A data classification model based on the average mutual information

Mikhail Lange, Andrey Lange

Federal Research Center «Computer Science and Control» of RAS,
Moscow, Russia

12-th International Conference IDP-2018, October 8-12, Gaeta, Italy

2. Statement of the problem

Main definitions

- $\Omega^c = \{\omega_i\}_{i=1}^c, c \geq 2$ is a set of classes of a given prior probability distribution $\mathrm{P} = \{P(\omega_i)\}_{i=1}^c$
- $\mathrm{X} = \{\mathrm{x}\}$ is a set of objects of a given classs-conditional probability densities $\mathrm{p} = \{p(\mathrm{x} \mid \omega_i)\}_{i=1}^c$
- $\Omega = \{\omega_j\}_{j=1}^c$ is a set of decisions on the objects of a conditional distribution $\mathrm{Q} = \{Q(\omega_j \mid \mathrm{x})\}_{i=1}^c$

Model for classification

- Scheme* $\Omega^c \Rightarrow \boxed{\text{observation channel p}} \Rightarrow \mathrm{X} \Rightarrow \boxed{\text{test-channel Q}} \Rightarrow \Omega$

- Average mutual information $I_{\mathrm{Q}}(\mathrm{X};\Omega) = \int_{\mathrm{X}} p(\mathrm{x})dx \sum_{j=1}^{c} Q(\omega_j \mid \mathrm{x}) \text{ In } \left(Q(\omega_j \mid \mathrm{x})/Q(\omega_j\right)$

- Average error rate $E_{\mathrm{Q}}(\mathrm{X},\Omega) = \int_{\mathrm{X}} p(\mathrm{x})dx \sum_{j=1}^{c} Q(\omega_j \mid \mathrm{x}) \sum_{i=1}^{c} P(\omega_i \mid \mathrm{x})[\omega_i \neq \omega_j]$

- Given $\varepsilon > 0$, the rate-distortion function $R(\varepsilon) = \min_{\mathrm{Q}:\, E_{\mathrm{Q}}(\mathrm{X},\Omega) \leq \varepsilon} I_{\mathrm{Q}}(\mathrm{X};\Omega)$

The goal is to calculate a lower bound $R_L(\varepsilon)$ to the rate - distortion function $R(\varepsilon)$

* Dobrushin R.L, Tsybakov B.S. Information transmission with additional noise / / *IRE Trans. Information Theory*, 1962, **8**(5), 293 – 304

3. Basic lower bounds to the functions $R(\varepsilon)$

[*]**Lower bound in the Shannon's scheme** $\Omega^c \to \Omega$ is given by the form

$$R_L(\varepsilon) = H(\Omega^c) - H(\Omega \mid \Omega^c),$$

where $H(\Omega^c) = -\sum_{i=1}^{c} P(\omega_i) \ln P(\omega_i)$ is the entropy of the set Ω^c,

$H(\Omega \mid \Omega^c) = -\sum_{i=1}^{c} P(\omega_i) \sum_{j=1}^{c} Q(\omega_j \mid \omega_i) \ln Q(\omega_j \mid \omega_i)$ is the conditional entropy of Ω subject to Ω^c

and $Q(\omega_j \mid \omega_i) = \begin{cases} \varepsilon/(c-1), & i \neq j \\ 1-\varepsilon, & i = j \end{cases}$ is the test-channel between Ω^c and Ω

Lower bound in the Dobrushin - Tsybakov's scheme $\Omega^c \to \mathrm{X} \to \Omega$ is calculated in the equivalent scheme $\Omega^c \to \mathrm{X} \to \Omega^* \to \Omega$, where $\mathrm{X} \to \Omega^*$ yields the error rate $\varepsilon_{\min} = 1 - \int_{\mathrm{X}} \left(\max_{i=1}^{c} P(\omega_i \mid \mathrm{x}) \right) p(\mathrm{x}) d\mathrm{x}$

Then the Shannon's bound $R_L(\varepsilon) = H(\Omega^*) - H(\Omega \mid \Omega^*),$

the test-channel $Q(\omega_j \mid \omega_k) = \begin{cases} (\varepsilon - \varepsilon_{\min})/(c-1), & k \neq j \\ 1-(\varepsilon - \varepsilon_{\min}), & k = j \end{cases}$ between Ω^* and Ω ,

and the inequality $H(\Omega^*) \geq I(\mathrm{X};\Omega^c)$ yield **the final lower bound as follows**

$$R_L(\varepsilon) = I(\mathrm{X};\Omega^c) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min}) \ln(c-1) ,$$

where $I(\mathrm{X};\Omega^c) = H(\Omega^c) - H(\Omega^c \mid \mathrm{X})$ is the average mutual information between X and Ω^c

and $h(z) = -z \ln z - (1-z)\ln(1-z)$ is the Shannon's binary entropy

[*] Gallager R.G. Information Theory and Reliable Communication / / *Wiley and Sons*, 1968.

4. Generalization for the scheme with reject

Main definitions

- $\Omega^{c+1} = \Omega^c \cup \omega_0, c \geq 1$ is a set of classes, where $\omega_0 \notin \Omega^c$ is a class for reject
- $\theta = P(\Omega^c)$ and $(1\text{-}\theta) = P(\omega_0)$ are unknown probabilities
- $\mathrm{X} = \{\mathrm{x}\}$ is a set of objects of the given classs-conditional probability densities $\mathrm{p} = \{p(\mathrm{x} \mid \omega_i)\}_{i=1}^{c}$
- $\Omega = \{\omega_j\}_{j=0}^{c}$ is a set of decisions on the objects
- $\mathrm{P} = \{P(\omega_i \mid \Omega^c)\}_{i=1}^{c}$ is a given prior distribution in the set of the positive classes Ω^c
- $\mathrm{Q} = \{Q(\omega_j \mid \mathrm{x}, \Omega^c)\}_{i=1}^{c}$ is a test-channel of the decisions in $\Omega^c \subset \Omega$ on a submitted object $\mathrm{x} \in \mathrm{X}$

Under some assumptions, the scheme $\Omega^{c+1} \to \mathrm{X} \to \Omega$ yields the following characteristics

- Average mutual information is given by $I_{\mathrm{Q},\theta}(\mathrm{X};\Omega) = \theta I_{\mathrm{Q}}(\mathrm{X};\Omega^c)$
- Average error rate is given by $E_{\mathrm{Q},\theta}(\mathrm{X},\Omega) = \theta^2 E_{\mathrm{Q}}(\mathrm{X},\Omega^c) + 2\theta(1-\theta)$
- Given $\varepsilon_\theta > 0,$ the rate-distortion function is defined by $\tilde{R}(\varepsilon_\theta) = \min\limits_{\mathrm{Q}:\, E_{\mathrm{Q},\theta}(\mathrm{X},\Omega) \leq \varepsilon_\theta} I_{\mathrm{Q},\theta}(\mathrm{X};\Omega)$

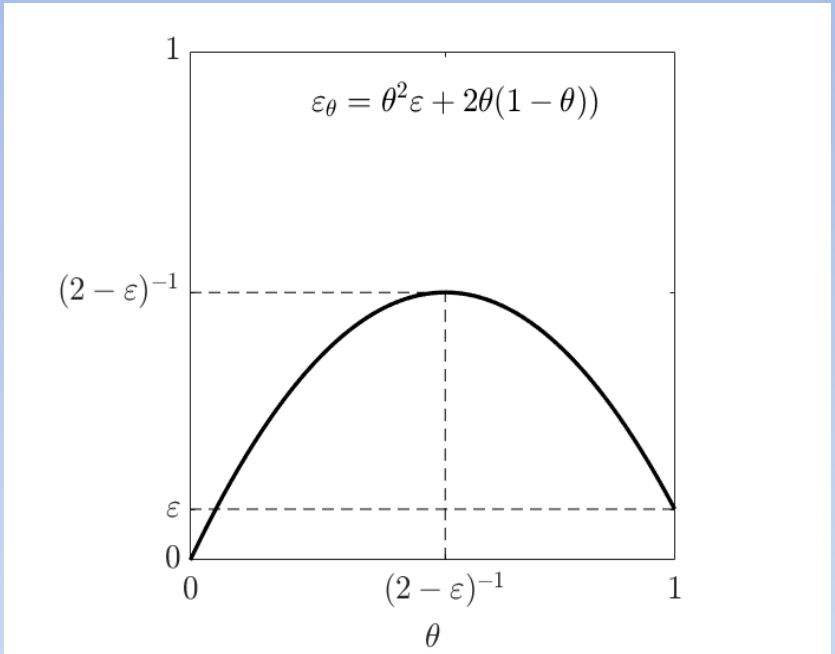
For $\theta \to 1$, there are valid the relations

$$\begin{cases} \tilde{R}(\varepsilon_\theta) = \theta R(\varepsilon) \\ \varepsilon_\theta = \theta^2 \varepsilon + 2\theta(1-\theta) \end{cases}$$

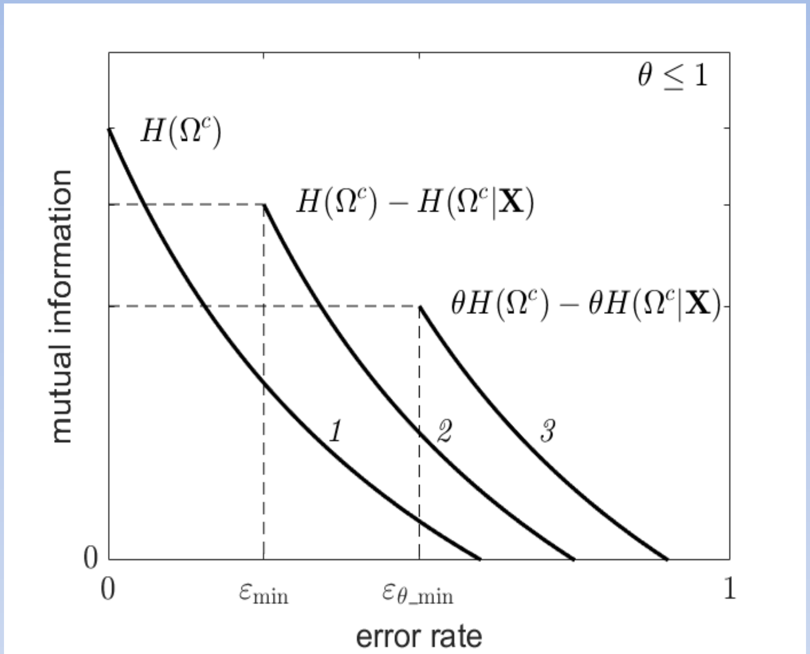
subject to the prior distribution P and the test-channel Q in the set of the positive classes are invariant for the schemes without and with reject, respectively.

5. Graphical interpretation

Error rates



Lower bounds



1. The Shannon's bound in the coding scheme $\Omega^c \to \Omega$
2. The bound in the classification scheme without reject $\Omega^c \to X \to \Omega$
3. The bound in the classification scheme with reject $\Omega^{c+1} \to X \to \Omega$ for $\theta \to 1$

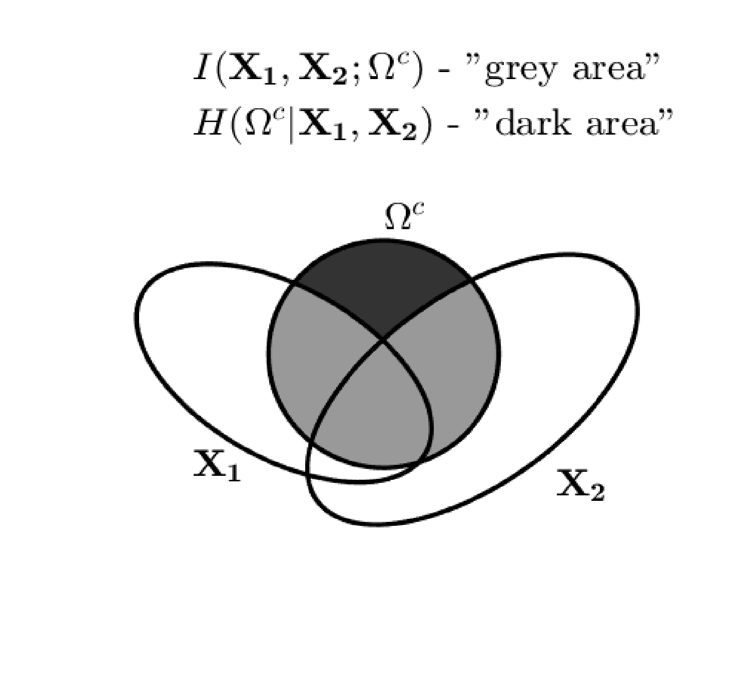
Given the entropy $H(\Omega^c)$, **the goal is to reduce the error rate** $\varepsilon_{\min}$ by decreasing the conditional entropy $H(\Omega^c \mid X)$ that is equivalent to increasing the average mutual information $I(X;\Omega^c) = H(\Omega^c) - H(\Omega^c \mid X)$

6. Ensemble of data sources

Let $X^M = (X_1,\dots,X_M)$ be an ensemble of the objects taken from M sources and $I(X^M;\Omega^c) = H(\Omega^c) - H(\Omega^c \mid X_1,\dots,X_M)$ be the average mutual information between X^M and Ω^c, where $H(\Omega^c \mid X_1,\dots,X_M) \le H(\Omega^c \mid X_m)$, $m = 1,\dots,M$

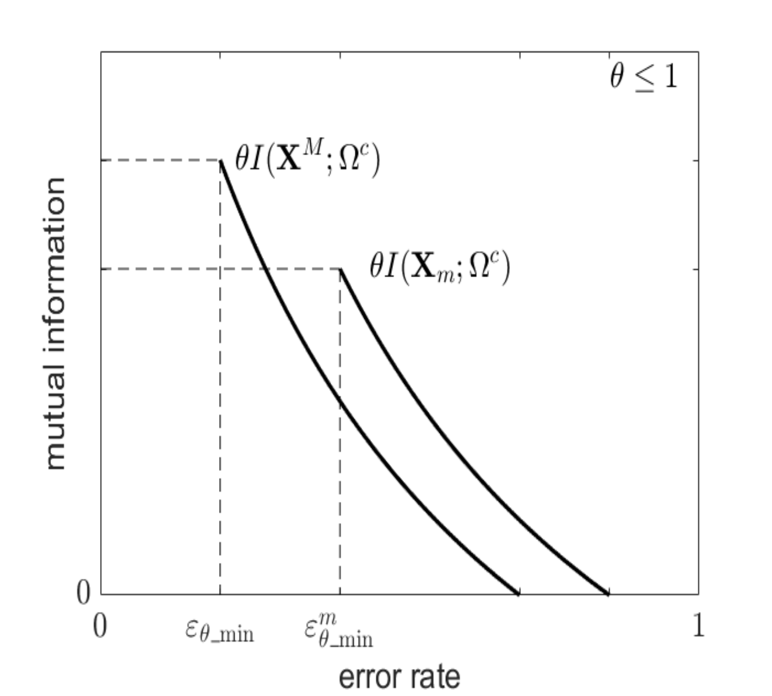
For the decorrelated sources, the last inequalities are strict $(<)$ and $I(X^M;\Omega^c) > \max\limits_{m=1}^{M} I(X_m;\Omega^c)$

Exsample of the ensemble of size $M = 2$



$I(\mathbf{X_1},\mathbf{X_2};\Omega^c)$ - "grey area"
$H(\Omega^c|\mathbf{X_1},\mathbf{X_2})$ - "dark area"
Ω^c
$\mathbf{X_1}$
$\mathbf{X_2}$

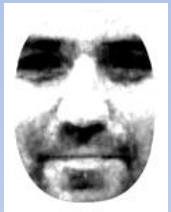
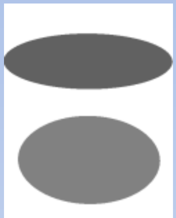
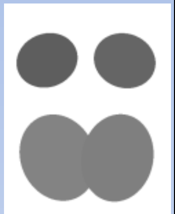
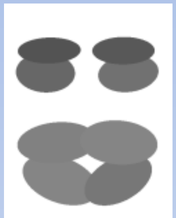
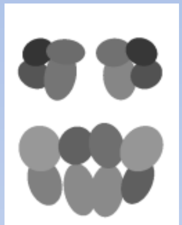
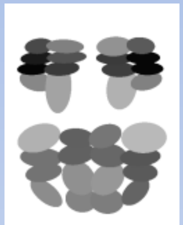
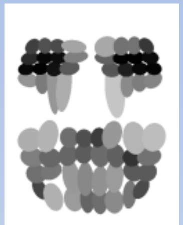
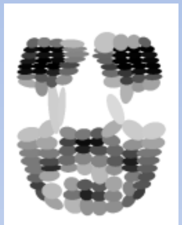
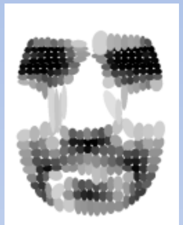
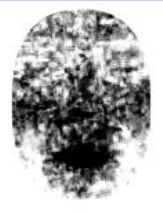
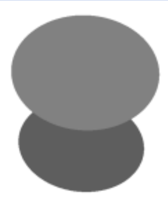
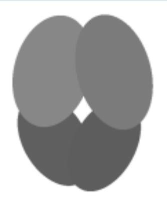
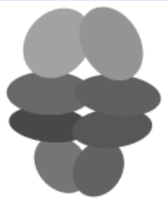
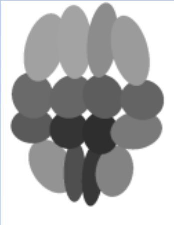
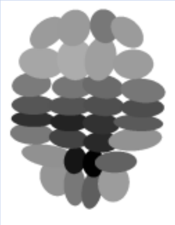
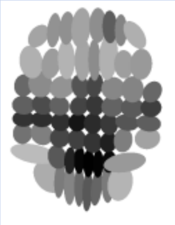
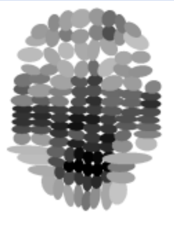
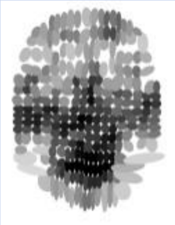
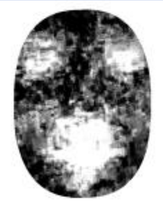
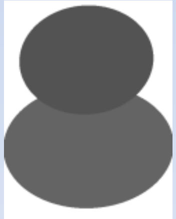
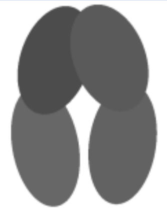
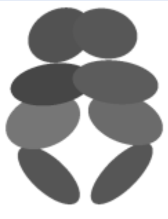
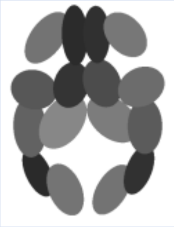
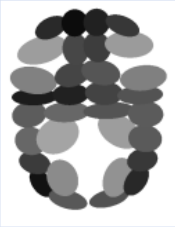
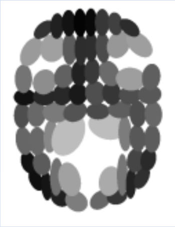
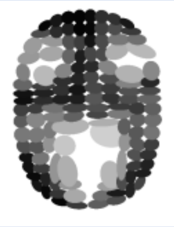
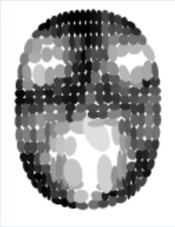
The curves of the lower bounds



7. Ensemble of tree-structured representations for decorrelated components of RGB face images*

Sources of images: $\mathbf{X}_1=\{\mathrm{I}_1(\mathrm{RGB})\}$, $\mathbf{X}_2=\{\mathrm{I}_2(\mathrm{RGB})\}$, $\mathbf{X}_3=\{\mathrm{I}_3(\mathrm{RGB})\}$

Ensemble of the sources: $\mathbf{X}^3=(\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3)$; **number of representation levels**: 8

								
I_1(R,G,B)	***l*=1**	***l*=2**	***l*=3**	***l*=4**	***l*=5**	***l*=6**	***l*=7**	***l*=8**
								
I_2(R,G,B)	***l*=1**	***l*=2**	***l*=3**	***l*=4**	***l*=5**	***l*=6**	***l*=7**	***l*=8**
								
I_3(R,G,B)	***l*=1**	***l*=2**	***l*=3**	***l*=4**	***l*=5**	***l*=6**	***l*=7**	***l*=8**

*Collection of color faces.URL: http://sourceforge.net/projects/colorfaces/ (accessed 04.07.2017)

8. Experimental results

For the multiclass NN and SVM classifiers, two fusion schemes have been investigated, namely:

WMV scheme by weighted majority voting the decisions for the objects $I_1 \in X_1$, $I_2 \in X_2$ and $I_3 \in X_3$;
GDM scheme by making a group decision for any submitted composite object $(I_1, I_2, I_3) \in X^3$
using the discriminant functions by the general dissimilarity measure in the ensemble of the sources.

Error rates without reject $(\theta = 1)$

	Sources			Fusion	
	X_1	**X_2**	**X_3**	WMV	GDM
NN	0,0042	0,0010	0,0016	0,00004	0,00002
SVM	0,0024	0,0027	0,0025	0,00020	0,00001

Error rates with reject $(\theta = 0{,}5)$

	Sources			Fusion	
	X_1	**X_2**	**X_3**	WMV	GDM
NN	0,0055	0,0145	0,0295	0,0035	0,0030
SVM	0,0190	0,0500	0,0540	0,0080	0,0025

The error rates in the ensemble are smaller than the error rates in the individual sources.

GDM fusion scheme yields a profit in classification fidelity with respect to WMV scheme.

9. Conclusion

- The minimal average mutual information between a set of the objects and a set of the classes has been defined as a function of a given admissible error rate.

- This function is similar to the rate-distortion function for the known source coding scheme with a given fidelity in a presence of the "noisy" observation channel.

- The lower bounds to the rate-distortion functions have been obtained for two classification schemes by making the dicision without or with a reject option.

- The obtained bounds show a possibility of reducing the error rate by increasing the average mutual information in the ensemble of data sources.

- The theoretical results have been supported by experimental error rates for face recognition without and with reject using the decorrelated components of RGB images and the ensemble of the components.

- Two fusion schemes have been investigated, namely the known WMV scheme based on Weighted Majority Vote of the decisions on the individual components and the new GDM scheme, which makes the group decision on any composite object using the General Dissimilarity Measure in the ensemble of the components.

- The experimental results show a profit in error rate for the GDM fusion scheme as compared with the WMV fusion scheme.