Multilayer network analysis for improved credit risk prediction

Rapport de lecture

Rédigé par : Nzekon Nzeko'o Armel Jacques

Contexte :

La plupart des travaux sur le *credit scoring* considèrent uniquement les attributs de description des prêts (caractéristiques de l'emprunteur, somme à prêter, …), et ne s'attardent pas particulièrement sur la modélisation explicites des relations entre les emprunteurs.

Ceci peut être une limite, car intuitivement les individus aux caractéristiques communes peuvent avoir les mêmes comportements de prêts et donc à partir des comportements connus d'un ensemble d'individus similaires à un individu cible, on peut déduire le comportement de ce dernier.

A cet effet, les auteurs utilisent la modélisation par des graphes multicouches (multilayer network) où un emprunteur a autant de nœuds qu'il y a de dimensions qui le caractérisent, et dans chaque dimension, il est relié à des attributs qui le définissent suivant cette dimension. Ainsi, plus les emprunteurs sont similaires, plus ils sont proches dans le graphe multicouche.

Une dimension peut-être la localisation géographique ou encore le type d'activité exercé.

Cette nouvelle modélisation peut permettre de déduire de nouvelles caractéristiques d'entrées sur les prêts et éventuellement permettre l'amélioration de prédiction des techniques classiques du *credit scroring*.

Problème :

On considère un ensemble de prêts décrits par les caractéristiques sur l'emprunteur et des caractéristiques propres au prêt, et même la classe du prêt (remboursé ou non-remboursé). On veut prédire la classe d'un nouveau prêt à partir de sa description.

La question est la suivante : étant donné la description d'un prêt donné, est-ce que l'emprunteur va rembourser ou pas ?

Ceci est un problème de classification binaire supervisé des prêts à partir de leur description.

Motivation :

Les prêts des emprunteurs qui ont un grand nombre de caractéristiques en commun (suivant l'ensemble des dimensions) doivent avoir de grandes probabilités d'être de la même classe.

C'est ce dont il est question dans cet article.

Pour ce faire, il se pose les difficultés suivantes :

- Comment établir les relations entre les emprunteurs ?
- Comment déduire des caractéristiques à exploiter à partir de la nouvelle représentation ?
- Comment prédire la classe d'un prêt ?

I- Comment établir les relations entre les emprunteurs ?

Pour établir les relations entre les emprunteurs, les auteurs construisent un graphe multicouche.

Pour construire un tel graphe, on fixe les dimensions considérées et les attributs associés à chacune de ces dimensions.

Dans le cas de l'article qui s'attarde sur les prêts agricoles, les deux dimensions choisis sont : la localisation géographique et les produits vendus par les agriculteurs.

Les attributs de la dimension localisation géographique peuvent être le district, l'arrondissement … Et concernant la dimension produit, les attributs peuvent être les différents produits répertoriés.

Dans le graphe multicouche :

- chaque emprunteur a autant de nœud qu'il y a de couches considérées
- les nœuds de chaque emprunteur sont tous reliés les uns aux autres
- chaque attribut d'une dimension a un nœud associé
- si un emprunteur est décrit par un attribut dans une dimension donnée, alors le nœud emprunteur de cette dimension est relié au nœud attribut associé
- la navigation d'une couche à une autre se fait en passant par les nœuds emprunteurs des différentes couches

Construction de la matrice multicouche

Un graphe multicouche M, ayant N nœuds, et L couches, correspond à une représentation de dimension N x N x L x L, ceci peut être résumé en une matrice carrée (N x L) x (N x L).

Dans l'article, les auteurs considèrent deux dimensions pour décrire les emprunteurs dans le graphe multicouche, à savoir la localité et les produits vendus par ces derniers.

Considérons un cas où nous avons 4 emprunteurs, 2 localités et 3 produits. Dans ce cas de figure, nous avons 3 couches (Emprunteur, Localité et Produit), et nous avons 9 nœuds (4 nœuds emprunteurs + 2 nœuds localités + 3 nœuds produits), et donc la matrice carrée qui permet de représenter le graphe multicouche est de taille (9 x 3) x (9 x 3)

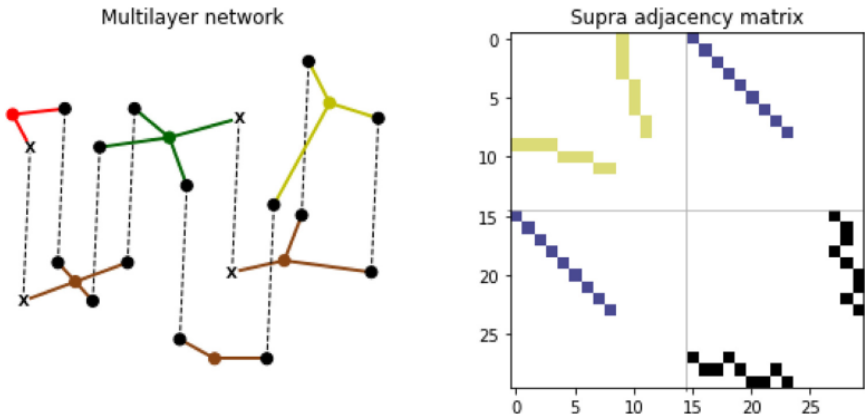

Fig. 1. A multilayer network (left) and its supra adjacency matrix (right). The multilayer network has nine borrower nodes (black), three location nodes (brown) and three product nodes (red, green, yellow) in two layers. The nodes in the first layer are connected with brown edges and the nodes in the second layer are connected with red, green and yellow edges. The inter edges (darkgray) connect a borrower node to itself in the other layer. In the supra adjacency matrix the upper left submatrix (yellow) denotes the adjacency matrix for the first layer, the lower right matrix (black) is the adjacency matrix for the second layer. The upper right and lower left submatrices (blue) are the adjacency matrices for the inter edges. They are diagonal as each node is only linked to itself in the other layer. The black nodes marked with **X** are the source of influence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

II- Comment déduire des caractéristiques à exploiter à partir de la nouvelle représentation ?

Lorsque le graphe multicouche est construit, les nouveaux descripteurs du prêt sont calculés suite à des applications du PageRank Personnalisé sur le graphe résultat.

Les auteurs proposent 03 façons différentes calculer les nouveaux descripteurs :

- Intra-influence :
- Inter-influence :
- Influence-combinée :

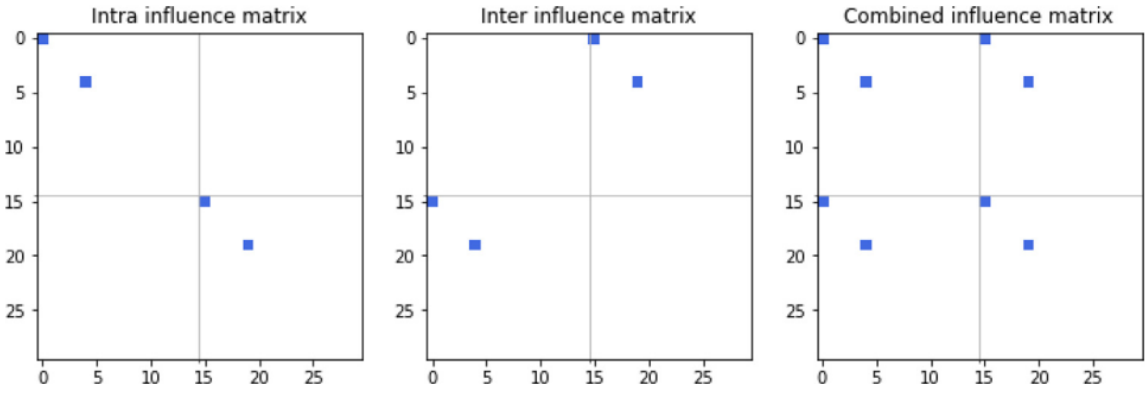

Fig. 2. Three scenarios for the influence matrix of a two layer multiplex network when computing a personalized PageRank.

III- Comment prédire la classe d'un prêt ?

Positionner les descripteurs classiques des prêts et les nouveaux descripteurs en entrée des modèles classiques (Régression logistique et XGBoost) de prédiction des risques de crédit.

Une fois que ces modèles sont construits, et dont utilisé pour prédire les classes des prêts du jeu de test, on évalue la contribution de chaque descripteur aux décisions du modèle.

Dans l'article, ils ont montré que les nouveaux descripteurs étaient parmi ceux qui contribuent le plus à la prise de décision des modèles de Régression logistique et XGBoost.

IV- Critiques :

1- Tous les attributs ne sont pas considérés dans le graphe multicouche.

2- Le choix des attributs catégorielles à considéré comme couche du graphe multicouche est fait de façon arbitraire. Il serait intéressant de proposer un protocole qui permet de faire des choix pertinents.

3- Les applications du PageRank Personnalisé sur le graphe multicouche ne sont pas suffisamment personnalisées, car le niveau de personnalisation reste à l'échelle intra et inter collectif (pour tous les nœuds concernés). Cependant, on pourrait avoir 03 applications différentes du PageRank Personnalisé pour chacun des prêts. Dans cette optique, seuls les nœuds qui sont liés au prêt courant peuvent recevoir le valeur 1.

IV- Extensions possibles :

1- Exploiter un graphe qui prend en compte tous les attributs descriptifs des prêts. On peut par exemple avoir un nœud pour chacune des modalités possibles de chaque attribut. Ensuite, relier tous les nœuds du même prêt ou alors incrémenter les poids des arcs qui relient tous les nœuds des modalités d'un prêt. Ensuite, appliquer le PageRank Personnalisé par chaque prêt, sur le graphe résultat, afin de ressortir avec de nouveaux descripteurs du prêt à l'exemple de l'estimer la classe de ce prêt par le PageRank.

2- Proposer un protocole qui permet de choisir efficacement les attributs catégoriels à considérer dans le graphe multicouche.

3- Personnaliser davantage les exécutions du PageRank Personnalisé de manière à affecter la valeur 1, uniquement aux nœuds associés au prêt courant pour lequel on calcul les valeurs des descripteurs.