Impact of contamination on empirical and theoretical error rates in classification

Ch. Ruwet, and G. Haesbroeck

Mathematics Department - University of Liège - Belgium

ICORS 2009

Outline

Outline

Statistical cluster analysis

Suppose

$X \sim F$ arises from G_1 or G_2 with $\pi_i(F) = \mathbb{P}_F[X \in G_i]$

then

F is a mixture of two distributions

$$F = \pi_1(F) F_1 + \pi_2(F) F_2$$

with density $f = \pi_1(F) f_1 + \pi_2(F) f_2$.

Additional assumption : one dimension !

The generalized 2-means clustering method

- Aim of clustering : Find estimations $C_1(F)$ and $C_2(F)$ of the two underlying groups.
- The clusters' centers $(T_1(F), T_2(F))$ are solutions of

$$\min_{\{t_1,t_2\}\subset\mathbb{R}} \int \Omega\left(\inf_{1\leq j\leq 2} |x - t_j|\right) dF(x)$$

for a suitable strictly increasing penalty function $\Omega : \mathbb{R}^+ \to \mathbb{R}^+$.

- Classical penalty functions :

$$\Omega(x) = x^2 \to \text{ 2-means method}$$
$$\Omega(x) = x \to \text{ 2-medoids method}$$

Classification rule

- The classification rule is

$$R_F(x) = C_j(F) \Leftrightarrow \Omega(|x - T_j(F)|) = \min_{1 \leq i \leq 2} \Omega(|x - T_i(F)|)$$

- The clusters are simply :

$$C_1(F) =]-\infty, C(F)[$$

$$C_2(F) =]C(F), +\infty[$$

where $C(F) = \dfrac{T_1(F) + T_2(F)}{2}$ is the cut-off point.

- $T_1(F)$ and $T_2(F)$ are the generalized Ω-means of the corresponding clusters.

Outline

Optimality in classification

- The error rate is defined as the probability to misclassify data ;
- A classification rule is optimal if the corresponding error rate is minimal ;
- The optimal classification rule is the Bayes rule (BR) :

$$x \in C_1 \Leftrightarrow \pi_1(F)f_1(x) > \pi_2(F)f_2(x)$$

  (Anderson, 1958) ;
- The 2-means procedure is optimal under the model

$$F_N = 0.5\,N(\mu_1, \sigma^2) + 0.5\,N(\mu_2, \sigma^2) \text{ with } \mu_1 < \mu_2$$

  (Qiu and Tamhane, 2007).

Simulation settings (1)

- $F_N = \pi_1 \, N(-\mu, 1) + (1 - \pi_1) \, N(\mu, 1)$;
- $m = 1000$ simulations ;
- Samples of size $n \Rightarrow {T_1}^k, {T_2}^k, \text{EER}^k \; (k = 1, \ldots, m)$

$$\Rightarrow \qquad \overline{\text{EER}} = \frac{1}{m} \sum_{k=1}^{m} \text{EER}^k \, ;$$

- $F_\varepsilon = (1 - \varepsilon) F_N + \varepsilon \Delta_x$ with $\varepsilon = 0.01$ and x coming from G_1.

Simulation results for $\pi_1 = 0.5$ (1)

μ	x	ER of BR	n	$\overline{\text{EER}}$ 0%	$\overline{\text{EER}}$ 1%
1	-4	0.1587	100	0.1618	0.1607
			500	0.1590	0.1579
			1000	0.1587	0.1574
1.5	-5	0.0668	100	0.0678	0.0676
			500	0.0676	0.0669
			1000	0.0671	0.0666

Simulation results for $\pi_1 = 0.5$ (1)

μ	x	ER of BR	n	$\overline{\text{EER}}$ 0%	$\overline{\text{EER}}$ 1%
1	-4	0.1587	100	0.1618	0.1607
			500	0.1590	0.1579
			1000	0.1587	0.1574
1.5	-5	0.0668	100	0.0678	0.0676
			500	0.0676	0.0669
			1000	0.0671	0.0666

Simulation settings (2)

- $F_N = \pi_1 \, N(-\mu, 1) + (1 - \pi_1) \, N(\mu, 1)$;
- $m = 1000$ simulations ;
- Training samples of size $n \Rightarrow {T_1}^k, {T_2}^k, \text{EER}^k$ $(k = 1, \dots, m)$;
- $F_\varepsilon = (1 - \varepsilon)F_N + \varepsilon\Delta_x$ with $\varepsilon = 0.01$ and x coming from G_1 ;
- Test sample of size $N = 100000 \Rightarrow \text{TER}^k$ $(k = 1, \dots, m)$

$$\Rightarrow \quad \overline{\text{TER}} = \frac{1}{m} \sum_{k=1}^{m} \text{TER}^k .$$

Simulation results for $\pi_1 = 0.5$ (2)

μ	x	ER of BR	n	$\overline{\text{TER}}$ 0%	$\overline{\text{TER}}$ 1%
1	-4	0.1587	100	0.1625	0.1632
			500	0.1595	0.1597
			1000	0.1604	0.1611
1.5	-5	0.0668	100	0.0697	0.0702
			500	0.0676	0.0678
			1000	0.0669	0.0672

Formal definitions

- Theoretical error rate : TER
  - Training sample according to F : estimation of the rule
  - Test sample according to F_m : evaluation of the rule
  - In ideal circumstances : $F = F_m$

$$\text{TER}(F, F_m) = \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m} \left[R_F(X) \neq C_j(F) | \, G_j \right]$$

- Empirical error rate : EER
  - Training sample according to F : estimation and evaluation of the rule

$$\text{EER}(F, F) = \sum_{j=1}^{2} \pi_j(F) \mathbb{P}_{F} \left[R_F(X) \neq C_j(F) | \, G_j \right]$$

Formal definitions

- Theoretical error rate : TER
  - Training sample according to F : estimation of the rule
  - Test sample according to F_m : evaluation of the rule
  - In ideal circumstances : $F = F_m$

$$\mathrm{TER}(F, F_m) = \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m}\left[R_F(X) \neq C_j(F) \middle| G_j\right]$$

- Empirical error rate : EER
  - Training sample according to F : estimation and evaluation of the rule

$$\mathrm{EER}(F, F) = \sum_{j=1}^{2} \pi_j(F) \mathbb{P}_{F}\left[R_F(X) \neq C_j(F) \middle| G_j\right]$$

Formal definitions

- Theoretical error rate : TER
  - Training sample according to F : estimation of the rule
  - Test sample according to F_m : evaluation of the rule
  - In ideal circumstances : $F = F_m$

$$\mathrm{TER}(F, F_m) = \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m}\left[R_F(X) \neq C_j(F) \mid G_j\right]$$

- Empirical error rate : EER
  - Training sample according to F : estimation and evaluation of the rule

$$\mathrm{EER}(F, F) = \sum_{j=1}^{2} \pi_j(F) \mathbb{P}_{F}\left[R_F(X) \neq C_j(F) \mid G_j\right]$$

In ideal circumstances, TER = EER.

Under contamination (1)

Now, the training sample is contaminated by a mass ε at the point x :

$$F \to F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$$

- Theoretical error rate :

$$\text{TER}(F_\varepsilon, F_m) = \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right]$$

- Empirical error rate :

$$\text{EER}(F_\varepsilon, F_\varepsilon) = \sum_{j=1}^{2} \pi_j(F_\varepsilon)\mathbb{P}_{F_\varepsilon}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right]$$

Under contamination (2)

Under $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, one has

$$\begin{aligned}\text{TER}(F_\varepsilon, F_m) &= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right] \\ &= \pi_1(F_m)\left\{1 - F_{m,1}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_m)F_{m,2}\left(C(F_\varepsilon)\right)\end{aligned}$$

Under contamination (2)

Under $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, one has

$$\begin{aligned}\mathrm{TER}(F_\varepsilon, F_m) &= \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right] \\ &= \pi_1(F_m)\left\{1 - F_{m,1}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_m)F_{m,2}\left(C(F_\varepsilon)\right)\end{aligned}$$

$$\begin{aligned}\mathrm{EER}(F_\varepsilon, F_\varepsilon) &= \sum_{j=1}^{2} \pi_j(F_\varepsilon)\mathbb{P}_{F_\varepsilon}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right] \\ &= \pi_1(F_\varepsilon)\left\{1 - F_{1,\varepsilon}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_\varepsilon)F_{2,\varepsilon}\left(C(F_\varepsilon)\right)\end{aligned}$$

Under contamination (2)

Under $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, one has

$$\begin{aligned} \text{TER}(F_\varepsilon, F_m) &= \sum_{j=1}^{2} \pi_j(F_m) \mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right] \\ &= \pi_1(F_m)\left\{1 - F_{m,1}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_m) F_{m,2}\left(C(F_\varepsilon)\right) \end{aligned}$$

$$\begin{aligned} \text{EER}(F_\varepsilon, F_\varepsilon) &= \sum_{j=1}^{2} \pi_j(F_\varepsilon) \mathbb{P}_{F_\varepsilon}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon) \middle| G_j\right] \\ &= \pi_1(F_\varepsilon)\left\{1 - F_{1,\varepsilon}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_\varepsilon) F_{2,\varepsilon}\left(C(F_\varepsilon)\right) \end{aligned}$$

$\pi_i(F_\varepsilon) = ?$ and $F_{i,\varepsilon} = ?$

Under $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, one has

- $\pi_i(F_\varepsilon) = \mathbb{P}_{F_\varepsilon}[X \in G_i] = (1-\varepsilon)\pi_i(F) + \varepsilon I\{x \in G_i\}$
- $F_{i,\varepsilon} = \left(1 - \frac{\varepsilon I\{x \in G_i\}}{\pi_i(F_\varepsilon)}\right) F_i + \frac{\varepsilon I\{x \in G_i\}}{\pi_i(F_\varepsilon)} \Delta_x$

$$\Rightarrow F_\varepsilon = \pi_1(F_\varepsilon)F_{1,\varepsilon} + \pi_2(F_\varepsilon)F_{2,\varepsilon}$$

$\pi_i(F_\varepsilon) = ?$ and $F_{i,\varepsilon} = ?$

Under $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, one has

- $\pi_i(F_\varepsilon) = \mathbb{P}_{F_\varepsilon}[X \in G_i] = (1-\varepsilon)\pi_i(F) + \varepsilon \mathrm{I}\{x \in G_i\}$
- $F_{i,\varepsilon} = \left(1 - \dfrac{\varepsilon \mathrm{I}\{x \in G_i\}}{\pi_i(F_\varepsilon)}\right) F_i + \dfrac{\varepsilon \mathrm{I}\{x \in G_i\}}{\pi_i(F_\varepsilon)} \Delta_x$

$$\Rightarrow F_\varepsilon = \pi_1(F_\varepsilon)F_{1,\varepsilon} + \pi_2(F_\varepsilon)F_{2,\varepsilon}$$

$\pi_i(F_\varepsilon) = ?$ and $F_{i,\varepsilon} = ?$

Under $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, one has

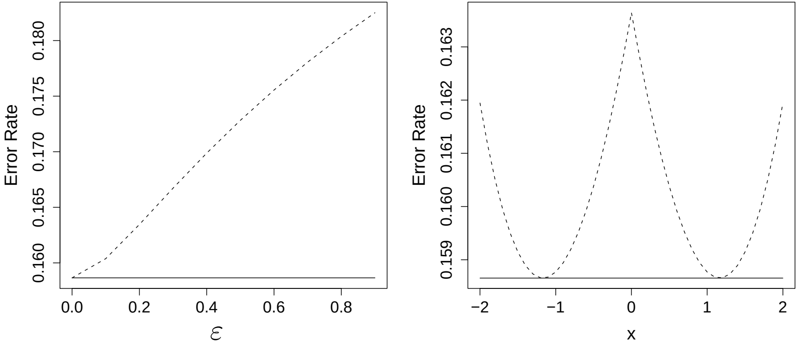
- $\pi_i(F_\varepsilon) = \mathbb{P}_{F_\varepsilon}[X \in G_i] = (1-\varepsilon)\pi_i(F) + \varepsilon \mathrm{I}\{x \in G_i\}$
- $F_{i,\varepsilon} = \left(1 - \frac{\varepsilon \mathrm{I}\{x \in G_i\}}{\pi_i(F_\varepsilon)}\right) F_i + \frac{\varepsilon \mathrm{I}\{x \in G_i\}}{\pi_i(F_\varepsilon)} \Delta_x$

$$\Rightarrow F_\varepsilon = \pi_1(F_\varepsilon)F_{1,\varepsilon} + \pi_2(F_\varepsilon)F_{2,\varepsilon}$$

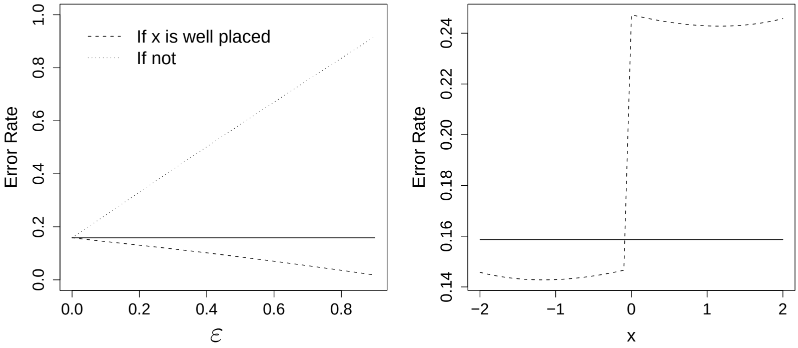
Graphs of TER and EER under contamination

- $F_m = F_N \equiv 0.5\,N(-1,1) + 0.5\,N(1,1)$ an optimal model ;
- Error rate of the Bayes rule : 0.1587 ;
- The 2-means procedure ;
- $C(F_N) = \frac{-1+1}{2} = 0$;
- $F_\varepsilon = (1-\varepsilon)F_m + \varepsilon\Delta_x$;
- $x = -0.5$ and ε varying ;
- $\varepsilon = 0.1$ and $x \in G_1$ varying.

Theoretical error rate under contamination



Empirical error rate under contamination



Outline

TER vs EER : Influence function

$$\text{TER}(F_\varepsilon, F) \approx \text{TER}(F, F) + \varepsilon \text{IF}(x; \text{TER}, F)$$

$$\text{EER}(F_\varepsilon, F_\varepsilon) \approx \text{EER}(F, F) + \varepsilon \text{IF}(x; \text{EER}, F)$$

where $\text{IF}(x; \text{ER}, F) = \frac{\partial}{\partial \varepsilon} \text{ER}((1-\varepsilon)F + \varepsilon \Delta_x)\Big|_{\varepsilon=0}$

(under condition of existence).

- Theoretical error rate :

$$\text{TER}(F_\varepsilon, F_N) \geq \text{TER}(F_N, F_N) \Rightarrow \text{IF}(x; \text{TER}, F_N) \equiv 0$$

- Empirical error rate : The IF of EER does not vanish!

TER vs EER : Influence function

$$\mathrm{TER}(F_\varepsilon, F) \approx \mathrm{TER}(F, F) + \varepsilon \mathrm{IF}(x; \mathrm{TER}, F)$$

$$\mathrm{EER}(F_\varepsilon, F_\varepsilon) \approx \mathrm{EER}(F, F) + \varepsilon \mathrm{IF}(x; \mathrm{EER}, F)$$

where $\mathrm{IF}(x; \mathrm{ER}, F) = \frac{\partial}{\partial \varepsilon} \mathrm{ER}((1-\varepsilon)F + \varepsilon \Delta_x)\Big|_{\varepsilon=0}$

(under condition of existence).

- Theoretical error rate :

$$\mathrm{TER}(F_\varepsilon, F_N) \geq \mathrm{TER}(F_N, F_N) \Rightarrow \mathrm{IF}(x; \mathrm{TER}, F_N) \equiv 0$$

- Empirical error rate : The IF of EER does not vanish!

TER vs EER : Influence function

$$\mathrm{TER}(F_\varepsilon, F) \approx \mathrm{TER}(F, F) + \varepsilon \mathrm{IF}(x; \mathrm{TER}, F)$$

$$\mathrm{EER}(F_\varepsilon, F_\varepsilon) \approx \mathrm{EER}(F, F) + \varepsilon \mathrm{IF}(x; \mathrm{EER}, F)$$

where $\mathrm{IF}(x; \mathrm{ER}, F) = \frac{\partial}{\partial \varepsilon} \mathrm{ER}((1-\varepsilon)F + \varepsilon \Delta_x)\Big|_{\varepsilon=0}$

(under condition of existence).

- Theoretical error rate :

$$\mathrm{TER}(F_\varepsilon, F_N) \geq \mathrm{TER}(F_N, F_N) \Rightarrow \mathrm{IF}(x; \mathrm{TER}, F_N) \equiv 0$$

- Empirical error rate : The IF of EER does not vanish!

Influence function of the empirical error rate

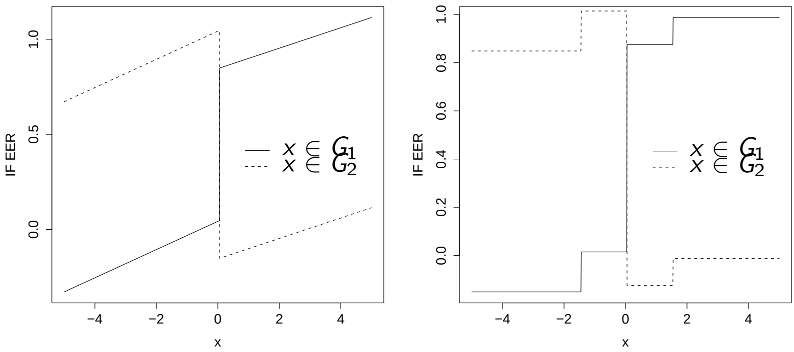
Proposition

For all $x \neq C(F)$,

$$\begin{aligned}\mathrm{IF}(x; \mathrm{EER}, F) = & -\mathrm{EER}(F, F) + \mathrm{I}\{x \in G_1\} \\ & + \mathrm{I}\{x \leq C(F)\}(1 - 2\,\mathrm{I}\{x \in G_1\}) \\ & + \frac{1}{2}(\mathrm{IF}(x; T_1, F) + \mathrm{IF}(x; T_2, F)) \\ & \{\pi_2(F) f_2(C(F)) - \pi_1(F) f_1(C(F))\}.\end{aligned}$$

Expressions of $\mathrm{IF}(x; T_1, F)$ and $\mathrm{IF}(x; T_2, F)$ were computed by García-Escudero and Gordaliza (1999).

Representation of the IF of the EER



IF of the EER under the optimal model

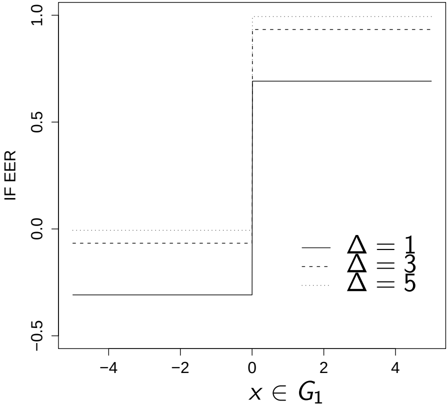
For all $x \neq C(F_N)$,

$$\begin{aligned} \mathrm{IF}(x; \mathrm{EER}, F_N) &= -\mathrm{EER}(F_N, F_N) + \mathrm{I}\{x \in G_1\} \\ &\quad + \mathrm{I}\{x \leq C(F_N)\}(1 - 2\,\mathrm{I}\{x \in G_1\}) \\ &= \begin{cases} \Phi(-\mu_1) - \mathrm{I}\{x < 0\} & \text{if } x \in G_1 \\ \mathrm{I}\{x < 0\} - \Phi(-\mu_2) & \text{if } x \in G_2 \end{cases} \end{aligned}$$

where Φ denotes the standard normal cumulative distribution function.

Representation under optimal model

$$F_N = 0.5\,N(-\Delta/2, 1) + 0.5\,N(\Delta/2, 1)$$



Outline

1. Classification based on clustering
2. Theoretical error rate vs empirical error rate
3. Influence function of the error rates
4. **Conclusions and future researches**

Conclusions

Under optimal generalized 2-means clustering rule,

- when working with a single sample, contamination may improve the quality of the clustering rule;
- when working with two samples, contamination make always the error rate on the test sample increase;

BUT when working with two samples, the property of the clusters'centers obtained by a generalized 2-means procedure is not true anymore on the test sample.

Future researches

- More than 1 dimension (work in progress) and more than 2 groups.
- Generalized trimmed 2-means : for $\alpha \in [0, 1]$, $(T_1(F), T_2(F))$ are solution of

$$\min_{\{A:F(A)=1-\alpha\}} \min_{\{t_1,t_2\}\subset\mathbb{R}} \int_A \Omega\left(\inf_{1\leq j\leq 2} |x - t_j|\right) dF(x)$$

(Cuesta-Albertos, Gordaliza, and Matrán, 1997).

- Nondecreasing penalty function, leading to a trimming procedure because observations far away from the two clusters'centers have the same Ω-distance from the centers.

Thank you for your attention!

Bibliography

- Croux C., Filzmoser P. and Joossens K. (2008), Classification efficiences for robust linear discriminant analysis, *Statistica Sinica* 18, pp. 581-599.
- Croux C., Haesbroeck G. and Joossens K. (2008), Logistic discrimination using robust estimators : an influence function approach, *The Canadian Journal of Statistics*, 36, pp. 157-174.
- Fernholz L. T., On multivariate higher order von Mises expansions in *Metrika* 2001, vol. 53, pp. 123-140.

Université de Liège

Bibliography

- García-Escudero L. A. and Gordaliza A., Robustness Properties of k Means and Trimmed k Means, *Journal of the American Statistical Association*, September 1999, Vol. 94, n°447, pp. 956-969.
- Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A., Robust Statistics : The Approach Based on Influence Functions, John Wiley and Sons, New-York, 1986.
- Anderson T.W., *An Introduction to Multivariate Statistical Analysis*, Wiley, New-York, 1958, pp. 126-133.

Université de Liège

Bibliography

- Pollard D., Strong Consistency of k-Means Clustering, *The Annals of Probability*, 1981, Vol.9, n°4, pp.919-926.
- Pollard D., A Central Limit Theorem for k-Means Clustering, *The Annals of Probability*, 1982, Vol.10, n°1, pp.135-140.
- Qiu D. and Tamhane A. C. (2007), A comparative study of the *k*-means algorithm and the normal mixture model for clustering : Univariate case, *Journal of Statistical Planning and Inference*, 137, pp. 3722-3740.