Improved nearest centroid classifier with shrunken distance measure for null LDA method on cancer classification problem

A. Sharma and K.K. Paliwal

Null linear discriminant analysis (LDA) is a well-known dimensionality reduction technique for the small sample size problem. When the null LDA technique projects the samples to a lower dimensional space, the covariance matrices of individual classes become zero, i.e. all the projected vectors of a given class merge into a single vector. In this case, only the nearest centroid classifier (NCC) can be applied for classification. To improve the classification performance of NCC in the reduced-dimensional space, a shrunken distance based NCC technique is proposed that uses class-conditional *a priori* probabilities for distance computation. Experiments on several DNA microarray gene expression datasets using the proposed technique show very encouraging results for cancer classification.

Introduction: Cancer classification using the DNA microarray data comes under the category of a small sample size (SSS) problem and consists of a large number of genes (dimensions) compared to the number of feature vectors available in the training set. The high dimensionality of the feature space degrades the generalisation performance of the classifier and increases its computational complexity. This situation (commonly known as the curse of dimensionality) can be overcome by first reducing the dimensionality of feature space, followed by classification in the lower-dimensional feature space.

Several dimensionality reduction methods for the SSS problem have been proposed in the literature (see [1] for details). Of these methods, the null LDA method [2–4] has recently attracted much attention. In this method, the data is transformed to the null space of the within-class scatter matrix. When the null LDA method is used for dimensionality reduction, all the training vectors of a given class get merged into a single vector in the reduced feature space (i.e. the class-conditional variances of the features in the reduced feature space are zero). As a result, the pattern classifiers that need class-conditional variance information (such as the Mahalanobis distance classifier [5], shrunken centroid classifier [6], etc.) cannot be used. In this case, only the nearest centroid classifier (NCC) can be used for classification (with any training vector in a given class defining the centroid of that class). Note that in this case the nearest neighbour classifier behaves similarly to NCC.

In this Letter, we attempt to improve the classification performance of NCC in the reduced dimensional space. For this, we propose a shrunken distance based NCC (SD-NCC) technique, where a shrunken distance measure is used for distance computation. In SD-NCC, we include the *a priori* probability information for computing the shrunken distance. Experiments on several microarray gene expression datasets using the SD-NCC technique show encouraging results for cancer classification.

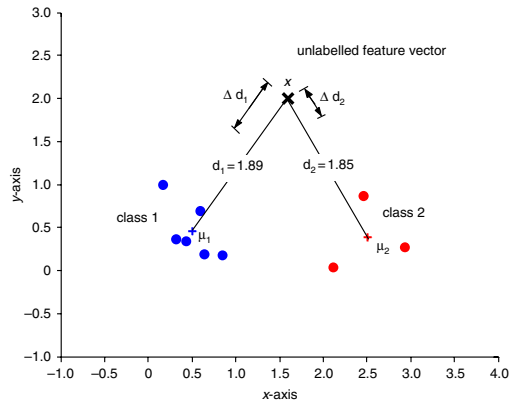

Fig. 1 *Classification using nearest shrunken distance classifier*
In Figure, $d_1 > d_2$, even though **x** belongs to class 1 since $d_1 - \Delta d_1 < d_2 - \Delta d_2$

Shrunken distance NCC technique: To explain the SD-NCC technique, let $\mathcal{X}$ be the d-dimensional set of n training vectors and $\mathbf{x} \in \mathcal{X}$. The technique is illustrated in Fig. 1. The Figure represents a two-class problem where the centroid of class 1 is $\boldsymbol{\mu}_1$ and of class 2 $\boldsymbol{\mu}_2$. The Euclidean distance from feature vector **x** to $\boldsymbol{\mu}_1$ is d_1 and from **x** to $\boldsymbol{\mu}_2$ is d_2. In the SD-NCC technique, the distances d_1 and d_2 are reduced by an amount that depends upon the *a priori* probability information. In the Figure, class 1 has more samples (almost twice) than class 2. Consequently, the amount of reduction of distance between **x** and the centroid of class 1 (Δd_1) will be more than the reduction of distance between **x** and the centroid of class 2 (Δd_2). The resultant distances from feature vector **x** to centroids will now be $d_1 - \Delta d_1$ and $d_2 - \Delta d_2$. Thus, the shrinkage in distance between the test vector **x** and the centroid of a class depends on the *a priori* probability of that class, i.e.

$$\Delta d_j \propto \lambda_j \|\mathbf{x} - \boldsymbol{\mu}_j\|, \quad \text{for } j = 1 \ldots c$$

or

$$\Delta d_j = p\lambda_j \|\mathbf{x} - \boldsymbol{\mu}_j\|, \quad \text{for } j = 1 \ldots c$$

where λ_j is the *a priori* probability of the jth class and can be given as $\lambda_j = n_j/n$ (number of samples in class j/ total number of training samples), c is the number of classes, and p denotes the proportionality constant which depends upon the type of training data used. The value of p can be evaluated using the cross-validation procedure on training data. The value for which the misclassification error is minimum in the cross-validation process is the desired p value for the SD-NCC technique. The shrunken distance d_j between feature vector **x** and the class centroid can now be given as:

$$\begin{aligned} d_j &= \|\mathbf{x} - \boldsymbol{\mu}_j\| - \Delta d_j \\ &= \|\mathbf{x} - \boldsymbol{\mu}_j\| - p\lambda_j \|\mathbf{x} - \boldsymbol{\mu}_j\| \text{ for } j = 1 \ldots c \end{aligned} \tag{1}$$

The cross-validation procedure (to find the optimum value of p) is as set out below. In the training phase parameter λ_j, $\boldsymbol{\mu}_j$ and p are computed and in the test phase a sample is labelled for which distance d_j is minimum.

Step 1: Given training data, partition it randomly into k roughly equal segments.

Step 2: Hold out one segment as validation data and the remaining $k - 1$ segments as learning data from the training data.

Step 3: Use the learning data for finding the null LDA transformation matrix, the centroids of individual classes and their *a priori* probabilities.

Step 4: Use validation data to compute misclassification error using shrunken distance (1) for a range of values of p. Store the obtained misclassification errors.

Step 5: Repeat steps 1–4 N times.

Step 6: Evaluate average misclassification error over N repetitions.

Step 7: Plot a curve of average misclassification error as a function of p.

Step 8: The argument of minimum average misclassification error will be the optimum p value.

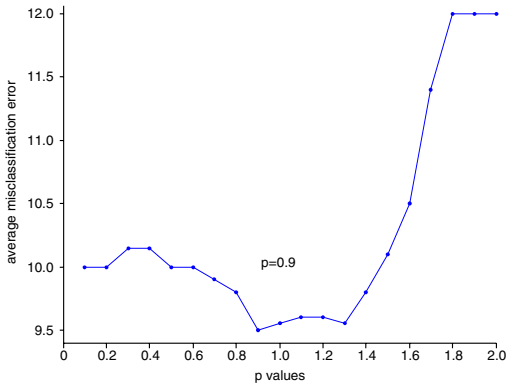

Fig. 2 *Average misclassification error against p values*

Results: Five DNA microarray gene expression datasets are used for the experimentation. The description of these datasets is given in Table 1, which refers to [7–11]. The optimum value of p is computed by the k-fold cross-validation procedure (described at the end of the preceding Section) with $k = 3$ and $N = 20$. The curve of the average misclassification error as a function of p-values for the breast cancer dataset [11] is shown in Fig. 2 for illustration purpose. The optimum value of p is the argument of minimum misclassification error. In Fig. 2, the optimum value of p is 0.9. For evaluating the performance of the procedure, we used an independent test set which was not used during

the tuning of parameter p. The results are presented in Table 2. The following techniques, namely null LDA [2] using NCC and regularised LDA [12], have been used for comparing the performance with the SD-NCC technique. We can observe from Table 2 that the SD-NCC technique performs better than the other techniques. In particular, SD-NCC shows improvement over the NCC technique.

Table 1: Datasets used in experimentation

Datasets	Class	Dimension	Number of training samples	Number of testing samples
Acute leukemia [7]	2	7129	38	34
ALL subtype [8]	7	12558	215	112
GCM [9]	14	16063	144	46
SRBCT [10]	4	2308	63	20
Breast cancer [11]	2	24481	78	19

Table 2: Classification accuracy (%) on DNA microarray gene expression datasets (value of proportionality constant p computed by cross-validation shown in brackets)

Database	Null LDA with NCC	Null LDA with SD-NCC	Regularised LDA
SRBCT	100	100 ($p = 0.4$)	100
Acute leukemia	97.1	100 ($p = 0.5$)	97.1
ALL subtype	86.6	90.2 ($p = 0.5$)	92.0
GCM	70.4	74.1 ($p = 3.4$)	74.1
Breast cancer	57.9	68.4 ($p = 0.9$)	47.4
Average	82.4	86.6	82.1

Conclusion: We have presented a shrunken distance nearest centroid classifier which utilises class-conditional *a priori* probabilities for distance computation. The null LDA technique is used for dimensionality reduction and the classifier is applied on reduced dimensional feature space. The SD-NCC is compared with other classifiers on several DNA microarray gene expression data and encouraging results have been noted.

13 July 2010
doi: 10.1049/el.2010.1927
One or more of the Figures in this Letter are available in colour online.

A. Sharma and K.K. Paliwal (*Signal Processing Lab, Griffith University, Brisbane, QLD-4111, Australia*)

E-mail: sharma_al@usp.ac.fj

References

1 Sharma, A., and Paliwal, K.K.: 'Rotational linear discriminant analysis for dimensionality reduction', *IEEE Trans. Knowl. Data Eng.*, 2008, **20**, (10), pp. 1336–1347
2 Chen, L.-F., Liao, H.-Y.M., Ko, M.-T., Lin, J.-C., and Yu, G.-J.: 'A new LDA-based face recognition system which can solve the small sample size problem', *Pattern Recognit.*, 2000, **33**, pp. 1713–1726
3 Cevikalp, H., Neamtu, M., Wlkes, M., and Barkana, A.: 'Discriminative common vectors for face recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (1), pp. 4–13
4 Ye, J., and Xiong, T.: 'Computational and theoretical analysis of null space and orthogonal linear discriminant analysis', *J. Mach. Learn. Res.*, 2006, **7**, pp. 1183–1204
5 Fukunaga, K.: 'Introduction to statistical pattern recognition' (Academic Press Inc., Hartcourt Brace Jovanovich Publishers, 1990)
6 Tibshiriani, R., Hastie, T., Narasimhan, B., and Chu, G.: 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci. USA*, 2002, **99**, (10), pp. 6567–6572
7 Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S.: 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, 1999, **286**, pp. 531–537
8 Yeoh, E.J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.H., Evans, W.E., Naeve, C., Wong, L., and Downing, J.R.: 'Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling', *Cancer*, 2002, **1**, (2), pp. 133–143
9 Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., and Golub, T.R.: 'Multiclass cancer diagnosis using tumor gene expression signatures', *Proc. Natl. Acad. Sci. USA*, 2001, **98**, (26), pp. 15149–15154
10 Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., and Meltzer, P.S.: 'Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural network', *Nature Medicine*, 2001, **7**, pp. 673–679
11 van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.M.H., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., and Friend, S.H.: 'Gene expression profiling predicts clinical outcome of breast cancer', Letters to Nature, *Nature*, 2002, **415**, pp. 530–536
12 Guo, Y., Hastie, T., and Tinshirani, R.: 'Regularized discriminant analysis and its application in microarrays', *Biostatistics*, 2007, **8**, (1), pp. 86–100