FEMDA : une méthode de classification robuste et flexible

Pierre HOUDOUIN[1], Matthieu JONCKHEERE[2], Frédéric PASCAL[1]

[1]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France

[2]LAAS-CNRS, Université de Toulouse, 31077 Toulouse, France
pierre.houdouin@centralesupelec.fr, frederic.pascal@centralesupelec.fr
mjonckheer@laas.fr

Résumé − L'analyse discriminante quadratique (QDA) est une méthode de classification qui suppose que les données sont issues de distributions gaussiennes. Cette hypothèse la rend peu robuste aux données bruitées. Le but de cet article est d'étudier la robustesse aux changements d'échelle dans les données d'une nouvelle méthode d'analyse discriminante où chaque point est modélisé par sa propre distribution elliptique symétrique avec son propre facteur d'échelle. Une telle modélisation fournit une grande flexibilité pour traiter des données hétérogènes et non identiquement distribuées. Cette nouvelle méthode s'avère plus robuste aux changements d'échelle dans les données que les autres méthodes de l'état de l'art.

Abstract − Linear and Quadratic Discriminant Analysis (LDA and QDA) are well-known classical methods but can heavily suffer from non-Gaussian distributions and/or contaminated datasets, mainly because of the underlying Gaussian assumption that is not robust. This paper studies the robustness to scale changes in the data of a new discriminant analysis technique where each data point is drawn by its own arbitrary Elliptically Symmetrical (ES) distribution and its own arbitrary scale parameter. Such a model allows for possibly very heterogeneous, independent but non-identically distributed samples. The new decision rule derived is simple, fast and robust to scale changes in the data compared to others state-of-the-art methods.

1 Introduction

L'analyse discriminante est un outil très utilisé pour les tâches de classification. La méthode historique [1] présuppose que les données sont issues de distributions gaussiennes et la règle de décision consiste à choisir le cluster qui maximise la vraisemblance de la donnée. Au début des années 80, [2] et [3] ont étudié l'impact de la contamination et du *mislabelling* sur les performances et concluent à une grande sensibilité. Pour traiter ce problème, [4] suggère l'utilisation de M-estimateurs qui sont robustes au bruit. Plus récemment, [5] a proposé de modéliser les données par une distribution de student multivariée, plus flexible. En 2015, [6] généralise même aux distributions elliptiques symétriques (ES). Cette nouvelle méthode, appelée *Generalized* QDA (GQDA) repose sur l'estimation d'un seuil, dont la valeur varie avec la forme de la distribution. Enfin, [7] a complété GQDA avec l'utilisation d'estimateurs robustes, pour obtenir RGQDA.

Toutes ces méthodes supposent que les points d'un même cluster sont issus de la même distribution, hypothèse qui n'est pas toujours valide. [8], inspiré par [9], propose une méthode alternative qui ne suppose aucun a priori sur les distributions, et permet à chaque point d'être issu de sa propre distribution elliptique symétrique. Les points d'un même cluster ne sont pas forcément identiquement distribués, seulement tirés indépendamment. La contrepartie d'une telle flexibilité réside dans les caractéristiques des clusters : au sein d'un même cluster, les points partagent seulement la même moyenne et la même matrice de dispersion. Nous allons étudier dans ce papier la robustesse aux changements d'échelle dans les données de cette nouvelle méthode.

Le modèle est présenté dans la section 2, la section 3 contient des expériences sur données simulées, la section 4 les expériences sur données réelles et les conclusions et remarques sont effectuées dans la section 5.

2 FEMDA : Flexible EM-inspired discriminant Analysis

Modèle statistique: On suppose que chaque donnée $\mathbf{x}_i \in \mathbb{R}^m$, $i \in [1, n]$ est tirée d'une distribution ES indépendante du cluster. La moyenne et la matrice de dispersion dépendent du cluster auquel le point appartient tandis que le facteur d'échelle $\tau_{i,k}$ peut dépendre de l'observation également. La donnée $\mathbf{x}_i$ du cluster $\mathcal{C}_k$, $k \in [1, K]$ est tirée selon la densité de probabilité suivante :

$$f(\mathbf{x}_i) = A_i \left|\mathbf{\Sigma}_k\right|^{-\frac{1}{2}} \tau_{i,k}^{-\frac{m}{2}} g_i\left(\frac{(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)}{\tau_{i,k}}\right)$$

Expression de la log-vraisemblance et des estimateurs du maximum de vraisemblance: Soient $\mathbf{x}_1, ..., \mathbf{x}_{n_k}$ des données indépendantes du cluster $\mathcal{C}_k$, la log-vraisemblance de l'échantillon peut s'écrire:

$$l(\mathbf{x}_1, ..., \mathbf{x}_{n_k}) = \sum_{i=1}^{n_k} \log\left(A_i \left|\mathbf{\Sigma}_k\right|^{-\frac{1}{2}} t_{i,k}^{-\frac{m}{2}} s_{i,k}^{\frac{m}{2}} g_i(s_{i,k})\right) \quad (1)$$

où $t_{i,k} = (\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)$ and $s_{i,k} = t_{i,k}/\tau_{i,k}$.

Maximiser le terme de l'équation (1) par rapport à $\tau_{i,k}$, avec $\boldsymbol{\mu}_k$ et $\mathbf{\Sigma}_k$ fixés mène à

$$\hat{\tau}_{i,k} = \frac{(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)}{\arg\max_{t\in\mathbb{R}^+}\{t^{\frac{m}{2}} g_i(t)\}}.$$

Les hypothèses sur g_i assurent la stricte positivité du dénominateur. Après avoir remplacé dans l'équation (1) $\tau_{i,k}$ par $\hat{\tau}_{i,k}$, on obtient:

$$l(\mathbf{x}_i) = \tilde{A}_i - \frac{1}{2}\log\left(\left|\mathbf{\Sigma}_k\right|\left((\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\right)^m\right)$$

où $\tilde{A}_i = \log(A_i) + \log(\max_{t\in\mathbb{R}^+}\{t^{\frac{m}{2}} g_i(t)\})$.

A cette étape, on comprend que la flexibilité dans le choix de l'échelle de la matrice de covariance nous permet de réduire l'impact de la fonction génératrice g_i dans la vraisemblance à une constante multiplicative indépendante de k. Enfin, l'utilisation de l'estimateur du maximum de vraisemblance permet d'obtenir les estimateurs robustes suivants pour la moyenne et la matrice de dispersion :

$$\begin{cases} \hat{\boldsymbol{\mu}}_k &= \dfrac{\sum_{i=1}^{n_k} w_{i,k}\mathbf{x}_i}{\sum_{i=1}^{n_k} w_{i,k}}, \\ \hat{\mathbf{\Sigma}}_k &= \dfrac{m}{n_k}\displaystyle\sum_{i=1} w_{i,k}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_k)^T \end{cases} \quad (2)$$

où $w_{i,k} = 1/t_{i,k}$.

Il est intéressant de noter que $\hat{\boldsymbol{\mu}}_k$ est insensible à l'échelle de $\hat{\mathbf{\Sigma}}_k$. Par conséquent, si $\hat{\mathbf{\Sigma}}_k$ est une solution à l'équation de point fixe, $\lambda\hat{\mathbf{\Sigma}}_k$ l'est également. Les estimateurs obtenus sont similaires aux M-estimateurs robustes, sauf que les poids $w_{i,k}$ sont proportionnels au carré de la distance de Mahalanobis. La convergence de ces deux équations de point-fixe couplées a été analysée par [9].

Règle de classification: Grâce à ces deux estimateurs, on utilise les données d'entraÃ®nement pour estimer les paramètres inconnus. On suppose le nombre de clusters connu. Il est maintenant possible de déduire la règle de classification. On a la proposition suivante :

Proposition 2.1. *La règle de décision pour Flexible EM-Inspired Discriminant Analysis (FEMDA) est :*

$$\mathbf{x}_i \in \mathcal{C}_k \iff \left(\forall j \neq k, \Delta_{jk}^2(\mathbf{x}_i) \geq \frac{1}{m}\lambda_{jk}\right) \quad (3)$$

avec $\Delta_{jk}^2(\mathbf{x}_i) = \log\left(\dfrac{(\mathbf{x}_i-\boldsymbol{\mu}_j)^T\mathbf{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)}{(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)}\right)$

et $\lambda_{jk} = \log\left(\dfrac{|\mathbf{\Sigma}_k|}{|\mathbf{\Sigma}_j|}\right)$.

Preuve: La preuve repose sur le fait que la log-vraisemblance ne dépend de k qu'à travers le terme

$$\frac{1}{m}\log\left(|\mathbf{\Sigma}_k|\right) + \log\left((\mathbf{x}_i-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\right)$$

Remarque 2.2. *Cette règle de décision est similaire à la version robuste de QDA. La différence est que nous comparons le logarithme des distances de Mahalanobis au carré plutôt que directement les distances de Mahalanobis au carré. Cela rend notre règle de décision également insensible à l'échelle de* $\mathbf{\Sigma}$.

3 Expériences sur données simulées

FEMDA, la méthode proposée, est comparée avec les méthodes suivantes : QDA classique modélisant les données par des distributions gaussiennes, QDA modélisant les données par des distributions de student (t-QDA, [5]), GQDA et GQDA [6], [7].

Paramètres de simulation: Les moyennes des clusters sont tirées aléatoirement sur la m-sphère unité. Les matrices de covariance sont générées avec des valeurs propres et une matrice orthogonale aléatoires. On choisit $m = 10$, $K = 5$, $N_{train} = 5000$, $N_{test} = 20000$ et $\tau \sim \mathcal{U}(1, m)$.

Scénarios considérés: On génère les points grâce à deux familles de distributions ES différentes.

Famille de distributions	Représentation stochastique
gaussienne généralisée	$\boldsymbol{\mu} + \Gamma(\frac{m}{2\beta}, 2)^{\frac{1}{2\beta}}\mathbf{\Sigma}^{\frac{1}{2}}\mathcal{U}(\mathcal{S}(0,1))$
t-distribution	$\boldsymbol{\mu} + \mathcal{N}(0,\mathbf{\Sigma})\sqrt{\frac{1}{\Gamma(\frac{\nu}{2},\frac{2}{\nu})}}$

$\mathcal{U}(\mathcal{S}(0,1))$ représente une distribution uniforme sur la m-sphère unité. Le paramètre de forme β (resp. ν) est tiré de manière uniforme dans $[0.25, 10]$ (resp. $[1, 10]$) pour les gaussiennes généralisées (resp. pour les t-distributions).

Les scénarios de génération de données sont définis comme suit : $0.6GG-0.4T$ correspond à 60% des points de chaque cluster générés avec une gaussienne généralisée et 40% avec une t-distribution.

On utilise le code couleur suivant pour la génération des paramètres: $0.6GG - 0.4T$ signifie que les mêmes β et ν sont utilisés pour les points d'un même cluster et $0.6GG-0.4T$ signigie qu'on utilise un paramètre différent pour chaque point de chaque cluster.

Résultats

Pour chaque scénario dans la première colonne, le tableau 1 présente les différences de taux de bonne classification entre la meilleure méthode et les autres :

Scénario	QDA	t-QDA	GQDA	FEMDA
GG - T				
$1-0$	-0.51	**76.27**	-0.47	-0.02
$0-1$	-0.64	**76.74**	-0.69	-0.16
$1-0$	-0.59	**76.39**	-0.58	-0.10
$0-1$	-1.24	**77.08**	-1.27	-0.21
$\frac{1}{2}-\frac{1}{2}$	-1.17	**80.85**	-1.13	-0.39
$\frac{1}{2}-\frac{1}{2}$	-1.31	-0.02	-0.87	**80.59**

Table 1: Précision de la classification

Dans le tableau 1, on remarque que GQDA et QDA obtiennent des performances plus faibles que FEMDA et t-QDA. t-QDA est la meilleure méthode dans la plupart des scénarios et surpasse légèrement FEMDA, au prix de l'estimation de plus de paramètres et donc d'une méthode plus lente. [8] a étudié plus en détails les vitesses de convergence de chaque estimateur et règle de décision. Dans le tableau 2, on bruite les données avec un changement d'échelle. Une fraction des données subit le changement suivant : $x \longleftarrow \mu+\lambda(x-\mu)$. On observe alors que FEMDA est la méthode la plus robuste au bruit, t-QDA est surpassée dans presque tous les scénarios lorsque la contamination atteint 25% avec $\lambda = 4$, et dans tous avec $\lambda = 8$. L'écart-type entre plusieurs simulations est faible, de l'ordre de 0.05%.

Scénario	t-QDA	FEMDA	t-QDA	FEMDA
Bruit	10%		25%	
GG - T				
$1-0$ - $\lambda=4$	**73.23**	-0.19	-0.37	**66.44**
$0-1$ - $\lambda=4$	**74.65**	-0.36	-0.15	**67.19**
$1-0$ - $\lambda=4$	-0.08	**72.98**	-0.22	**66.48**
$0-1$ - $\lambda=4$	**73.93**	-0.24	-0.04	**66.70**
$\frac{1}{2}-\frac{1}{2}$ - $\lambda=4$	**77.14**	-0.61	**70.61**	-0.21
$\frac{1}{2}-\frac{1}{2}$ - $\lambda=4$	**76.28**	-0.41	-0.13	**69.74**
$1-0$ - $\lambda=8$	-0.11	**72.87**	-0.67	**64.79**
$0-1$ - $\lambda=8$	**74.11**	-0.29	-0.45	**65.98**
$1-0$ - $\lambda=8$	-0.31	**71.93**	-0.33	**65.49**
$0-1$ - $\lambda=8$	-0.08	**73.22**	-0.24	**64.29**
$\frac{1}{2}-\frac{1}{2}$ - $\lambda=8$	**76.36**	-0.44	-0.14	**68.69**
$\frac{1}{2}-\frac{1}{2}$ - $\lambda=8$	**75.56**	-0.37	-0.32	**67.61**

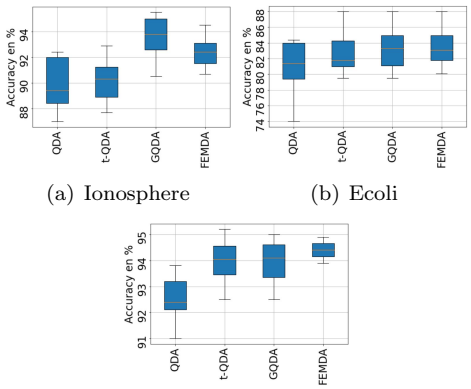
Table 2: Précision en présence de bruit

4 Expériences sur données réelles

Les jeux de données sont issus de l'UCI machine learning repository [10]. Trois datasets sont utilisés : **Ionosphere** avec 351 données de dimension 34, **Ecoli** avec 336 données de dimension 8 et **Breast cancer** avec 699 données de dimension 10.

4.1 Résultats de classification

Pour obtenir ces résultats, 100 simulations ont été effectuées, et après 10 simulations successives, les *train* et *test set* sont recomposés (70% train set et 30% test set).



(a) Ionosphere (b) Ecoli

(c) Breast cancer

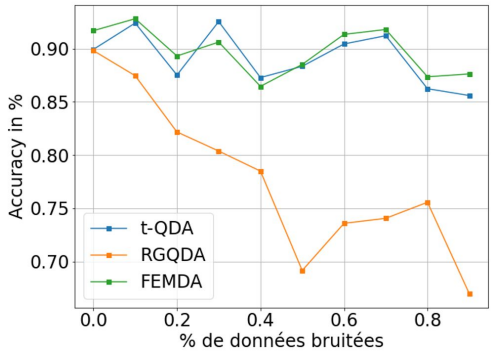
Figure 1: Précision médiane

On peut voir sur les figures 1(a) et 1(b) que GQDA surperforme d'au moins 1% les autres méthodes, suivi par FEMDA et ensuite par t-QDA pour le dataset Ionosphere. Les écarts sont plus resserrés pour le dataset Ecoli. Sur la figure 1(c), on remarque que FEMDA devient la meilleure méthode avec une précision proche de 95%, suivi de près par t-QDA puis GQDA. La variance dans les résultats est plutôt faible. Pour conclure sur ces trois datasets, les performances de FEMDA sont légèrement supérieures à celles de t-QDA, et souvent inférieures à celles de GQDA, qui sont cependant plus variables.

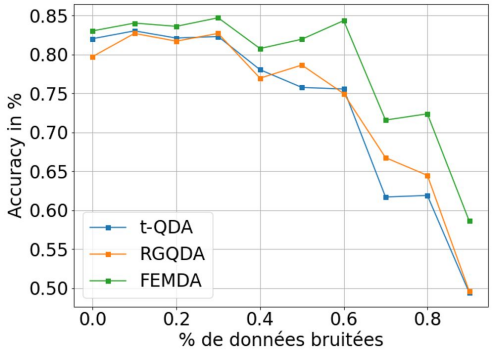
4.2 Résultats après changements d'échelle

On va maintenant bruiter les données d'une manière similaire à ce qui a été effectué pour les données simulées. On choisit $\lambda = 5$. On trace ensuite l'évolution de la précision des trois méthodes robustes selon le taux de contamination.

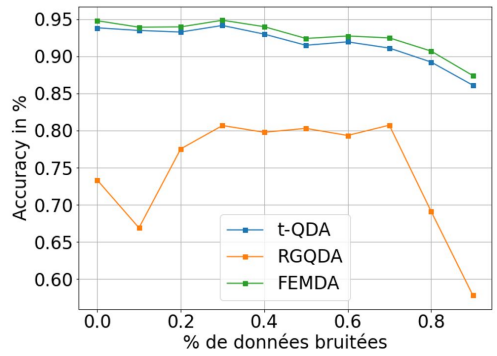
On remarque sur la figure 2 que même avec des taux de bruit très élevés, t-QDA et FEMDA conservent de très bons résultats pour Spambase et Ionosphere. En revanche, les performances de RGQDA chutent beaucoup plus rapidement lorsque le taux de contamination augmente. Pour le dataset Ecoli, le comportement est beaucoup plus uniforme, les trois méthodes voient leurs performances baisser, surtout lorsqu'on dépasse un taux de 40% de contamination. FEMDA affiche malgré tout une résilience légèrement supérieure pour les hauts taux de contamination, mais les performances restent très proches de celles de t-QDA. La robustesse de FEMDA aux changements d'échelle dans les données d'entraînement peut être expliquée par l'expression des estimateurs, qui sont intrinsèquement insensibles aux changements d'échelle. En-



(a) Ionosphere



(b) Ecoli



(c) Breast cancer

Figure 2: Données bruitées par changement d'échelle avec $\lambda = 5$

fin, la différence de sensibilité à l'augmentation de la contamination pour Ecoli peut s'expliquer par la faible dimension des données par rapport aux autres datasets. En effet, en grande dimension, la direction de la matrice de covariance est beaucoup plus discriminante pour séparer les données.

5 Conclusion

Dans ce papier, nous avons présenté une nouvelle méthode d'analyse discriminante robuste aux changements d'échelle dans les données d'entraînement. Elle surpasse toutes les méthodes de l'état de l'art en présence de données contaminées, et se comporte de manière similaire à t-QDA sans bruit, tout en étant plus rapide. FEMDA est donc une méthode rapide et très résiliente face aux données bruitées. Dans ce nouveau paradigme, les clusters ne partagent plus la même matrice de covariance, mais seulement la même matrice de dispersion. Permettre à chaque point d'avoir son propre facteur d'échelle entraîne un gain de flexibilité qui permet de traiter des jeux de données contaminées et non nécessairement identiquement distribuées. On peut donc considérer que FEMDA est une amélioration de t-QDA : performances similaires sans contamination, mais plus robuste et plus rapide.

References

[1] Carl J. Huberty. *Discriminant Analysis*. Review of Educational Research, 1975.

[2] Peter A. Lachenbruch. *Discriminant Analysis When the Initial Samples Are Misclassified*. Technometrics, 1996.

[3] P. A. Lachenbruch and M. Goldstein. *Discriminant Analysis*. Biometrics, 1979.

[4] Huber Peter J. *Robust covariances*. Statistical decision theory and related topics, 1977.

[5] Andrews Jeffrey L and McNicholas Paul D and Subedi Sanjeena. *Model-based classification via mixtures of multivariate t-distributions*. Computational Statistics & Data Analysis, 2011.

[6] Bose Smarajit and Pal Amita and SahaRay Rita and Nayak Jitadeepa. *Generalized quadratic discriminant analysis*. Pattern Recognition, 2015.

[7] Ghosh Abhik and SahaRay Rita and Chakrabarty Sayan and Bhadra Sayan. *Robust generalised quadratic discriminant analysis*. Pattern Recognition, 2021.

[8] Houdouin Pierre and Pascal Frédéric and Jonckheere Matthieu and Wang Andrew. *Robust classification with flexible discriminant analysis in heterogeneous data*. https://arxiv.org/abs/2201.02967, 2022.

[9] Roizman Violeta and Jonckheere Matthieu and Pascal Frédéric. *A flexible EM-like clustering algorithm for noisy data*. arXiv preprint arXiv:1907.01660, 2019.

[10] Dua Dheeru and Graff Casey. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017.