose eigenvectors of \hat{S}_W as $U_w = [U_{wr}, U_{wn}]$, where $U_{wr} \in R^{t \times w}$ represents the range space, $U_{wn} \in R^{t \times (t-w)}$ represents the null space and $w = rank(S_W)$. To obtain the classification accuracy using the range space, the data is first transformed to w -dimensional space by transform U_{wr} and then the nearest neighbour classifier (NNC) (i.e., k -NN classifier with $k = 1$) is used to classify test feature vectors. For null space classification accuracy, the data is first transformed to $(t - w)$ -dimensional space by U_{wn} and then NNC is used to classify test feature vectors. It can be seen from Table 1 that both the spaces (null space and range space) of S_W contain information useful for classification. However, the range space of S_W is less effective for classification than its null space. Therefore, utilizing both the spaces of S_W could further improve the classification performance.

6. Approximate LDA (ALDA) technique

The rationale behind the proposed ALDA technique is to use both the null space and range space information of S_W and at the same time avoid the need of any heuristic method for computing the perturbation term.

In order to present the ALDA technique, we introduce first some notations. Let X be a set of n training vectors (samples or patterns) in a d -dimensional feature space, $\Omega = \{\omega_i : i = 1, 2, \dots, c\}$ be the finite set of c states of nature or class labels where X_i denotes the i^{th} class label. The set X can be subdivided into c subsets X_1, X_2, \dots, X_c ; i.e., $X_i \subset X$ and $X_1 \cup X_2 \cup \dots \cup X_c = X$, where each subset X_i belongs to X_i and consists of n_i number of

Table 1: A comparison of performance between the null space of S_W and the range space of S_W .

Database	Classification accuracy in percentage using the range space of S_W	Classification accuracy in percentage using the null space of S_W
Acute Leukemia [13]	76.5	97.1
ALL subtype [14]	80.4	86.6
Breast Cancer [15]	47.4	57.9
Lung Cancer [16]	89.9	98.0
MLL [17]	80.0	100.0
SRBCT [18]	40.0	100.0
average	69.0	89.9

samples such that:

$$n = \sum_{i=1}^c n_i.$$

The samples or patterns of set X can be written as $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ where $\mathbf{x}_j \in R^d$. Let μ_j be the centroid of X_j and μ be the centroid of X , then matrix S_T can be formed as follows:

$$S_T = A_t A_t^T, \quad (1)$$

where $A_t \in R^{d \times n}$ is a rectangular matrix given by

$$A_t = [(x_1 - \mu), (x_2 - \mu), \dots, (x_n - \mu)].$$

The singular value decomposition (SVD) of matrix A_t can be given as

$$A_t = U_t \Sigma_t V_t^T. \quad (2)$$

By using equation 2, equation 1 can be written as $S_T = U_t \Sigma_t^2 U_t^T$. If the $\text{rank}(S_T) = t$ (where $t < d$) then the size of matrix U_t would be $R^{d \times t}$ and the size of diagonal matrix Σ_t would be $R^{t \times t}$.

In order to reduce the computational complexity, a pre-processing step is used where the feature vectors are projected onto the range space of S_T . Let $Y = U_t^T X$ be the reduced dimensional feature vectors in the range space of S_T . Then reduced dimensional within-class scatter matrix \hat{S}_W , of rank w , can be formed as

$$\hat{S}_W = \hat{A}_w \hat{A}_w^T,$$

where rectangular matrix $\hat{A}_w \in R^{t \times n}$ can be defined as

$$\hat{A}_w = [(y_1 - \hat{\mu}_1), \dots, (y_{n_1} - \hat{\mu}_1), (y_{n_1+1} - \hat{\mu}_2), \dots, (y_n - \hat{\mu}_c)],$$

where $y_i \in Y$ (for $i = 1 \dots n$) is the reduced dimensional feature vector, $\hat{\mu}_j$ (for $j = 1 \dots c$) is the mean of samples of class j in the reduced feature space and $\hat{\mu}$ is the mean of all the training data Y . The reduced dimensional between-class scatter matrix \hat{S}_B (of rank b) can be given as

$$\hat{S}_B = \sum_{j=1}^c n_j (\hat{\mu}_j - \hat{\mu})(\hat{\mu}_j - \hat{\mu})^T.$$

Table 2: ALDA technique.

-
1. Pre-processing step: compute SVD of rectangular matrix A_t to obtain its eigenvectors $U_t \in R^{d \times t}$ then project the data sample X into the range space of total-scatter matrix, i.e., $Y = U_t^T X$.
 2. Evaluate eigenvectors U_w and square root of eigenvalues D_w of the reduced dimensional within-class scatter matrix \hat{S}_W (equation 3).
 3. Find the maximum of square root of eigenvalue of \hat{S}_W , i.e., $\alpha = \max(\text{diag}(D_w))$ to get $D_\alpha = \alpha I_{t \times t} - D_w$ and $\hat{S}_\alpha^{-1} = U_w D_\alpha^2 U_w^T$.
 4. Solve $\hat{S}_\alpha^{-1} \hat{S}_B w_i = \lambda_i w_i$ for h leading eigenvalues and corresponding eigenvectors to get orientation W where $1 \leq h \leq c - 1$.
-

The singular value decomposition (SVD) of matrix \hat{A}_w will give

$$\hat{S}_W = U_w D_w^2 U_w^T, \quad (3)$$

where $D_w = \begin{bmatrix} \Lambda_w & 0 \\ 0 & 0 \end{bmatrix}$, $\Lambda_w \in R^{w \times w}$ and $U_w \in R^{t \times t}$. If \hat{S}_W is a full rank matrix (i.e. $w = t$) then the inverse of \hat{S}_W can be written as:

$$\hat{S}_W^{-1} = U_w D_w^{-2} U_w^T. \quad (4)$$

But since $w < t$, then D_w^{-1} is not possible. However, we can substitute it with D_α , where D_α is defined as $D_\alpha = \alpha I_{t \times t} - D_w$ and $\alpha = \max(\text{diag}(D_w))$ [11]. Then equation 4 can be written as:

$$\hat{S}_\alpha^{-1} = U_w D_\alpha^2 U_w^T.$$

The orientation matrix W can be obtained by solving the following conventional eigenvalue problem:

$$\hat{S}_\alpha^{-1} \hat{S}_B w_i = \lambda_i w_i, \quad (5)$$

where w_i are the column vectors of W that correspond to the λ_i (eigenvalues) in equation 5. The implementation of the ALDA technique is summarized in Table 2.

7. Datasets and experimentation

Several datasets have been used for the experimentation purpose, including DNA microarray gene expression data, face recognition data and text classification data. The description of these datasets is given in Table 3. It can be seen from the table that the dimensionality of the dataset is very large compared to the number of samples which leads to the SSS problem.

The datasets used in the experimentation have two distinct sets namely, training set and test set. The model parameters of all the algorithms experimented in this work are computed using training set only. Once the model parameters are measured then an independent test set is used to evaluate the performance in terms of classification accuracy. The classification accuracy is computed on test set only. The classification accuracy defined in this paper is the percentage of the number of samples correctly labelled in the test set over the total number of samples in the test set. This will give us unbiased generalization error.

Table 3: Datasets used in the experimentation.

Datasets	Class	Dimension	No. training samples	No. test samples
Acute Leukemia [13]	2	7129	38	34
ALL subtype [14]	7	12558	215	112
Breast Cancer [15]	2	24481	78	19
Lung Cancer [16]	2	12533	32	149
MLL [17]	3	12582	57	15
SRBCT [18]	4	2308	63	20
Face ORL [19]	40	10304	200	200
Dexter [20]	2	20000	300	300

Table 4: Classification accuracy (in percentage) on datasets.

Database	Null LDA	Reg. LDA	ULDA	Pseudoinverse technique	ALDA
Acute Leukemia	97.1	97.1	100.0	73.5	100.0
ALL subtype	86.6	87.5	82.1	62.5	92.9
Breast Cancer	57.9	47.4	57.9	47.4	73.7
Lung Cancer	98.0	98.0	98.0	92.0	97.3
MLL	100.0	100.0	100.0	80.0	100.0
SRBCT	100.0	100.0	100.0	85.0	100.0
Face ORL	91.0	90.5	85.0	89.5	90.0
Dexter	93.7	93.7	93.7	67.3	93.0
average	90.5	89.3	89.6	74.7	93.4

The ALDA technique is utilized to reduce the dimensionality while the nearest neighbour classifier (i.e., k-NN classifier with $k = 1$) is used for classifying the data. For the regularized LDA technique, the regularization parameter was estimated by using leave-one out cross-validation procedure on training set. Table 4 illustrates the classification accuracy of the ALDA technique on all the datasets. For all the techniques, except ULDA [21], the pre-processing step is used to reduce the computational complexity by projecting the data using the range space of S_T on the lower dimensional space. It can be observed from the table that for three cases ALDA is beating Null LDA significantly: in Breast Cancer by 15.8%; in ALL subtype by 6.3%; and, in Acute Leukemia by 2.9%. For two datasets (MLL and SRBCT), ALDA, Null LDA, regularized LDA and ULDA techniques are giving perfect classification accuracy. In another three cases (Lung Cancer, face ORL and Dexter) Null LDA is getting the highest classification accuracy. However, the classification performance is equal or less than 1% higher when compared with ALDA technique. The overall or average classification accuracy of ALDA technique is the highest which is 93.4%. The Null LDA technique is also competing well at an average classification accuracy of 90.5% followed by ULDA (89.6%), regularized LDA (89.3%) and pseudoinverse technique (74.7%). It can be observed from the results that the ALDA technique outperforms the null LDA technique, regularized LDA technique, ULDA technique, and pseudoinverse technique. The average classification accuracy obtained by ALDA is higher than the other techniques.

8. Conclusion

The approximate LDA (ALDA) technique has been proposed in this paper. This technique avoids any heuristic approach to evaluate the regularization parameter (as required in the regularized LDA technique) and retains both the null space and range space of S_W efficiently. The technique has been experimented on several datasets and promising results have been obtained.

References

- [1] R.O. Duda and P.E. Hart, *Pattern classification and scene analysis Related Fields*. Wiley, 1973.
- [2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Acad. Press Inc., HBJ Pub., 1990.
- [3] D.L. Swets and J. Weng, *Using discriminative eigenfeatures for image retrieval*. IEEE Trans. PAMI, vol. 18, no. 8, pp. 831-836, 1996.
- [4] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, *Eigenfaces vs. fisherfaces: recognition using class specific linear projection*. IEEE Trans. PAMI, vol. 19, no. 7, pp. 711-720, 1997.
- [5] R. Huang, Q. Liu, H. Lu and S. Ma, *Solving the Small Sample Size Problem of LDA*. Proceedings of ICPR 2002, vol. 3, pp. 29-32, 2002.
- [6] L.-F. Chen, H.-Y.M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, *A new LDA-based face recognition system which can solve the small sample size problem*. Patt. Recog., vol. 33, pp. 1713-1726, 2000.
- [7] J. Yang and J. Yang, *Why can LDA be performed in PCA transformed space?* Patt. Recog., vol. 36, pp. 563-566, 2003.
- [8] Q. Tian, M. Barbero, Z.H. Gu and S.H. Lee, *Image classification by the Foley-Sammon transform*. Optical Eng., vol. 25, no. 7, pp. 834-840, 1986.
- [9] J.H. Friedman, *Regularized discriminant analysis*. Journal of the American Statistical Association, vol. 84, pp. 165-175, 1989.
- [10] Y. Guo, T. Hastie and R. Tibshirani, *Regularized discriminant analysis and its application in microarrays*. Biostatistics, vol. 8, no. 1, pp. 86-100, 2007.
- [11] K.K. Paliwal and A. Sharma, *Improved direct LDA and its application to DNA microarray gene expression data*. Pattern Recognition Letters, vol. 31, issue 16, pp. 2489-2492, 2010.
- [12] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, *Discriminative common vectors for face recognition*. IEEE Trans. PAMI, vol. 27, no. 1, pp. 4-13, 2005.
- [13] 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, 531-537, 1999.
- [14] E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.H. Pui, W.E. Evans, C. Naeve, L. Wong, J.R. Downing, *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling*. Cancer, vol. 1, no. 2, pp 133-143, 2002.
- [15] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D He, A.M.H. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards and S.H. Friend, *Gene expression profiling predicts clinical outcome of breast cancer*, *Letters to Nature*. Nature, vol. 415, pp. 530-536, 2002.
- [16] G.J. Gordon, R.V. Jensen, L.-L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker and R. Bueno, *Translation of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma*. Cancer Research, vol. 62, pp 4963-4967, 2002.
- [17] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia*. Nat. Gen., vol. 30, pp 41-47, 2002.
- [18] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson and P.S. Meltzer, *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network*. Nat. Med., vol. 7, pp. 673-679, 2001.

- [19] F. Samaria and A. Harter, *Parameterization of a stochastic model for human face identification*. Proc. Second IEEE Workshop Applications of Comp. Vision, pp. 138-142, 1994.
- [20] C.L. Blake and C.J. Merz, *UCI repository of machine learning databases*, <http://www.ics.uci.edu/~mllearn>. Irvine, CA, Uni. of Calif., Dept. of Info. and Comp. Sci., 1998.
- [21] J. Ye, R. Janardan, Q. Li, and H. Park, *Feature extraction via generalized uncorrelated linear discriminant analysis*. The 21st Int. Conf. on Machine Learning, pp. 895-902, 2004.