International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:5, No:7, 2011

Self Organizing Mixture Network in Mixture Discriminant Analysis: An Experimental Study

Nazif Çalış, Murat Erişoğlu, Hamza Erol and Tayfun Servi

***Abstract*—**In the recent works related with mixture discriminant analysis (MDA), expectation and maximization (EM) algorithm is used to estimate parameters of Gaussian mixtures. But, initial values of EM algorithm affect the final parameters' estimates. Also, when EM algorithm is applied two times, for the same data set, it can be give different results for the estimate of parameters and this affect the classification accuracy of MDA. Forthcoming this problem, we use Self Organizing Mixture Network (SOMN) algorithm to estimate parameters of Gaussians mixtures in MDA that SOMN is more robust when random the initial values of the parameters are used [5]. We show effectiveness of this method on popular simulated waveform datasets and real glass data set.

***Keywords*—**Self Organizing Mixture Network, Mixture Discriminant Analysis, Waveform Datasets, Glass Identification, Mixture of Multivariate Normal Distributions

I. INTRODUCTION

MIXTURE discriminant analysis (MDA) is a method for classifying observations into known pre-existing non-normal class. This method firstly proposed by Hastie and Tibshirani [1] in which Gaussian mixtures is used to obtain density estimation for each non-normal class. For the Gaussian mixtures, expectation and maximization (EM) algorithm [6] is used to determine number of components and estimate parameters. Xu and Jordan[11], improved the EM algorithm for the Gaussian mixtures and demonstrated its advantages and disadvantages over other algorithms. Choosing initial values of parameters in EM algorithm are very important [12]. EM algorithm can be give different results for the same initial values. In the MDA, classifications of the training observations to the true classes are very important for classification of the test data which is affect the classification rate. Therefore, parameters of the training data must be estimated truly. As an alternative to the EM algorithm Yin and Allinson [13], proposed Self Organizing Mixture Network (SOMN) algorithm for density modelling. Also, Yin and Allinson [5] showed that initial conditions effects convergent results of EM algorithm greater than the SOMN algorithm which is more robust when random initial values are used.

N.Ç Department of Statistics, Faculty of Science and Letters, Çukurova University, 01330 Adana, Turkey (corresponding author to provide phone: 90-338 60 84; fax: 90-338 60 70; e-mail: ncalis@ cu.edu.tr).
M.E. Department of Statistics, Faculty of Science and Letters, Çukurova University, 01330 Adana, Turkey (e-mail: merisoglu@cu.edu.tr).
H.E. Department of Statistics, Faculty of Science and Letters, Çukurova University, 01330 Adana, Turkey (e-mail: herol@cu.edu.tr).
T.S. Elementary Mathematics Education, Adıyaman University, 02040, Adıyaman, Turkey(e-mail: tservi@adiyaman.edu.tr)

In this study when applying MDA, we estimate parameters of Gaussians mixtures using EM and SOMN algorithms for simulation and real data sets. Then we compare results according to classification accuracy rate. In Section 2 and Section 3, we give some notations and estimation of parameters' with MDA and SOMN algorithm, respectively. In Section 4 we apply these algorithms to simulation and reel data sets; also we give the comparison results in Section 4.

II. MIXTURE DISCRIMINANT ANALYSIS

In the mixture discriminant analysis, suppose we have training observation n_j from population j for $j = 1,...,G$. Each class j is divided into R_j artificial subclasses denoted by c_{jr}. According to this clustered approach, each subclass has a multivariate normal distribution $x_i \sim \mathrm{N}(\mu_{jr}, \Sigma_{jr})$ with its own mean vector μ_{jr} and Σ_{jr} is covariance matrix for the r th subclass n j th class. The prior probability for class j is π_j and π_{jr} is the mixing probability for the r th subclass in j th class, such that $\sum_{r=1}^{R_j} \pi_{jr} = 1$. Then mixture density for class j is

$$m_j(x) = P(X = x|G = j) = \left|2\pi\Sigma_j\right|^{-\frac{1}{2}} \sum_{r=1}^{R_j} \pi_{jr} \exp\left[-D(x-\mu_{jr})/2\right] \quad (1)$$

where $D(x-\mu_{jr}) = (x-\mu_{jr})^T \Sigma_{jr}^{-1} (x-\mu_{jr})$ is Mahalanobis distance. The posterior probabilities are obtained, base on Bayes rule, such that

$$P(G = j|X = x) \sim \pi_j \mathrm{Prob}(x|j) \sim \pi_j \sum_{r=1}^{R_j} \pi_{jr} \exp\left[-D(x-\mu_{jr})/2\right] \quad (2)$$

where π_j is the prior probability for class j. An observation is classified into the class j which has the highest posterior probability. The discrimination rules depend on the unknown parameters which are to be estimated from the training data.

III. SELF ORGANIZING MIXTURE NETWORK

Extending self organizing map (SOM) to a mixture density model, Yin and Allinson[13] proposed the self organizing mixture network (SOMN) algorithm. Yin and Allinson [13]describe a new parameter estimation technique minimizing Kullback-Leibler information metric [7] by using Robins-

Monro stochastic approximation method [8]. Structure of SOMN algorithm based on mixture distribution is illustrated in Figure 1.

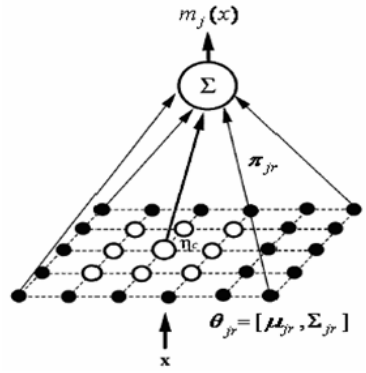


Fig.1. Structure of SOMN algorithm based on mixture distribution [13].

In SOMN algorithm, mean vector μ_{jr} and covariance matrix Σ_{jr} are also called learning weights. The output of neural Network or upper level is equal to summation of prior probabilities or learning weights and component densities weighted with π_{jr}. The updating process of SOM algorithm is described below.

$$\hat{\mu}_{jr}(n+1) = \hat{\mu}_{jr}(n) + \alpha(n)\hat{z}_{jr}(n)[x(n) - \hat{\mu}_{jr}(n)] \quad (3)$$

$$\hat{\Sigma}_{jr}(n+1) = \hat{\Sigma}_{jr}(n) + \alpha(n)\hat{z}_{jr}(n)\left\{[x(n) - \hat{\mu}_{jr}(n)][x(n) - \hat{\mu}_{jr}(n)]^T - \hat{\Sigma}_{jr}(n)\right\} \quad (4)$$

$$\hat{\pi}_{jr}(n+1) = \hat{\pi}_{jr}(n) + \alpha(n)\left\{\hat{z}_{jr}(n) - \hat{\pi}_{jr}(n)\right\} \quad (5)$$

where η_c is neighborhood of cth winner component, $0 < \alpha(n) < 1$ monotone decreasing $\alpha(n)$ is learning rate in the nth iteration and $\hat{z}_{jr}$ is defined as follows

$$\hat{z}_{jr}(n) = \frac{\hat{\pi}_{jr}(n)\hat{m}_{jr}(x(n))}{\sum_{r=1}^{R_j}\hat{\pi}_{jr}(n)\hat{m}_{jr}(x(n))} \quad (6)$$

IV. Applications

4.1. Simulation with Random Waveform Datasets

We show effectiveness of SOMN algorithm in mixture model discriminant analysisis method on popular simulated dataset, taken from Breiman et al. [10]. It is three- class problem with 21 variables and is considered to be a difficult pattern recognition problem. The predictors are defined by

$$x_i = uh_1(i) + (1-u)h_2(i) + \varepsilon_i \quad \text{(class 1)}$$
$$x_i = uh_1(i) + (1-u)h_3(i) + \varepsilon_i \quad \text{(class 2)}$$
$$x_i = uh_2(i) + (1-u)h_3(i) + \varepsilon_i \quad \text{(class 3)}$$

where $i = 1,2,...,21$, u is uniformly distributed on $(0,1)$, ε_i are standard normal variates and the h_i are the shifted triangular waveforms: $h_1(i) = \max(6 - |i-11|, 0)$, $h_2(i) = h_1(i-4)$ and $h_3(i) = h_1(i+4)$. Each training sample has 600 observations, and equal priors were used, so there are 200 observations in each class. We used different 10,000 test samples of size 600. Firstly each classes of a training data is divided into 3 subgroups. In other words each class of a training sample is modeled by a Gaussian mixture model with 3 components. Parameters of the mixture model are calculated by SOMN. The evaluated mixture models are used for discrimination. Each of the observation in test sample is classified into one of 3 classes which has the highest probability in. Classification error rates according to discriminant functions obtained with SOMN-MDA for simulated 10000 independent test data sets are shown by Figure 2 and also the descriptive statistics of them are summarized on Table 1. Mean of misclassification rate after 10000 independent simulations is 21.23%. The minimum and maximum misclassification rates after 10000 independent simulations are 15% and 29% respectively.

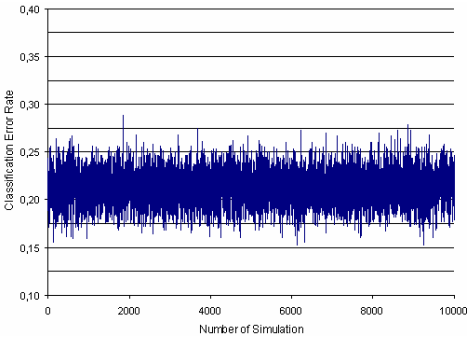


Fig. 2. Misclassification error rates for 10000 independent simulated data sets

TABLE I
DESCRIPTIVE STATISTICS OF MISCLASSIFICATION RATES AFTER 10000 INDEPENDENT TEST SAMPLES

	Statistic	Std. Error
Mean	0.2123	0.00017
Median	0.2117	
Variance	0.000	
Std. Deviation	0.01729	
Minimum	0.15	
Maximum	0.29	
Range	0.14	
Skewness	0.066	0.024
Kurtosis	-0.014	0.049

4.2. Glass Data

This example is from forensic testing of glass. The glass data were obtained from the UCI Machine Learning Repository maintained by Murphy and Aha [9]. A subset of the original data set was used for convenience. The training

data consisted of two groups and seven predictors. The two groups are window float glass and window non-float glass. The variables measured are weight proportions of different oxides. A sample of 80 observations with equal priors for the 2 groups was chosen as the training set, while the test data were of size 83. Variables 8 and 9 are not used in analysis because these variables consist of lots of zero values, moreover five of the other variables which coefficient of variation (CV) is high are selected to analysis. Bashir and Carter [4] compared full rank and reduced rank discriminant method for glass data with three subgroups. The solution of their work on misclassification error rates of Glass data given in Table 2.

TABLE II
MISCLASSIFICATION ERRORS, GLASS DATA THREE SUBGROUPS/GROUP [4]

v	Full rank mda	Reduced rank mda
0.05	0.4167	0.4167
0.10	0.4167	0.3167
0.15	0.4167	0.4167
0.20	0.4167	0.4167
0.25	0.4167	0.3333
0.30	0.4167	0.3333
mda	0.4000	0.4000

Bashir and Carter [4] found the robust reduced rank MDA at $v = 0.10$ was the best model with minimum error of 0.3167. We evaluate the misclassification error rate based on linear discriminant analysis is 0.245 for Glass data with 5 variables chosen for the analysis.

Three subgroups per group of the training data taken from Glass data is used for the analysis. Each group of training data is divided into 3 subgroups. Mixture of normal distribution model is constructed for each group. Form of mixture models for float processed and non float processed groups are given by

Float:

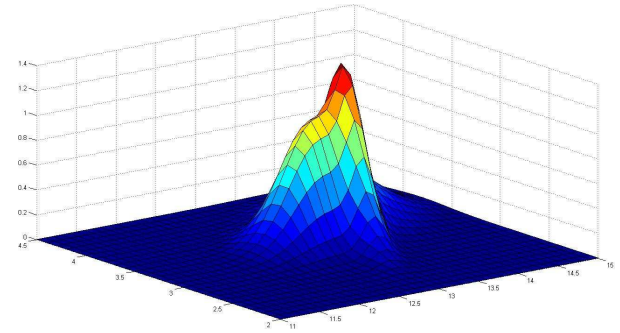
$$m_1(\mathbf{x};\boldsymbol{\theta}) = \hat{\pi}_{11} m_{11}(\mathbf{x};\hat{\mu}_{11},\hat{\Sigma}_{11}) + \hat{\pi}_{12} m_{12}(\mathbf{x};\hat{\mu}_{12},\hat{\Sigma}_{12}) + \hat{\pi}_{13} m_{13}(\mathbf{x};\hat{\mu}_{13},\hat{\Sigma}_{13})$$

Non Float:

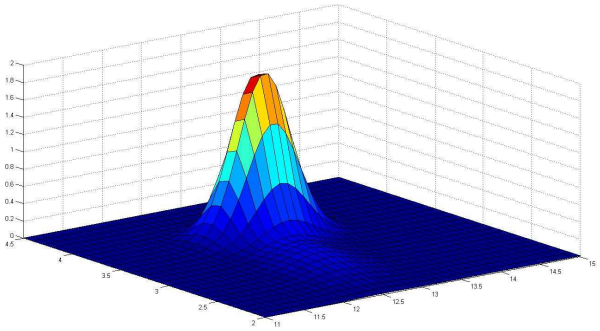
$$m_2(\mathbf{x};\boldsymbol{\theta}) = \hat{\pi}_{21} m_{21}(\mathbf{x};\hat{\mu}_{21},\hat{\Sigma}_{21}) + \hat{\pi}_{22} m_{22}(\mathbf{x};\hat{\mu}_{22},\hat{\Sigma}_{22}) + \hat{\pi}_{23} m_{23}(\mathbf{x};\hat{\mu}_{23},\hat{\Sigma}_{23})$$

where μ_{jr} and Σ_{jr} is mean vector and covariance matrix of j th class and r th subclass for $j=1$ float process, $j=2$ non-float process $j=1,2,3$ respectively. π_{jr} is mixing weight.

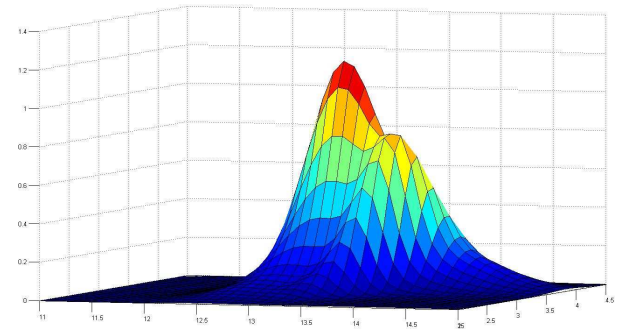
In Glass data set, number of float processed observations is 87 out of 163 so mixture weight of float processed is $\pi_1 = 87/163$ and for non float processed is $\pi_2 = 76/163$. Mixture model for train data is given by $m(\mathbf{x},\boldsymbol{\theta}) = \hat{\pi}_1 m_1(\mathbf{x},\boldsymbol{\theta}) + \hat{\pi}_2 m_2(\mathbf{x},\boldsymbol{\theta})$. Parameter estimations which computed with SOMN in mixture models for two groups of training data are given by Table 3. The graphs of Mixture pdf of float processed window glasses, non window glasses and mixture form of them is given Figure 3 (a-c) respectively.



(a) $m_1(\mathbf{x};\boldsymbol{\theta})$



(b) $m_2(\mathbf{x};\boldsymbol{\theta})$



(c) $m(\mathbf{x};\boldsymbol{\theta})$

Fig. 3. (a) The pdf of mixture of normal with 3 components of float processed group in train dataset (b) The pdf of mixture of normal with 3 components of non-float processed group in train dataset (c) The pdf of mixture of two mixture normal with 3 components of train dataset.

TABLE III
PARAMETER ESTIMATIONS OF THE MIXTURE MODELS FOR EACH GROUP IN TRAIN DATASET.

			Subgroup 1					Subgroup 2					Subgroup 3				
Group 1 (Float processed)	$\hat{\pi}_{1\nu}$		0.267					0.510					0.224				
	$\hat{\mu}_{1\nu}$		13.94	3.78	0.87	0.16	9.43	13.15	3.41	1.26	0.55	8.55	13.56	3.39	1.38	0.38	8.76
			X_2	X_3	X_4	X_6	X_7	X_2	X_3	X_4	X_6	X_7	X_2	X_3	X_4	X_6	X_7
	$\hat{\Sigma}_{1\nu}$	X_2	0.21	0.01	-0.04	-0.03	-0.03	0.17	0.02	-0.02	0.00	-0.04	0.06	0.07	0.00	-0.03	-0.08
		X_3	0.01	0.07	0.02	-0.01	-0.08	0.02	0.05	0.00	0.00	-0.03	0.07	0.13	0.00	-0.05	-0.09
		X_4	-0.04	0.02	0.03	0.00	-0.04	-0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.03	0.02	0.00
		X_6	-0.03	-0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	-0.05	0.02	0.05	0.03
		X_7	-0.03	-0.08	-0.04	0.00	0.13	-0.04	-0.03	0.00	0.00	0.05	-0.08	-0.09	0.00	0.03	0.18

			Subgroup 1					Subgroup 2					Subgroup 3				
Group 2 (Non-float processed)	$\hat{\pi}_{2\nu}$		0.217					0.639					0.145				
	$\hat{\mu}_{2\nu}$		13.58	0.40	1.23	0.22	12.50	12.96	3.65	1.47	0.60	8.16	12.71	2.95	1.15	0.48	9.06
			X_2	X_3	X_4	X_6	X_7	X_2	X_3	X_4	X_6	X_7	X_2	X_3	X_4	X_6	X_7
	$\hat{\Sigma}_{2\nu}$	X_2	1.25	0.45	-0.23	-0.23	-1.21	0.11	0.00	0.00	0.01	-0.04	0.12	0.04	0.04	-0.02	-0.04
		X_3	0.45	1.69	-0.14	-0.06	-0.97	0.00	0.03	-0.02	-0.01	0.02	0.04	0.16	0.00	0.02	-0.03
		X_4	-0.23	-0.14	0.33	0.16	0.02	0.00	-0.02	0.05	0.02	-0.04	0.04	0.00	0.09	-0.02	-0.10
		X_6	-0.23	-0.06	0.16	0.10	0.06	0.01	-0.01	0.02	0.02	-0.03	-0.02	0.02	-0.02	0.05	0.03
		X_7	-1.21	-0.97	0.02	0.06	2.81	-0.04	0.02	-0.04	-0.03	0.09	-0.04	-0.03	-0.10	0.03	0.13

In test data set, each observation is assigned into one of groups namely float processed window and non-float processed window. After classification, 41 out of 47 observations in float processed window group and 30 out of 36 observations in non-float processed window group are correctly assigned. So general misclassification rate for test data is 14.5%. After the classification, scatter plot of observations according to magnesium variable and silicon variable is given Figure 4.

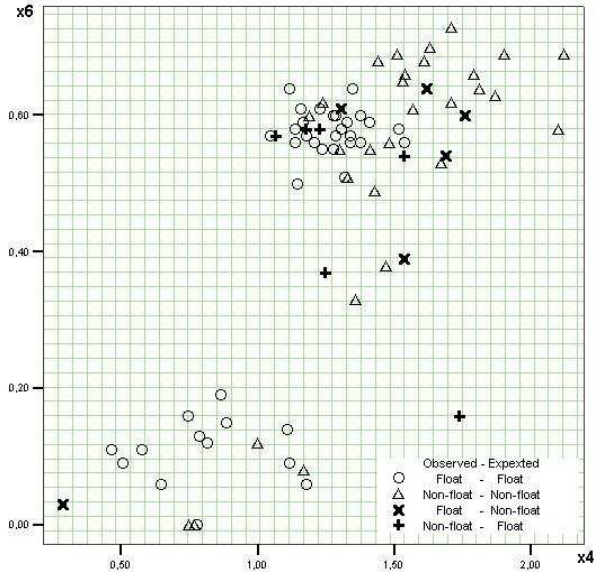


Fig. 4. Scatter plot of observations in test data after the classification.

V. Conclusions

In this study, SOMN algorithm is suggested for parameter estimation in mixture discriminant analysis. We show effectiveness of this method on popular simulated waveform datasets and real glass data set. Although for glass data set, Bashir and Carter [4] found the robust reduced rank MDA at $\nu = 0.10$ was the best model with minimum error of 0.3167, we evaluate misclassification rate is 0.145 in this approach. Classification results of proposed approximation is much better than Bashir and Carter for glass data.

Acknowledgements

We would like to thank the editor and referees whose comments significantly improved this manuscript.

References

[1] Hastie T. and Tibshirani R. (1996), Discriminant Analysis by Gaussian Mixtures, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1, pp. 155-176
[2] Zohar Halbe, Mayer Aladjem(2005),Model-based mixture discriminant analysis—an experimentalstudy, Pattern Recognition, 38, 437-440
[3] Bashir S. and Carter E. M. (2005),Robust Reduced Rank Mixture Discriminant Analysis, Communications in Statistics Theory and Methods, 34, 135-145
[4] Bashir S. and Carter E.M. (2005),High breakdown mixture discriminant analysis, Journal of Multivariate Analysis, 93, 102-11
[5] Yin H. and Allinson N. M. (2001),Self-Organizing Mixture Networks for Probability Density Estimation, IEEE Transactions on Neural Networks, Vol. 12, No. 2, pp. 405-411
[6] P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete date via the EM algorithm," J. Roy. Statist. Soc. B, vol.39, pp. 1–38, 1977.
[7] S. Kullback and R. A. Leibler, "On information and sufficiency," Ann. Math. Statist., vol. 22, pp. 79–86, 1951.
[8] H. Robbins and S. Monro, "A stochastic approximation method," Ann. Math. Statist., vol. 22, pp. 400–407, 1951.
[9] Murphy, P. M., Aha, D. W. (1995). UCI Repository of Machine Learning Databases. Irvine, CA: University of California, Dept. of Information and Computer Science.
[10] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), Classification and Regression Trees Belmont; Wadsworth
[11] Xu, L. and Jordan, M. I. (1993). Unsupervised learning by EM algorithm based on finite mixture of Gaussians. In Proc. World Congress Neural Networks, vol. 2, pp. 431-434, Portland, OR, USA.
[12] McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. Journal of Statistical Software 4, No. 2.
[13] Yin, H. and Allinson, N. M. (1997). Comparison of a Bayesian SOM with the EM algorithm for Gaussian mixtures. Proc. Workshop on Self-Organising Maps (WSOM'97), pp. 118-123.