MDI343 Data Challenge

Rapport

Introduction

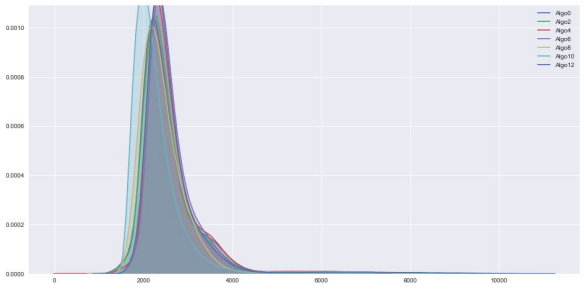
À mon sens, tout exercice de Machine Learning doit se baser sur un examen préalable des données, une étude de la façon dont ces données ont été collectées ainsi qu'une familiarisation avec le contexte métier de l'exercice. Mon travail a donc débuté par des recherches sur le web concernant les méthodes d'identification biométrique et la problématique de fusion des scores. Voici la liste des sources qui m'ont été le plus utiles :

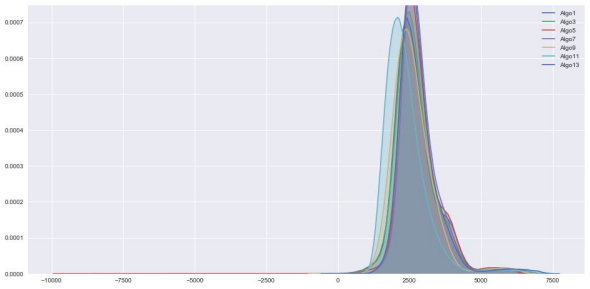
- N. Poh and S. Bengio, *"Towards Predicting Optimal Subsets of Base-Experts in Biometric Authentication Tasks"*, IDIAP Research Report 04-17, 2004.
- N. Poh, *"Multi-system Biometric Authentication: Optimal Fusion and User-Specific Information"*, Thèse doctorat, EPFL, 2006.
- M. Belahcene, *"Authentification et Identification en Biométrie"*, Thèse de doctorat, Université Mohamed Khider Biskra, 2013.

Analyse exploratoire des données

Nombre initial d'observations : 2.048.852
Nombre d'observations après nettoyage des données : 2.048.839

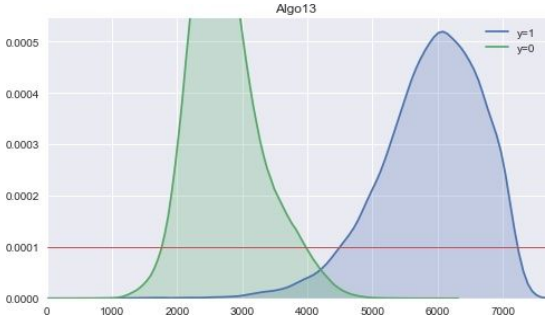
Les classes positives et négatives sont déséquilibrées dans les données d'apprentissage : 97,51% d'observations positives et 2,49% d'observations négatives.





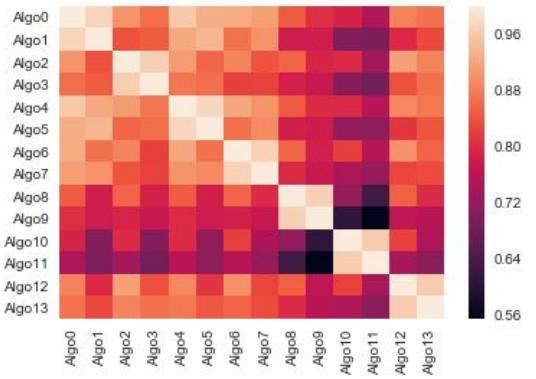
Les scores des algorithmes sont tous dans le même ordre de grandeur (avec une moyenne autour de 2.500), même si uniquement la moitié d'entre eux comportent des valeurs négatives. Les distributions sont également semblables et semblent suivre une loi normale (cf. courbes ci-dessus).

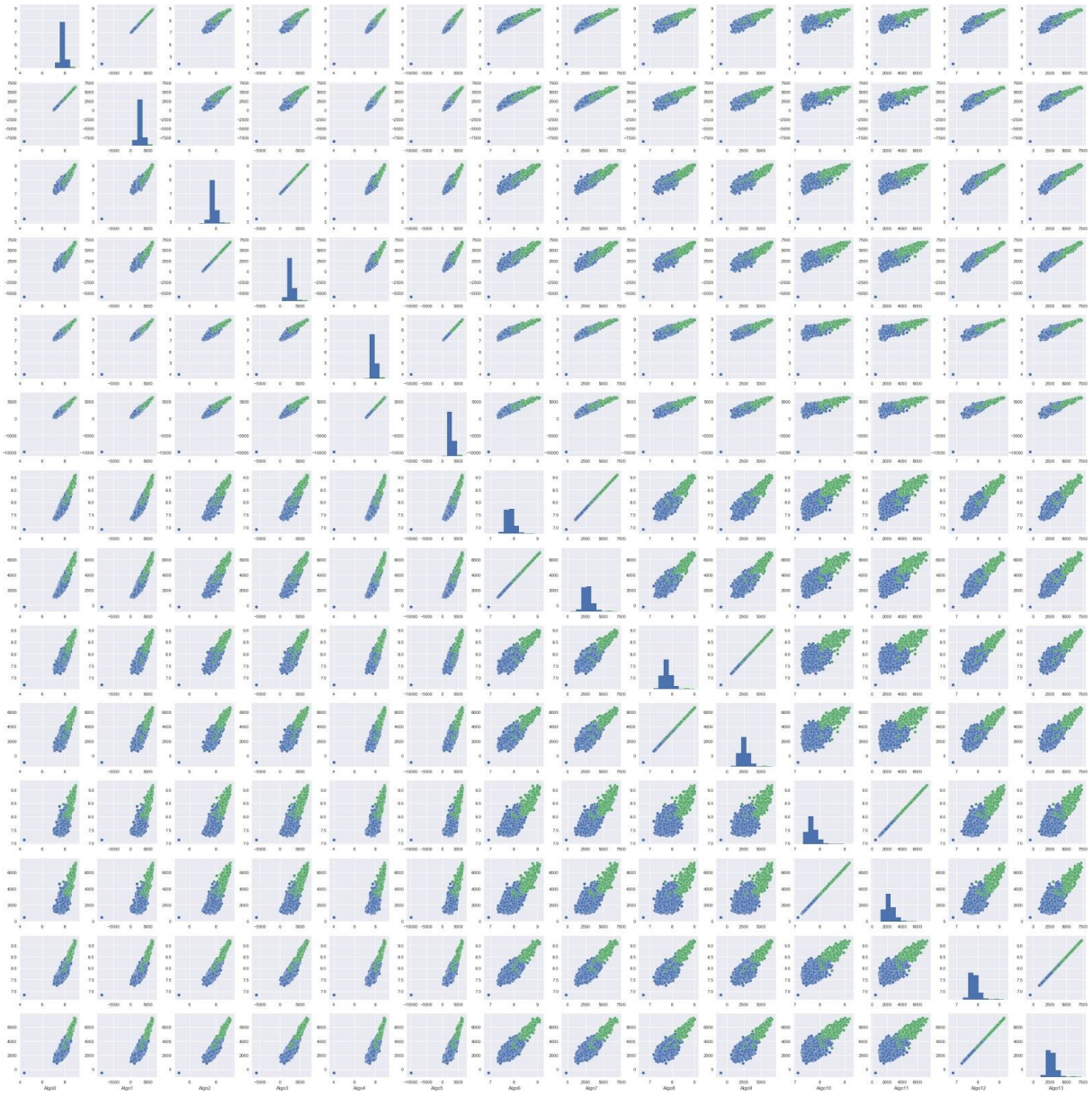
Cependant, lorsque l'on affiche celles-ci en distinguant les classes positives des classes négatives, on s'aperçoit que les distributions sont en réalité constituées d'un mélange, comme dans l'exemple ci-contre. Rappelons que dans le cadre de fusion des scores, nous cherchons à minimiser l'*Error Equal Rate*, fortement liée à la zone de superposition entre les distributions des classes.



Jean-Marc SEVIN - 06/02/2018

Un autre point important réside dans la corrélation entre les algorithmes ; ils sont fortement correlés par paires (cf. heatmap) et cela se précise en affichant un corrélogramme entre les logarithmes des algorithmes pairs et les algorithmes impairs. Les scores des algorithmes impairs correspondent donc aux logarithmes des scores pairs, après normalisation ; ceci explique notamment les temps de traitement identiques dans chaque paire.





Les conclusions issues de cette phase d'exploration des données sont les suivantes :

Jean-Marc SEVIN - 06/02/2018

- On peut se restreindre à n'employer qu'une combinaison d'algorithmes pairs ou impairs, afin de de limiter la redondance des informations, d'optimiser la diversité des sources et de réduire le champ d'exploration pour la phase suivante.
- Les algorithmes 9 et 11 sont les plus décorrélés deux à deux.
- Les algorithmes 7 et 13 semblent être ceux dont les distributions entre les classes positives et négatives se chevauchent le moins.

Recherche de modèles

Pour cette étape, le mode opératoire retenu est le suivant :

1. On définit une fonction permettant de calculer l'EER afin d'évaluer la performance des modèles,
2. En se restreignant aux algorithmes impairs, on établit une liste des combinaisons de quatre algorithmes dont le total des temps de traitement n'excède pas 600 millisecondes,
3. On teste successivement plusieurs modèles pour chaque combinaison d'algorithmes.

Régression logistique

En effectuant une première approche de référence, basée sur la régression logistique, on obtient des performances médiocres.

Analyse discriminante

L'approche suivante se base sur une analyse discriminante, linéaire dans un premier temps, puis quadratique. Elle offre des performances acceptables, mais présente l'inconvénient de limiter la construction de la matrice de fusion à des combinaisons linéaires des scores d'algorithmes.

Arbres de décision

Les arbres de décision et le modèle Extra Trees tombent rapidement dans le surapprentissage, mais les forêts aléatoires semblent offrir les meilleures performances par défaut.

Densités de probabilités

Compte tenu de la nature des distributions observées lors de l'analyse exploratoire des données et d'après la littérature sur la fusion des scores qui suggère d'enrichir les données avec des informations sur la probabilité d'appartenir aux clients ou aux imposteurs, j'ai tenté d'employer une approche basée sur des modèles de mélange gaussien, mais je me suis très vite heurté à l'incapacité d'exploiter ces résultats pour construire la matrice de fusion.

Modèle	**Meilleure combinaison**	**Meilleur score (ERR)**
Régression logistique	(1,3,9,11)	0.499524
LDA	(3,9,11,13)	0.024386
QDA	(3,9,11,13)	0.030866
Arbres de décision	(1,9,11,13)	0.124626
Forêts aléatoires	(3,9,11,13)	0.007560
Extra Trees	(1,3,7,9)	0.000078
GMM	(1,5,7,11)	0.093322
Bayesian GMM	(1,5,7,9)	0.035012

Optimisation

L'étape précédente m'a permis de sélectionner les forêts aléatoires comme modèle de référence. En ajoutant un pré-traitement en amont du modèle, permettant de calculer les combinaisons quadratiques des scores, on évalue les différentes combinaisons d'algorithmes afin de retenir les meilleures et d'obtenir ensuite les coefficients qui vont pondérer la somme.

C'est finalement une combinaison quadratique des algorithmes 7, 9, 11 et 13 qui m'a permis d'atteindre un FFR de 0.0719769673704 sur les données de test.