REPUBLIQUE DU CAMEROUN
Paix-Travail-Patrie
UNIVERSITE DE YAOUNDE 1
DEPARTEMENT
D'INFORMATIQUE
BP/P.O.Box 812
Yaounde-Cameroun

REPUBLIC OF CAMEROON
Peace-Work-Fatherland
UNIVERSITY OF YAOUNDE 1
COMPUTER SCIENCES
DEPARTMENT
BP/P.O.Box 812
Yaounde-Cameroun

Prédiction du risque de crédit à base de descripteurs issus de la modélisation des données en graphes

Noms et prénoms : Victor Nico DJIEMBOU TIENTCHEU
Matricule : 17T2051
Niveau : Master 2
Spécialité : Sciences de Données (DS)
Encadreur : Dr. Armel Jacques NZEKON NZEKO'O

Superviseur : Pr. Maurice TCHUENTE

Table des matières

1 Introduction

Le trimestre dernier nous avons implémenté les concepts d'extraction de nouveaux descripteurs issus des graphes de *"Multilayer network analysis for improved credit risk prediction"* sur le jeu de données Afriland First Bank. Suite à des limites observé à l'article lû, des extensions ont été pensé notamment

- Construire un graphe multicouches avec toutes les variables qualitatives (sans sélectionner arbitrairement 2 comme les auteurs),
- Construire un graphe multicouche à une couche à variable qualitative pour pouvoir découler un protocole de selection d'attribut.
- Appliquer un PageRank personnalisé pour chaque exemple pour lequel on doit faire une prédiction, contrairement aux auteurs qui applique un PageRank personnalisé global dont les résultats sont répercuté sur tous les exemples (En d'autre termes, personnaliser l'exécution du PageRank pour chaque exemple du jeu de données).

Nous nous sommes concentré ce trimestre dans l'implementation de ses extensions appliqué à 5 jeux de données benchmark.

2 Positionnement dans le projet

Dans cette section du projet, il est question de modéliser les données d'une base de prêts bancaires par des graphes dont la définition des noeuds et des arcs est suffisamment pertinente pour que les nouveaux descripteurs extraits de ces graphes contribuent fortement à la décision des modèles d'apprentissage automatique pour la prédiction du risque de crédit.

3 Jeux de données

Pour mieux apprécier les extensions définit, 5 jeux de données très fréquemment utilisé dans le domaine de la prediction du risque de crédit ont été utilisé.

Nom du jeu de données	Nombre de variable catégorielle	Nombre de variable Numérique
AFB	04	08
AER de Kagle	03	08
Credit Risk	03	07
German	13	07
Japan	09	06

TABLE 1 – Description des jeux de données Crédit Risk benchmark utilisés

4 PageRank personnalisé

PageRank personnalisé :

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \left(\frac{PR(j)}{L(j)} \cdot P(j) \right) \tag{1}$$

où :

$PR(i)$ est le PageRank de l'emprunteur i,
d est le facteur d'amortissement (typiquement $d = 0,85$),
N est le nombre total d'emprunteur dans le réseau,
$M(i)$ est l'ensemble des emprunteurs qui ont un lien entrant vers l'emprunteur i,
$L(j)$ est le nombre total de liens sortants de l'emprunteur j,
$P(j)$ est le vecteur de probabilités représentant la probabilité de démarrer à l'emprunteur j.

La personnalisation du PageRank en attribuant des probabilités initiales plus élevées à certains noeuds permet d'identifier les noeuds clés qui ont le plus d'impact sur la propagation d'information ou d'influence dans le réseau. L'article à effet de pouvoir déterminer ses noeuds d'importances, propose trois types de personnalisation : **Combine, inter et intra**.

La logique Intra voudrait que seul les noeuds de modalité d'attributs soit pointé comme potentielle noeuds de démarrage de l'algorithme de PageRank. Celle d'inter se concentre plutôt sur le démarrage sur les noeuds emprunteurs dans le réseau. La version combine est tout simplement une exécution de l'algorithme de PageRank classique sur un réseau donné.

Nous nous proposons donc d'étendre la personnalisation du PageRank aux noeuds informatives d'un emprunteur ceci pour chaque emprunteur du réseau.

4.1 PageRank personnalisé à un seul emprunteur dans un graphe multicouche

Nous nous intéressons donc dans cette section à personnaliser le démarrage aux noeuds associés à la l'emprunt de l'emprunteur i.

4.1.1 Identification des noeuds associés à l'emprunteur

1: **procedure** GETUSERNODESLABEL(graph, borrowers, layers)
2: $linkTab \leftarrow \emptyset$
3: **for** chaque $borrower \in borrowers$ **do**

4: $edges \leftarrow (A, U_{borrower}^{layer \in layers}) \in graph,$
5: $linkTab \leftarrow \{A, U_{borrower}^{layer \in layers} \in edges\} \cup linkTab$
6: **end for**
7: **return** linkTab
8: **end procedure**

4.1.2 Génération du vecteur de personnalisation du PageRank

1: **procedure** COMPUTEPERSONALIZATION(nodeList, graph)
2: $perso \leftarrow \emptyset$
3: **for** chaque $noeud \in graph$ **do**
4: **if** $noeud \in nodeList$ **then**
5: $perso \leftarrow (noeud, 1) \cup perso$
6: **else**
7: $perso \leftarrow (noeud, 0) \cup perso$
8: **end if**
9: **end for**
10: **return** perso
11: **end procedure**

Les deux étapes ci-dessus nous montre comment personnaliser le PageRank à un seul emprunteur afin d'évaluer l'influence des informations d'un prêts dans le réseau des emprunteurs.

5 Graphes multicouches

L'article de base propose une seul modélisation des données, celle en graphe multicouche à 2 couches biparti avec les attributs formant les deux couches étant des descriptions catégorielles dans les informations de prêt prisent de façon pseudo-aléatoire. Toutefois, nous nous sommes intéressé à des questions telles que comment définir un mécanisme de selection des attributs devant servir pour les couches ? et est-ce que l'analyse d'un graphe multicouche à k dimensions des emprunteurs ne serait pas significative par rapport au problème de la prédiction du risque de crédit ?

Alors pour découler un protocole de selection d'attributs, nous nous sommes intéressé à l'étude des contribution d'attributs dans une modélisation de données de prêts en graphe multicouche à une couche. Puis nous avons pour répondre à la question de limitation de couche, conçu un graphe multicouche avec tous les attributs catégoriels.

5.1 Attributs extraits du graphe multicouche

Étant donné une modélisation des données en graphes multicouche à k couche pour la variable **case_k**, nous nous intéressons à l'extraction de neuf descripteurs :

- **MLN_{case_k}_degree** : le nombre d'emprunteur qui possède les informations dans les k couches.
- **MLN_bipart_intra_{case_k}** : le score maximale de PageRank sur les noeuds emprunteurs des k couches pour un PageRank avec initialisation probable uniquement des noeuds intra-couche (modalités).
- **MLN_bipart_inter_{case_k}** : le score maximale de PageRank sur les noeuds emprunteurs des k couches pour un PageRank avec initialisation probable uniquement des noeuds inter-couche (emprunteurs).
- **MLN_bipart_combine_{case_k}** : le score maximale de PageRank sur les noeuds emprunteurs des k couches pour un PageRank avec initialisation probable de tous les noeuds.
- **MLN_bipart_ultra_{case_k}** : le score maximale de PageRank sur les noeuds emprunteurs des k couches pour un PageRank avec initialisation probable uniquement des noeuds d'information d'un emprunteur.
- **MLN_bipart_intra_max_{case_k}** : le score maximale de PageRank sur les noeuds modalités des k couches associés à un emprunteur pour un PageRank avec initialisation probable uniquement des noeuds intra-couche.
- **MLN_bipart_inter_max_{case_k}** : le score maximale de PageRank sur les noeuds modalités des k couches associés à un emprunteur pour un PageRank avec initialisation probable uniquement des noeuds inter-couche .
- **MLN_bipart_combine_max_{case_k}** : le score maximale de PageRank sur les noeuds modalités des k couches associés à un emprunteur pour un PageRank avec initialisation probable de tous les noeuds.
- **MLN_bipart_ultra_max_{case_k}** : le score maximale de PageRank sur les noeuds modalités des k couches associés à un emprunteur pour un PageRank avec initialisation probable uniquement des noeuds d'information d'un emprunteur.

6 Résultats d'expérimentations

6.1 Meilleurs métriques

7 Conclusion

8 Discussion