Information criterion of comparing metric classifiers in ensemble of sources

Mikhail Lange

Federal Research Center "Computer Science and Control" of RAS

The 11th International Conference "Intelligent Data Processing"
(IDP - 11, Spain, Barcelona, October, 10 - 14 , 2016)

1. Contents

- Metric schemes of classification based on **Majority Voting** source decisions and using **General Measure** in an ensemble of sources.

- Average mutual information between an ensemble of sources and a set of classes as a characteristic of efficiency of a classifier.

- Connection of a classification task with a problem of source encoding in a presence of a noisy channel between source and coder.

- Dissimilarity measures in sets of objects of individual sources and in a set of composite objects produced by the ensemble of the sources. Class-conditional densities by these measures.

- Average mutual informations and their estimations for MV и GM classifiers.

- Main result as a relation of estimations of average mutual information for MV и GM classifiers

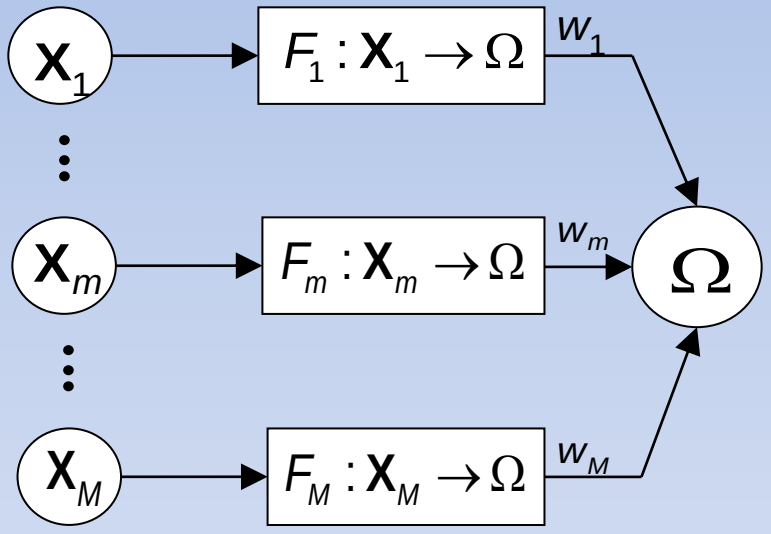
- Experimental results on recognition of faces via color images given by triple component HSI model (ensemble of three sources)

2. Schemes of classifiers in ensemble of sources

A set of classes: $\Omega = \{\omega_1,...,\omega_c\}, c \geq 2$ with a priori probabilities $p(\omega_i): \sum_{i=1}^{c} p(\omega_i) = 1$

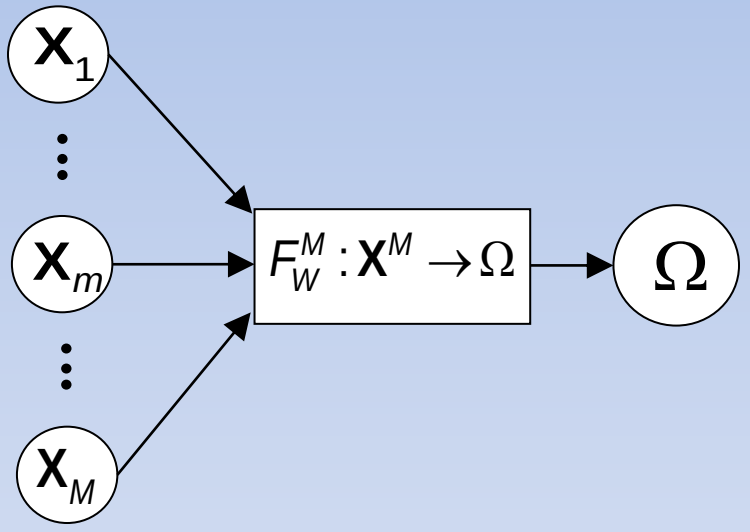
An ensemble of sources : $\mathbf{X}^M = (\mathbf{X}_1,...,\mathbf{X}_M)$ with weights $W = \{w_m > 0, m = 1,...,M\}$

Majority Voting Classifier



Voting decisions given by individual sourses *

General Measure Classifier



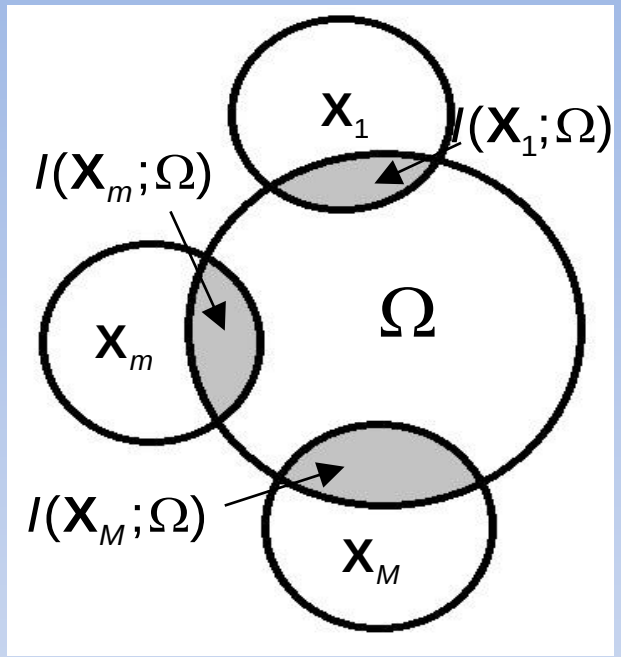
Combining the sources by general measure in the ensemble **

* L.I.Kuncheva. *Combining Pattern Classifiers* // Wiley and Sons, 2004

** M.M. Lange, S.N. Ganevnykh, A.M. Lange. *Multiclass Pattern* Re*cognition in a Space of Tree-structured Multiresolution* Re*presentations* // Machine Learning and Data Analysis, 2016, 2(1), c.70-88

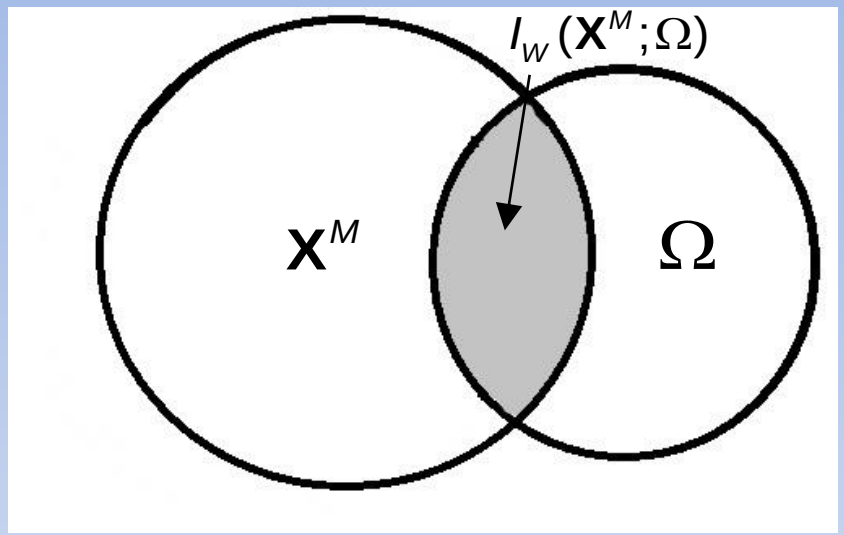
3. Average mutual informations of the classifiers

Majority Voting (MV classifier)



$$I_W^{\mathrm{MV}}(\mathbf{X}^M;\Omega) = \sum_{m=1}^{M} I(X_m;\Omega)\frac{w_m}{\sum_{m=1}^{M} w_m}$$

General Measure (GM classifier)



$$I_W^{\mathrm{GM}}(\mathbf{X}^M;\Omega) = \frac{1}{M}\, I_W(\mathbf{X}^M;\Omega)$$

Goal is to show inequality

$$\max_W I_W^{\mathrm{MV}}(\mathbf{X}^M;\Omega) \le I_{W^*}^{\mathrm{GM}}(\mathbf{X}^M;\Omega)\ , \qquad \text{where} \qquad W^* = \arg\max_W I_W^{\mathrm{MV}}(\mathbf{X}^M;\Omega)$$

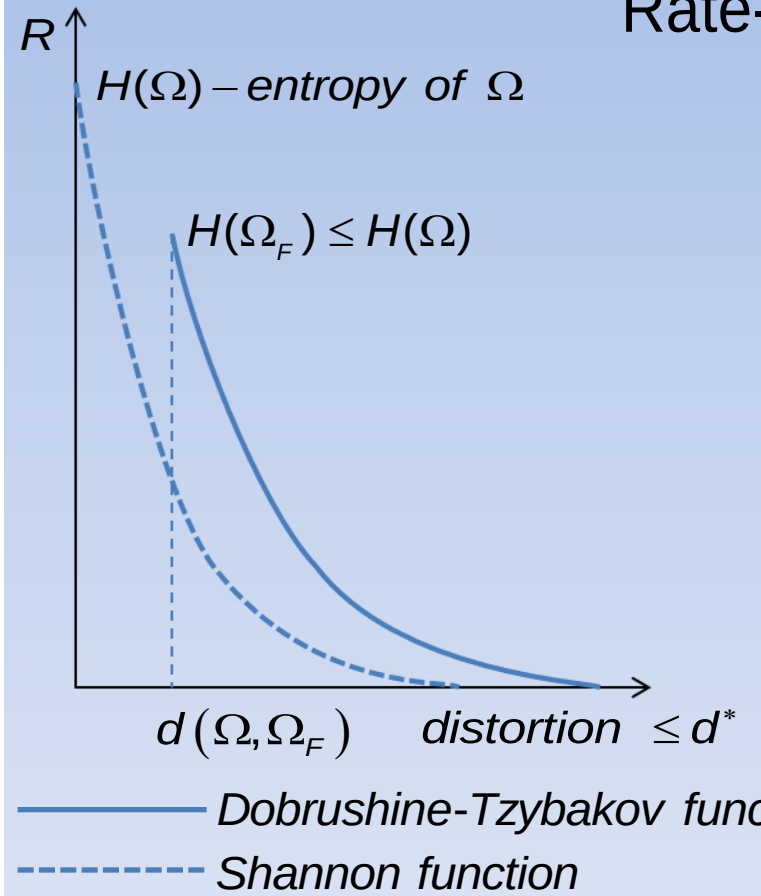
Information-Theoretic Models - In book Simon Haykin. *Neural Networks. A Comprehensive Foundation* // Practice Hall International, Inc., 1998

4. Connection with source coding problem

$$\Omega \Rightarrow \boxed{\text{noisy channel}} \Rightarrow \mathbf{X} \Rightarrow \boxed{\text{estimator}} \Rightarrow \Omega_F \Rightarrow \boxed{\text{coder}} \Rightarrow \hat{\Omega}_F$$

$\|\Omega_F\| = \|\Omega\|, \quad \|\hat{\Omega}_F\| \le \|\Omega_F\|$ noisy encoding (<) or noiseless encoding (=)

Rate-distortion function*



Average distortion:

$$d(\Omega, \hat{\Omega}_F) = \frac{1}{N} \underset{\Omega, \hat{\Omega}_F}{E} \sum_{n=1}^{N} [\omega(n) \neq \hat{\omega}(n)], \omega \in \Omega, \hat{\omega} \in \hat{\Omega}_F$$

$$d(\Omega, \hat{\Omega}_F) \le d(\Omega, \Omega_F) + d(\Omega_F, \hat{\Omega}_F)$$

Average mutual unformation : $I(\mathbf{X}; \hat{\Omega}_F) \ge 0$

Rate-distortion function for a given admissible distortion value $d^* \ge d(\Omega, \Omega_F)$:

$$R(d^*) = \min_{\hat{\Omega}_F} I(\mathbf{X}; \hat{\Omega}_F)$$

$$\hat{\Omega}_F : d(\Omega_F, \hat{\Omega}_F) \le d^* - d(\Omega, \Omega_F)$$

* R.L. Dobrushin, B.S.Tsybakov. *Information transmission with added noise //* Problems of information transmission , 1963, 14, *c*. 21 - 42 (in Russian)

5. Dissimirarity measures of objects

Measure in a set $\mathbf{X}_m$, $\mathbf{x}_m = (x_{m1},...,x_{mN_m}) \in \mathbf{X}_m$, $\hat{\mathbf{x}}_m = (\hat{x}_{m1},...,\hat{x}_{mN_m}) \in \mathbf{X}_m, m = 1,...,M$:

$$d(\mathbf{x}_m, \hat{\mathbf{x}}_m) = \sum_{n=1}^{N_m} \frac{(x_{mn} - \hat{x}_{mn})^2}{\sigma_{mn}^2} \ , \ 0 < \sigma_{mn}^2 < \infty$$

Measure in ensemble $\mathbf{X}^M = (\mathbf{X}_1,...,\mathbf{X}_M)$, $\mathbf{x}^M = (\mathbf{x}_1,...,\mathbf{x}_M) \in \mathbf{X}^M$, $\hat{\mathbf{x}}^M = (\hat{\mathbf{x}}_1,...,\hat{\mathbf{x}}_M) \in \mathbf{X}^M$:

$$D(\mathbf{x}^M, \hat{\mathbf{x}}^M) = \sum_{m=1}^{M} w_m d(\mathbf{x}_m, \hat{\mathbf{x}}_m) \ , \ w_m > 0$$

Kernels by the mesures $d(\mathbf{x}_m, \hat{\mathbf{x}}_m)$ and $D(\mathbf{x}^M, \hat{\mathbf{x}}^M)$:

simple kernel:
$$K^{(d)}(\mathbf{x}_m, \hat{\mathbf{x}}_m) = \frac{e^{-d(\mathbf{x}_m, \hat{\mathbf{x}}_m)}}{\int_{\mathbf{x}_m} e^{-d(\mathbf{x}_m, \hat{\mathbf{x}}_m)} d\mathbf{x}_m} \ , \qquad 0 < \int_{\mathbf{x}_m} e^{-d(\mathbf{x}_m, \hat{\mathbf{x}}_m)} d\mathbf{x}_m < \infty$$

$$K_W^{(D)}(\mathbf{x}^M, \hat{\mathbf{x}}^M) = \frac{e^{-D(\mathbf{x}^M, \hat{\mathbf{x}}^M)}}{\int_{\mathbf{x}^M} e^{-D(\mathbf{x}^M, \hat{\mathbf{x}}^M)} d\mathbf{x}^M} = \prod_{m=1}^{M} K_{w_m}^{(d)}(\mathbf{x}_m, \hat{\mathbf{x}}_m)$$

weighted kernel:
$$K_{w_m}^{(d)}(\mathbf{x}_m, \hat{\mathbf{x}}_m) = \frac{e^{-w_m d(\mathbf{x}_m, \hat{\mathbf{x}}_m)}}{\int_{\mathbf{x}_m} e^{-w_m d(\mathbf{x}_m, \hat{\mathbf{x}}_m)} d\mathbf{x}_m} \ , \qquad 0 < \int_{\mathbf{x}_m} e^{-w_m d(\mathbf{x}_m, \hat{\mathbf{x}}_m)} d\mathbf{x}_m < \infty$$

6. Class-conditional densities

Collections of template objects in classes (basic objects):

$$\mathbf{X}_{mi} = \{\mathbf{x}_{mik} \in \mathbf{X}_m, k = 1,...,L_{mi}\}, i = 1,...,c,$$

where L_{mi} is a number of the templates of the *m-th* sourse in the *i-th* class.

Class-conditional densities given by the kernel mixtures for the *m-th* sourse:

$$g(\mathbf{x}_m \mid \omega_i) = \sum_{k=1}^{L_{mi}} \theta_{mik} K^{(d)}(\mathbf{x}_m, \mathbf{x}_{mik}),\ i = 1,...,c,$$

where $\theta_{mik} > 0 : \sum_{k=1}^{L_{mi}} \theta_{mik} = 1$ are the weights of the template objects.

Class-conditional densities given by *M*-tuple mixtures for the ensemble:

$$g_W(\mathbf{x}^M \mid \omega_i) = \prod_{m=1}^{M} \sum_{k=1}^{L_{mi}} \theta_{mik} K_{w_m}^{(d)}(\mathbf{x}_m, \mathbf{x}_{mik}) = \prod_{m=1}^{M} g_{w_m}(\mathbf{x}_m \mid \omega_i), \quad i = 1,...,c,$$

where $g_{w_m}(\mathbf{x}_m \mid \omega_i)$ is the weighted density of the *i-th* class for the *m-th* sourse

7. Functional of the average mutual information*

A priori probability distribution in the set of classes Ω :

$$p(\omega_i): \sum_{i=1}^{c} p(\omega_i) = 1$$

The weighted density by w_m in the set $\mathbf{X}_m$, m=1,...,M :

$$p_{w_m}(\mathbf{x}_m) = \sum_{i=1}^{c} p(\omega_i)\, g_{w_m}(\mathbf{x}_m \mid \omega_i)$$

The average mutual information between the set of objects $\mathbf{X}_m$ and the set of classes Ω by the weighted densities $p_{w_m}(\mathbf{x}_m)$ and $g_{w_m}(\mathbf{x}_m \mid \omega_i)$:

$$I_{w_m}(\mathbf{X}_m; \Omega) = H_{w_m}(\mathbf{X}_m) - H_{w_m}(\mathbf{X}_m \mid \Omega)$$

Differential entropies by the densities $p_{w_m}(\mathbf{x}_m)$ and $g_{w_m}(\mathbf{x}_m \mid \omega_i)$:

$$H_{w_m}(\mathbf{X}_m) = -\frac{1}{N_m} \int_{\mathbf{x}_m} p_{w_m}(\mathbf{x}_m)\, \log p_{w_m}(\mathbf{x}_m) d\mathbf{x}_m$$

$$H_{w_m}(\mathbf{X}_m \mid \Omega) = -\frac{1}{N_m} \sum_{i=1}^{c} p(\omega_i) \int_{\mathbf{x}_m} g_{w_m}(\mathbf{x}_m \mid \omega_i)\, \log g_{w_m}(\mathbf{x}_m \mid \omega_i) d\mathbf{x}_m$$

* R.G. Gallager. *Information Theory and* Re*liable Communication* // Wiley and Sons, 1968.

8. The average mutual information for the classifiers

For MV classifier

$$I_W^{\mathrm{MV}}(\mathbf{X}^M;\Omega) = \sum_{m=1}^{M} I(\mathbf{X}_m;\Omega)\frac{w_m}{\sum_{m=1}^{M} w_m}$$

where $I(\mathbf{X}_m;\Omega) = I_{w_m=1}(\mathbf{X}_m;\Omega),\ m = 1,....,M$

For GM classifier

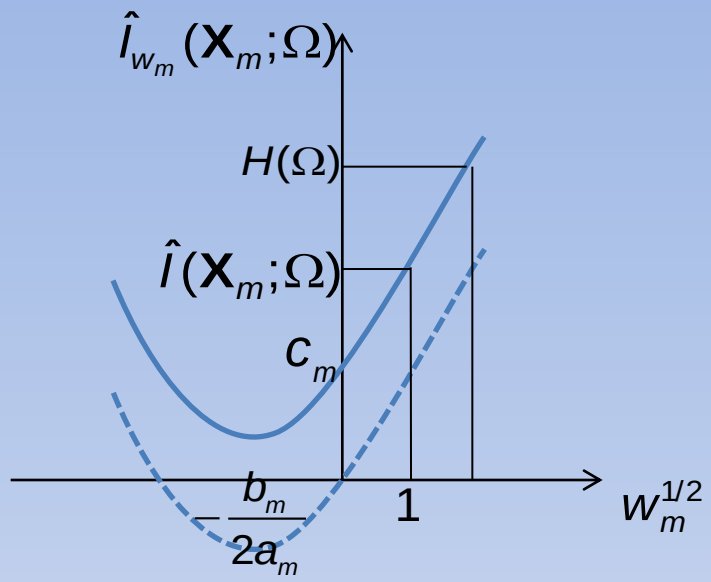
$$I_W^{\mathrm{GM}}(\mathbf{X}^M;\Omega) = \frac{1}{M}\sum_{m=1}^{M} I_{w_m}(\mathbf{X}_m;\Omega)$$

9. Estimations for the functionals of the mutual information

Estimations over the individual sources ($m = 1,\ldots,M$)

$$I_{w_m}(\mathbf{X}_m;\Omega) \le \hat{I}_{w_m}(\mathbf{X}_m;\Omega) = a_m w_m + b_m w_m^{1/2} + c_m$$

$$I(\mathbf{X}_m;\Omega) \le \hat{I}(\mathbf{X}_m;\Omega) = a_m + b_m + c_m$$



The coefficients $a_m > 0, b_m > 0$ and $c_m \ge 0$ depend on the parameters of the mixtures over the classes (collections of templates, dispersions, and weights of the mixture components)

Weights of the sources $w_m(s,v) = (1+v)^2 \, \mathrm{e}^{\,s \frac{\hat{I}(\mathbf{X}_m;\Omega)}{H(\Omega)}}$, $0 \le s \le 1$, $v \ge 0$

Estimations of the average mutual information over the ensemble of sources

For MV classifier: $$\hat{I}_s^{\mathrm{MV}}(\mathbf{X}^M;\Omega) = \sum_{m=1}^{M} \hat{I}(\mathbf{X}_m;\Omega) \frac{w_m(s,v)}{\sum_{m=1}^{M} w_m(s,v)}$$

For GM classifier: $$\hat{I}_{s,v}^{\mathrm{GM}}(\mathbf{X}^M;\Omega) = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_{s,v}(\mathbf{X}_m;\Omega)$$

10. Main result

Lemma. For $s \to 0$, there exists $s^* \geq 0$ that gives the maximal value

$$\max_s \hat{I}_s^{\mathrm{MV}}(\mathbf{X}^M;\Omega) = \hat{I}_{s^*}^{\mathrm{MV}}(\mathbf{X}^M;\Omega) \approx (\mu + s^*\Delta)H(\Omega),$$

where

$$\mu = \frac{1}{M}\sum_{m=1}^{M}\frac{\hat{I}(\mathbf{X}_m;\Omega)}{H(\Omega)} \quad , \quad \Delta = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{I}(\mathbf{X}_m;\Omega)}{H(\Omega)}\right)^2 - \left(\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{I}(\mathbf{X}_m;\Omega)}{H(\Omega)}\right)^2 \geq 0$$

and $s^* = 0$ when $\Delta = 0$ (for the same values of the source mutual informations).

Theorem. For $s \to 0$, the values $s^* \geq 0$ и $v^* \geq s^*\Delta\, H(\Omega) \Big/ \frac{1}{M}\sum_{m=1}^{M}(2a_m + b_m)$,

subject to $a_m w_m(s^*,v^*) + b_m w_m^{1/2}(s^*,v^*) + c_m \leq H(\Omega),\ m = 1,...,M,$

yield the inequality

$$\hat{I}_{s^*}^{\mathrm{MV}}(\mathbf{X}^M;\Omega) \leq \hat{I}_{s^*,v^*}^{\mathrm{GM}}(\mathbf{X}^M;\Omega),$$

which is the equality when $w_m(s^*,v^*) = 1$, $m = 1,...,M$.

11. Experimental results of face recognition via colour HSI images

Error rates and standard deviations

sources ensemble	error rate deviation	NN	MT	SVM
face: H	ε	0.015	0.008	0.006
	σ	0.005	0.006	0.003
face: S	ε	0.017	0.012	0.009
	σ	0.006	0.003	0.004
face: I	ε	0.022	0.012	0.017
	σ	0.006	0.005	0.006
face: HSI (GM)	ε	0.007	0.002	0.001
	σ	0.005	0.002	0.001
face: HSI (MV)	ε	0.010	0.005	0.005
	σ	0.005	0.003	0.005

Dataset :

sources H,S,I; 1000 objects for each source;

25 classes, 40 images per class for each source.

Scheme of experiment :

200 times, 2 fold cross-validation

Discriminant functions given by

NN - nearest neighbor

MT - mixture of templates

SVM - support vector machine

12. Final remarks

- For metric classification schemes in a given ensemble of sources, a comparative criteria of their efficiency in terms of average mutual information between the ensemble and a set of classes is suggested.

- Two schemes are investigated, namely the first is based on majority voting (MV) decisions over individual sources and the second one combines the sources using a general measure (GM) in the ensemble.

- In the frame of these schemes, the functionals of the average mutual information of the MV and GM classifiers are defined and their estimations are obtained.

- It is shown that the maximal average mutual information of the MV classifier does not exceed the average mutual information of the GM classifier subject to the same collections of the source weights.

- For a set of person faces given by color HSI images that yield an ensemble of three sources, experiments on face recognition were performed. For different discriminant functions, the results shown the less error rates of the GM classifier with respect to the error rates of the MV classifier.