From Fisher's Linear Discriminant Analysis to NLDA or the Story of the Solution of a Very Difficult Nonlinear Classification Problem

JOSÉ BARAHONA DA FONSECA
Department of Electrical Engineering and Computer Science
New University of Lisbon
Monte de Caparica, 2829-516 Caparica
PORTUGAL

Abstract- This work deals with pattern classification of single pap-smear cells from an existing database developed on Herlev University Hospital [1]-[2] with 917 cells characterized by 20 numerical features and classified over 7 classes by Human experts. Medical, the method can be used for detecting pre-malignant cells in uterine cervix before the progress into cancer. Available cell features like area, position and brightness of nucleus and cytoplasm are used for the classification into normal and abnormal cells. We began to solve this problem with a modified Kohonnen neural network that took into account the classification errors, but even after long hours of fine tuning of a set of parameters we only got 66.7% of good classifications. Then using Fisher's linear discriminant analysis we also got a similar result, 66.8% of good classifications. So we reached the conclusion that our classification problem is nonlinear and that our modified Kohonen network was essentially equivalent to LDA. Then we implement NLDA with a very simple feedforward neural network and after only 50 epochs of training with BP and varying the number of sigmoidal neurons in the first hidden layer we got a surprising result of 98.3% of good classifications in the best of five successive runs of BP over 50 epochs with random weights initialization and 60 sigmoidal neurons in the first hidden layer. Next we formatted the input data such that all variables have unit variance and we obtained 99.1% of good classifications after 1,000 epochs of training and forcing also zero mean in all variables we got an even better result of 99.8%, i.e. 2 errors in 917 classifications. Finally we compare our solution to recent works and our implementation of NLDA to more sophisticated neural networks that also approximate LDA.

Keywords- Kohonnen Neural Network, LDA, NLDA, Backpropagation, Overfitting, Influence of Initial Weights in BP.

1 Introduction

The first statistical test that answered the question of the significance of a given classification based on a set of numerical features was Fisher's linear discriminant analysis [3]. This test has two major drawbacks. The first is the small sample size (SSS) problem which imposes that the sample has to increase when the number of variables or features increases. And the second problem is LDA can give a low separation score but the classes may be separated by nonlinear frontiers. It seems that the first proposal of the extension of the LDA to the nonlinear case (NLDA) was done by Nobuyuki Otsu [4]. And fifteen years later he co-authored a paper where it is proposed a simple feedforward neural network to approximate NLDA [5]. We implement this network for the database developed on Herlev University Hospital [1]-[2] with 20 numerical features and obtained very good classification results, the best of them was 98.3%. Recently the LDA and NLDA are being revisited in the solution of difficult pattern recognition problems like Human face recognition [6]-[16].

2. Application of Linear Discriminant Analysis

After we had failed to obtain good results with a preliminary experiment based on a modified Kohonnen network that took into account its own errors in the training process we tried LDA using SPSS software. In the following tables and figures we summarize the main outputs of LDA which also gave a bad classification percentage of good classifications of 66.8%.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	12.655[a]	82.2	82.2	.963
2	1.646[a]	10.7	92.8	.789
3	.710[a]	4.6	97.4	.644
4	.287[a]	1.9	99.3	.472
5	.066[a]	.4	99.7	.248
6	.042[a]	.3	100.0	.201

a. First 6 canonical discriminant functions were used in the analysis.

Table 1- Summary of Canonical Discriminant Functions.

.

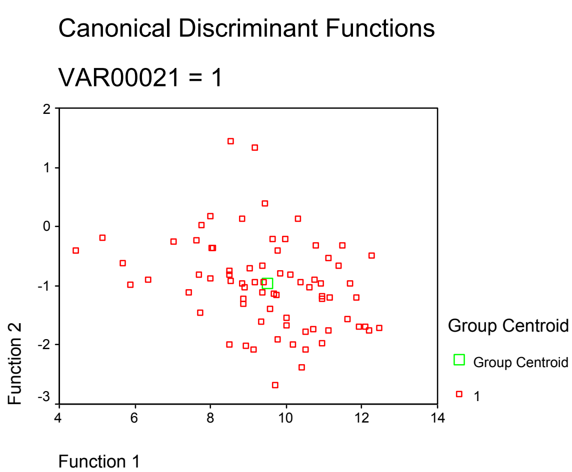
Standardized Canonical Discriminant Function Coefficients

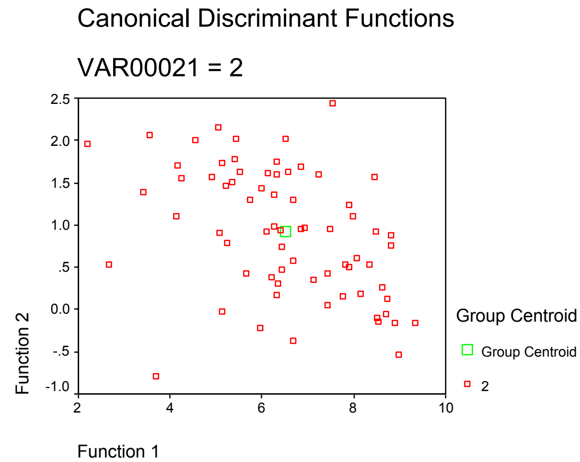
	Function					
	1	2	3	4	5	6
VAR00001	1.570	-1.513	-2.129	1.596	.458	-1.356
VAR00002	-2.049	-.219	-1.139	.523	-.207	-.157
VAR00003	1.735	-.632	1.746	-.473	.321	.141
VAR00004	-.066	.044	-.587	-.071	-.571	-1.208
VAR00005	.020	-.373	.255	-.005	1.110	.915
VAR00006	-2.250	2.080	1.451	-1.554	-1.027	.755
VAR00007	-1.954	1.208	.532	.484	-1.144	1.516
VAR00008	1.628	-1.571	-1.279	.808	1.415	-.886
VAR00009	-1.458	1.272	.853	.044	-1.303	1.304
VAR00010	1.949	.000	.394	-.089	-.408	.047
VAR00011	2.665	.707	1.987	-1.239	1.668	-.792
VAR00012	-1.077	.272	.012	.374	-.404	-.728
VAR00013	1.647	.153	1.138	-.798	.834	.682
VAR00014	.302	-.465	.064	.171	1.305	-.960
VAR00015	-.592	-.350	-.302	.581	-1.519	.417
VAR00016	-.035	-.085	-.315	-.029	-.139	.217
VAR00017	.218	.234	-.281	-.570	.913	.198
VAR00018	-.172	-.612	-.005	.129	-.405	.264
VAR00019	.321	.387	-.102	-5.524	-.033	-2.042
VAR00020	-.167	-.527	.184	5.757	.437	2.159

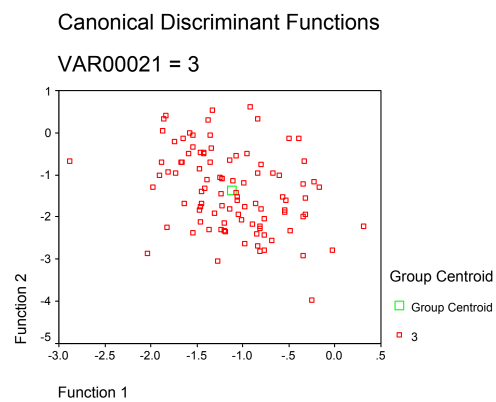
Functions at Group Centroids

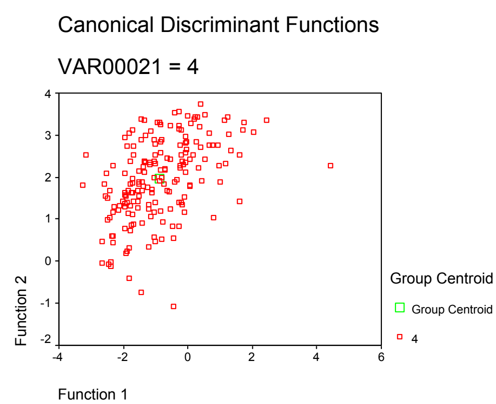
	Function					
VAR00021	1	2	3	4	5	6
1.00	9.499	-.972	-.076	1.005	-.022	.052
2.00	6.506	.915	.449	-1.499	-.022	-.034
3.00	-1.116	-1.362	-2.109	-.295	-.058	-.137
4.00	-.866	1.965	-.179	.304	.203	-.160
5.00	-1.719	.880	.007	.112	-.472	.215
6.00	-1.745	-.740	.251	-.090	.296	.260
7.00	-1.977	-1.326	1.086	.037	-.116	-.277

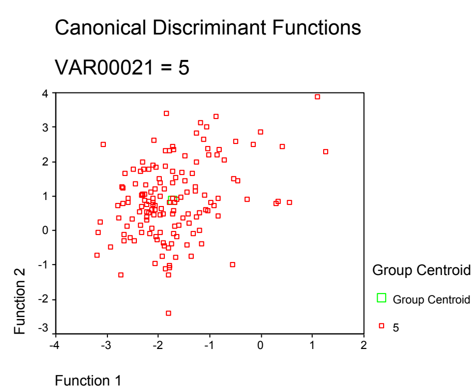
Unstandardized canonical discriminant functions evaluated at group means

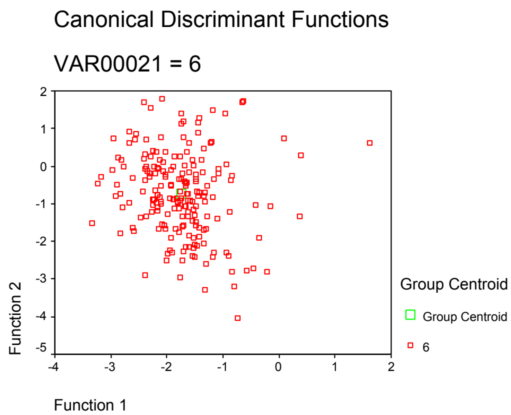


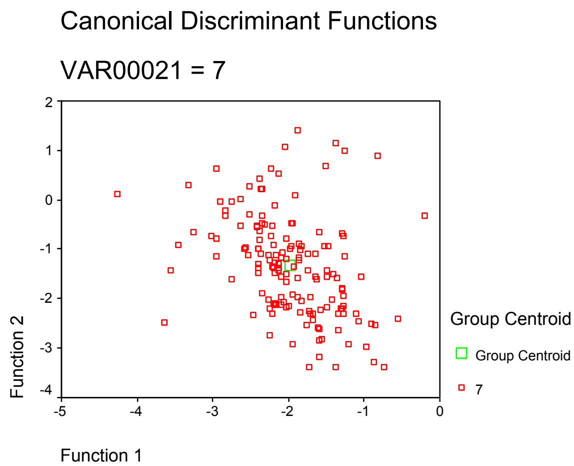


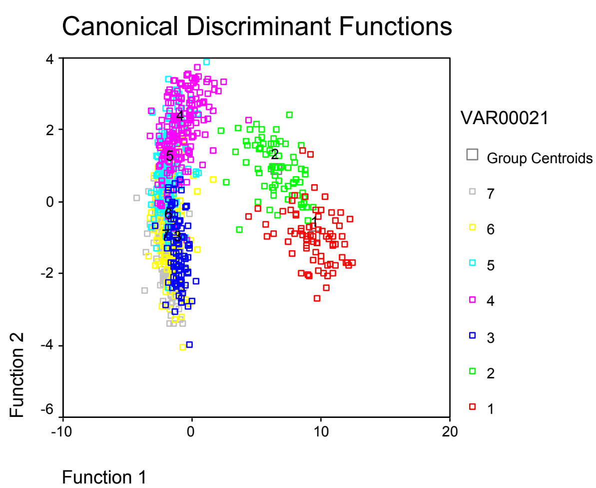












Classification Results[a]

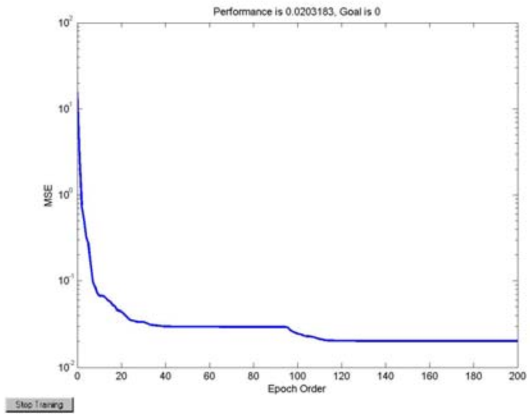
		VAR00021	Predicted Group Membership 1.00	2.00	3.00	4.00	5.00	6.00	7.00	Total
Original	Count	1.00	67	7	0	0	0	0	0	74
		2.00	6	63	0	1	0	0	0	70
		3.00	0	0	73	0	1	22	2	98
		4.00	0	0	1	135	33	13	0	182
		5.00	0	0	7	33	71	27	8	146
		6.00	0	0	12	11	18	109	47	197
		7.00	0	0	6	3	10	36	95	150
	%	1.00	90.5	9.5	.0	.0	.0	.0	.0	100.0
		2.00	8.6	90.0	.0	1.4	.0	.0	.0	100.0
		3.00	.0	.0	74.5	.0	1.0	22.4	2.0	100.0
		4.00	.0	.0	.5	74.2	18.1	7.1	.0	100.0
		5.00	.0	.0	4.8	22.6	48.6	18.5	5.5	100.0
		6.00	.0	.0	6.1	5.6	9.1	55.3	23.9	100.0
		7.00	.0	.0	4.0	2.0	6.7	24.0	63.3	100.0

a. 66.8% of original grouped cases correctly classified.

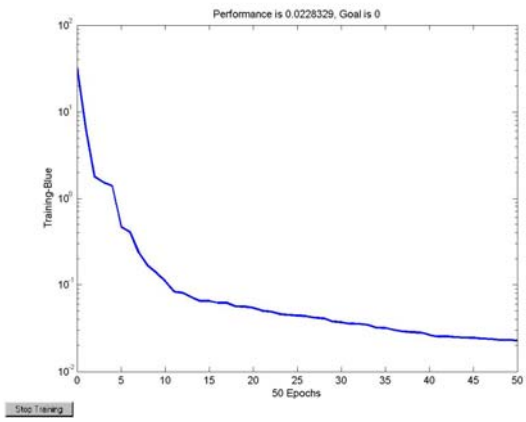
Table 2- Main classification results based on LDA.

3 Application of NLDA based on a 4 Layer Feedforward Neural Network

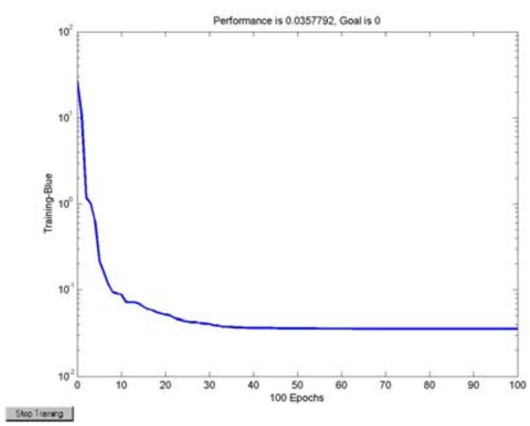
Finally we approximate NLDA by the 4 layer feedforward proposed in [5], with sigmoidal neurons in the first hidden layer and linear neurons in the remaining 2 layers. After having had some problems with overfitting due to too much sigmoidal neurons we reached the conclusion, by trial and error, that the ideal number was 60. In the following figures and tables we present the MSE through 50 epochs and the number of classification errors in each class, respectively for five successive runs of BP with random initialized weights.



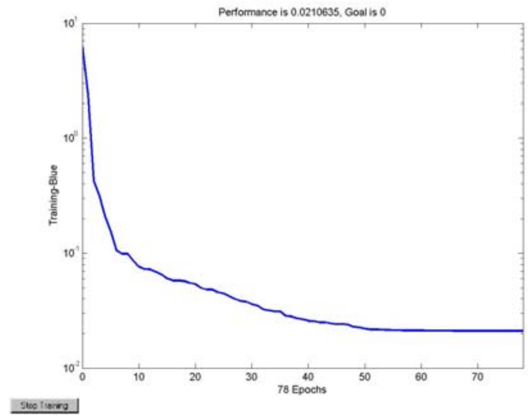
3 2 3 3 2 2 2=17 ERRORS=1.8%



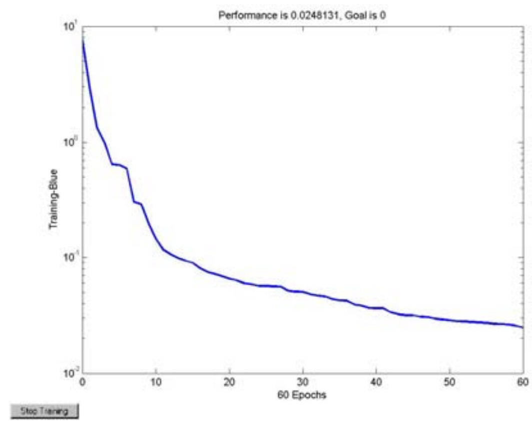
2 1 3 4 2 2 2=16 ERRORS=1.7%



4 0 8 17 17 16 15=77 ERRORS=8.4%



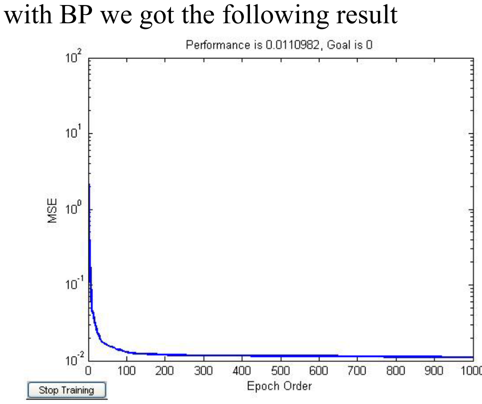
3 0 4 5 5 3 5=25 ERRORS=2.7%



3 3 6 4 2 2 1=21 ERRORS=2.3%

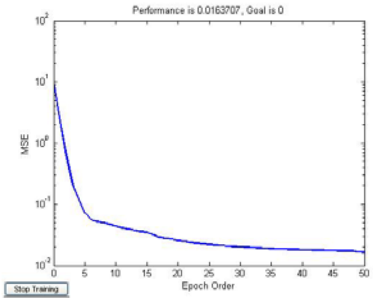
5 Formatting the Input Data

Formatting the input data such that all 20 variables have unity variance, after 1,000 epochs of training with BP we got the following result

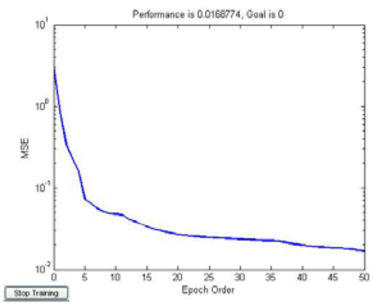


1 0 1 2 3 0 1=8 ERRORS=0.87%

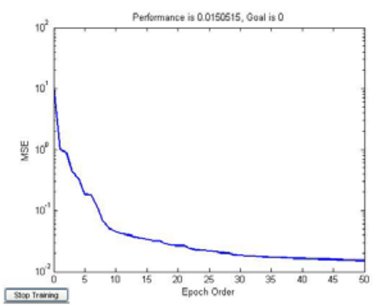
In five successive runs of BP with random initialized weights after 50 epochs we got the following results



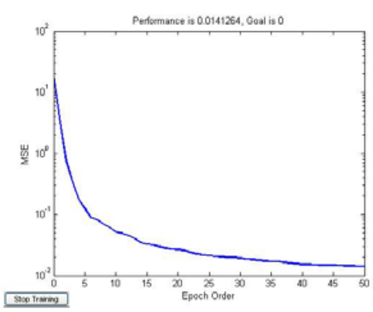
0 0 5 5 6 0 3=19 ERRORS=2.1%



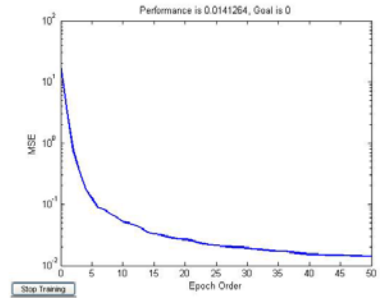
0 0 3 8 6 3 2=22 ERRORS=2.4%



5 0 4 1 3 1 3=17 ERRORS=1.9%

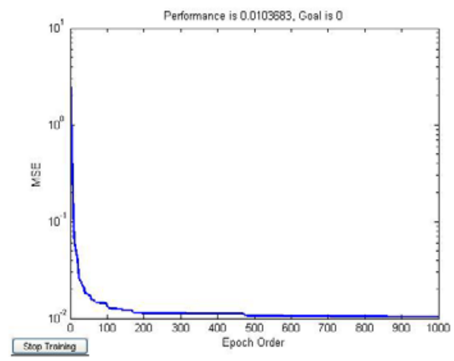


0 1 2 2 1 0 1=7 ERRORS=0.76%



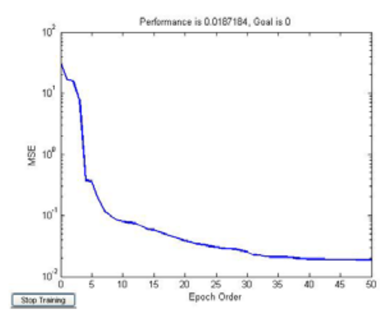
1 0 2 2 3 0 2=10 ERRORS=1.1%

And formatting the input data such that all 20 variables have unity variance and zero mean, after 1,000 epochs of training with BP we got the following result

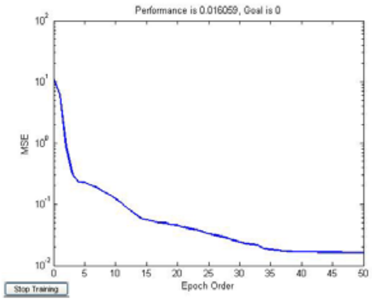


1 0 0 0 1 0 0=2 ERRORS=0.2%

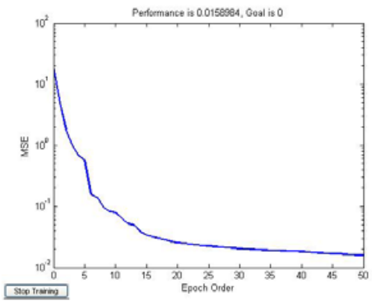
In five successive runs of BP with random initialized weights after 50 epochs we got the following results



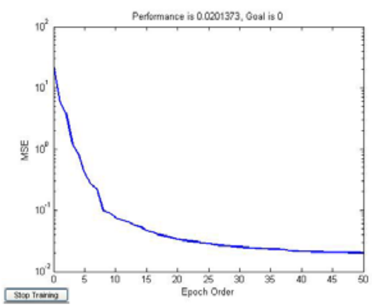
2 2 1 1 4 0 1=11 ERRORS=1.2%



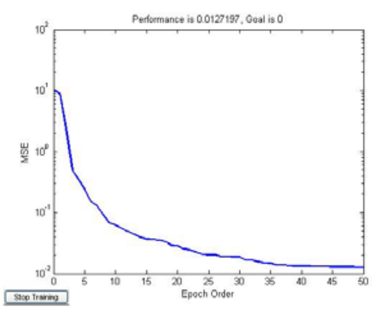
3 0 3 0 2 0 1=9 ERRORS=0.98%



2 0 5 2 1 3 3=16 ERRORS=1.7%



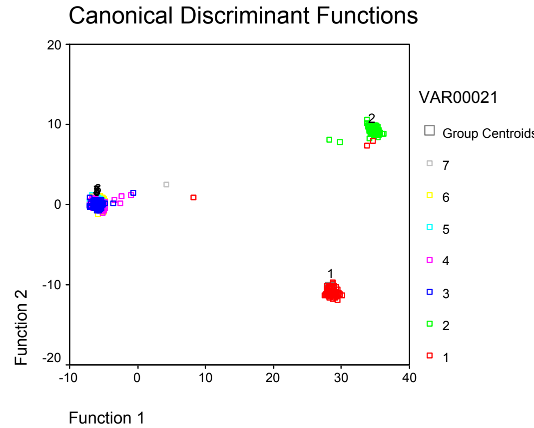
2 0 3 4 0 1 1=11 ERRORS=1.2%



2 0 1 0 0 0 2=5 ERRORS=0.5%

6 Application of LDA to the Output of Hidden2 Layer

Since the output layer is linear, we expect that the LDA applied to the 6 outputs of the second layer will give good results. Indeed using SPSS we got 97.4% of good classifications. In the following figure and table we summarize the main result of this discriminant analysis.



Classification Results[a]

			Predicted Group Membership							Total
		VAR00021	1.00	2.00	3.00	4.00	5.00	6.00	7.00	
Original	Count	1.00	72	2	0	0	0	0	0	74
		2.00	0	70	0	0	0	0	0	70
		3.00	0	0	96	0	0	0	2	98
		4.00	0	0	1	178	1	0	2	182
		5.00	0	0	2	0	140	3	1	146
		6.00	0	0	1	0	1	193	2	197
		7.00	0	0	3	0	1	2	144	150
	%	1.00	97.3	2.7	.0	.0	.0	.0	.0	100.0
		2.00	.0	100.0	.0	.0	.0	.0	.0	100.0
		3.00	.0	.0	98.0	.0	.0	.0	2.0	100.0
		4.00	.0	.0	.5	97.8	.5	.0	1.1	100.0
		5.00	.0	.0	1.4	.0	95.9	2.1	.7	100.0
		6.00	.0	.0	.5	.0	.5	98.0	1.0	100.0
		7.00	.0	.0	2.0	.0	.7	1.3	96.0	100.0

a. 97.4% of original grouped cases correctly classified.

7 Discussion of Results

Although much slower and computationally more expensive the NLDA revealed to be a much more powerful statistical test and pattern recognition algorithm. We believe we could get even better results if we train the network during more epochs and make some fine tuning of the number of sigmoidal neurons in the first hidden layer.

8 Conclusions and Future Work

Although the results are very good we are aware that our approach to NLDA is not good for those not skilled in the art of neural computing. Since even the software package SPSS does not have NLDA, we are preparing a proposal of research project to insert NLDA in SPSS and other statistical software packages.

References:

[1] Erik Martin, *Pap-Smear Classification*, MSc Thesis, Technical University of Denmark – DTU, Oersted Dept. of Automation, 2003.

[2] Jonas Norup, *Classification of pap-smear data by transductive neuro-fuzzy methods*, MSc Thesis, Technical University of Denmark – DTU, Oersted, Dept. of Automation, 2005.

[3] R. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", *Ann. Eugenics, 7,* pp. 178-188, 1936.

[4] N. Otsu, "Nonlinear Discriminant Analysis as a Natural Extension of the Linear Case", *Behaviormetrika, 2,* pp. 45-59, 1975.

[5] Asoh, H. and Otsu, N., "An approximation of nonlinear discriminant analysis by multilayer neural networks", in *Proceedings of IEEE IJCNN '90, International Joint Conference on Neural Networks*, 17-21 June 1990 Page(s):211 - 216 vol. 3

[6] Guang Dai and Yuntao Qian, "Kernel generalized nonlinear discriminant analysis algorithm for pattern recognition" in *Proceedings of ICIP '04, 2004 International Conference on Image Processing*, 24-27 Oct. 2004, Vol. 4, Page(s):2697 – 2700.

[7] Guang Dai, Yuntao Qian and Sen Jia, "A kernel fractional-step nonlinear discriminant analysis for pattern recognition", in *Proceedings of the 17th International Conference on Pattern Recognition*, 2004, *ICPR 2004*, 23-26 Aug. 2004,Vol. 2, Page(s):431 – 434.

[8] Guang Dai and Yuntao Qian; "A robust feature extraction framework for face recognition", in *Proceedings of 2004 International Conference on Image Processing, 2004. ICIP'04*, Vol. 2, 24-27 Oct. 2004 Page(s):1393 – 1396.

[9] Hyunsoo Kim, Drake, B.L. and Haesun Park; "Adaptive nonlinear discriminant analysis by regularized minimum squared errors", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 18, No. 5, May 2006, pp. 603 – 612.

[10] Tsujitani, M. and Koshimizu, T., "Neural discriminant analysis", *IEEE Transactions on Neural Networks*, Vol. 11, No. 6, Nov. 2000, pp. 1394 – 1401.

[11] Tae-Kyun Kim and Kittler, J., "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 27, No. 3, March 2005, pp. 318 – 327.

[12] Guang Dai and Dit-Yan Yeung, "Nonlinear dimensionality reduction for classification using kernel weighted subspace method", in *Proceedings of IEEE International Conference on Image Processing, 2005. ICIP 2005.* 11-14 Sept. 2005, Vol. 2, pp. 838-841.

[13] Shu Yang, Shuicheng Yan, Dong Xu, Xiaoou Tang and Chao Zhang, "Fisher+Kernel criterion for discriminant analysis", in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, CVPR 2005,* 20-25 June 2005, Vol. 2, pp. 197–202.

[14] Baochang Zhang, Xilin Chen, Shiguang Shan and Wen Gao, "Nonlinear face recognition based on maximum average margin criterion", in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.* 20-25 June 2005, Vol. 1, pp.554 – 559.

[15] Cheong Hee Park, Haesun Park and Pardalos, P., "A comparative study of linear and nonlinear feature extraction methods", in *Proceedings of Fourth IEEE International Conference on Data Mining, 2004. ICDM 2004.* 1-4 Nov. 2004, pp. 495–498.

[16] Abdallah, F., Richard, C. and Lengelle, R., "A sequential approach for multi-class discriminant analysis with kernels", in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 200,ICASSP '04,* 17-21 May 2004, Vol. 5, pp. 453-456.