Extraction de caractéristiques avec l'algorithme Correlation Explanation : application en risque de crédit financier

Sung-Hyuk PANG[1,2], Alban GOUPIL[1], Valeriu VRABIE[1], Loïc KOLODZIEJCZAK[2]

[1]CReSTIC, Université de Reims Champagne-Ardenne
UFR Sciences Exactes et Naturelles - Moulin de la Housse - BP 1039 - 51687 Reims CEDEX 2, France

[2]CMG Conseil
160 Boulevard Haussmann, 75008 Paris, France
sung-hyuk.pang@etudiant.univ-reims.fr

Résumé – Parmi les méthodes de réduction de dimension (RD), les transformations non supervisées permettent d'interpréter les données et de fournir un pré-traitement souvent nécessaire pour un apprentissage efficace. Si les méthodes classiques telles que la PCA ou l'ICA se basent sur des projections, l'algorithme Correlation Explanation (CorEx) permet d'analyser les données en fonction de la dépendance entre elles en s'appuyant sur des notions de la théorie de l'information et offre un moyen de représentation très efficace. Cet algorithme est appliqué sur des données de défaut de crédit et les résultats montrent qu'il est possible d'estimer les caractéristiques explicatives discriminant le défaut de crédit.

Abstract – Among the different dimensionality reduction method (DR), the unsupervsied transformations can be used to interpret the data and to provide a preprocessing step often necessary for an efficient learning. Compared to the classical methods, like PCA or ICA which are based on projection, the Correlation Explanation (CorEx) algorithm provides a data analysis by measuring the dependecy between the variables with an information-theoretic point of view. The algorithm is applied on a credit default dataset and the results show that we can estimate relevant features to discriminate the credit default.

1 Introduction

Compte tenu des risques systémiques, la gestion des risques financiers attire davantage l'attention. En gestion des risques de crédit, les approches statistiques fournissent des estimations de risque comme des indicateurs de défaut ou la perte en cas de défaut. Ces approches s'appuyant sur un nombre important de données très hétérogènes, une analyse préalable peut s'avérer une étape cruciale pour identifier des caractéristiques relatives au risque. Cela permettrait aussi de retenir ou de projeter les données les plus pertinentes afin d'assurer une stabilité des modèles d'apprentissage automatique pour l'estimation des risques [1]. A l'aide des méthodes de réduction de dimension (RD), il est possible d'obtenir une représentation des données pertinentes aussi bien pour l'analyse que pour l'apprentissage.

La sélection de variables, visant à trouver un sous-ensemble parmi les données d'entrée, est une des deux approches de RD. Si la corrélation est largement utilisée en pratique, des critères basés sur l'entropie peuvent être une alternative pour ne pas se restreindre aux dépendances linéaires [2]. Une autre approche de la RD largement utilisée est l'extraction ou la transformation de variables qui consiste à transformer les variables tout en préservant au mieux l'ensemble des informations. Contrairement à la sélection, la transformation permet la combinaison des variables. On retrouve couramment la PCA (Principal Component Analysis) ou encore l'ICA (Independent Component Analysis) comme méthodes de projections. Bien que ces méthodes soient applicables dans de nombreux cas, il n'est pas toujours garanti d'obtenir des résultats satisfaisants dû à la non-linéarité des données, même en utilisant les méthodes à noyau. Dans ce cas, la projection n'étant pas complètement représentative, l'analyse de données serait peu pertinente.

Lorsque la variable cible est disponible, une RD supervisée est envisageable par exemple à l'aide des algorithmes génétiques ou de sélection séquentielle [3]. L'analyse discriminante linéaire permet aussi de réaliser une projection supervisée. Puisque ces approches sont sensibles à la variable cible, les réductions obtenues peuvent varier selon le régime de la cible. Dans notre cas applicatif, cela revient à estimer des indicateurs de défaut spécifique à une banque, une région, etc. chaque changement dans les données nécessitant une mise à jour des estimateurs.

Basé sur les notions d'information mutuelles, l'algorithme de Correlation Explanation (CorEx) [4] propose une représentation des variables en les regroupant en fonction de leurs dépendances. En établissant ce regroupement, le CorEx peut être aussi vu comme une RD. Dans cet article, après avoir introduit les notions nécessaires, nous proposons de mesurer sa performance en tant que méthode de RD par rapport à d'autres méthodes classiques. Nous allons l'appliquer dans le cadre d'un problème concret d'analyse de défaut et nous allons montrer qu'en plus des résultats numériques intéressants, cette méthode

offre la possibilité d'identifier et de résumer les caractéristiques pour expliquer le défaut, les représentations que l'on peut obtenir étant informatives pour les collaborateurs du domaine.

2 Principe du CorEx pour une RD basée sur la dépendance

L'algorithme du CorEx transforme le vecteur X des p variables d'entrée $X_1, \ldots, X_p$ en un vecteur stochastique $Z|X$ de dimension réduite m de composantes $Z_1|X, \ldots, Z_m|X$. Cette transformation a des caractéristiques comparables à la PCA :

— de façon analogue aux composantes de la PCA qui résument la variance significative, le vecteur $Z|X$ résume les dépendances au sein de X ;

— les composantes $Z_j|X$ sont indépendantes entre elles. L'orthogonalité de la base des composantes en est la version linéaire dans la PCA ;

— chaque X_i est associé à une unique composante Z_j. Cela permet de quantifier facilement la contribution de la variable pour la projection. La variance expliquée remplit ce rôle pour la PCA.

Pour pouvoir mesurer la dépendance au sein des variables initiales X, le CorEx se base sur la Corrélation Totale (TC pour *Total Correlation*) [5], une généralisation à plusieurs variables de l'information mutuelle,

$$TC(X) = \sum_{i=1}^{p} H(X_i) - H(X), \tag{1}$$

où $H(\cdot)$ mesure l'entropie de Shannon. Si l'information mutuelle mesure la dépendance entre deux variables, la corrélation totale mesure la dépendance au sein de multiples variables. Elle peut être aussi vue comme l'éloignement par rapport à l'indépendance en la réécrivant comme une divergence de Kullback-Leibler, $TC(X) = D_{\mathrm{KL}}\big(p(x)||\prod_i p(x_i)\big)$ en notant $p(x)$ la distribution de probabilité de X. On définit aussi la TC de X conditionnellement à une variable Z avec l'entropie conditionnelle $TC(X|Z) = \sum_i H(X_i|Z) - H(X|Z)$. Cette valeur peut-être interprétée comme la dépendance moyenne interne à X en fonction de la réalisation de Z. À partir de ces deux notions, on mesure la dépendance capturée par Z avec l'Explication de la Corrélation Totale (TCE, *Total Correlation Explanation*),

$$TC(X;Z) = TC(X) - TC(X|Z). \tag{2}$$

L'objectif du CorEx est alors de trouver une transformation $Z|X$, ou plus précisément sa distribution $p(z|x)$, optimisant le critère de TCE. Dans le cas général où la dimension résultante m est supérieure à 1, la méthode du CorEx maximise la somme des TCE sous la contrainte d'indépendance entre les $Z_i|X$,

$$\max_{p(z|x)} \sum_{j=1}^{m} TC(X;Z_j) \quad \text{avec} \quad p(z|x) = \prod_{j=1}^{m} p(z_j|x). \tag{3}$$

La condition d'indépendance évite des chevauchements entre les informations résumées par les variables Z_j.

Aussi, toujours dans le souci d'interprétabilité de la représentation, on associe chaque variable initiale X_i à un unique Z_j et, par conséquent, les variables X_i sont partitionnées en groupes $G_j = \{X_i | X_i \text{ est liée à } Z_j\}$. Cette affectation unique permet de regrouper les X_i ayant les plus fortes dépendances entre elles qu'on notera X_{G_j}. Z_j est alors en quelque sorte le centre du groupe, capturant les dépendances des X_i associées. Et $I(X_i, Z_j)$ quantifie la contribution de X_i pour la composante Z_j. La contribution de X_i pour la composante Z_j se mesure alors avec leur information mutuelle $I(X_i, Z_j)$. L'introduction des variables indicatrices $\alpha_{ij} = \mathbb{1}(X_i \in G_j)$, et l'expression de la TC comme une différence d'informations mutuelles ainsi que l'expression des principes de représentation ci-dessus aboutissent à la nouvelle écriture de (3),

$$\max_{\alpha, p(z|x)} \sum_{j=1}^{m} \sum_{i=1}^{p} \alpha_{i,j} I(X_i, Z_j) - \sum_{j=1}^{m} I(X, Z_j) \tag{4}$$

Sous cette forme, ce critère reste applicable pour les variables continues ; l'approche gaussienne en est un exemple [6], mais nous nous intéressons au cas discret, ne nécessitant pas d'*a priori*. Dans le cas discret, cette optimisation permet de déduire un algorithme de point fixe pour la recherche de la solution en introduisant les multiplicateurs de Lagrange. Cette solution s'approche de son majorant $TC(X)$, sans toutefois une preuve d'optimalité du résultat. Car plusieurs points fixes peuvent coexister et leur stabilité est inconnue.

3 CorEx et comparaison à d'autres méthodes de RD

Dans cette partie, on s'intéresse au calcul du CorEx et des résultats obtenus en l'appliquant sur des données de crédit financier. Ensuite, la pertinence du CorEx en tant que RD est comparée par rapport à des méthodes classiques de RD.

Le calcul du CorEx se base sur la recherche de point fixe qui découle de la solution de Lagrange de (4). À chaque étape, la solution précédente $p_t(z_j|x)$ est améliorée par

$$p_{t+1}(z_j|x) \propto p_t(z_j) \prod_{i=1}^{p} \left[\frac{p_t(z_j|x_i)}{p_t(z_j)} \right]^{\alpha_{ij}^t}. \tag{5}$$

ainsi que l'estimation de α est mise à jour par

$$\alpha_{ij}^{t+1} = (1-\lambda)\alpha_{ij}^t + \lambda \alpha_{ij}^*, \tag{6}$$

où $\alpha_{ij}^* = \exp(\gamma I(X_i, Y_j) - \max_j I(X_i, Y_j))$ qui a pour limite α_{ij} quand $\gamma \to \infty$ [4]. Les deux attributs $p(z_j|x)$ et α sont mis à jour en alternance, jusqu'à ce que le critère (4) se stabilise.

Cet algorithme a été appliqué sur les données des cartes de crédit à Taïwan [7, 8]. Les informations disponibles sont très hétérogènes : le statut du détenteur (âge, genre, statut marital, niveau d'éducation) et l'historique des paiements (montant autorisé, retards de paiements les 6 derniers mois, factures des 6 derniers mois, paiements des 6 derniers mois). Ce qui nous amène à considérer $p = 23$ variables initiales X. Parmi les

échantillons de taille $N = 30000$, 22,12% représentent un défaut de paiement. Nous allons utiliser 70% des échantillons tirés aléatoirement pour estimer les paramètres nécessaires, les 30% étant utilisés par la suite pour mesurer la performance. Cette séparation a été effectuée avec une stratification du défaut, permettant de respecter le taux de défaut dans les deux jeux de données.

En pratique, le calcul de Z à l'aide de (5), (6), nécessite une étape préliminaire qui consiste à établir la distribution de l'entrée $p(x)$ à partir des N échantillons notés $\tilde{x}_n$. Cette distribution est alors estimée par histogramme $\hat{p}(x) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\tilde{x}_n = x}$. Étant donné que le CorEx est basé sur des variables discrètes, chaque variable continue (comme les montants) a été discrétisée en 4 quartiles. Après avoir donné en entrée la dimension m et la taille du support k de Z, la table de distribution $p(z|x)$ se représente par un tableau de taille $m \times k \times p \times l$, où l est la taille du support des X_i, quitte à fixer des probabilités nulles. Dans toute notre étude, on fixe $k = 2$ pour que Z_j soit interprété comme une sortie binaire. La distribution $p(z_j|x)$ est alors mise à jour grâce (5). La distribution marginale $p(z_j)$ est obtenue en passant par la distribution jointe $p(z_j, x) = p(z_j|x).\hat{p}(x)$. La matrice α est de taille $p \times m$ dont la somme des éléments de chaque colonne vaut 1. Le pas d'estimation est fixé à $\lambda = 0.3$. Les étapes précédentes sont répétées 10 fois et le CorEx ayant le critère le plus élevé est retenu.

Une première analyse des résultats du CorEx est possible à partir de la structure résultante. La figure 1 illustre une visualisation des variables d'entrées (X_i colorées en gris) et des composantes de Z pour une dimension $m = 3$ (Z_1, Z_2, Z_3 colorés). Les variables ayant la plus grande dépendance entre elles sont regroupées autour d'un Z_j. La dépendance au sein d'un groupe est caractérisée en calculant la somme des informations mutuelles, $\sum_{i \in G_j} I(X_i, Z_j)$. La longueur des arrêtes entre un X_i et un Z_j calculée par $\log \frac{H(X_i)}{I(X_i, Z_j)}$ mesure la dépendance entre les deux variables.

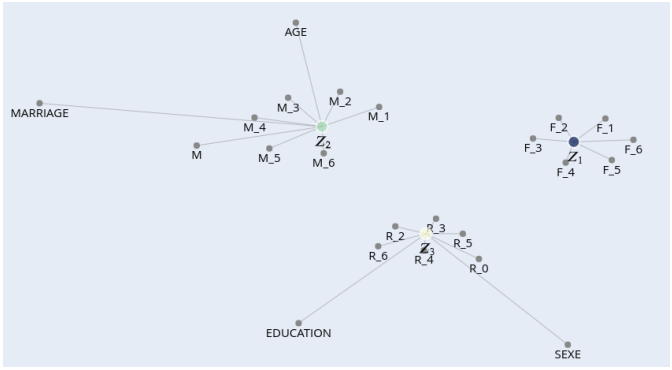

FIGURE 1 – Représentation visuelle du CorEx, pour une dimension de réduction à 3.

Dans un second temps, on s'intéresse à la pertinence de la représentation du CorEx comme RD. Pour ce faire, une régression logistique (RL) est entraînée pour classer le défaut. Pour chaque échantillon d'apprentissage $\tilde{x}_n$, on récupère les probabilités $\mathbb{P}(Z_j = 1|X = \tilde{x}_n)$ depuis le CorEx. Ces dernières servent alors de l'entrée pour l'apprentissage de la RL. Les groupes de variables et les coefficients de la RL correspondants sont présentés dans le tableau 1. Ces résultats montrent que la transformation du CorEx peut être significative dans le cadre d'une RD. Par exemple, la composante Z_3 regroupant entre autres les variables de retards de paiements est celle qui affecte le plus le défaut.

TABLE 1 – Groupe de variables associé pour chaque projection Z_j et le coefficient de la RL associé à Z_j

	$\mathbf{X_i}$ **associées**	**Coef. RL**
Z_1	Factures des 6 derniers mois ($F\_1, \ldots, F\_6$)	-0.0289
Z_2	Âge, Mariage, Montant actuel (M), Montants payés les 6 derniers mois ($M\_1, \ldots, M\_6$)	-0.5904
Z_3	Sexe, Éducation, Retards de paiements des 6 derniers mois ($R\_1, \ldots, R\_6$)	2.1215

Enfin, la pertinence du CorEx est comparée par rapport aux autres méthodes de RD non-supervisées, à savoir la PCA et l'ICA. Pour ces méthodes, à l'inverse du CorEx, les variables discrètes soient significatives. En utilisant un encodage 0 - 1, les catégories sont transformées en de nouvelles variables binaires représentant la catégorie correspondante. Cette transformation amène à considérer 70 variables contre les 23 initiales. Les données sont ensuite centrées et réduites pour éviter l'effet d'échelle. La moyenne et l'écart-type sont calculés sur les données d'apprentissage et seront utilisés pour le traitement des données de validation. Après avoir projeté ces données avec différentes méthodes de RD, une RL est entraînée sur le même jeu de données d'apprentissage. Les performances des RL sont comparées en utilisant la métrique mesurant l'aire sous la courbe ROC (AUC). Cette procédure a été testée en choisissant différentes dimensions, c.-à-d. $m \in [1, 3, 5, 10, 15, 20]$. Les résultats obtenus sur le jeu de données de validation sont présentés dans le tableau 2, les performances étant similaires entre les deux jeux de données.

En termes de performance, la RL sur les 70 variables reste opérationnelle avec un AUC de 0,755. En revanche, les performances similaires peuvent être atteintes pour des dimensions réduites. C'est le cas avec les réductions linéaires PCA et ICA, où on atteint un AUC de 0,747 pour une dimension $m = 15$. Enfin, les performances à partir du CorEx restent légèrement en dessous des autres méthodes, bien que ces résultats restent significatifs avec un AUC de 0,733 pour une dimension $m = 15$. Ces résultats peuvent s'expliquer par le fait que le défaut est séparable linéairement dans cette application. En revanche, le CorEx est la seule méthode capable de fournir une représentation graphique sur la façon dont les X_i peuvent être regroupées pour expliquer la cible Y. Cette analyse de données a permis d'identifier les variables permettant de discriminer le défaut.

Les données traitées par l'approche PCA/ICA sont par na-

TABLE 2 – AUC des méthodes de RD en fonction de la dimension réduite $m \in [1, 3, 5, 10, 15, 20]$

Méthodes RD	$\mathbf{m = 1}$	$\mathbf{m = 3}$	$\mathbf{m = 5}$	$\mathbf{m = 10}$	$\mathbf{m = 15}$	$\mathbf{m = 20}$	$\mathbf{m = 70}$
Sans RD	-	-	-	-	-	-	0.755
PCA	0.563	0.736	0.744	0.746	0.747	0.741	-
ICA	0.563	0.736	0.744	0.746	0.747	0.743	-
CorEx	0.543	0.725	0.717	0.723	0.733	0.738	-

ture continues contrairement au CorEx. Pour la première approche, les variables catégorielles sont pré-traitées par un encodage 0/1, dit one-hot, tandis que le CorEx discrétise les variables continues. La dimension des données se retrouve augmentée pour la PCA/ICA mais sans perte d'information alors que le CorEx subit une perte qui reste minime d'après les résultats obtenus.

Pour étudier la robustesse sur d'autres données réelles, la même méthodologie a été testée sur les données d'une banque comprenant plus de variables d'entrées, $p = 94$, ayant pour objectif de classifier le défaut d'un crédit. Les résultats pour une réduction en $m = 10$ variables sont présentés dans la figure 2.

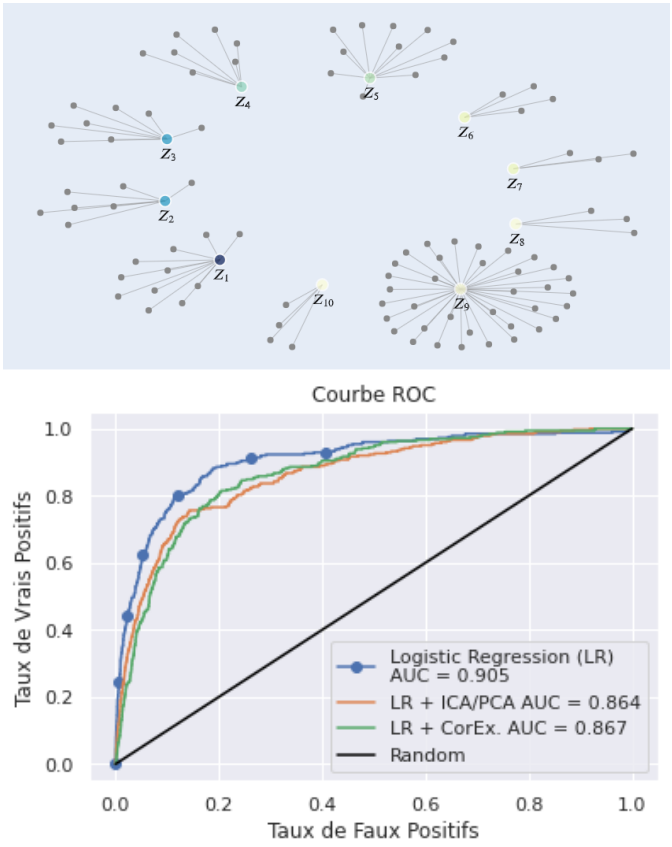

FIGURE 2 – En haut, représentation de la réduction par CorEx des 94 variables vers la dimension 10 ; en bas, courbe ROC des différentes méthodes.

4 Conclusion

Nous avons présenté l'algorithme CorEx en tant que méthode de RD et nous l'avons appliqué au défaut de crédit financier. Cet algorithme s'avère être un outil de RD efficace avec une partie interprétable qui peut apporter des informations utiles par rapport à l'application visée. Sans connaissance *a priori* des données, les résultats de classification à partir du CorEx s'approchent des méthodes linéaires de RD qui se sont avérées adaptées pour l'identification au défaut. En revanche, la structure obtenue avec le CorEx offre une représentation graphique permettant d'identifier les caractéristiques discriminantes par rapport au défaut. Il serait intéressant de comparer les méthodes de RD dans d'autres cas de figure où les jeux de données ne seraient pas linéaires. En particulier, on peut s'attendre à ce que le CorEx soit plus robuste dans ce contexte. Entre autres, nous envisageons de comparer le CorEx à d'autres méthodes de RD adaptées aux jeux de données.

Références

[1] N. Chen, B. Ribeiro, A. Chen, *Financial credit risk assessment : a recent review*, Artif Intell Rev 45, 1–23 (2016). https ://doi.org/10.1007/s10462-015-9434-x

[2] R. S. Ramya, S. Kumaresan, *Analysis of feature selection techniques in credit risk assessment*, 2015 International Conference on Advanced Computing and Communication Systems, 2015, pp. 1-6, doi : 10.1109/ICACCS.2015.7324139.

[3] H. Kim, C. Park, H. Yang, K. Sim, *Genetic Algorithm Based Feature Selection Method Development for Pattern Recognition*, 2006 SICE-ICASE International Joint Conference, 2006, pp. 1020-1025, doi : 10.1109/SICE.2006.315742.

[4] G. Ver Steeg, A. Galstyan, *Discovering structure in high-dimensional data through Correlation Explanation*, Proc. of the 27th International Conference on Neural Information Processing Systems, December 2014, pp 577–585

[5] S. Watanabe, *Information theoretical analysis of multivariate correlation*, IBM Journal of Research and Development 4, 1960 pp. 66–82.

[6] G. Steeg, H. Harutyunyan, D. Moyer, A. Galstyan, *Fast structure learning with modular regularization*, Advances in Neural Information Processing Systems, 2019

[7] *default of credit card clients Data Set*, UCI Machine Learning Repository, `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

[8] I. Yeh, C. Lien,*The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*, Expert Systems with Applications, Volume 36, Issue 2, Part 1, 2009, Pages 2473-2480, https ://doi.org/10.1016/j.eswa.2007.12.020.