Une comparaison de quatre techniques d'inférence des refusés dans le processus d'octroi de crédit

Asma Guizani, Besma Souissi, Salwa Benammou, Gilbert Saporta

UNE COMPARAISON DE QUATRE TECHNIQUES D'INFÉRENCE DES REFUSÉS DANS LE PROCESSUS D'OCTROI DE CRÉDIT

Asma Guizani[1] & Besma Souissi[2] & Salwa Ben Ammou[3,4] & Gilbert Saporta[4]

[1]*Institut Supérieur de Gestion de Sousse, rue Abdlaaziz il Behi . Bp 763. 4000 Sousse Tunisie. asmaguizani@gmail.com*
[2]*Computational Mathematics Laboratory, Faculté des Sciences, Avenue de l'Environnement, Université de Monastir, 5000, Tunisie. Besma.swissi@yahoo.fr*
[3]*Faculté des Sciences Economiques et de gestion de Sousse Cité Erriadh - 4023 Sousse Tunisie. Saloua.benammou@fdseps.rnu.tn*
[4]*Laboratoire Cédric - CNAM, 292 rue Saint Martin, 75141 Paris cedex 03, France. gilbert.saporta@cnam.fr*

Résumé. L'objectif principal des techniques d'inférence des refusés est de corriger le biais de sélection résultant d'un modèle construit sur la base d'un échantillon non représentatif de la population globale. Cette communication a pour but de présenter une comparaison expérimentale de quatre techniques (La repondération, le parceling, la classification mixte et la reclassification itérative) pour remédier au problème du biais de sélection.

Mots clés. Credit scoring, repondération, reclassification itérative, parceling, classification mixte, courbe ROC.

Abstract. The main objective of reject inference techniques is to correct the selection bias which is the result of a model built with a non representative sample of the entire population. Reweighting, parceling, mixed classification and iterative reclassification are the four techniques that we expose in this paper to deal with the problem of selection bias.

Keywords. Credit scoring, reweighting, iterative reclassification, parceling, mixed classification, ROC curve.

1 Introduction

Le credit scoring est une méthode d'évaluation du niveau du risque associé à un dossier de crédit potentiel. Cette méthode implique l'utilisation de différentes techniques statistiques pour aboutir à un modèle de scoring. Basé sur les caractéristiques du dossier du client, ce modèle estime le risque de crédit en prévoyant la solvabilité du demandeur de crédit.

Les institutions financières utilisent ce modèle pour estimer la probabilité de défaut qui va être utilisée pour affecter chaque client à la catégorie qui lui correspond le mieux : bon payeur (remboursement du crédit à temps) ou mauvais payeur (impayés).

Dans le contexte des accords de Bâle, les institutions financières cherchent constamment à améliorer ces modèles de score pour arriver à identifier les caractéristiques nécessaires pour mieux distinguer les bons des mauvais payeurs dans le futur.

Les seules données disponibles pour construire le modèle de scoring sont un historique de dossiers acceptés dont la variable à expliquer est connue (bon/mauvais). La probabilité de défaut ne peut être estimée que pour les dossiers acceptés.

Le modèle de score d'octroi ne tient donc pas compte des demandeurs rejetés dès le départ ce qui implique qu'on ne pourra pas en estimer la probabilité de défaut (données manquantes). Le modèle donne donc des résultats biaisés à cause de la non-représentativité de l'échantillon (biais de sélection [1]).

L'inférence des refusés (« reject inference » en anglais) tente de remédier au problème de biais de sélection et de le corriger en réintégrant les dossiers refusés dans l'échantillon initial pour le rendre représentatif de la population globale (les acceptés et les refusés). Plusieurs techniques de réintégration des refusés ont été développées dans la littérature. Dans un travail antérieur [5] nous nous étions basés sur la méthode de l'augmentation simple comme méthode d''inférence des refusés pour aboutir à un modèle de score construit sur la base d'un échantillon représentatif de la population globale. Nous avions mis en œuvre deux techniques statistiques pour construire notre modèle de score : la discrimination PLS-DA et l'analyse factorielle discriminante. Les deux méthodes donnaient pratiquement les mêmes résultats.

Nous présentons dans la section 2 quatre autres techniques. Nous les appliquons dans la section 3 et nous comparons la performance des modèles de score obtenus avec chacune des ces quatre méthodes. La section 5 est consacrée aux conclusions et perspectives de recherches pour la mise en œuvre de nouvelles méthodes plus performantes.

2 Les techniques d'inférence des refusés

2.1 La repondération

La repondération intitulée aussi « reweighting » se déroule en trois étapes [6]:

- Tout d'abord, on construit un score d'acceptation qui sera appliqué à la population totale pour obtenir la probabilité d'acceptation pour chaque individu.
- Ensuite, on pondère chaque dossier accepté par l'inverse de la probabilité d'acceptation.
- Enfin, on construit le modèle de score de défaut « bon/mauvais » sur les dossiers acceptés ainsi pondérés.

Cette pondération permet en quelque sorte de « compenser » l'absence des refusés [7].

2.2 Reclassification itérative

Cette méthode commence par construire un modèle de score de défaut « bon/mauvais » sur les dossiers acceptés. Par la suite, on applique ce modèle de score sur les dossiers refusés pour les affecter à la catégorie correspondante : bon payeur ou mauvais payeur. Ensuite, on construit ce qu'on appelle l'ensemble augmenté (dite « augmented data set » [9]) qui consiste à ajouter aux acceptés « bon » (respectivement « mauvais ») les dossiers rejetés prédits par le modèle précédent comme « bon » (respectivement « mauvais »).

On construit un nouveau modèle de score sur l'ensemble augmenté. On applique alors ce nouveau modèle aux refusés pour les classer en « bon » ou « mauvais ». On réitère le processus jusqu'à stabilisation des scores obtenus.

2.3 Parceling [7]

Cette méthode est considérée comme une amélioration de la précédente. Elle consiste :

- Tout d'abord à construire un modèle de score de défaut « bon/mauvais » sur les dossiers acceptés.
- Ensuite on répartit la population en intervalles de score et on calcule le taux d'impayés (défaut) dans chaque tranche de score. On applique le modèle de score aux refusés, de

façon à les placer chacun dans un intervalle de score sous l'hypothèse que le taux d'impayés est le même que celui des acceptés.

- On classe ensuite aléatoirement, les refusés de chaque intervalle en deux catégories « bon » et « mauvais » tout en respectant la proportion de « bon » et de « mauvais » payeurs calculé pour chaque intervalle.
- On regroupe alors les acceptés initiaux et les refusés étiquetés et on construit un nouveau modèle de score de défaut « bon/mauvais ».

2.4 Classification mixte [8]

L'algorithme de cette méthode se déroule en deux étapes :

- La première consiste à utiliser la méthode des « k-means » pour un partitionnement initial sur l'ensemble des observations (acceptés et refusés) de façon à obtenir « k » groupes homogènes avec « k » déterminé par le critère de Wong qui suggère que $k \geq n^{0.3}$, n étant le nombre d'individus. Chaque classe contient une forte proportion de bons ou de mauvais dossiers en plus des dossiers refusés.
- Ces « k » classes seront, dans un second temps, réduites à « q » classes ($q<k$) par une Classification Ascendante Hiérarchique (CAH) appliquée aux centres de gravités des groupes de la première classification. Le nombre « q » de classes est déterminé par analyse des indicateurs statistiques comme R-Square (RSQ), Semi-Partial R-Square (SPRSQ), Root-Mean-Square Standard Deviation (RMSSTD), Pseudo F, Cubic Clustering Criterion (CCC).

Une fois les « q » classes définies, les refusés, appartenant à chacune de ces classes, seront affectés à la catégorie « bon » ou « mauvais » en fonction de la catégorie dominante de la classe. À l'issue de cette méthode, tous les dossiers refusés se voient attribuer une étiquette, un modèle de score est alors construit sur la base des dossiers acceptés et des refusés ainsi étiquetés.

3 Données et méthodes

Les données utilisées proviennent de l'agence de notation externe « Experian ». Ses données appartiennent en réalité à la société « Financo » qui est un organisme de crédit à la consommation dans l'automobile, la moto, les véhicules de loisirs, l'habitat et l'équipement général des ménages. L'échantillon se compose de 13 319 dossiers de crédit sur une période de production correspondante aux années 2000 et 2001. Le comportement de remboursement du client est observé sur une durée minimale de 18 mois. Les dossiers avec une règle de refus conduisant à un refus systématique ne relèvent pas de la procédure de réintégration et sont exclus de l'échantillon. En effet, la question de l'inférence des refusés ne se pose que pour les dossiers refusés susceptibles d'être acceptés.

La variable « BM » définie le comportement de remboursement du client.

Code	Étiquette	Description
1	Bon	Pas de défaut (bon payeur)
2	Intermédiaire	1 ou 2 défauts de paiement
3	Mauvais	Plus que 3 défauts de paiement
98	Sans suite	Dossiers classes sans suite
99	Refusé	Les dossiers refuses

Tableau 1 : Description des attributs de la variable « BM »

Les dossiers intermédiaires ainsi que ceux classés sans suite sont exclus de l'échantillon. Au final, on a un total de 9 892 dossiers de crédit composés de 7 986 dossiers acceptés dont la variable à expliquer est connue (bon payeurs et mauvais payeurs) et 1 906 dossiers refusés dont la variable à expliquer n'est pas définie.

La variable dépendante est une variable binaire qui indique si le client a fait défaut (BM=3) sinon il est classé comme bon payeur (BM=1).

Nous comptons 15 variables indépendantes dont 12 variables sont quantitatives et 3 sont qualitatives. Nous nous trouvons dans un contexte de données mixtes (mélanges de données qualitatives et quantitatives), nous avons transformé chaque variable qualitative à « r » modalités en « r » variables dichotomiques.

Afin de remédier au problème d'inférence des refusés nous avons appliqué les quatre méthodes évoquées précédemment. Toutes les quatre utilisent la régression logistique pour la construction de leurs modèles de score. Leur objectif commun est de réintégrer les 1 906 dossiers refusés ceci en les affectant à l'une des catégories « bon » ou « mauvais » et avoir ainsi à la fin un modèle de score construit sur la base de la population entière pour remédier au biais de sélection. Afin de valider les modèles de scores obtenus par les quatre méthodes, nous avons simulé un processus de rejet sur les 7986 observations selon l'algorithme : à l'aide d'un tirage suivant la loi uniforme entre [0, 1] pour chaque observation, on compare la variable uniforme U_i obtenue à la probabilité de refus (Pr) obtenue selon la discrimination acceptés-refusés. Si $U_i < Pr(_i)$ alors l'observation i est rejetée de l'échantillon sinon elle est conservée. Au final, nous avons un échantillon de 4784 observations à partir duquel nous avons tiré aléatoirement 3000 observations composé de 90,87% bons payeurs et 9,13% mauvais payeurs. Cet échantillon est utilisé pour la construction de la matrice de confusion pour comparer entre la qualité prédite et la qualité réelle.

Pour étudier la performance de nos modèles, nous avons utilisé la courbe ROC (Receiver Operating Characteristics) qui relie la proportion de vrais positifs (bons dossiers classés tels) à la proportion de faux négatifs (mauvais dossiers classés bons) lorsqu'on fait varier le seuil du score d'acceptation. L'aire sous la courbe (Area Under the Curve – AUC) est un indice synthétique de performance. L'AUC correspond à la probabilité pour qu'un événement positif ait une probabilité donnée par le modèle plus élevée qu'un événement négatif. L'AUC appartient à l'intervalle [0,1], le modèle est parfait si l'AUC est égal à 1[7].

4 Résultats et interprétations

La figure 1 représente les courbes ROC pour les quatre méthodes après traitement du problème de l'inférence des refusés. Nous constatons, d'après cette figure, que la technique de reclassification itérative (AUC=0.8404) est la plus performante de toutes les techniques. La convergence a été obtenue en 8 itérations. La repondération est classée en seconde place avec un AUC égal à 0.7362. La courbe ROC de la méthode de parceling est presque identique à celle de la repondération (AUC=0.7082).

La méthode la moins performante est celle de la classification mixte avec un AUC égal à 0.6654.

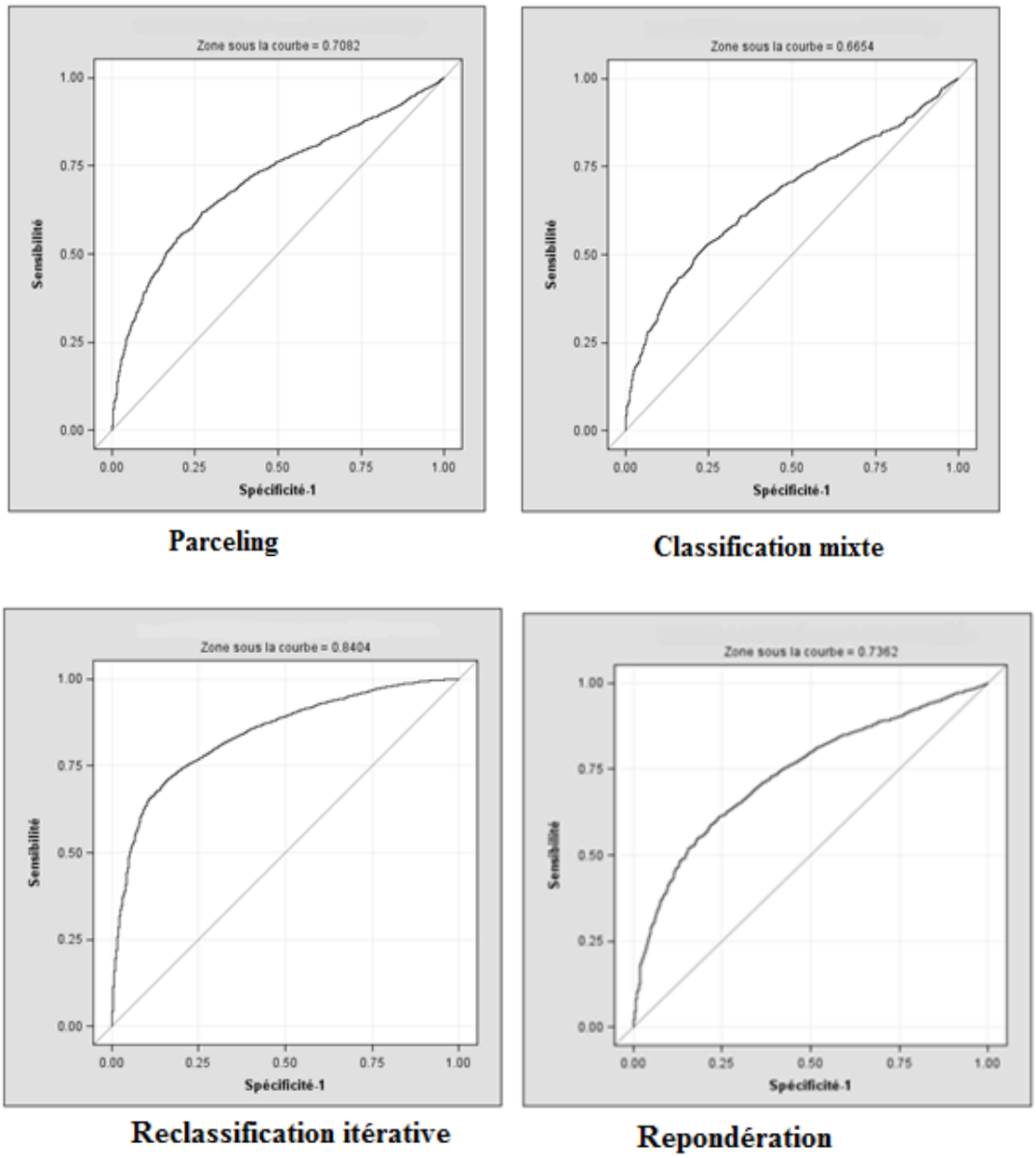

Figure 1 : courbes ROC après réintégration des refusés

Pour valider les modèles de score obtenus après traitement de l'inférence des refusés, nous avons calculé le taux de bons classements des individus appartenant à l'échantillon simulé. Les résultats sont les suivants : pour la repondération, la classification mixte et le parceling, toutes les observations réelles ont été affectées, par le modèle, à la catégorie des bons payeurs avec un taux de bien classées de 91,67%. Concernant la méthode de reclassification itérative, le taux de bons classements pour la catégorie des bons payeurs (respectivement mauvais payeurs) est de 86,87% (respectivement 1,53%).

5 Conclusions et perspectives

Nous avons exposé dans ce papier quatre techniques d'inférence des refusés, dont les résultats sont prometteurs, La méthode de la reclassification itérative semblant meilleure que les autres et infirme les conclusions pessimistes de [4]. Les prochains travaux porteront sur des techniques spécifiques d'apprentissage semi-supervisé [2] car le problème qui se pose ici relève de la classification où l'on dispose à la fois d'un ensemble de données étiquetées (les dossiers acceptés) et d'un ensemble de données non-étiquetées (les dossiers refusés). Le but sera d'atteindre un taux de classification élevé

en combinant l'information contenue dans les données étiquetées et celle contenue dans les données non-étiquetées[3].

Remerciements

Les auteurs tiennent à remercier MM. Sylvain Chamley (Experian) et Christophe Conq (Financo) pour avoir fourni la base de données.

Bibliographie

[1] Banasik, J. et Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, 183, 1582-1594.

[2] Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning*. The MIT Press, London

[3] Chawla, N.V. and Karakoulas, G. (2005). Learning from Labeled and Unlabeled Data: An Empirical Study Across Techniques and Domains. *Journal of Artificial Intelligence*, 23, 331-366.

[4] Hand, D.J. et Henley ,W.E. (1993). Can reject inference ever work? *IMA J Maths Appl Bus Ind* 5: 45–55.

[5] Guizani A., Benammou S. et Saporta G. (2011). Une méthode de traitement des refusés dans le processus d'octroi de crédit. 43ème édition des Journées de Statistique.

[6] Siddiq, N. (2006). *Credit risk scorecards developing and implementing intelligent credit scoring*. John Wiley & Sons, Inc., New Jersey.

[7] Tufféry, S. (2012). *Data mining et statistique décisionnelle : l'intelligence des données*. Editions Technip, 4ème édition.

[8] Tufféry, S. (2009). *Une étude de cas en statistique décisionnelle*, Editions Technip.

[9] Viennet, E., et Fogelman Soulié, F. (2007). Le traitement des refusés dans le risque crédit. *Revue des Nouvelles Technologies de l'Information (RNTI-A-1)*, 23-45.